

HathiTrust and Collective Stewardship at Scale

Washington Research Library Council
March 10, 2015
Mike Furlough
Executive Director, HathiTrust

Mission

To contribute to research, scholarship, and the common good by collaboratively collecting, organizing, preserving, communicating, and sharing the record of human knowledge.

...building comprehensive collections and infrastructure co-owned and managed by partners.

...infrastructure for digital content of value to scholars and researchers

...enabling access by users with print disabilities.

...supporting research with the collections.

...stimulating shared collection storage strategies among libraries.

Timeline: Highlights

- Google Library Project announced (2004)
- Launch (2008)
- TRAC certification (2011)
- Constitutional convention (2011)
- 10 million volumes (2012)
- New governance established (2012)
- Current bylaws and fee structure (2013)
- 13 million volumes (2014)



Today's Conversation

- HathiTrust Today
 - Collections
 - Organization
 - Current Initiatives
 - Short term plans
- HathiTrust Tomorrow
 - How has the world changed?
 - How should we change it?



Collections



Preservation with Access

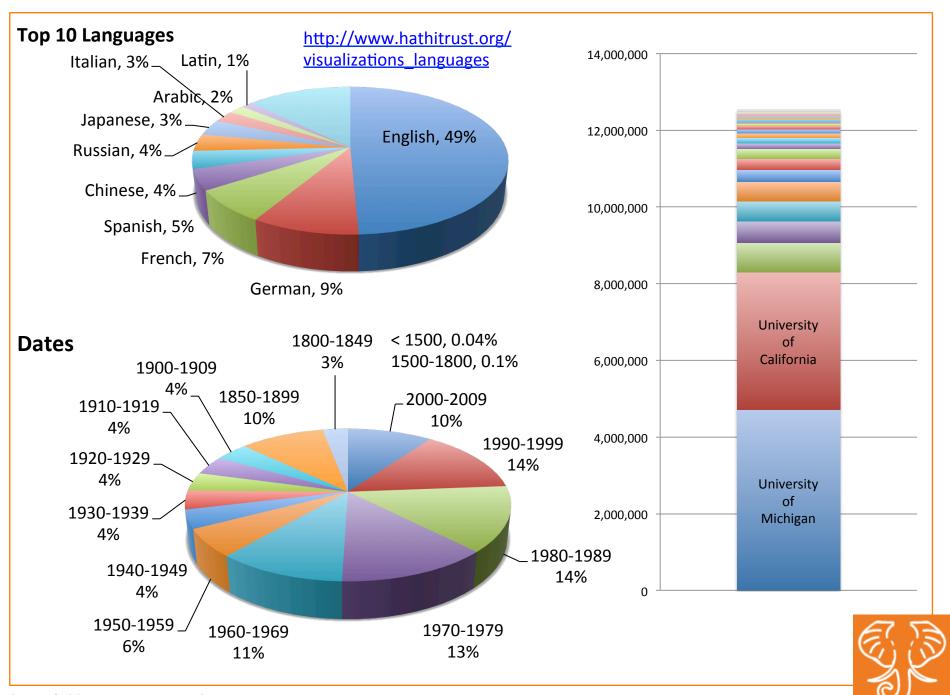
- Preservation
 - TRAC-certified
 - Long-term commitments on digital content facilitate planning, decision-making
- Discovery
 - Bibliographic and full-text search of all materials
 - Mechanisms for local loading of records
- Access and Use
 - Full text search (all users)
 - Public domain and open access works (all users)
 - Print on demand (all users, selected works)
 - Collections and APIs (all users)
 - Lawful uses of in-copyright works (members)



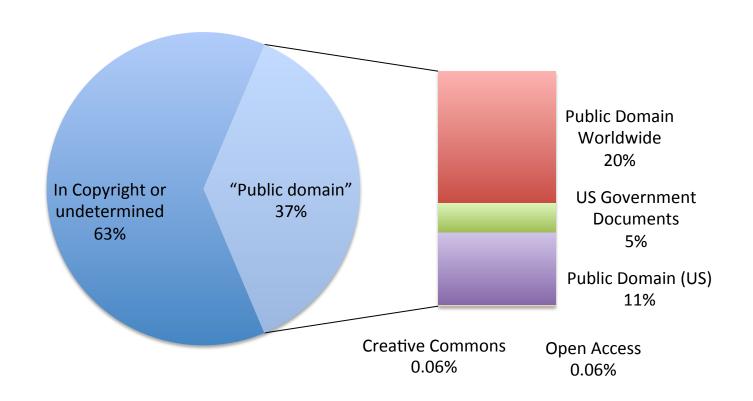
HathiTrust in 2015

- 13.2 million total items
 - 6.7 million book titles
 - 352,000 serial titles
 - 604,000 US federal government documents
 - 4.98 million items open (public domain & CC-licenses)
 - A handful of images and thimbleful of audio files





Copyright Distribution





Type of work	Searchable (bibliographic and full-text)	Viewable*	Full-PDF download	Print on Demand	Print disabilities*	Preservation uses (Section 108)*
Public domain worldwide	Worldwide	Worldwide	Partners only if 3 rd -party restrictions, if not, worldwide.	Worldwide	Worldwide	N/A
Public domain (US) – Non-US works published between 1873 and 1923.	Worldwide	When accessed from with the United States	Partners in the US if 3 rd party restrictions, if not, anyone in the US	Available within the United States	Partners in the US; partners worldwide where laws permit	N/A
Works that rights holders have opened access to in HathiTrust	Worldwide	Worldwide	Worldwide (if digitized by Google, full-PDF only available if opened with CC license)	Worldwide with permission	Worldwide	N/A
Works that are in-copyright or of undetermined status	Worldwide	Not available	Not available	Not available	Partners in the US; partners worldwide where laws permit	Partners in the US; partner worldwide where laws permit

^{*} Note: Access to in-copyright works is subject to conditions listed in HathiTrust's policies on <u>Access and Use</u>.



Access: Lawful uses of in-copyright works

- Sensitive to multiple legal regimes
 - Full-text search (everyone everywhere)
 - Access to users who have print disabilities (through member proxy in US, and where law permits)**
 - Access works that are damaged or missing and also out of print and unavailable (members in US only)

**Terms and conditions at http://www.hathitrust.org/access use#ic-access



Collective Action: Copyright Review

- Copyright Review Management System
 - Systematic manual review of copyright registrations to determine status of portions of the HathiTrust Collection
 - CRMS US: Published in US, 1923-1963
 - 318,887 reviewed / 168,248 PD (~53%)
 - CRMS-World: Published in UK (1874-1944),
 Canada, Australia (1894-1964)
 - 175,681 reviewed / 92,919 PD-world 9 (~53%)

Supported generously by IMLS







Library Catalog What happened to the Library Catalog? Tell us what you think of the Library Catalog Login nella larsen guicksand Keyword \$ Local Catalog Only \$ Find Feedback Advanced Search | Classic Search | Course Reserves | E-Reserves | Search History Showing 1 - 20 of 33 Results for nella larsen quicksand Sort Relevance Results for Library: "University of Illinois at Urbana-Champaign" [show all Libraries] Add to favorites Quicksand by Larsen, Nella Published 1928 **Narrow Your Search** HathiTrust eResources (public domain) Get it online **Format** Book (33) Electronic (2) Location Main Stacks Add to favorites Passing novels in the Harlem Renaissance: identity Literatures and Languages politics and textual strategies / Oak Street Facility by Gallego Durán, María del Mar Published c2003 Call Number: 810.9 G135p Undergraduate Location: Oak Street Facility [request only] History, Philosophy and Available Newspaper University of Illinois at Urbana-Champaign more... Author Larsen, Nella (7)

10 March 2015 13

"The changing same": black women's literature,

by McDowell, Deborah E. Published c1995 Call Number: PS374.N4 M37 1995

University of Illinois at Urbana-Champaign

Location: Literatures & Languages

criticism, and theory /

Not Available

Add to favorites

Ahad, Badia Sahar (1) Ahlin, Lena (1)

Berg, Allison (1)

