HathiTrust: Reviewing Goals, Accomplishments, and Opportunities for Collective Action

CNI Fall 2011
December 13, 2011
Jeremy York
Project Librarian, HathiTrust
# Short-term objectives (1)

<table>
<thead>
<tr>
<th>Then</th>
<th>Now</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PageTurner</strong></td>
<td>Multiple views, embeddable</td>
</tr>
<tr>
<td><strong>Branding</strong></td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Format validation, migration, and error-checking</strong></td>
<td>Quarterly audits of all content (two instances with balancing and failover)</td>
</tr>
<tr>
<td><strong>APIs (access and integrate information)</strong></td>
<td>OAI</td>
</tr>
<tr>
<td><strong>Users who have print disabilities</strong></td>
<td>UM-only</td>
</tr>
<tr>
<td><strong>All institutions, keyed off of holdings database</strong></td>
<td>All institutions, keyed off of holdings database</td>
</tr>
</tbody>
</table>
## Short-term objectives (2)

<table>
<thead>
<tr>
<th></th>
<th>Then</th>
<th>Now</th>
</tr>
</thead>
<tbody>
<tr>
<td>Public Discovery Interface</td>
<td>No</td>
<td>Bibliographic Catalog (April 2009)</td>
</tr>
<tr>
<td>Virtual Collections</td>
<td>Yes</td>
<td>Much improved interface, collections of arbitrary size</td>
</tr>
<tr>
<td>Mechanisms for direct ingest of non-Google content</td>
<td>No</td>
<td>Yes, IA + framework for scalable ingest of non-Google</td>
</tr>
</tbody>
</table>
## Long-term objectives

<table>
<thead>
<tr>
<th>Then</th>
<th>Now</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compliance with TRAC</td>
<td>Yes!</td>
</tr>
<tr>
<td>Robust discovery like full-text search</td>
<td>Full-text search (November 2009)</td>
</tr>
<tr>
<td>Open service definition (for development of access and discovery tools)</td>
<td>Data API + Development environment</td>
</tr>
<tr>
<td>Support beyond books and journals</td>
<td>Pilots with images, audio, MLibrary working system for born-digital</td>
</tr>
<tr>
<td>Development of data mining tools</td>
<td>Plans for</td>
</tr>
<tr>
<td></td>
<td>1. Data distribution</td>
</tr>
<tr>
<td></td>
<td>2. SEASR integration</td>
</tr>
<tr>
<td></td>
<td>3. Research Center</td>
</tr>
<tr>
<td></td>
<td>Data Distribution Research Center (July 2011)</td>
</tr>
</tbody>
</table>
## Goals

<table>
<thead>
<tr>
<th>Then</th>
<th>Now</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reliable and increasingly comprehensive digital archive</td>
<td>~2 million (MLibrary, Wisc)</td>
</tr>
<tr>
<td>Co-owned</td>
<td>24 partners</td>
</tr>
<tr>
<td>Dramatically improve access...first and foremost meet needs of partners</td>
<td>See above</td>
</tr>
<tr>
<td>Preserve materials</td>
<td>Digital materials</td>
</tr>
<tr>
<td>Coordinated print storage</td>
<td>No</td>
</tr>
<tr>
<td>Create and sustain “public good”... mitigate free riders</td>
<td>-</td>
</tr>
<tr>
<td>Technical framework...centralized...open</td>
<td>Modular infrastructure</td>
</tr>
</tbody>
</table>
More

<table>
<thead>
<tr>
<th></th>
<th>Then</th>
<th>Now</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bibliographic Data Management</td>
<td>Yes</td>
<td>New system under development by CDL</td>
</tr>
<tr>
<td>Rights determination</td>
<td>Yes bibliographically- and manually-determined</td>
<td>CRMS</td>
</tr>
<tr>
<td>Holdings database</td>
<td>No</td>
<td>Structure in place, gathering data</td>
</tr>
</tbody>
</table>
Constitutional Convention (1)

• 7 ballot initiatives
• 5 passed
  – Print monograph storage
    • To establish a print monograph archiving program
  – Approval Process for development initiatives
    • To invite, evaluate, rank, launch development initiatives
  – Governance
    • Establishes a 12-member Board of Governors
  – U.S. Government Documents
    • Coordinated and collective action to expand and enhance access to U.S. federal publications
  – Fee for service content deposit
What’s next?

