HathiTrust and Your Library: 2016 Update

Mike Furlough
13 January 2016
University of Houston
Latest @hathitrust find: Anthropometry of Airline Stewardesses
hdl.handle.net/2027/len.35556...

wow, @hathitrust provides us with this outlandishly lovely gem: a coloring book for dentists! bit.ly/1AjHlHo

The saga of Dr. Anton Pazu: a coloring book for dentists.
babet.hathitrust.org

I found one of my key documents digitized by @hathitrust from @nypl - that's saved me a lot of research time and train fare!

Hathitrust: an amazing digital library and resource collection: hathitrust.org #hstr912 #hstr400 #hc12181

Happy to see a fully digitised version of my dissertation online, thanks to google + Hathitrust hdl.handle.net/2027/uc1.b2714...

Einfluss des werbegesangs von gomphocerus rufus I. auf...
Einfluss des werbegesangs von gomphocerus rufus I. auf das web...
babet.hathitrust.org

To see how somethings never change, read John Cotton Dana's "Library Problems"
hdl.handle.net/2027/njp.32101... Thank you @hathitrust.

This historian loves tech! Thanks @internetarchive and @hathitrust who saved my paper from death by early closing hours. #yesallmedievalists
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 100 research libraries partnering to develop its programs.
• A range of transformative programs enabled by working at a very large scale.
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.
Today’s Discussion

• The organization
• What we’ve collected
• Services and activities
• Current initiatives

• Your questions answered
The Organization
HathiTrust Members

Allegheny College
American University of Beirut
Arizona State University
Auburn University
Baylor University
Boston College
Boston University
Brandeis University
Brown University
Bryn Mawr College
Carnegie Mellon University
Case Western Reserve
Colby College
Columbia University
Cornell University
Dartmouth College
Duke University
Emory University
Getty Research Institute
Georgetown University
Georgia Tech
Harvard University Library
Haverford College
Indiana University
Iowa State University
Johns Hopkins University
Kansas State University
Lafayette College
Library of Congress
Massachusetts Institute of Technology
McGill University
Michigan State University
Montana State University
Mount Holyoke College
New Mexico State University
New York Public Library
New York University
North Carolina Central University
North Carolina State University
Northern Illinois University
Northeastern University
Northwestern University
Oklahoma State University
The Ohio State University
The Pennsylvania State University
Princeton University
Purdue University
Rutgers University
Smith College
Stanford University
State University System of Florida
Swarthmore College
Syracuse University
Temple University
Texas A&M University
Texas Christian University
Texas Tech University
Tufts University
Universidad Complutense de Madrid
University of Alabama
University of Alberta
University of Arizona
University of British Columbia
University of Calgary
University of California Berkeley
University of California Davis
University of California Irvine
University of California Los Angeles
University of California Merced
University of California Riverside
University of California San Diego
University of California San Francisco
University of California Santa Barbara
University of California Santa Cruz
University of California Digital Library
The University of Chicago
University of Connecticut
University of Delaware

University of Houston
University of Illinois
University of Illinois at Chicago
The University of Iowa
University of Kansas
University of Maine
University of Maryland
University of Massachusetts, Amherst
University of Miami
University of Michigan
University of Minnesota
University of Missouri
University of Nebraska-Lincoln
University of Nevada-Las Vegas
University of New Mexico
The University of North Carolina at Chapel Hill
University of Notre Dame
University of Oklahoma
University of Pennsylvania
University of Pittsburgh
University of Queensland
University of Rochester
University of Tennessee, Knoxville
University of Texas
University of Utah
University of Vermont
University of Virginia
University of Washington
University of Wisconsin-Madison
University of Wyoming
Utah State University
Vanderbilt University
Virginia Tech
Wake Forest University
Washington University
Washington State University
Yale University Library
HathiTrust remains one of the best things that has happened to our library in the last 10 years. Thank you.

One of our motivations is that HathiTrust is a COLLECTIVE solution with access.

...the real interest is in providing much higher quality services and access at much lower cost. Even if we were more generously funded to do something local, doing so would not advance our Library's interests as much as leveraging HathiTrust and doing more appropriate local things.

I hope to see my library increase its engagement with HathiTrust over time to collaborate on shared challenges and provide improved services to our users.