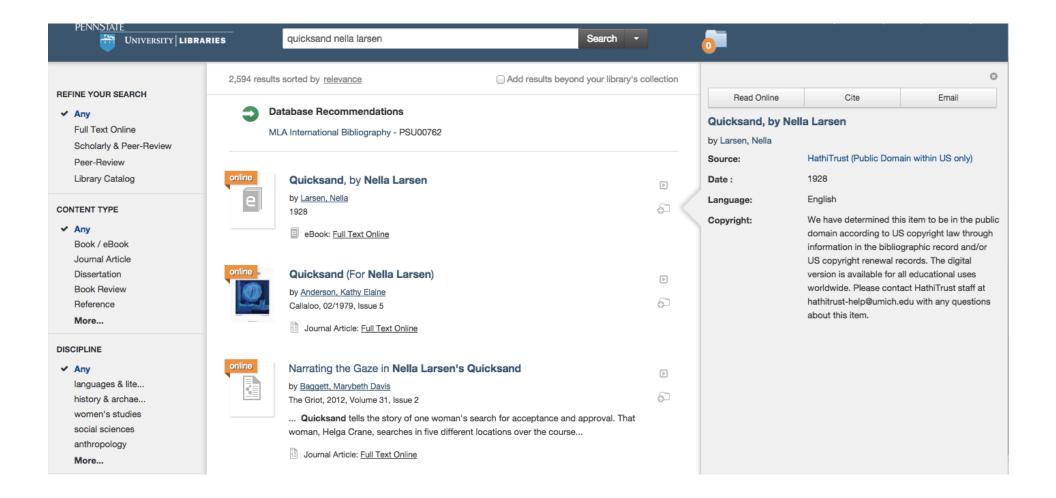
more...

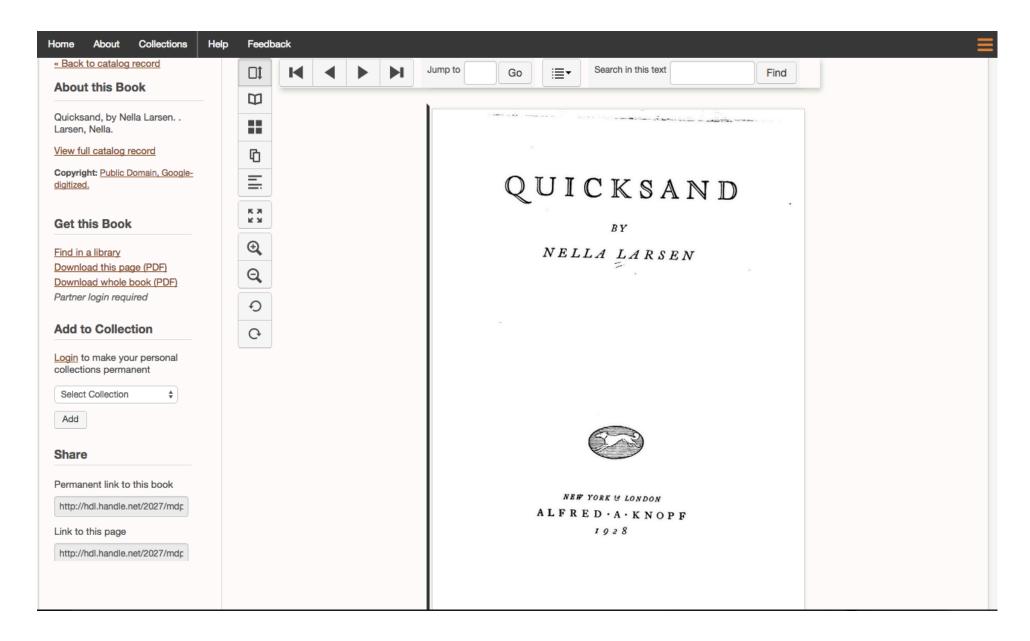
Topic

Calverton, V. F. (1)

African American authors (16)

Hitean /10\









Advanced catalog search Search tips

■ Full view only

Similar Items

Passing /

By: Larsen, Nella. Published: (2000)

Quicksand,

By: Larsen, Nella. Published: (1928)

Quicksand; and, Passing /

By: Larsen, Nella. Published: (1986)

Quicksand.

By: Larsen, Nella. Published: (1969)

The complete fiction of Nella Larsen

By: Larsen, Nella. Published: (2001)

Passing strangers,

By: Riesenberg, Felix, 1879-1939.

Published: (1932)

Passing by,

By: Baring, Maurice, 1874-1945. Published: (1928)

Napoleon passes

← Back to Search Results

← Previous record Showing record 5 of 9 Next record →

Cite this

Export to Endnote

LOG IN *



Main Author

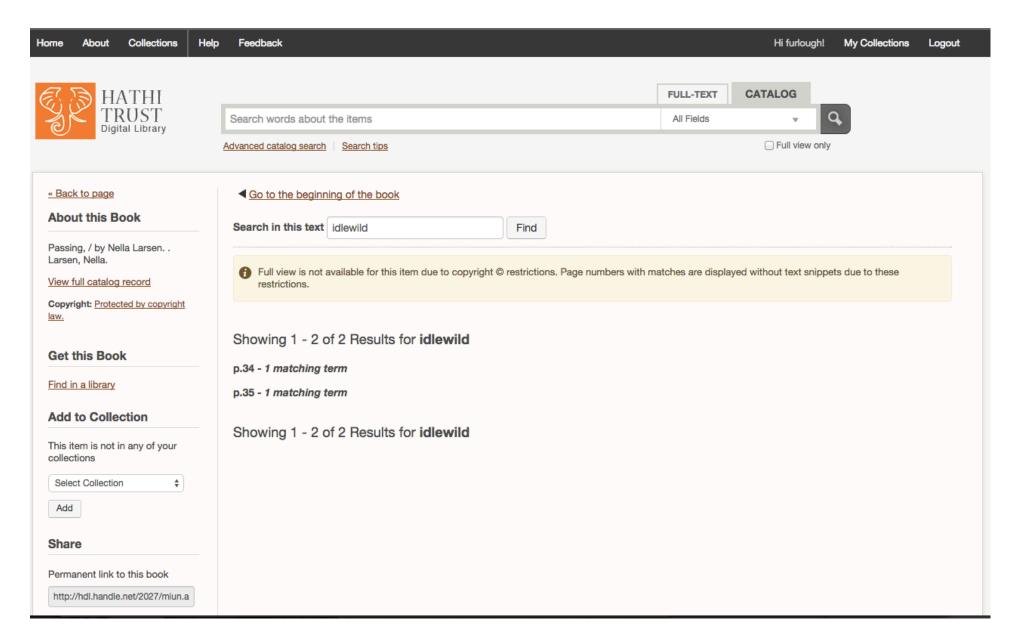
Passing, / by Nella Larsen.

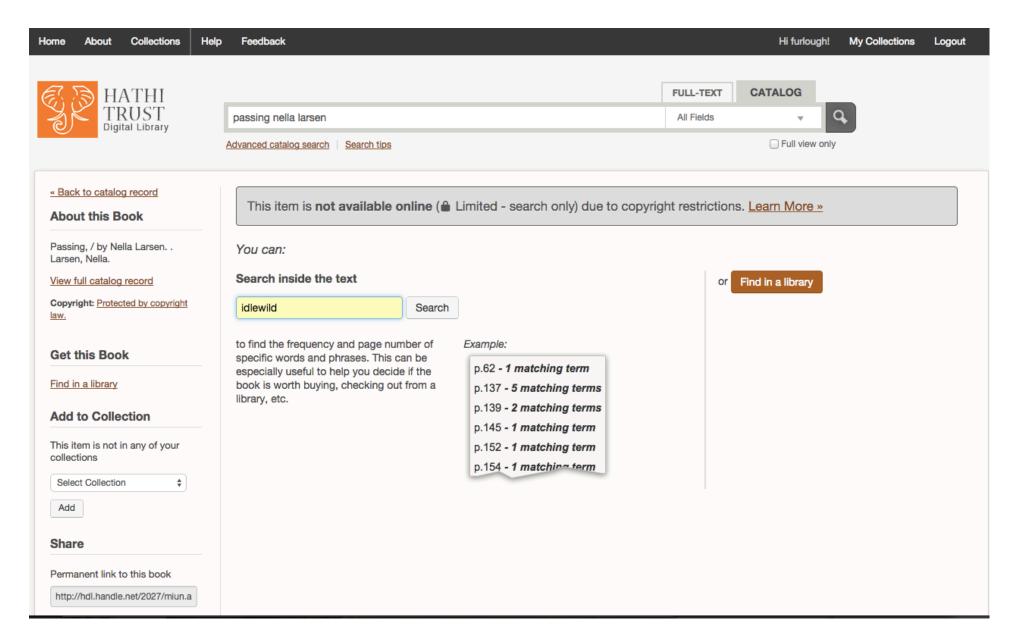
Maili Addioi.	Laisen, Nella.
Language(s):	English
Published:	New York, A. A. Knopf, 1929.
Physical Description:	6 p. L., 3-215, [1] p., 1 L. ; 20 cm.
Locate a Print Version:	Find in a library

