HATHITRUST

A Shared Digital Repository

HathiTrust: Partner Update

Indiana University Bloomington
April 16, 2015
Mike Furlough
Executive Director, HathiTrust

HathiTrust is...

- A trusted digital preservation service enabling the broadest possible access worldwide.
- An organization with over 100 research libraries making up its membership.
- A distributed set of services operated by different members (California Digital Library, Illinois, Indiana, Michigan).
- A range of programs enabled by the large scale collection of digitized materials.

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.

Today's Conversation

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

Shared Stewardship

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)

HathiTrust Members

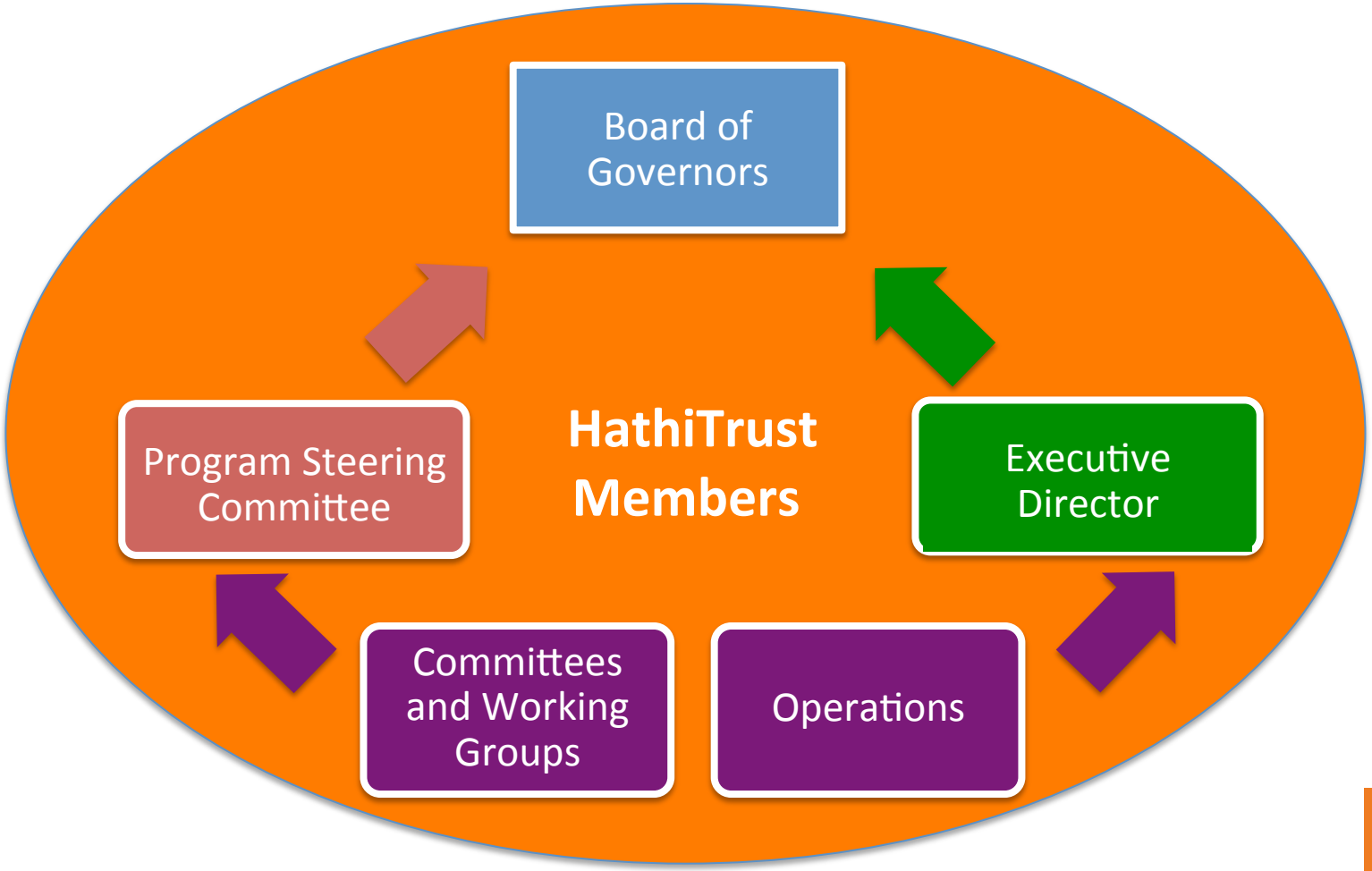
Allegheny College
American University of Beirut
Arizona State University
Baylor University
Boston College
Boston University
Brandeis University
Brown University
Carnegie Mellon University
Case Western Reserve
Colby College
Columbia University
Cornell University
Dartmouth College
Duke University
Emory 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
Lafayette 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
Berkeley
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