We see the advantages of working to scale even if our local infrastructure is adequate.
Cooperative Work

We draw upon distributed expertise

<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;</td>
<td>✔️</td>
<td>✔️</td>
<td></td>
<td></td>
</tr>
<tr>
<td>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>
Governance

Board of Governors

HathiTrust Members

Program Steering Committee

Executive Director

Committees and Working Groups

Operations

13 January 2016
Currently Active Committees

• Board of Governors
• Program Steering Committee
• Collections Committee
• Metadata Policy, Strategy Use and Advisory Group
• Zephir Advisory Group
• Rights & Access Working Group
• Ad Hoc Quality Working Group
• User Support Working Group
• 3 search committees

• Recently completed:
  – Print Monographs Archive Planning Task Force
• Momentary Pause:
  – Government Documents Initiative Planning and Advisory Group
Collections
HathiTrust in January 2016

- 13.8 million total items
  - 6.9 million book titles
  - 366,000 serial titles
  - 657,000 US federal government documents
  - 5.4 million items open (public domain & CC-licenses)

The collection primarily includes published materials in bound form, *digitized from library collections.*
# Collection Contributors

<table>
<thead>
<tr>
<th>Boston College</th>
<th>New York Public Library</th>
<th>University of Chicago</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brooklyn Museum</td>
<td>North Carolina State University</td>
<td>University of Connecticut</td>
</tr>
<tr>
<td>Clark Art Institute Library</td>
<td>Northwestern University</td>
<td>University of Delaware</td>
</tr>
<tr>
<td>Columbia University</td>
<td>Pennsylvania State University</td>
<td>Univ of Illinois at Urbana-Champaign</td>
</tr>
<tr>
<td>Cornell University</td>
<td>Princeton University</td>
<td>University of Iowa</td>
</tr>
<tr>
<td>Duke University</td>
<td>Private Donor(s)</td>
<td>University of Massachussetts Amherst</td>
</tr>
<tr>
<td>Emory University</td>
<td>Purdue University</td>
<td>University of Michigan</td>
</tr>
<tr>
<td>Harvard University</td>
<td>State University System of Florida</td>
<td>University of Minnesota</td>
</tr>
<tr>
<td>Indiana University</td>
<td>Technical Report Archive &amp; Image Library</td>
<td>University of Missouri - Columbia</td>
</tr>
<tr>
<td>Keio University</td>
<td>Texas A&amp;M University</td>
<td>Univ of North Carolina at Chapel Hill</td>
</tr>
<tr>
<td>Knowledge Unlatched</td>
<td>The Getty Research Institute</td>
<td>University of Virginia</td>
</tr>
<tr>
<td>Library of Congress</td>
<td>The Ohio State University</td>
<td>University of Washington</td>
</tr>
<tr>
<td>McGill University</td>
<td>Tufts University</td>
<td>University of Wisconsin - Madison</td>
</tr>
<tr>
<td>Millennium Project</td>
<td>Universidad Complutense de Madrid</td>
<td>Utah State University Press</td>
</tr>
<tr>
<td>Minnesota Digital Library</td>
<td>University of Alberta</td>
<td>Yale University</td>
</tr>
<tr>
<td></td>
<td>University of California</td>
<td></td>
</tr>
</tbody>
</table>

13 January 2016
The top 10 languages make up ~87% of all content.
Quality Management

- Validation upon ingest
- Updated versions from Google
- Proactive work by partners
- User reports

- “Quality in HathiTrust” by Jeremy York and Kat Hagedorn
  – https://www.hathitrust.org/quality-in-hathitrust
Where are HathiTrust’s Users

Top 10 countries account for 74% of users.
231 countries account for remaining 26% of users.

Period: 12/1/2014 to 11/30/2015.
Data taken from Google Analytics.
# Top Referring Sites

<table>
<thead>
<tr>
<th>Website</th>
<th>Sessions</th>
<th>% New Sessions</th>
</tr>
</thead>
<tbody>
<tr>
<td>google.com</td>
<td>4,723,224</td>
<td>68.62%</td>
</tr>
<tr>
<td>onlinebooks.library.upenn.edu</td>
<td>911,699</td>
<td>19.88%</td>
</tr>
<tr>
<td>worldcat.org</td>
<td>262,240</td>
<td>5.72%</td>
</tr>
<tr>
<td>dp.la</td>
<td>60,117</td>
<td>1.31%</td>
</tr>
<tr>
<td>mirlyn.lib.umich.edu</td>
<td>54,180</td>
<td>1.18%</td>
</tr>
<tr>
<td>clio.columbia.edu</td>
<td>40,821</td>
<td>0.89%</td>
</tr>
<tr>
<td>en.wikipedia.org</td>
<td>38,806</td>
<td>0.85%</td>
</tr>
<tr>
<td>psu.summon.serialssolutions.com</td>
<td>31,452</td>
<td>0.69%</td>
</tr>
<tr>
<td>vufind.carli.illinois.edu</td>
<td>31,078</td>
<td>0.68%</td>
</tr>
<tr>
<td>oskicat.berkeley.edu</td>
<td>30,510</td>
<td>0.27%</td>
</tr>
<tr>
<td>lccn.loc.gov</td>
<td>29,289</td>
<td>0.27%</td>
</tr>
</tbody>
</table>