Viewability:

<u> Limited (search only)</u> (original from University of Michigan)

Larcon Nolla





Top Ten Titles January 2015

- 1. The Human Figure, by John H. Vanderpoel
- 2. Quicksand, by Nella Larsen.
- 3. Godey's Magazine, v.40-41, 1850.
- 4. Pennsylvania German pioneers: A Publication of the Original Lists of Arrivals in the Port of Philadelphia from 1727 to 1808, by Ralph Beaver Strassburger.
- 5. The Book of a Hundred Hands, by George Brant Bridgman.
- 6. Indian boyhood, by Charles A. Eastman.
- 7. Roster of the Confederate soldiers of Georgia, 1861-1865, v.2.
- 8. Solid mensuration, by Willis F. Kern and James R. Bland.
- 9. The Five Laws of Library Science, by S. R. Ranganathan.
- 10. Roster of the Confederate soldiers of Georgia, 1861-1865, v.1.



Shared Stewardship



HathiTrust Members

Allegheny College American University of Beirut Arizona State University **Baylor University Boston College Boston University** Brandeis University **Brown University** Carnegie Mellon University Case Western Reserve Colby College Columbia University Cornell University Dartmouth College **Duke University Emory University** Florida State University Getty Research Institute Georgetown University Georgia Tech Harvard University Library Indiana University Iowa State University Johns Hopkins University Kansas State University Lafavette College Library of Congress Massachusetts Institute of Technology McGill University` Michigan State University Montana State University Mount Holyoke College New York Public Library **New York University** North Carolina Central University

North Carolina State University Northeastern University Northwestern University Oklahoma State University The Ohio State University The Pennsylvania State University Princeton University Purdue University Rutgers University Stanford University State University System of Florida Syracuse University Temple University Texas A&M 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 Berkelev Davis Irvine Los Angeles Merced Riverside San Diego San Francisco Santa Barbara Santa Cruz California Digital Library The University of Chicago University of Connecticut

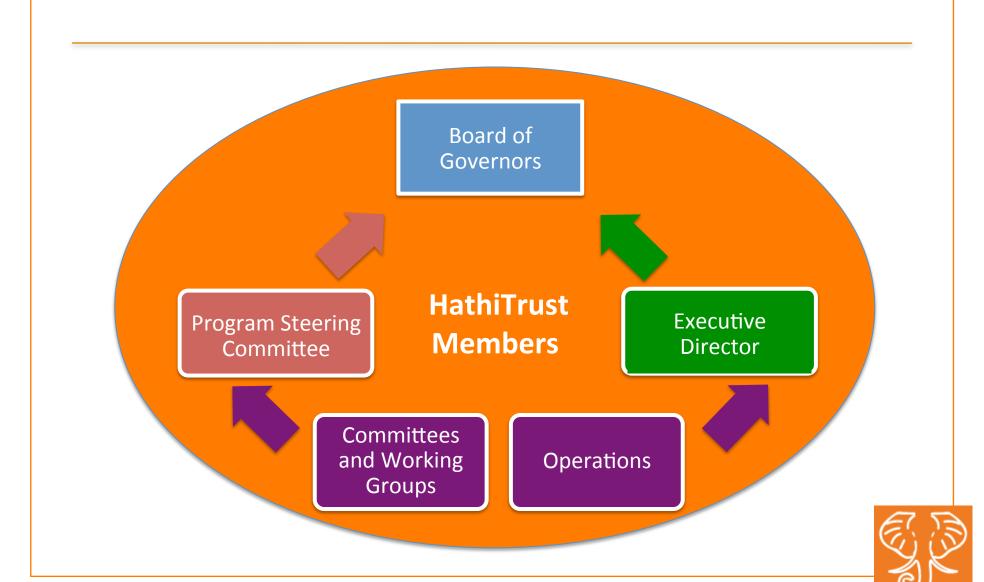
University of Delaware University of Houston University of Illinois University of Illinois at Chicago The University of Iowa University of Kansas University of Maine University of Maryland University of Massachusetts, Amherst University of Miami University of Michigan University of Minnesota University of Missouri University of Nebraska-Lincoln University of 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 Tennessee, Knoxville University of Texas University of Utah University of Vermont University of Virginia University of Washington University of Wisconsin-Madison Utah State University Vanderbilt University Virginia Tech Wake Forest University Washington University Yale University Library

Cooperative Work

- Draw upon knowledge across institutions
- Distributed Functions and Services
 - Preservation repository and access services
 - University of Michigan
 - Mirror site: Indiana University
 - Metadata management services
 - California Digital Library
 - HathiTrust Research Center
 - Indiana University and University of Illinois



Governance



Committees and Working Groups

- Program Steering Committee
- Collections Committee
- Zephir Advisory Group
- User Support Working Group
- Rights and Access Working Group
- Government Documents Initiative Planning and Advisory Group
- Print Monographs Archive Planning Task Force
- On Hiatus
 - Communications
 - User Experience



Requirements

- Non-profit libraries or non-profit institutions with libraries
- Partnership agreement
- Print holdings information
- Shibboleth

http://www.hathitrust.org/eligibility agreements
http://www.hathitrust.org/partnership checklist



Checklist

- 1. Partnership contract (Required)
- 2. Print holdings information (Required)
- Member Representative (Required)
- 4. Shibboleth (Required to receive authenticated partner services)
- 5. Primary contact (Required)
- 6. Press release
- 7. Institutional description
- 8. Outages contact
- 9. Communications contact
- 10. Google Analytics contact (optional)
- 11. Institutional IP addresses We are using IP ranges to partner name in HathiTrust interface



Financial Model

All partners share in operations and infrastructure costs for **public domain** volumes:

(Public Domain * Cost per volume * X) / Total Members

Share in operations and infrastructure costs for in copyright volumes based on holdings:

For a given in-copyright volume: (Cost per volume * X) / Holding Members

Cost per volume = ~\$0.167**

X = 1.5 (multiplier for programmatic funds)**

**2015 cost per volume and multiplier values



HathiTrust overall benefits to libraries

Digital Curation

- Drive costs down
- Reduce "bibliographic indeterminacy"
- Make meaningful decisions about formats and quality
- Increase discoverability, use
- Consolidate development talent
- Improve strength of archiving
- Print Curation
 - Means to associate our print holdings
 - Coordinated record-keeping
- Subsidiary benefits
 - Quantify problems
 - Collective attention to solving shared problems
 - Understanding relationship between collective and local



Current Initiatives



Shared Print Monographs Archive

- Ballot Initiative passed at the 2011 HT Constitutional Convention (Con-Con)
 - "To develop a print monographs archive corresponding to volumes represented within the HathiTrust"
- Focus
 - Ensure preservation of print and digital collections
 - Catalyze national/continental collective management of collections

Why A Shared Print Archive Program

- Creation of the digital corpus provides significant overlap with research collections
- Significant need and desire to reduce costs of collection management and associated footprint
- Many regional efforts, but limited national/ international coordination
- Strengthens preservation commitments
 - Connects both print and digital preservation



Task Force proposal...

- Defines the character of the repository as...
 - A collection that mirrors HT's monographic holdings, is distributed and "light"
 - A repository that is governed, managed, and supported by the HT as a whole, not a subset of members
 - A repository that is relatively lightweight and focused on lowering barriers for early participation



Task Force proposal...

