HathiTrust Update

Google European Libraries Summit
October 18, 2012
Heather Christenson, California Digital Library
Overview

- Introduction
- Services
  - Preservation
  - Discovery
  - Access
- Impact
  - Copyright review
  - Collection management
- Governance/sustainability
- Future
Introduction
The Name

• The meaning behind the name
• Hathi (hah-tee)--Hindi for elephant
• Big, strong
• Never forgets, wise
• Secure
• Trustworthy
## Partnership

Arizona State University  
Baylor University  
Boston College  
Boston University  
California Digital Library  
Columbia University  
Cornell University  
Dartmouth College  
Duke University  
Emory University  
Florida State University  
Getty Research Institute  
Harvard University Library  
Indiana University  
Johns Hopkins University  
Lafayette College  
Library of Congress  
Massachusetts Institute of Technology  
McGill University  
Michigan State University  
New York Public Library  
New York University  
North Carolina Central University  
North Carolina State University  
Northwestern University  
The Ohio State University  
The Pennsylvania State University  
Princeton University  
Purdue University  
Stanford University  
Texas A&M University  
Universidad Complutense de Madrid  
University of Arizona  
University of Calgary  
University of California Berkeley  
University of California Davis  
University of California Irvine  
The University of Chicago  
University of Connecticut  
University of Delaware  
University of Florida  
University of Illinois  
University of Illinois at Chicago  
The University of Iowa  
University of Maryland  
University of Miami  
University of Michigan  
University of Minnesota  
University of Missouri  
University of Nebraska-Lincoln  
The University of North Carolina at Chapel Hill  
University of Notre Dame  
University of Pennsylvania  
University of Pittsburgh  
University of Utah  
University of Virginia  
University of Washington  
University of Wisconsin-Madison  
Utah State University  
Washington University  
Yale University Library
HathiTrust Mission

• To contribute to the common good by collecting, organizing, preserving, communicating, and sharing the record of human knowledge
Digital Repository

• Launched 2008
• Initial focus on digitized book and journal content
  – 10.5 million total volumes
  – 5.5 million book titles
  – 270,000 serial titles
  – 3.2 million public domain (~30%)
Initial Goals

• Reliable and comprehensive archive of materials converted from print...co-owned
• Improve access ...to meet the needs of the co-owning institutions
• Ensure the long-term preservation of content
• Coordinate shared storage strategies
• “public good” ...sustaining the historical record
• Simultaneously ...centralized ...open
Content Distribution

- In-copyright or undetermined: 70%
- "Public Domain": 30%
- Public Domain (worldwide): 15%
- Public Domain (US): 10%
- U.S. Federal Government Documents (worldwide): 4%
- Open Access: .1%
- Creative Commons: .01%
Content Sources

- Michigan: 45%
- California: 33%
- Other sources:
  - Cornell: 4%
  - Wisconsin: 5%
  - NYPL: 3%
  - Princeton: 3%
  - Indiana: 2%
  - Harvard: 1%
  - Columbia: 1%
  - LC: 1%
  - Minnesota: 1%
  - Virginia: 0%
  - Yale: 0%
  - Utah State: 0%
  - UNC-Chapel Hill: 0%
  - Duke: 0%
  - Illinois: 0%
  - NCSU: 0%
  - Chicago: 0%
  - Purdue: 0%
  - Penn State: 0%
Dates

- 1990-1999: 14%
- 1980-1989: 15%
- 1970-1979: 13%
- 1960-1969: 11%
- 1950-1959: 6%
- 1940-1949: 4%
- 1930-1939: 4%
- 1920-1929: 4%
- 1910-1919: 4%
- 1900-1909: 4%
- 1850-1899: 8%
- 1800-1849: 3%
- 1600-1699: 0%
- 1500-1599: 0%
- 0-1500: 0%
- 2000-2009: 10%
Language Distribution (1)

The top 10 languages make up ~86% of all content

- English: 48%
- German: 9%
- French: 7%
- Spanish: 5%
- Chinese: 4%
- Russian: 3%
- Japanese: 3%
- Italian: 3%
- Arabic: 2%
- Latin: 1%
- Remaining Languages: 14%
Language Distribution (2)

The next 40 languages make up ~13% of total
Services
Services

• Long-term preservation
• Bibliographic search
• Full-text search
• Reading and download capabilities
• Print on demand
• Collections
• Datasets, Research Center
Preservation with Access

- Cost effective preservation and access services
- Preservation
  - TRAC-certified
  - Robust infrastructure
  - Long-term commitments on digital content facilitate planning, decision-making
Discovery and Use

• Discovery
  – Bibliographic and full-text search of all materials
  – Extended discovery (ProQuest, EBSCO, OCLC, Ex Libris)

• Use
  – Multiple viewing mechanisms
  – Mobile interface
  – “Collection builder”
  – Print disabled access
Discovery and Use

• APIs and data feeds
  – Data API
  – Bibliographic API
  – “Hathifiles” inventory files
  – OAI

• Computational Research
  – Distribution of datasets
  – Protocol-based access
  – Research Center
15

CH.