Period: 12/1/2014 to 11/30/2015. Data taken from Google Analytics.
Top Ten Titles 2015

1. *Roster of the Confederate soldiers of Georgia, 1861-1865* (v.1-5)
4. *Godey's magazine*, v.40-41 1850
5. *A Short Guide to New Zealand*, [Special Services Division, US Army]
7. *Pennsylvania German Pioneers: a publication of the original lists of arrivals in the port of Philadelphia from 1727 to 1808*
10. *Consumption of the lungs and kindred diseases, treated and cured by kerosene*, by Charles Oscar King
Services and Activities
Preservation with Access

- Preservation
  - TRAC-certified

- Discovery
  - Bibliographic and full-text search of all materials (all)
  - Bibliographic data API enables local catalog integration (all)

- Access and Use
  - Varies
Modes of Access & Use

• Full text search (all users)
  – Text and data mining (all users)
    • Services require additional registration

• Public domain and open access works (all users)
  – Full-item download (members only)

• Collection building and APIs (all users)

• Lawful uses of in-copyright works (members only)
  – Access for users with print disabilities
  – Preservation replacement
Bringing Buried Treasure to Light
article-level discovery metadata for HathiTrust

Michelle Paolillo presenting for John Cline, Steven Folsom, Jaron Porciello

HathiTrust Annual Meeting
December 9, 2015
Adding Collections

• Different process for different sources
• 3 requirements: submission of documentation, metadata, and content
• Locally-digitized content must be prepared to our specifications – see https://www.hathitrust.org/ingest_tools
  – “Beyond Google Books: Getting Locally-Digitized Material into HathiTrust” by Aaron Elkiss http://is.gd/9btJk6
Submitting Print Holdings

- Requirement of membership to submit updated holdings annually
- Fees are based on local holdings and overlap with the HathiTrust collection
- Print holdings data is used to drive lawful access to in-copyright content
- Potential future use in the shared print repository, to drive federal docs scanning, analysis

More information at
https://www.hathitrust.org/print_holdings
Current Initiatives
Current Major Cooperative Initiatives

• Distributed manual copyright reviews.
• Establishing a distributed shared print monograph archive.
• Expanding services of the HathiTrust Research Center.
• Expanding and enhancing access to US Federal Government Documents
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

<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>329,520</td>
<td>175,794 (53%)</td>
</tr>
<tr>
<td>CRMS World Pub in UK (1875-1944), Canada and Australia (1894-1964)</td>
<td>244,469</td>
<td>130,374 (53%)</td>
</tr>
</tbody>
</table>
Shared Print Monographs Archive

• Ballot Initiative passed at the 2011 HT Constitutional Convention

• Focus
  – Ensure preservation of print and digital collections
  – Provide leadership in the area of monographic, print retention.
  – Catalyze national/continental collective management of collections

• Proposed Features:
  – Secure retention commitments for all monograph titles in HathiTrust. (50% within two years or less.)
  – Ensure that the retained volumes are accessible (will circulate).
  – Support work to disclose commitments and discover content.
  – Support members’ efforts to make local collection management decisions.

• Original proposal, task force charge and preliminary recommendations:
  – [https://www.hathitrust.org/print_monograph_archiving](https://www.hathitrust.org/print_monograph_archiving)
December 2015
45.68% match with HathiTrust
Federal Documents Initiative

- Ballot Initiative passed at the 2011 HT Constitutional Convention

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

- Near term activities:
  - Bibliographic and collections analysis
    - Developing a registry of US Federal Government Documents
      - https://www.hathitrust.org/usdocs_registry
    - Locate materials for inclusion in the collections
  - Focus first on known and cataloged materials
    - Prioritize print, post-1976 materials
    - Identify collections for inclusion (and get them)
    - Digitize where needed
  - Publicize the efforts
    - Within the library community and the general public
“Non-consumptive” Research: The HathiTrust Research Center

• HathiTrust distributes public domain datasets
  – [https://www.hathitrust.org/datasets](https://www.hathitrust.org/datasets)

• HathiTrust Research Center
  – Developed collaboratively by Indiana University and University of Illinois.
  – Funding from the Sloan Foundation, Andrew W. Mellon Foundation, and NEH Office of Digital Humanities as well as HathiTrust.
  – Secure HathiTrust Analytics Research Commons
    • [https://sharc.hathitrust.org/](https://sharc.hathitrust.org/)
HTRC Extracted Features Dataset

Page-level features from 4.8 million volumes [v.0.2]

Note that this is a beta data release. Please send feedback to htrc-support-l@list.indiana.edu.

