Building the Research Library at Web Scale
Heather Christenson, California Digital Library
The Rotunda of Bacon Library, c. 1890. Joseph Rowell at desk at far left. Rowell's flat-topped desk was under the dome. His author catalog was in a card cabinet on one corner of the desk, and his index on the other. He served as the loan and reference desk librarian, as well as the ordering and cataloging departments. He mixed his own paste and mounted bookplates and date slips, and he affixed the call numbers to each volume. In a large ledger book, he entered new volumes as they came in.

http://www.lib.berkeley.edu/give/historyroom/panel8.html

Photo: Library, University of California, Berkeley
10 campuses

100 libraries

California Digital Library

38,000,000 print volumes

3,740,672 digitized volumes

“Today's undergraduates--most born between 1986 and 1990--have come of age with the Internet, Google, cell phones, e-mail, RSS feeds, blogs, podcasts, file sharing, text messaging, laptops, PDAs, and wireless access. And university libraries must adapt to meet the demands of this new, digital age--and the expectations of a new, hyperconnected user.”

"When I became a librarian, you pretty much had to go to a library to find out what they had," says UCSC University Librarian Virginia Steel. "Now, with resources like Google and other online services, you can just link to the contents of a library from anywhere."

http://news.ucsc.edu/2008/10/2474.html

Photo: Phil Carter, University of California, Santa Cruz
UC University Librarians’ **Priority**

Capitalize on technological opportunities to accelerate the transition to a primarily digital environment

---

University of California Libraries, Systemwide Plan and Priorities, FY 2013-2016
http://libraries.universityofcalifornia.edu/about/vision-and-priorities

---

Photo: Library, University of California, Berkeley
California Digital Library programs support UC’s research & scholarship lifecycle

- **Access & Publishing**
  - Online Archive of California
  - Calisphere
  - eScholarship

- **Discovery & Delivery**
  - Shared discovery platforms
  - Resource sharing

- **UC Curation Center**
  - Web Archiving Service
  - Data management
  - Digital preservation

- **Collections**
  - Licensed Content
  - Shared Print
  - Mass Digitization
Mass Digitization: Digital Reformatting on a Massive scale

- Improve discovery
- Enhance student and faculty research
- Fulfill our public service mission
- Preserve and protect our collections
- Support collection management
- “A place at the table”
- Carpe diem!
UC Libraries’ Mass Digitization Timeline

- **UC Internet Archive digitization**
  - Oct 2005
- **UC Google digitization**
  - Aug 2006
  - Oct 2008
- **UC in HathiTrust**

- OCA founding member
- Google library partner
- Microsoft digitization funding
- HathiTrust founding member
Internet Archive / Open Content Alliance Projects

- ~200,000 books digitized
- Out-of-copyright works only
- Primarily English language, some romance languages
- Foldouts are included
- Scanning was primarily done on-site at UC
- Funded by Microsoft, Yahoo, Sloan Foundation, CDL, others
- Requires library and/or external funding
What have we digitized with the Internet Archive?

~200,000 volumes

- Pre-1923 English language books from **Northern Regional Library Facility** and **Southern Regional Library Facility**
- Pre-1923 foreign language books at **SRLF**
- **UC Davis**: California documents
- **UC Berkeley**: cookbooks, mathematics, selected Bancroft collections
- **UCLA**: children’s books, Italian comedies, rare business and economics texts
UC Google Projects

- ~3.6 million books digitized to date
- In-copyright and public domain works (currently only public domain)
- All languages
- Foldouts are skipped
- Scanning done at a Google off-site facility
- **Funded almost entirely by Google**
What have we digitized?

~3.6 million volumes

- **UC Northern Regional Library Facility**: all subjects, all languages
- **UC Santa Cruz**: Humanities / Social Sciences
- **UC San Diego**: East Asian collection, IRPS, Scripps, Music, Science/Engineering
- **UCLA**: E. Asia library, Management, Chem / Biomed, Arts Library
- **UC San Francisco**: Medical library collections, special collections
- **UC Davis**: Main library collection, wine collections, special collections
- **UC Southern Regional Library Facility**: all subjects, all languages
The California Digital Library’s role

- “11th UC Library”
- Maintain & develop primary relationships with external partners
- Project management & coordination across UC campuses
- Technical leadership, data management, and stewardship
UC Libraries’ Strategic Priorities

- **Access**: Provide access to UC’s mass digitized collections. Bring together our collections.