- Defines the national role of the repository as...
 - Providing leadership in the area of monographic, print retention.
 - Supporting the development of the technical infrastructure necessary to disclose commitments and discover content.
 - Providing services to members that support their efforts to make local collection management decisions

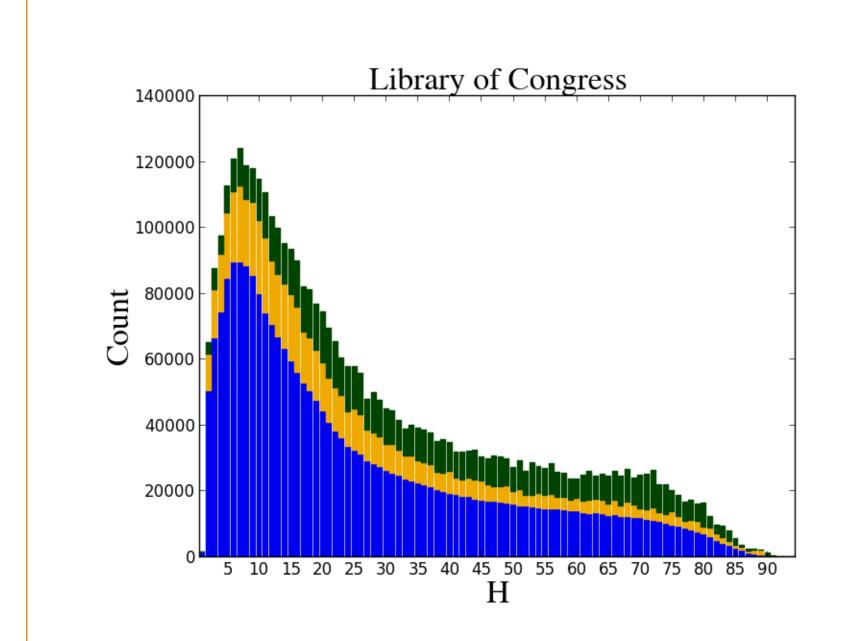


Selected WRLC Collection Overlap

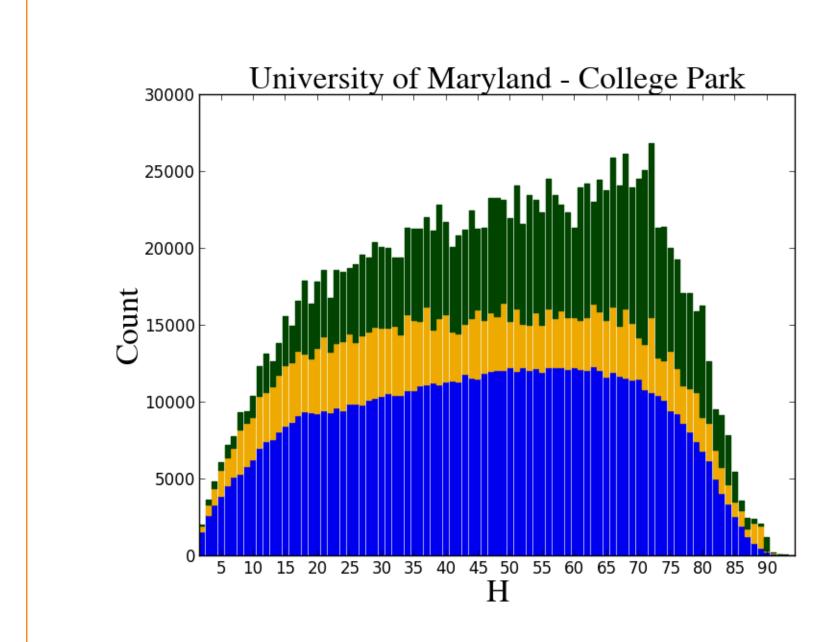
University	Holdings provided			OCLC#	HathiTrust Match (vols)	Matched Public Domain	Matched In Copyright
Howard	682,620	633,907	51.10%	326,537	757,852	24.04%	75.96%
GWU	961,506	831,939	45.84%	408,834	999,566	22.93%	77.07%
Georgetown	1,293,235	1,292,190	40.00%	516,830	1,153,606	21.25%	78.75%

Data is based on most recently provided holdings data, which may not be current.