• What problems?
  – Identification
  – Description
  – Rights
  – Preservation
  – Discovery and use
Approach

- Collective problems as collective
- Web of relationships
Bibliographic Data

- Normalization of bibliographic data
  - University of Michigan
    - Efficiency
  - California Digital Library
Copyright Review

• IMLS Grant awarded to University of Michigan 2008 to determine copyright status of books published in US between 1923 and 1963
  – 18 staff members, 4 institutions
    • Indiana University
    • University of Michigan
    • University of Minnesota
    • University of Wisconsin
  – 170k reviewed through CRMS (as of November 2011)
  – 87,000 (51%) in public domain

• Second grant to investigate non-U.S. works
  – 15 partner institutions involved
Breakdown of HathiTrust book corpus by publication date
Breakdown of HathiTrust book corpus by publication date
Copyright status of books published pre-1923 and US works published 1923-1963
Copyright status of books published pre-1923 and US works published 1923-1963

- **Pre-1923**: 100% Public Domain
- **1923-1963**
  - 1923-1963: 55% Public Domain, 35% Orphans, 10% In Copyright
- **1964-1977**
  - 1964-1977: 20% Public Domain, 35% Orphans, 40% In Copyright
- **1978-**
  - 1978-: 40% Public Domain, 60% In Copyright
Holdings Database

• Database will
  – Serve as basis for new pricing model
  – Support expansion of legal uses of materials: preservation uses, access for users who have print disabilities, access to orphan works
  – Facilitate individual and collaborative collection development and management operations
  – Will also benefit efforts in de-duplication
Print Holdings Database

• Volumes institutions own or **have owned**
  – For monographic holdings
    – Only print volumes (not microform, etc.)
    – OCLC number [required]
    – Bib record ID [required]
    – Enumeration/chronology, if available
    – Condition (e.g., brittle) [optional]
    – Holding Status (e.g., current holding, withdrawn, missing, etc.) [optional]
  – For serial holdings
    – OCLC number [required]
    – Bib record ID [required]
    – ISSN, if available
Columbia University

Number of volumes vs. Number of holding institutions

Columbia University - Serials

Number of titles vs. Number of holding institutions
Preservation Infrastructure

- Digital and print materials
- Definitional elements
- Relationships
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%

Rank in 2008 ARL Investment Index
Digitized Books in Shared Repositories

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

~3.5M titles

~2.5M

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
Discovery and Use

• Ability to find materials
• Situating HathiTrust holdings in broader landscape, working with OCLC
• APIs
• Assembling corpus for computational research
How does work get done?

• Collective work
  – e.g., working groups
  – Perform the work of the partnership
  – Now 40+ people across partner institutions

• Distributed work
  – Driven by needs of institutions – able to leverage across the partnership
  – Projects, e.g. grant work, ingest specifications, page-turner, bibliographic data management

• Leverage expertise across institutions
Emerging Governance

• Elections Committee (January 1)
• Nominations
  – Elections Committee select 12 (for 6 seats)
• Voting (March 1 – March 15)
• 6 seats to founding institutions
  – 2 California, 2 CIC (minus Indiana and Michigan)
  – 1 Indiana, 1 Michigan
• Begin work April 15, 2012
Work going forward

• Definitional elements
  – Identification
  – Description
  – Rights
  – Holdings
• Print archiving, management
• Government documents
• Discovery and use
  – Lawful uses
• Quality
• Research Center
• 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

• Twitter:   http://twitter.com/hathitrust

• Monthly newsletter:  
  http://www.hathitrust.org/updates
  • RSS: http://www.hathitrust.org/updates_rss

• Contact us: feedback@issues.hathitrust.org

• Soon: Facebook, blog
Thank you very much!