- **Preservation**: Store as much of UC mass digitized content as possible in the HathiTrust repository. Contribute to a robust preservation environment to ensure safety of UC content.

- **Collection development and management**: Position the UC Libraries for collection decisions by via HathiTrust collection development and management activities.

- **Rights**: Deposit as much public domain content as possible into HathiTrust. Lawfully open as much of the content as possible to end users.
What is HathiTrust?

• A shared digital repository founded in October 2008.
• A partnership of over 90 research libraries. UC is a founding partner.
• Dual focus on preservation and access.
• Research center for computational research using text.
• Very large scale.
HathiTrust Digital Library

Currently Includes
10,922,113 total volumes
5,718,881 book titles
285,713 serial titles
3,822,739,550 pages
490 terabytes
129 miles
8,874 tons

3,563,589 volumes (~33% of total) in the public domain

Volumes contributed from 23 partners, including UC

Brings together outputs of major mass digitization initiatives
Language Distribution

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

- English: 48%
- German: 9%
- French: 7%
- Spanish: 5%
- Chinese: 4%
- Russian: 4%
- Japanese: 3%
- Arabic: 2%
- Italian: 3%
- Latin: 1%
- Remaining Languages: 14%
A visualization of HathiTrust works by call number (Library of Congress classification). View statistics information >>

<table>
<thead>
<tr>
<th>LC Classification</th>
<th>Count</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>A - General Works</td>
<td>336</td>
<td>0.01</td>
</tr>
<tr>
<td>AG - Collections. Series</td>
<td>8,269</td>
<td>0.22</td>
</tr>
<tr>
<td>AE - Encyclopedias</td>
<td>779</td>
<td>0.02</td>
</tr>
<tr>
<td>AG - Dictionaries and other reference</td>
<td>593</td>
<td>0.02</td>
</tr>
<tr>
<td>AI - Indexes</td>
<td>313</td>
<td>0.01</td>
</tr>
<tr>
<td>AM - Museums, Collectors and collecting</td>
<td>1,337</td>
<td>0.04</td>
</tr>
<tr>
<td>AN - Newspapers</td>
<td>10</td>
<td>0.00</td>
</tr>
<tr>
<td>AP - Periodicals</td>
<td>4,827</td>
<td>0.13</td>
</tr>
<tr>
<td>QE - Geology Count: 22838</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Collection Development & Management

HathiTrust increasing coverage of ARL libraries
More than 50% median overlap with ARL institutions; higher for small liberal arts colleges

Significance of the collection
“...HathiTrust, an organization not yet five years old, already holds the greatest concentration of titles on the topic of marine biology, surpassing the Library of Congress...”

--Constance Malpas, OCLC Research

http://hangingtogether.org/?p=2680
HathiTrust Services

- Long-term preservation
- Bibliographic search
- Full-text search
- Reading and download capabilities for public domain
- Print on demand for public domain
- Mobile interface
- APIs and Data feeds
- Collection building tools
- Search widgets
- Datasets, Research Center
Everyone

- View PD content
- Search PD and in-copyright materials
- View public collections
- Create private or public collections
- Download single PD pages of Google-digitized books
- Link to Find in A Library (OCLC)
- Link to reprints (from UC and UM services)

HathiTrust Partners

Everything on the left, plus:

- Download *entire volumes* of PD Google-digitized books
- *Download PDF and EPUB* of PD Google-digitized books *from mobile interface*
- Potential access to in-copyright for print disabled
- Potential Section 108 uses
- Have a voice in the future of HathiTrust
What are people doing with HathiTrust?

• Building collections of full-text works
• Integrating HathiTrust sources into library guides
• Reading books online at HathiTrust or downloading them whenever it's more convenient than a physical copy
• Linking to collections
• Requesting limited view items from ILL
• Downloading relevant pages
• Searching full-text and collections
• Restricting to full-view items
• Consulting useful volumes on a mobile device
• Using full-text search to find phrases
• Using quotation marks and booleans to conduct advanced searches

Source: http://www.hathitrust.org/personas
This copy is from The University of California Libraries collection.


This book has been digitally re-mastered to remove artifacts that occur during aging; faded colors are restored to clarity and vibrancy; pages are straightened and trimmed; blurred text is sharpened and images are enhanced and scaled. Even with the most sophisticated processing rare books may contain some permanent flaws so we recommend you view the online copy prior to purchase. Read this book online, or purchase a custom-printed paperback copy.