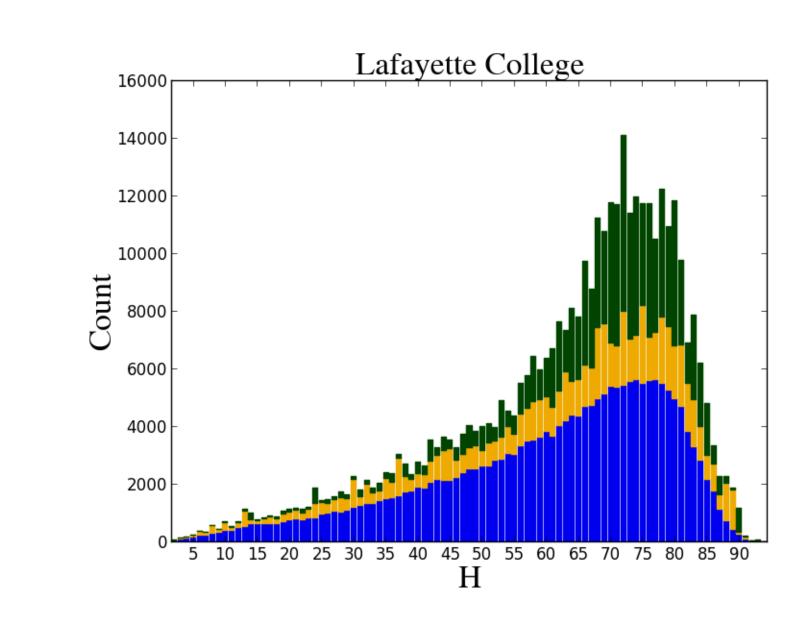




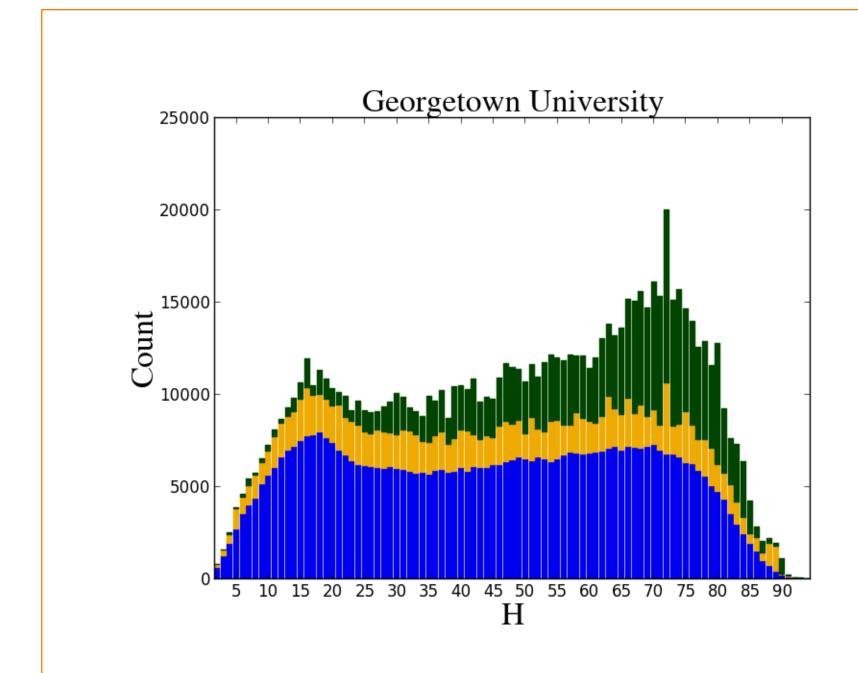




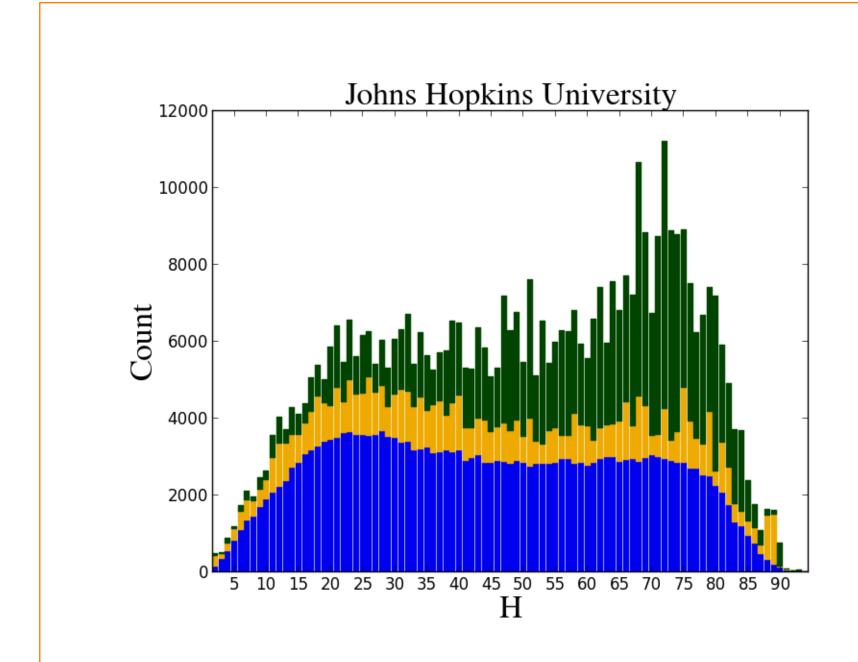




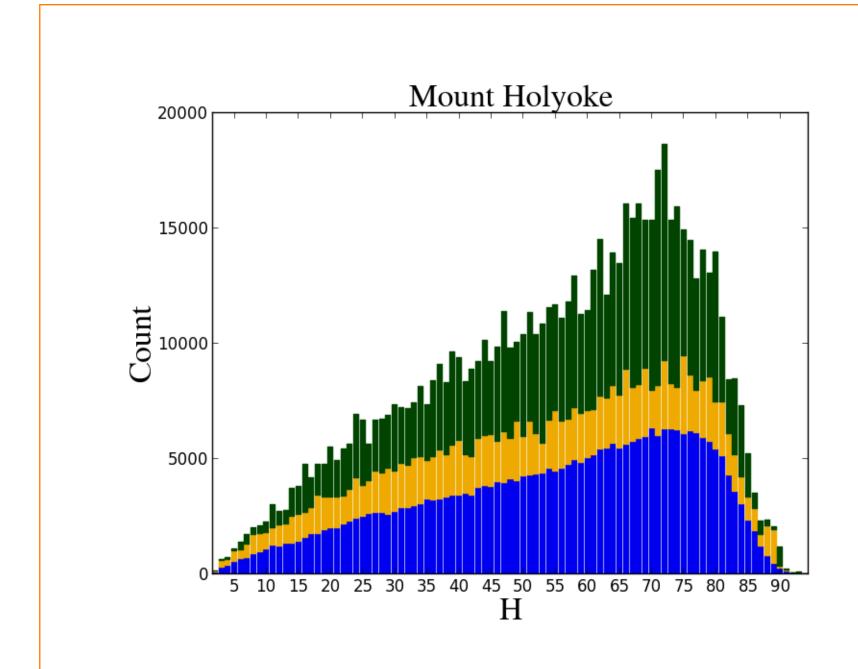












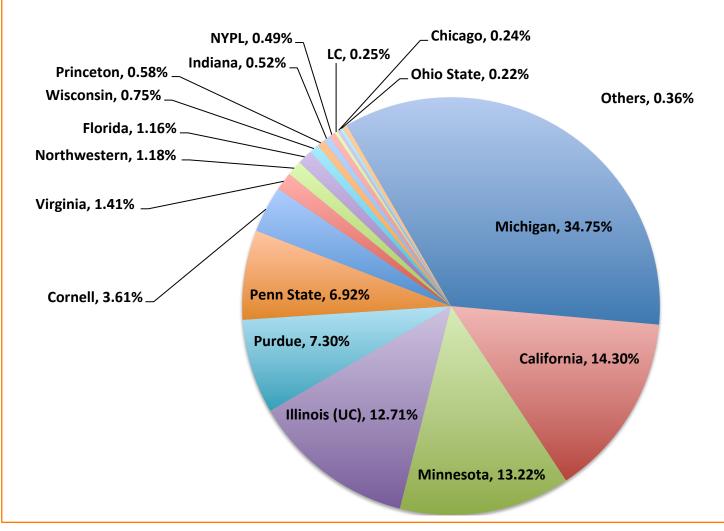


Government Documents Initiative

- Ballot Initiative: provide "expanded coverage & enhanced access to U.S. Government Documents."
- Activities:
 - Developing a registry of US Federal Government
 Documents
 - Locate materials for inclusion in the collections
 - Improve search and discovery

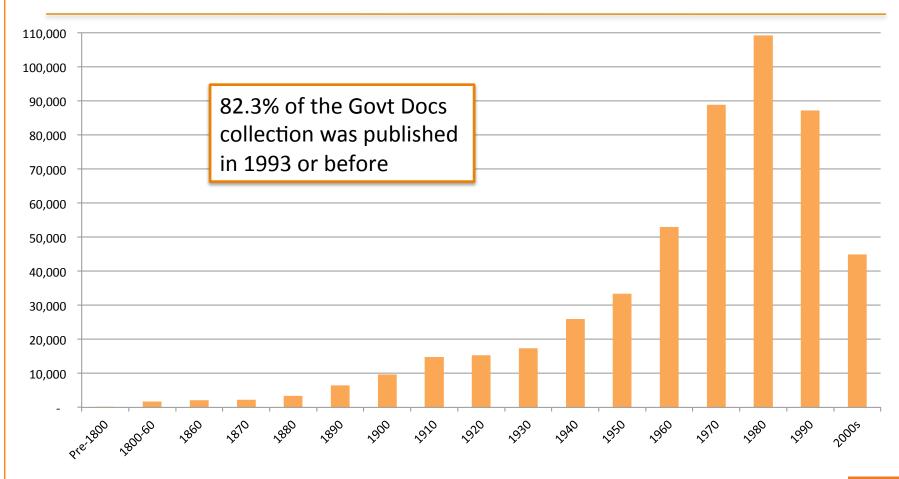


US Gov't Publications by Source Library





US Gov't Publications by Date





The Registry

- Goal: "....include metadata for the comprehensive corpus of U.S. federal documents. This will include materials produced at U.S. government expense, in all formats, at the item level, from 1789 to the present."
- Why?
 - Limited knowledge of this corpus.
 - Collection gap analysis
 - Digitization sourcing



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



Computational Access Initiatives

- HathiTrust distributes public domain datasets
- HathiTrust Research Center
 - Developed collaboratively by Indiana University and University of Illinois; launched July 2011
 - Funding from the Sloan Foundation, Andrew W.
 Mellon Foundation, and NEH Office of Digital Humanities.
 - Partially Funded by HathiTrust (2014-2018)

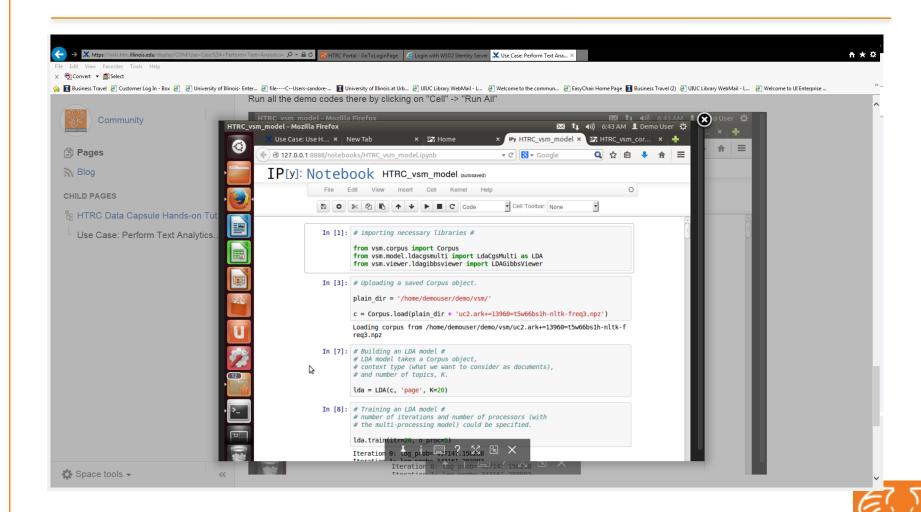




Goals for the Research Center

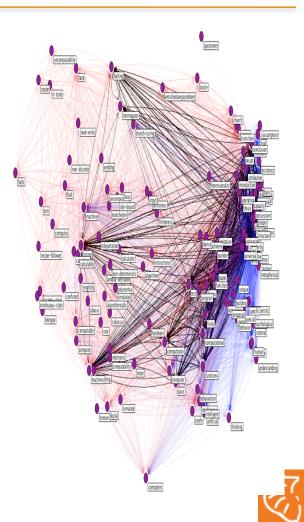
- Research arm of HathiTrust
- Provide a persistent and sustainable structure to enable original and cutting edge research.
 - Leverage data storage and computational infrastructure at Indiana & Illinois
 - Stimulate community development of new functionality and tools
 - Use tools to enable discoveries that would not be possible without the HTRC
- Enable scholars to fully utilize content of HathiTrust Library while preventing intellectual property misuse within U.S. copyright law.
 - Provision secure computational and data environment for scholars to perform research using HathiTrust Digital LibraryIndiana University and University of Illinois