We draw upon widely distributed expertise

	Michigan	Indiana	Illinois	California
Administration	✔			
Preservation & Access Repository	✔	✔		
Research Center		✔	✔	
Metadata Management (Zephir)				✔

Governance


Board of Governors
Executive Director
Operations
Committees and Working Groups
Program Steering Committee
HathiTrust Members

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

Collections

Preservation with Access

- Preservation
  - TRAC-certified
- Discovery
  - Bibliographic and full-text search of all materials
- Access and Use
  - Full text search (all users)
  - Public domain and open access works (all users)
  - Collections and APIs (all users)
  - Lawful uses of in-copyright works (members)

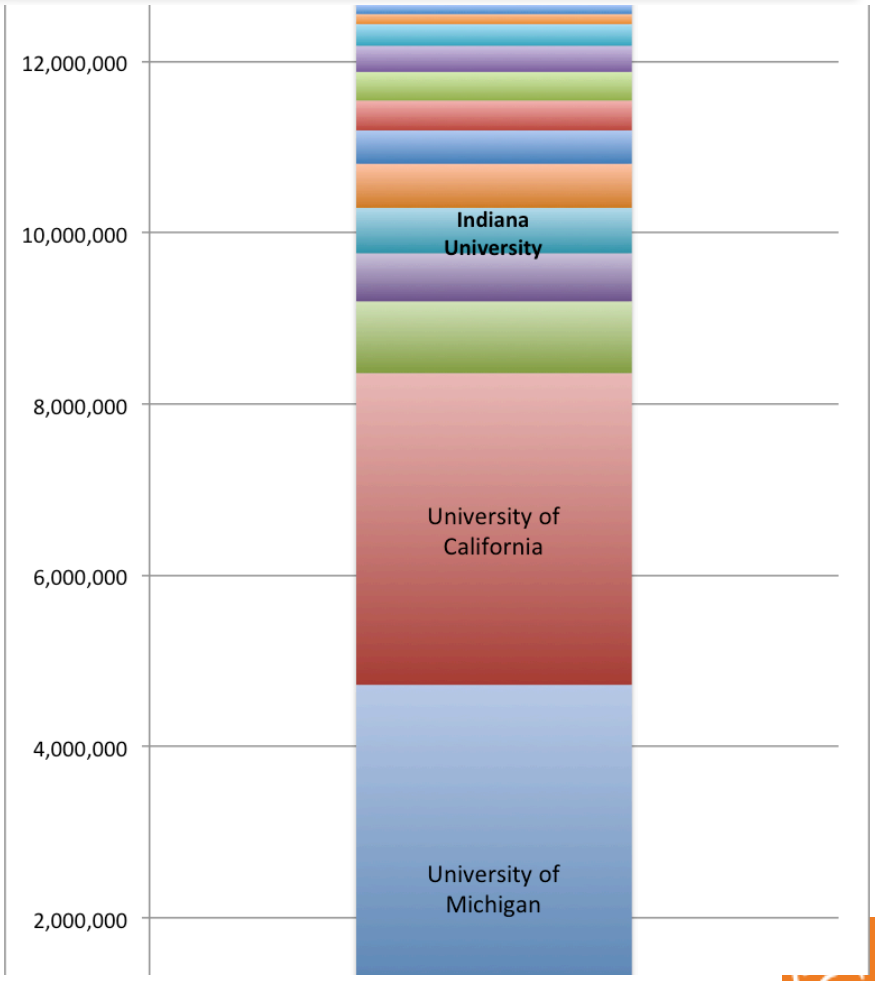
HathiTrust in April 2015

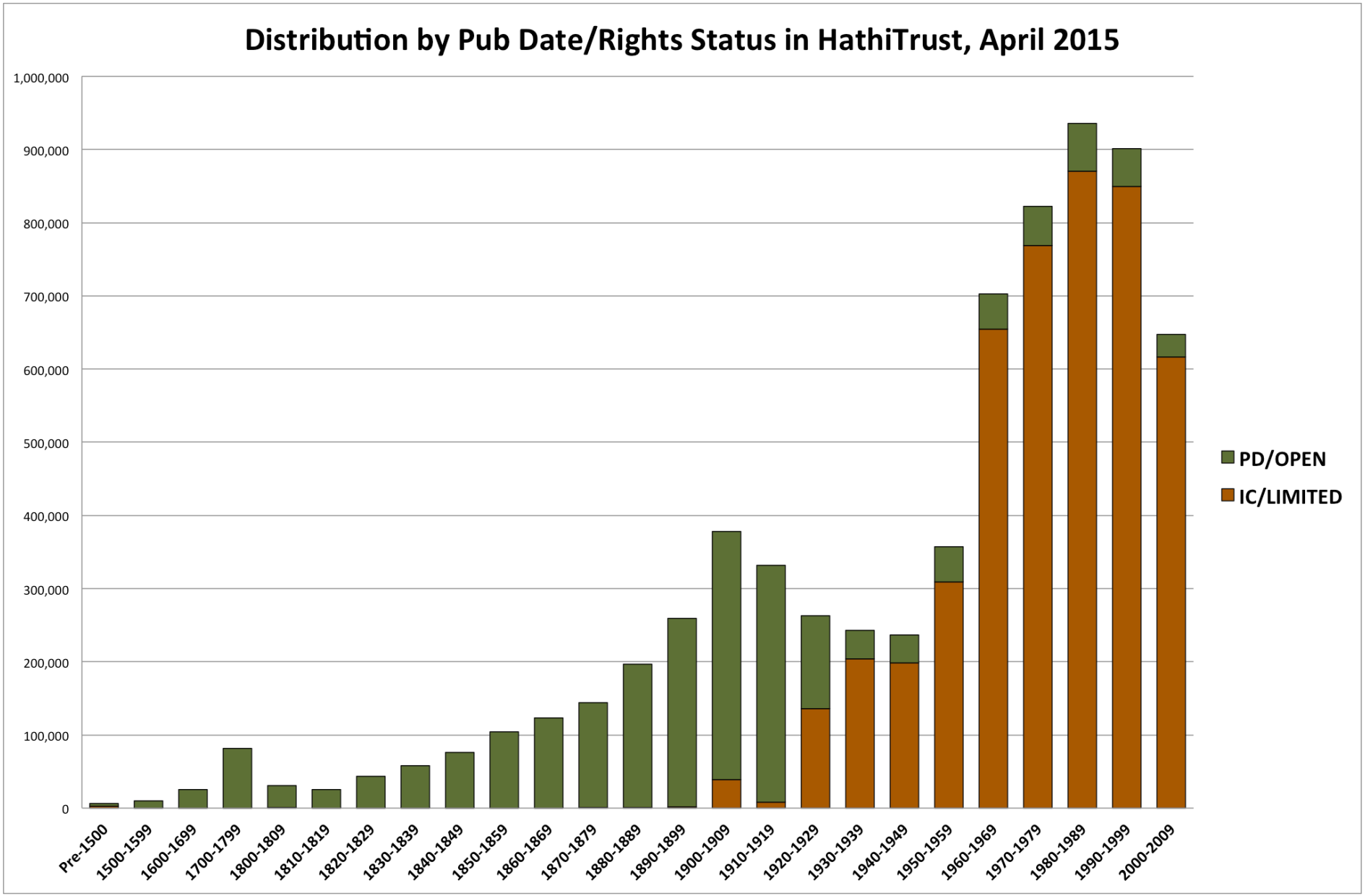
- 13.3 million total items
  - 6.8 million book titles
  - 355,000 serial titles
  - 612,000 US federal government documents
  - 5.03 million items open (public domain & CC-licenses)

The collection primarily includes published materials in bound form, digitized from library collections.