Read online
(free)

Buy BookPrep quality paperback
Approximately $11.99 (you will be redirected to Amazon.com)

• Partner with HP; sell via Amazon

Over 200,000 public domain books
HathiTrust Print Disabilities Access

• “Through a designated Proxy, Eligible Patrons at HathiTrust Supporting Institutions can receive special access to in-copyright materials in HathiTrust. The materials must be held currently, or have been held previously, by the Supporting Institution’s library, as indicated through print holdings information submitted to HathiTrust. “

• https://docs.google.com/document/d/1eU8n0lfBFdyxNhtDWTCqzJt0ucnA9karp-tyiQDuawA/edit?pli=1
HathiTrust Research Center

Bring researchers to the data
• Starting with public domain materials
• Support “non-consumptive” research
• Evolve around user demand

Elements
• Portal: browse volume lists and algorithms, execute algorithms, view results
• Catalog (Blacklight): assemble worksets
• Sandbox: can run own algorithms

Listserv: http://bit.ly/12KB3t7
How is HathiTrust working to make more content available to users?

- Grant-funded copyright review projects
- Rights holders opening up works
- Bringing in public domain content from partners
Some CDL Contributions

- Development of standards and process to transfer digitized volumes into HathiTrust from both Google & the Internet Archive
- Development for HathiTrust “Pageturner” interface
- Links in UC catalog Melvyl
- Link resolver “target”
- Search widget
- Zephir: new HathiTrust Metadata Management System
Significance of HathiTrust

- Continues and extends the historical role of libraries
- Services not driven by commercial purposes
- Draws upon contributions and expertise from across the partnership
- Standardization enables programmatic activities
- Scale enables savings
- Collaboration empowers the group as a whole
What we’ve learned about collaboration

Challenges

• Expanding the boundaries of existing work
• Geography & travel constraints
• Communication & gaining a common understanding

What has worked

• Starting with what we have in common, which is a lot!
• Tapping front-line experts; minds of the best and brightest across institutions
• In the process of collaboration, participants forced to solidify their own institutional goals
Google Books Litigation Family Tree

2004
- Google announces Library Project: scan and snippet display

2005
- 5 Publishers sue Google
- Authors Guild (AG) sues Google in class action

2008
- Parties agree to global class action settlement with publisher sub-class and author sub-class; settlement allows preview, consumer purchase of full text, institutional subscription
- Settlement authorizes Research Corpus, which is included in HathiTrust (HT)

2009
- After opposition from Justice Department and foreign rights holders, Parties agreed to Amended Settlement Agreement, which eliminates most foreign books
- American Society of Media Photographers sues Google (ongoing)

2010
- Judge Chin holds fairness hearing

2011
- Judge Chin rejects settlement because of opposition of class members
- AG sues HT

May 2012
- Judge Chin certifies author class and finds associational standing
- Judge Baer finds that HT is fair use

October 2012
- 5 Publishers settle with Google

June 2013
- Second Circuit reverses class certification
- AG appeal pending in Second Circuit

November 2013
- Judge Chin rules that Library Project is a fair use
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
What is the relationship between HathiTrust and the DPLA?
HathiTrust Going Forward

- Transitioning to next phase of partnership
- Print archiving, management
- Research Center
- Quality certification
- Government documents
- Continue to expand lawful uses
- Content continues to flow in

Mass Digitization Continues!
Top three UC books!

Last year (Jan – Dec 2013), the top three most viewed titles from UC collections in HathiTrust were:

Les Misérables / by Victor Hugo; translated from the French by Isabel F. Hapgood. [c1887]

Secrets from Lupe's kitchen: dishes of Aztec land, Stoker, Catharine Ulmer (1938)

The curriculum, by Franklin Bobbitt [c1918]
For more information

UC Mass Digitization Projects
http://www.cdlib.org/services/collections/massdig/

HathiTrust
www.hathitrust.org

HathiTrust Research Center
www.hathitrust.org/htrc/

UC Library Reprints Service
http://reprints.universityofcalifornia.edu/
Questions?
Heather Christenson, Mgr. Digital Content Development & Strategy
California Digital Library
heather.christenson@ucop.edu

Thank you!