HTRC DataCapsule: Secure Access



Scholarly Commons User Support Services

- Develop training materials
- Educational workshops
- Tool and workset support
- Collaborate with librarians and DH centers at HT institutions
- Assist researchers in HTRC text data mining research projects
- Collaboration: University
 Libraries, Illinois and Indiana



Advanced Collaborative Support Awards

- **Detecting Literary Plagiarisms: The Case of Oliver Goldsmith.**Douglas Duhaime. University of Notre Dame:developing tools for detecting plagiarisms...to detect the literary thefts of Goldsmith.
- Taxonomizing the Texts: Towards Cultural-Scale Models of Full Text. Colin Allen, Jaimie Murdock. Indiana University Bloomington. ...a cultural-scale investigation and topic modeling....random sampling to select collections according to the Library of Congress Subject Headings (LCSH).
- The Trace of Theory. Geoffrey Rockwell, Laura Mandell, Stefan Sinclair, Matthew Wilkens, Susan Brown. University of Alberta, Texas A&M University, University of Notre Dame. ...aim to subset theoretical subsets from the HT public corpus and apply large-scale topic modeling... develop tools and computational methods for tracking the concept of "theory".
- **Dr. Michelle Alexopolous**, University of Toronto...tracking technology diffusion through time using the HT corpus.



HTRC UnCamp

- MAR 30-31, 2015, Ann Arbor, MI
 - Workshops, speakers, demonstrations
- Keynotes

Michelle Alexopoulos, Professor, University of Toronto March 30, 2015, 8:45 to 9:45 am

Erez Lieberman Aiden, Assistant Professor, Baylor College of Medicine

March 31, 2015, 11:00 am to 12:00 pm

http://www.hathitrust.org/htrc_uncamp2015



Some Thoughts on the Present and Future



How are we positioned?

- Our mission, collection, and the repository operations are all strong.
- Our brand reputation is outstanding.
- Our work is solidly supported by the law.
- We have expanded access in unprecedented ways.
- The partnership provides a solid base for action.
- We have very important programs underway

Drivers to Set Immediate Repository Priorities

- Improved user experience
 - Backlog of requested/vetted enhancements
 - Regular upgrades, problem fixes, etc
- Supports current initiatives
 - Registry project
 - GPO partnership
 - Research Center data transfers
- Positions us to improve service for users with Print Disabilities
 - Two factor authentication
- Positions us to expand service portfolio and the universe of what we collect.
 - Support for additional text formats, e.g., PDF, Epub, TEI.

Some Pending Issues

- Metadata policy and strategy
- Quality metrics and assessment
- Additional content-types (non-text)?
- Methods to solicit and evaluate proposals for development
- Analytics services
- Translating HTRC research into operations.



What needs thought?

- Strategy, mission, and role in the future
 - (Inter)National digital infrastructure
 - Public policy
 - Membership growth
 - Collections program
 - Services portfolio
- 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
 - We should pursue complementarity and cooperation, not competition and duplication.
 - Scale will continue to drive our strategies
 - Potential partners are not just other libraries and library organizations, but also readers, authors, publishers.

How to find out more

- 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
 - RSS http://www.hathitrust.org/updates rss
- Contact us: <u>feedback@issues.hathitrust.org</u>
- Blogs: http://www.hathitrust.org/blogs
 - Large-scale Search
 - Perspectives from HathiTrust



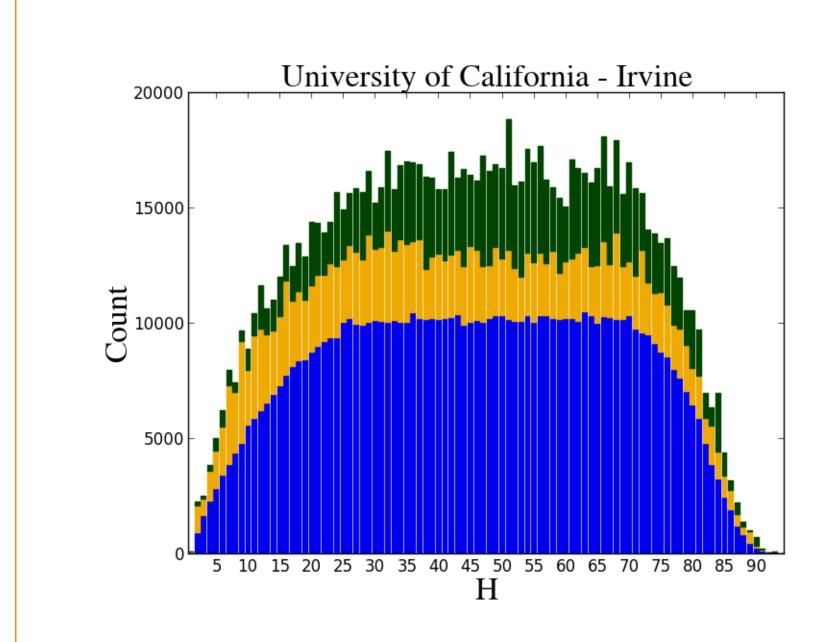
Thank you!