A great deal of useful research can be performed non-consumptively with pre-extracted features. For this reason, we've prepared a data export of features for the public domain volumes of the HathiTrust Digital Library.

Features are notable or informative characteristics of the text. Also, we have processed a number of useful features, including part-of-speech tagged token counts, header and footer identification, and various line-level information. These are provided per-page. Providing token information at the page-level makes it possible to separate text from paratext; for instance, a researcher may use the information to identify publishers' ads at the back of a book. For cleaner text, headers and footers are also identified distinctly from page content. The specific features that we extract for each page are described in more detail below.

The most useful extracted feature that we are provide is the token (unigram) count, on a per-page basis. Term counts are specific to the part-of-speech usage for that term, so that a term used as both a noun and a verb, for example, will have separate counts provided for both these modalities of its use. We also include line information, such as the number of lines with text on each page, and a count of characters that start and end lines on each page. This information can illuminate genre and volume structure: for instance, it helps distinguish poetry from prose, or body text from an index.

The present release is a beta release, and we would love to hear about how you use it, or what else you would like to see!

Review the documentation below or jump straight to the downloads.

Data Stats

<table>
<thead>
<tr>
<th># of volumes</th>
<th>4,801,237</th>
</tr>
</thead>
<tbody>
<tr>
<td># of pages</td>
<td>1,825,317,899</td>
</tr>
<tr>
<td>Median pages/volume</td>
<td>330</td>
</tr>
</tbody>
</table>

More information about HathiTrust datasets.

More Resources

- Use Cases
- User Guide

Attribution


This feature dataset is released under a Creative Commons Attribution 4.0 International License.
2016 Agenda: A Sampler

• Launch of Shared Print Monographs Program.
• Active collection development in Federal Documents effort.
• Improving engagement across the membership.
  – Input into development agenda (ballot initiative)
• Transitioning copyright review to ongoing activity.
• Assessment of print disability services.
• Issues referred from Program Steering Committee and others (more later).
What needs thought?

• Strategy, mission, and role in the future
  – Membership growth
  – Financial model
  – Collections program
  – Public policy
  – (Inter)National digital infrastructure
  – Services for members and the public

• Organizational
  – Engagement with researchers and libraries
  – Enabling more participation in plans and action
  – Standing on our own
Assumptions

• Our actions must align with the mission, goals, and purpose across our partnership.

• A few additional assumptions
  – Internationally we should pursue complementarity and cooperation, not competition and duplication.
  – Scale will continue to drive our work
  – Potential partners are not just other libraries and library organizations, but also readers, authors, publishers.
How to find out more

• Contact us: feedback@issues.hathitrust.org

• About: http://www.hathitrust.org/about
• Resources: http://www.hathitrust.org/resources
• Twitter: http://twitter.com/hathitrust
• Facebook: http://www.facebook.com/hathitrust
• Monthly newsletter:
  – http://www.hathitrust.org/updates
• Blogs: http://www.hathitrust.org/blogs
  – Large-scale Search
  – Perspectives from HathiTrust
Thank You!
Your Questions
Would Shakespeare ride a bicycle? All answers in this 1896 @hathitrust pamphlet hd.handle.net/2027/udel.3174... ht @AngelinaNZ

Another day, another reason to be thankful for @hathitrust.

Things I learned this week: HathiTrust is not pronounced "Hathy" or "Hothy" -- it's "Hah-tee Trust"

Love that HathiTrust is an integral part of UC's plan. #gpodlc15

Amo HathiTrust y que me pueda descargar libros como "Memorias de las reynas catedráticas" de Enrique Flórez que lo llevaba mucho buscando : ()

A British report from 1794 on rhinoplasty in India babel.hathitrust.org/cgi/imsrv...
# Contributing Institutions: Top 20