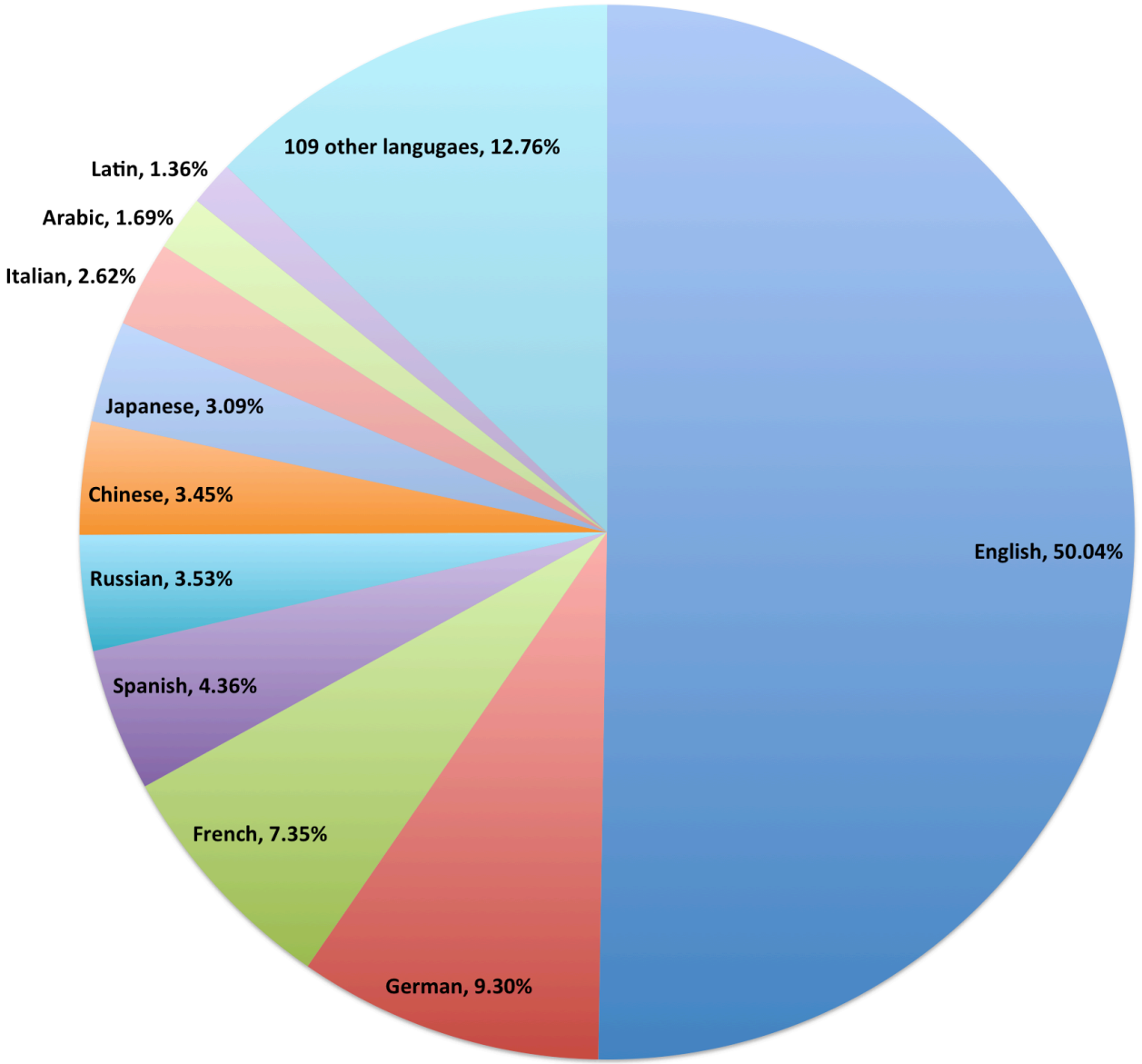
Contributions by Library, Apr 2015

Institution	Volumes
University of Michigan	4,722,050
University of California	3,639,937
Harvard University	838,122
University of Wisconsin	561,534
Indiana University	**529,798**
Cornell University	515,753
Penn State	389,247
University of Illinois	348,946
University of Minnesota	334,249
New York Public Library	304,610
Princeton University	252,841
Universidad Complutense	117,322
Library of Congress	108,892
Keio University	90,122
University of Alberta	76,106
Ohio State	74,525
Columbia University	73,396
Northwestern University	57,000
University of Chicago	56,981
University of Virginia	51,207