furlough@hathitrust.org @MikeFurlough

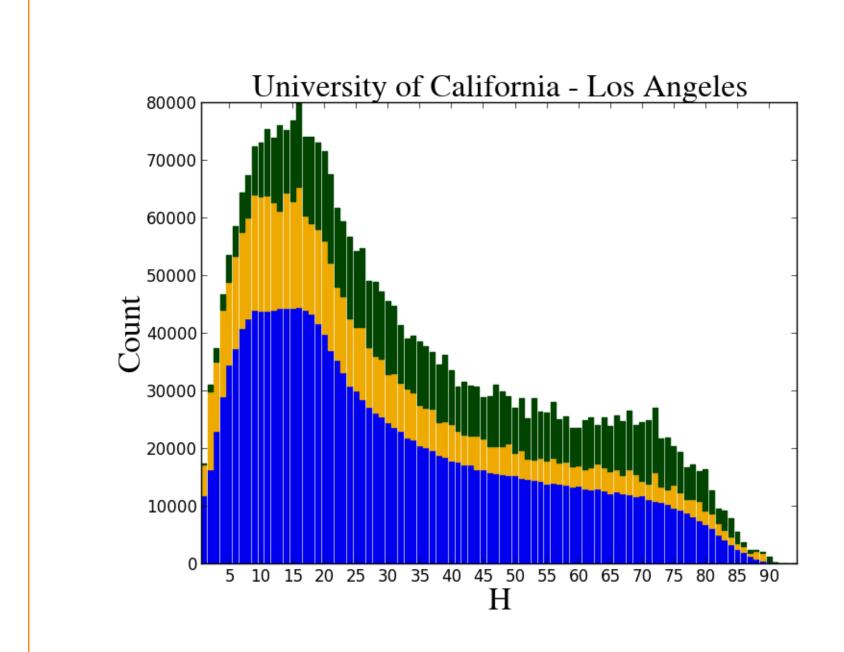


Partnership

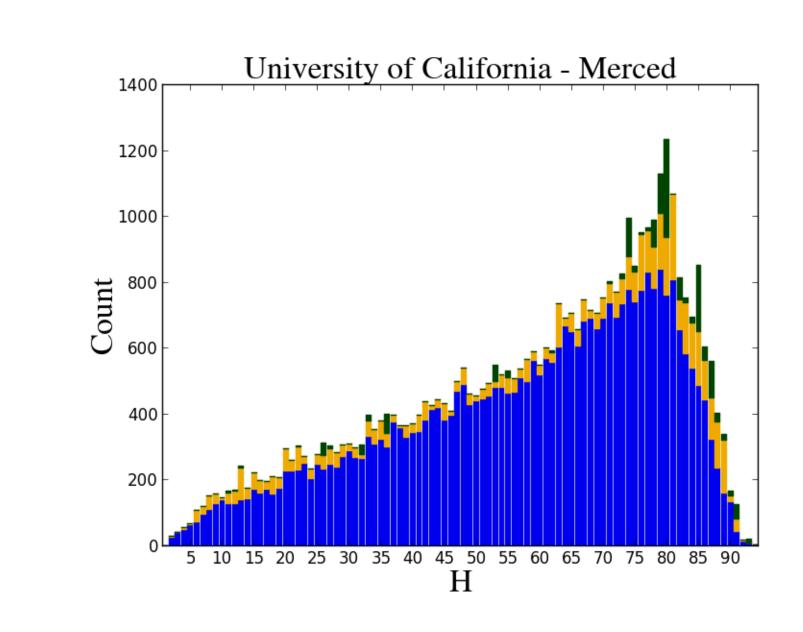




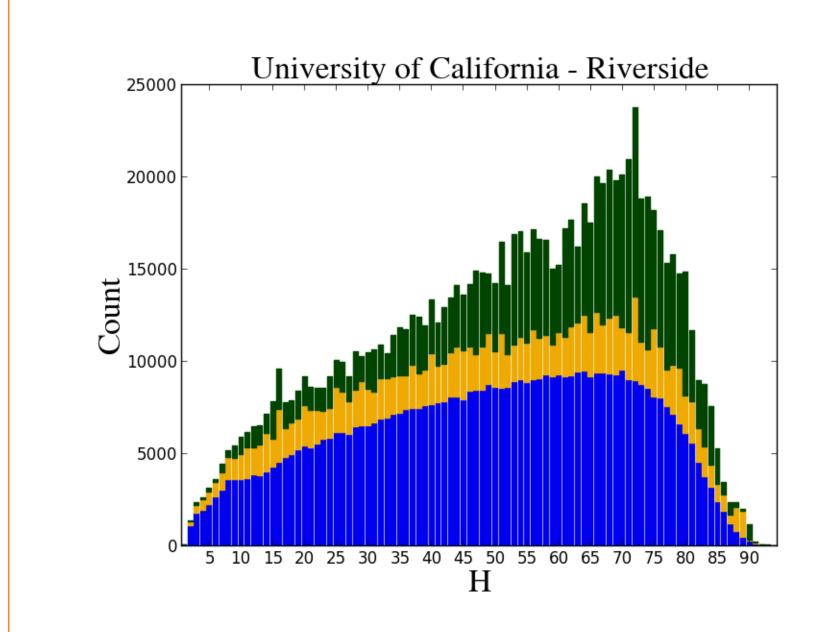




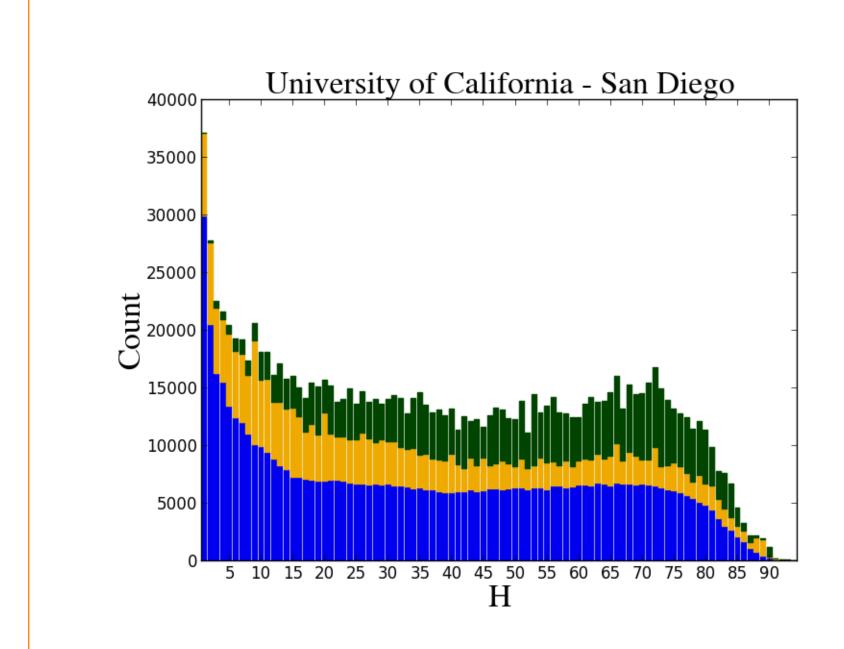




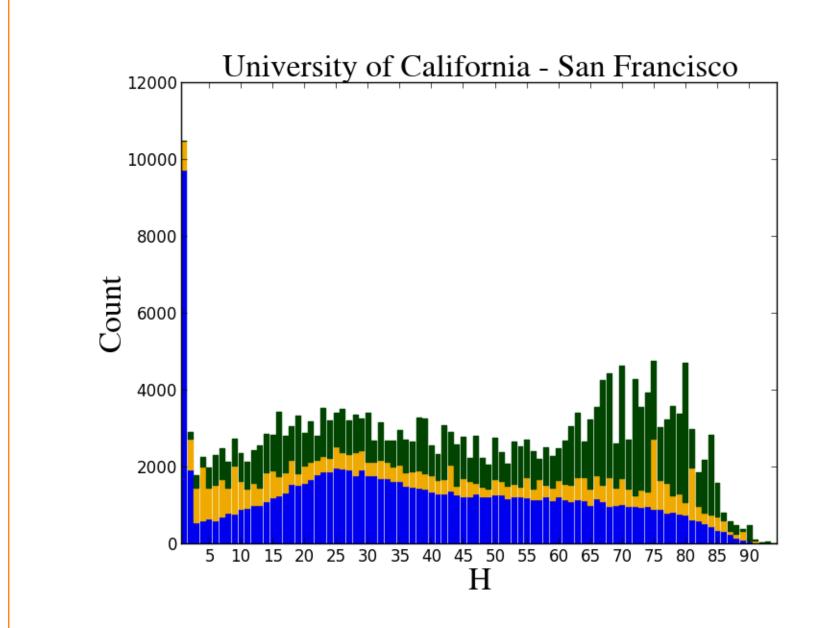




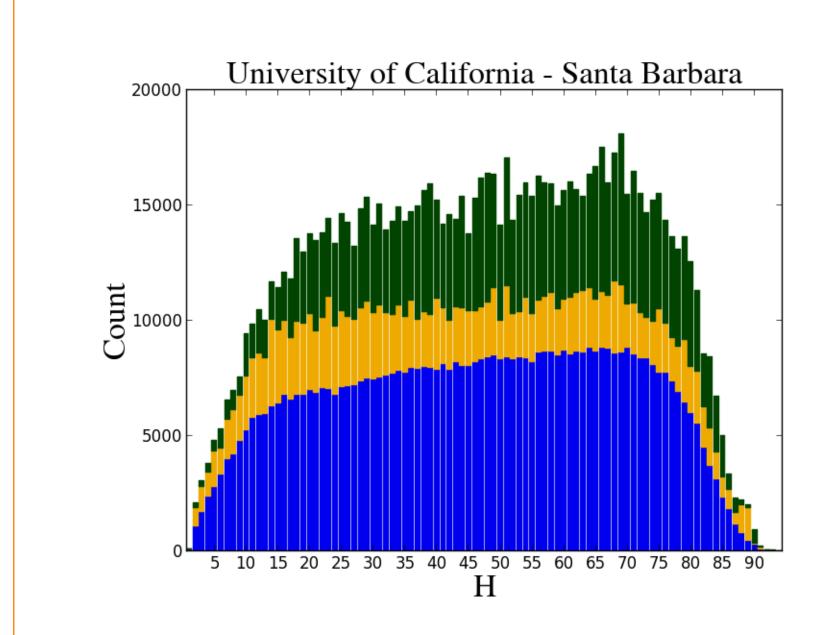




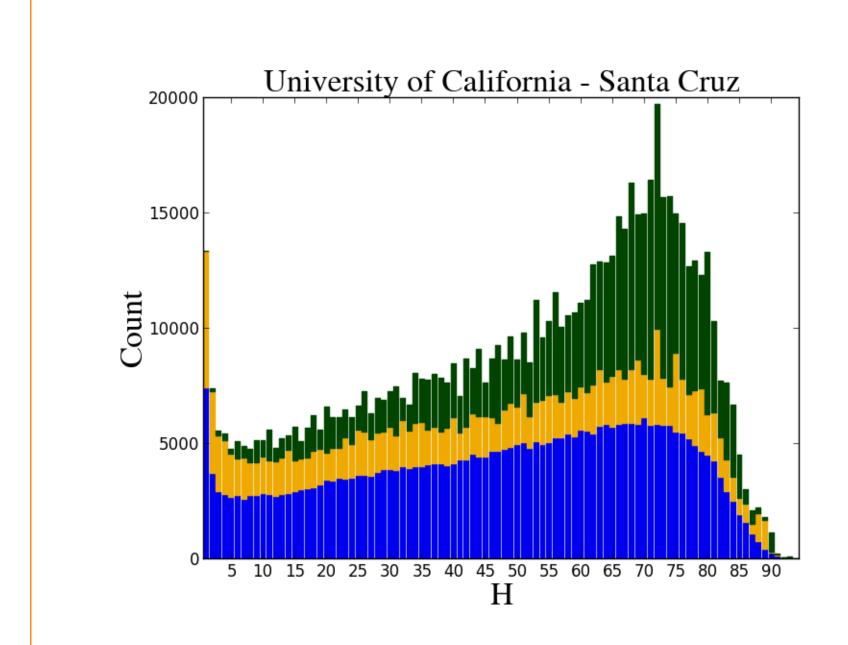




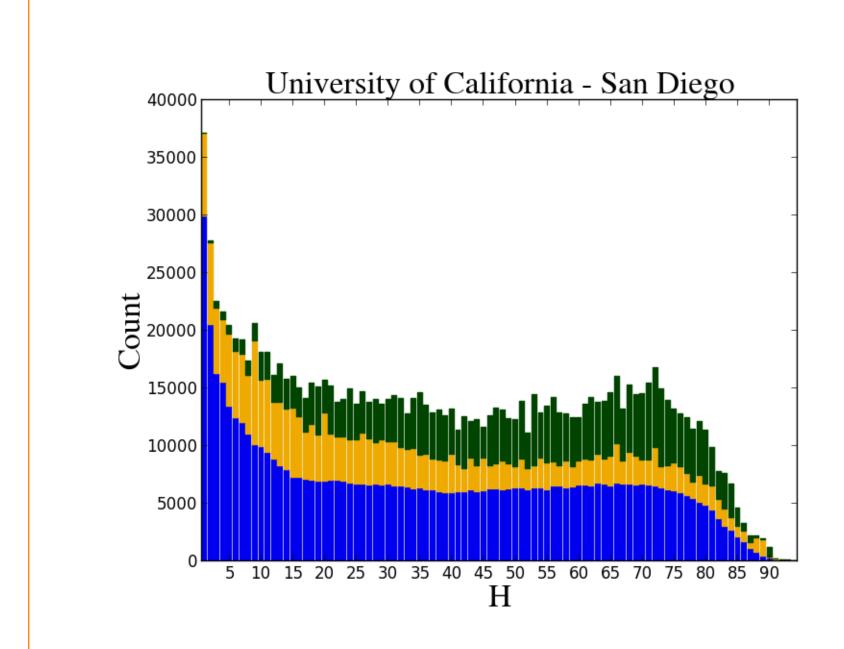