<table>
<thead>
<tr>
<th>CONTRIBUTOR</th>
<th>VOLUMES</th>
<th>PCT</th>
</tr>
</thead>
<tbody>
<tr>
<td>University of Michigan</td>
<td>4,698,900</td>
<td>33.99%</td>
</tr>
<tr>
<td>University of California</td>
<td>3,716,137</td>
<td>26.88%</td>
</tr>
<tr>
<td>Harvard University</td>
<td>838,526</td>
<td>6.06%</td>
</tr>
<tr>
<td>Cornell University</td>
<td>584,901</td>
<td>4.23%</td>
</tr>
<tr>
<td>University of Wisconsin - Madison</td>
<td>561,704</td>
<td>4.06%</td>
</tr>
<tr>
<td>Indiana University</td>
<td>530,632</td>
<td>3.84%</td>
</tr>
<tr>
<td>University of Minnesota</td>
<td>442,696</td>
<td>3.20%</td>
</tr>
<tr>
<td>University of Illinois at Urbana-Champaign</td>
<td>437,567</td>
<td>3.16%</td>
</tr>
<tr>
<td>Pennsylvania State University</td>
<td>390,164</td>
<td>2.82%</td>
</tr>
<tr>
<td>New York Public Library</td>
<td>310,761</td>
<td>2.25%</td>
</tr>
<tr>
<td>Princeton University</td>
<td>252,890</td>
<td>1.83%</td>
</tr>
<tr>
<td>The Ohio State University</td>
<td>118,569</td>
<td>0.86%</td>
</tr>
<tr>
<td>Universidad Complutense de Madrid</td>
<td>117,509</td>
<td>0.85%</td>
</tr>
<tr>
<td>Library of Congress</td>
<td>108,892</td>
<td>0.79%</td>
</tr>
<tr>
<td>University of Chicago</td>
<td>99,186</td>
<td>0.72%</td>
</tr>
<tr>
<td>Keio University</td>
<td>90,126</td>
<td>0.65%</td>
</tr>
<tr>
<td>University of Alberta</td>
<td>76,114</td>
<td>0.55%</td>
</tr>
<tr>
<td>Columbia University</td>
<td>74,514</td>
<td>0.54%</td>
</tr>
<tr>
<td>Northwestern University</td>
<td>57,156</td>
<td>0.41%</td>
</tr>
<tr>
<td>University of Virginia</td>
<td>51,220</td>
<td>0.37%</td>
</tr>
<tr>
<td>25 Other Contributors</td>
<td>268,138</td>
<td>1.94%</td>
</tr>
</tbody>
</table>

47 13 January 2016
Near/Intermediate Term Activity

• Bibliographic and collections analysis
  – Registry and holdings work
• Focus first on known and cataloged materials
  – Prioritize print, post-1976 materials
  – Identify collections for inclusion (and get them)
  – Digitize where needed
• Publicize the efforts
  – Within the library community
  – To the general public
Non-Consumptive Research

*From the proposed/rejected Google Book Settlement:*

“Non-Consumptive Research’ means research in which computational analysis is performed on one or more Books, but not research in which a researcher reads or displays substantial portions of a Book to understand the intellectual content presented within the Book.”

(a) Image Analysis and Text Extraction  
(b) Textual Analysis and Information Extraction  
(c) Linguistic Analysis  
(d) Automated Translation.  
(e) Indexing and Search
HTRC Non-Consumptive Research Paradigm

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.
• Differentiates human researcher from proxy: users are human beings.
Goals for the Research Center

• Research arm of HathiTrust
• Provide a persistent and sustainable structure to enable original and cutting edge research.
  • Leverage data storage and computational infrastructure at Indiana & Illinois
  • Stimulate community development of new functionality and tools
  • Use tools to enable discoveries that would not be possible without the HTRC
• Enable scholars to use the HathiTrust collections in new ways within US Copyright law.
  – Infrastructure and services for non-consumptive research.
Welcome to the HathiTrust Research Center!

The HathiTrust Research Center (HTRC) provides research access to the public domain corpus of the HathiTrust Digital Library. The HTRC is a collaborative research center launched jointly by Indiana University and the University of Illinois, along with the HathiTrust Digital Library, to help meet the technical challenges of dealing with massive amounts of digital text that researchers face by developing cutting-edge software tools and cyberinfrastructure to enable advanced computational access to the growing digital record of human knowledge. The HTRC provides an infrastructure to search, collect, analyze, and visualize the full text of nearly 3 million public domain works and is intended for nonprofit and educational researchers.

What would you like to do today?

- Create Workset
  - Create workset using our workset builder.

- Upload Workset
  - Upload a workset by specifying the necessary data about its volumes through a text file.

- Browse Workset
  - Browse through already created worksets.

- Execute Algorithms
  - Select and execute text analysis algorithms for word count to more sophisticated approaches.

SHARC = Secure HathiTrust Analytic Research Commons