AMERICAN PUBLISHER'S NOTICE

This work, although based upon Chambers's Encyclopaedia, whose distinguished merit is widely known, differs from it in important respects. It could scarcely be expected that an Encyclopaedia, edited and published for a foreign market, would give as much prominence to American topics as American readers might desire. To supply these and other deficiencies the American Editors have inserted about 15,000 titles, go arranging the whole, including Chambers's Supplement, in a single alphabet. The total number of titles is now about 40,000. The additions give greater fullness in the departments of biography, geography, history, natural history, and general and applied science. Scrupulous care has been taken not to mutilate or modify the original text of each edition of 1880: no changes have been made except such verbal alterations as are required by the omission of the wood-cuts. The titles of articles from Chambers's Encyclopaedia, either from the main work or from the Supplement, are printed in bold-faced type AMERICA. The titles of the American additions, whether of new topics or of enlargements of the old, are printed in plain capitals AMERICA. Should it appear that an article from the English work and its American continuation disagree in any points, the reader will readily refer the conflicting statements to their proper sources.

The labor of consultation will be much reduced by the catch-words in bold-faced type at the top of the page, being the first and last titles of the pages which face each other; and by the full title-words on the back of the volume, being the first and last titles contained therein.

[30114]
Famous Characters of History

HERNANDO CORTEZ

VOLUME XV.

ILLUSTRATED

1906

THE ST. HUBERT GUILD
Impact
A global change in the library environment

Academic print book collection already substantially duplicated in mass digitized book corpus

June 2010
Median duplication: 31%

June 2009
Median duplication: 19%
Collection Management, Development

- Overlap
  - More than 50% median overlap with ARL institutions; higher for small liberal arts colleges
- Print monographs archiving
- Pricing model based on Print holdings
  - Requires print holdings database
  - Also support expansion of legal uses, efforts in de-duplication
  - Facilitate individual and collaborative collection development and management operations
Copyright

- Automatic rights determination
- Manual rights determination
- Rights holder permissions
- [http://www.hathitrust.org/copyright](http://www.hathitrust.org/copyright)
Automatic Rights Determination

• Conducted on all works at time of ingest and when records are modified
  – Public domain worldwide
    • US works published before 1923, US federal government publications, non-US works published prior to 1872
  – Public domain in the United States
    • Non-US works published prior to 1923
Manual Rights Determination

• IMLS-funded CRMS project
  – US-published works 1923-1963
  – Double-blind review with expert review for conflicts
  – Staff at 4 HathiTrust partner institutions
  – Recently expanded to non-US works
  – Staff at 15 HathiTrust partner institutions
  – As of February 2012 ~190,000 reviewed, more than 100,000 opened

• Rights Holder Permissions
Governance/
Sustainability
Constitutional Convention

- October 2011
- 52 partners
- 3-year review overseen by Strategic Advisory Board
- Ballot Proposals
  - Print monograph storage
  - Approval Process for development initiatives
  - U.S. Government Documents
  - Fee-for-service content deposit
  - Governance
Collaborative Support

• New pricing model
• Base infrastructure costs
  – Public domain
  – In-copyright/undetermined
• Funds for programmatic initiatives
How much does it cost?

**Infrastructure Costs**

- **Storage, sync, maintenance**: 47.30%
- **Tape Backup**: 13.51%
- **Machine room**: 6.21%
- **Staff**: 24.77%
- **Servers**: 8.21%
Who can become a partner?

- Institutions worldwide
- Libraries with print holdings
The Future
Work Going Forward

• Print archiving, management
• Discovery and use
  – Lawful uses
• Research Center
• Quality: http://hathitrust-quality.projects.si.umich.edu/index.htm
• Government documents
• Beyond books and journals
• Publishing
• Transitioning to next phase of partnership
How to find out more

- Web site “About” section
  - [http://www.hathitrust.org/about](http://www.hathitrust.org/about)
- HathiTrust Research Center
  - [http://www.hathitrust.org/htrc](http://www.hathitrust.org/htrc)
- Twitter
  - [http://twitter.com/hathitrust](http://twitter.com/hathitrust)
- Monthly newsletter
  - [http://www.hathitrust.org/updates](http://www.hathitrust.org/updates)
  - RSS: [http://www.hathitrust.org/updates_rss](http://www.hathitrust.org/updates_rss)
- Contact us: [feedback@issues.hathitrust.org](mailto:feedback@issues.hathitrust.org)
- Blogs: [http://www.hathitrust.org/blogs](http://www.hathitrust.org/blogs)
  - Large-scale search
  - Perspectives from HathiTrust
CALIFORNIA

AN ELEPHANT REMEMBERS YOU,
AND IF YOU FEED HIM HAY
HE'S PLEASED, BUT IF YOU'RE MEAN TO HIM,
HE'LL PAY YOU BACK SOME DAY.

AN ELEPHANT.
Thank you!

Heather Christenson
California Digital Library
heather.christenson@ucop.edu
Main points in Authors Guild v. HathiTrust decision

- Section 108 on library privileges doesn’t limit the scope of fair use.
- A search index and access for the print-disabled are both fair uses.
- Search indexing is a transformative use.
- The libraries aren’t making commercial uses, even though they partnered with Google to get the scans.
- The plaintiffs haven’t proven that HathiTrust is creating any security risks.
- There is no market for scanning and print-disabled access, nor is one likely to develop.
- UM is required under the ADA to provide equal access to the print-disabled, and is allowed to under Section 121 of the Copyright Act