HathiTrust: Putting Research in Context

HTRC UnCamp
September 10, 2012
John Wilkin, Executive Director, HathiTrust
Introduction
<table>
<thead>
<tr>
<th>Arizona State University</th>
<th>North Carolina State University</th>
<th>University of Delaware</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baylor University</td>
<td>Northwestern University</td>
<td>University of Florida</td>
</tr>
<tr>
<td>Boston College</td>
<td>The Ohio State University</td>
<td>University of Illinois</td>
</tr>
<tr>
<td>Boston University</td>
<td>The Pennsylvania State University</td>
<td>University of Illinois at Chicago</td>
</tr>
<tr>
<td>California Digital Library</td>
<td>Princeton University</td>
<td>The University of Iowa</td>
</tr>
<tr>
<td>Columbia University</td>
<td>Purdue University</td>
<td>University of Maryland</td>
</tr>
<tr>
<td>Cornell University</td>
<td>Stanford University</td>
<td>University of Miami</td>
</tr>
<tr>
<td>Dartmouth College</td>
<td>Texas A&amp;M University</td>
<td>University of Michigan</td>
</tr>
<tr>
<td>Duke University</td>
<td>Universidad Complutense de Madrid</td>
<td>University of Minnesota</td>
</tr>
<tr>
<td>Emory University</td>
<td>University of Arizona</td>
<td>University of Missouri</td>
</tr>
<tr>
<td>Florida State University</td>
<td>University of California</td>
<td>University of Nebraska-Lincoln</td>
</tr>
<tr>
<td>Getty Research Institute</td>
<td>University of California Berkeley</td>
<td>The University of North Carolina at Chapel Hill</td>
</tr>
<tr>
<td>Harvard University Library</td>
<td>University of California Davis</td>
<td>University of Notre Dame</td>
</tr>
<tr>
<td>Indiana University</td>
<td>University of California Irvine</td>
<td>University of Pennsylvania</td>
</tr>
<tr>
<td>Johns Hopkins University</td>
<td>University of California Merced</td>
<td>University of Pittsburgh</td>
</tr>
<tr>
<td>Lafayette College</td>
<td>University of California Riverside</td>
<td>University of Utah</td>
</tr>
<tr>
<td>Library of Congress</td>
<td>University of California San Diego</td>
<td>University of Virginia</td>
</tr>
<tr>
<td>Massachusetts Institute of Technology</td>
<td>University of California San Francisco</td>
<td>University of Washington</td>
</tr>
<tr>
<td>McGill University</td>
<td>University of California Santa Barbara</td>
<td>University of Wisconsin-Madison</td>
</tr>
<tr>
<td>Michigan State University</td>
<td>University of California Santa Cruz</td>
<td>Utah State University</td>
</tr>
<tr>
<td>New York Public Library</td>
<td>The University of Chicago</td>
<td>Washington University</td>
</tr>
<tr>
<td>New York University</td>
<td>University of Connecticut</td>
<td>Yale University Library</td>
</tr>
</tbody>
</table>
Mission

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

Universal Library

Common Goal

Single Entity, Many Partners
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%)
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%

U.S. Federal Government Documents (worldwide): 4%

Public Domain (US): 10%

Open Access: .1%

Creative Commons: .01%
Content Sources

- Michigan: 45%
- California: 33%
- Wisconsin: 5%
- Cornell: 4%
- NYPL: 3%
- Princeton: 3%
- NYPL: 3%
- Columbia: 1%
- Indiana: 2%
- Harvard: 1%
- Madonna: 1%
- Northwestern: 1%
- Duke: 0%
- Yale: 0%
- Illinois: 0%
- Utah State: 0%
- UNC-Chapel Hill: 0%
- NCSU: 0%
- Chicago: 0%
- Penn State: 0%
- Purdue: 0%
- UC-Berkeley: 0%
- CU-Boulder: 0%
- UC-Davis: 0%
- UC-Irvine: 0%
- UC-Riverside: 0%
- UC-San Diego: 0%
- UC-Santa Barbara: 0%
Dates

- 1900-1909: 4%
- 1910-1919: 4%
- 1920-1929: 4%
- 1930-1939: 4%
- 1940-1949: 4%
- 1950-1959: 6%
- 1960-1969: 11%
- 1970-1979: 13%
- 1980-1989: 15%
- 1990-1999: 14%
Language Distribution (1)

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

- English: 48%
- German: 9%
- French: 7%
- Spanish: 5%
- Chinese: 4%
- Russian: 4%
- Japanese: 3%
- Italian: 3%
- Arabic: 2%
- Latin: 1%

Remaining Languages: 14%
The next 40 languages make up ~13% of total
Services

• Long-term preservation
  – Bit-level and migration
• Bibliographic search
• Full-text search
• Reading and download capabilities
• Print on demand
• Collections
• Datasets, Research Center
Impact
A global change in the library environment

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

June 2009
Median duplication: 19%

June 2010
Median duplication: 31%
Digitized Books in Shared Repositories

~3.5M titles

~75% of mass digitized corpus is ‘backed up’ in one or more shared print repositories

Unique Titles

Sep-09 Oct-09 Nov-09 Dec-09 Jan-10 Feb-10 Mar-10 Apr-10 May-10 Jun-10

Mass digitized books in Hathi digital repository

Mass digitized books in shared print repositories

~2.5M
Collection Management, Development

• Overlap
  – More than 50% median overlap with ARL institutions; higher for small liberal arts colleges
• 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
• Print monographs archiving
Discovery and Use

- Search, collections, online access
- APIs and data feeds
  - Data API
  - Bibliographic API
  - “Hathifiles” inventory files
  - OAI
- Computational Research
  - Distribution of datasets
  - Protocol-based access
  - Research Center
Research Center in Context
Institutional Support / Sustainability
Constitutional Convention

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

Strategic Advisory Board

Guidance on Policy, Planning

Executive Committee

Budget/Finances Decision-making

HathiTrust

- 12-member Board of Governors
- Executive Committee
- Executive Director
Collaborative Support

• New pricing model
• Base infrastructure costs
  – Public domain
  – In-copyright/undetermined
• Funds for programmatic initiatives
The Future
Concluding thoughts
Thank you!