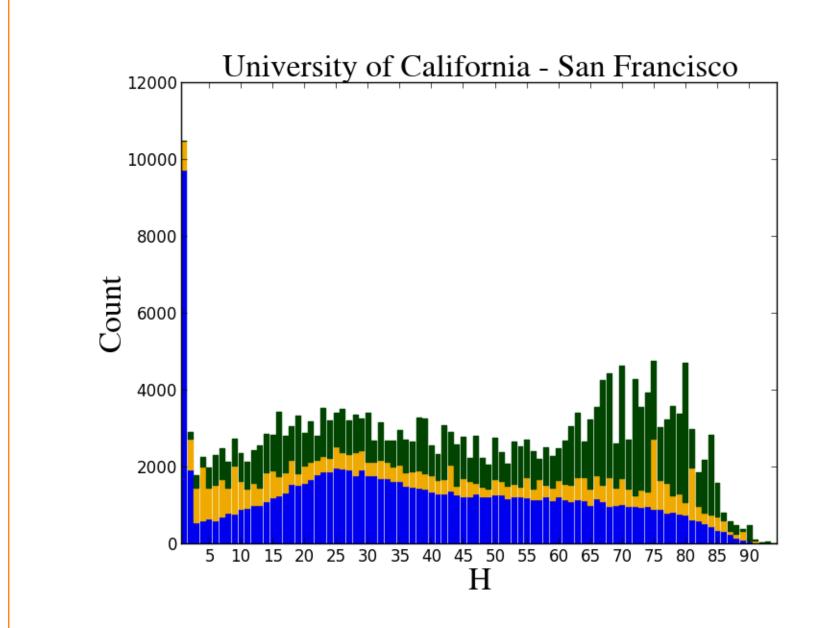




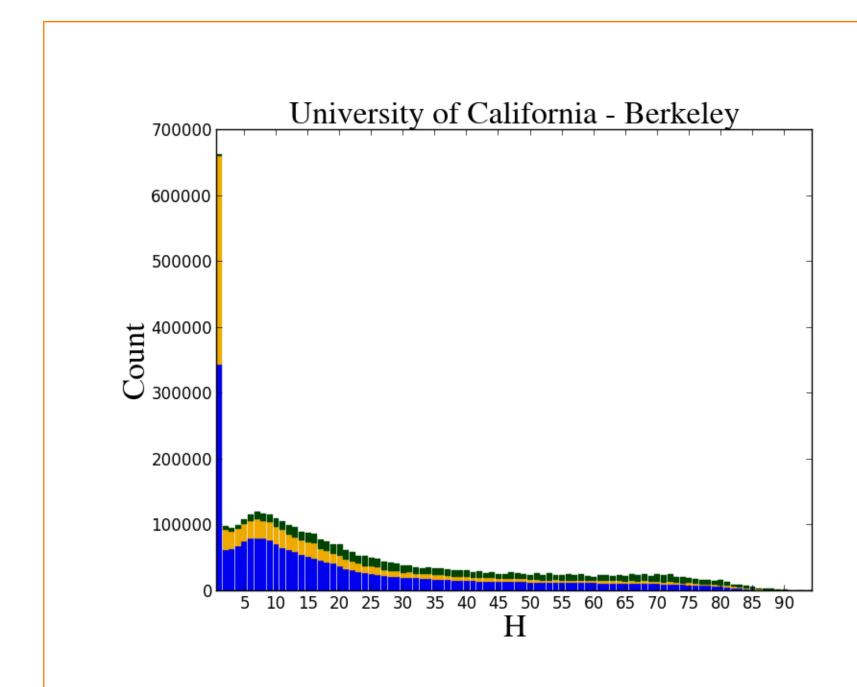




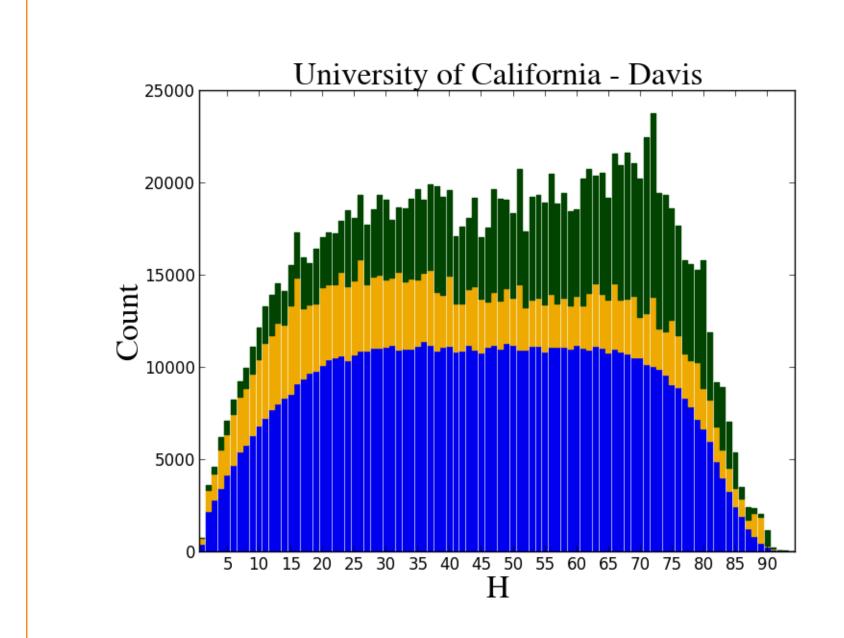














UC Collection Overlap (by titles)

	Submitted I	n Hathi I	Percent
single-part monographs	9,766,951	3,332,926	34.1%
multi-part monographs	985,087	467,318	47.4%
serials	349,422	89,545	25.6%
TOTAL	11,101,460	3,889,789	35.0%