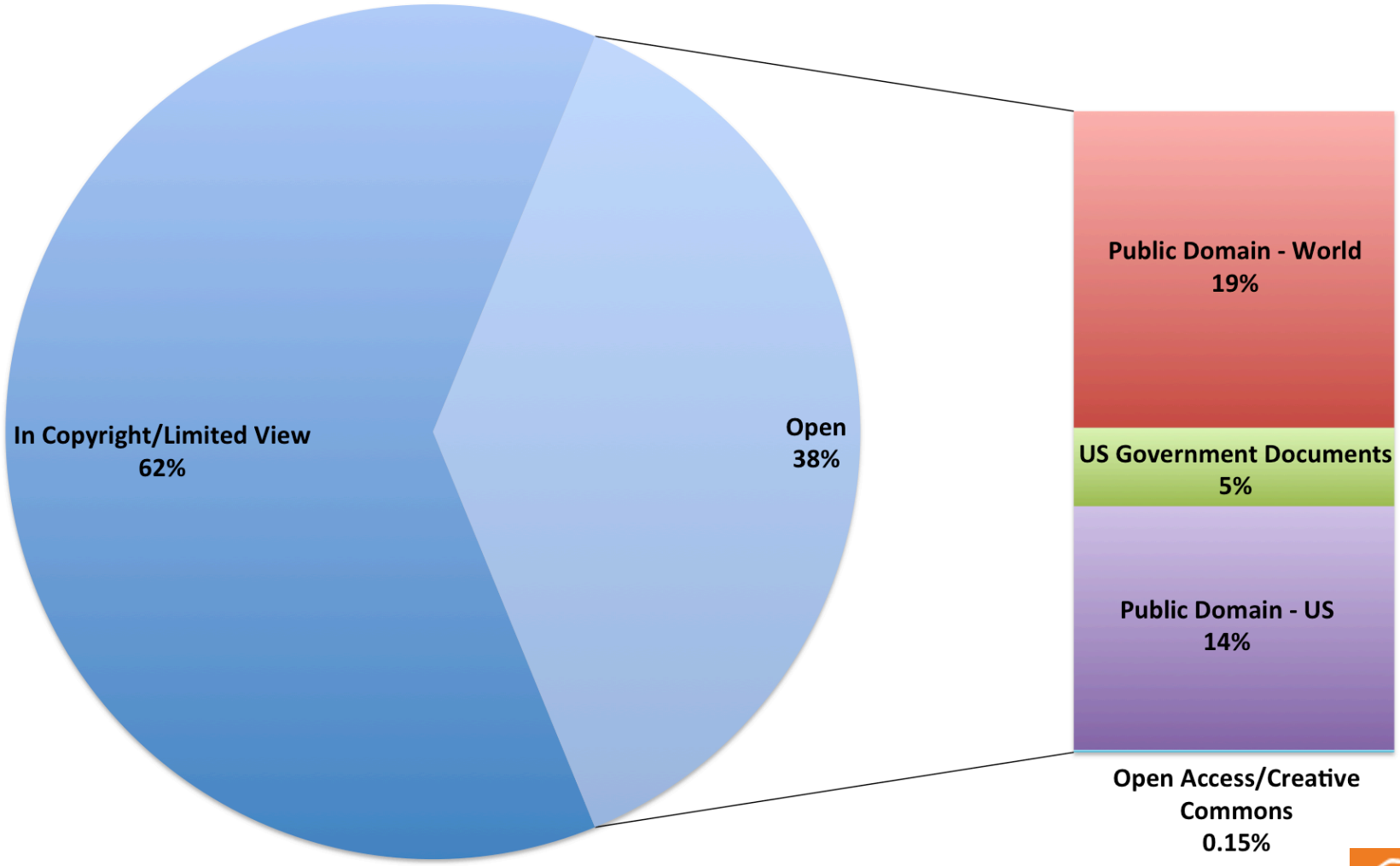



Distribution by Pub Date/Rights Status in HathiTrust, April 2015
1,000,000
900,000
800,000
700,000
600,000
500,000
400,000
300,000
200,000
100,000
0
Pre-1500
1500-1599
1600-1699
1700-1799
1800-1809
1810-1819
1820-1829
1830-1839
1840-1849
1850-1859
1860-1869
1870-1879
1880-1889
1890-1899
1900-1909
1910-1919
1920-1929
1930-1939
1940-1949
1950-1959
1960-1969
1970-1979
1980-1989
1990-1999
2000-2009
PD/OPEN
IC/LIMITED

Language Distribution of All Works in HathiTrust, April 2015



Distribution by Rights and Display Status in HathiTrust, April 2015



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 3rd-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 3rd 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 Access and Use.

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
    - **321,945 reviewed / 170,239 PD (53%)**
  - CRMS-World: Published in UK (1874-1944), Canada, Australia (1894-1964)
    - **185,575 reviewed / 100,740 PD (53%)**

Supported generously by IMLS

Top Ten Titles March 2015

1. Quicksand, by Nella Larsen.
2. Solid Mensuration, by Willis F. Kern and James R. Bland.
3. The Human Figure, by John H. Vanderpoel
4. Roster of the Confederate soldiers of Georgia, 1861-1865, v.1.
5. History of wages in the United States from Colonial times to 1928, United States Department of Labor.
6. Roster of the Confederate soldiers of Georgia, 1861-1865, v.2.
7. Abstracts of old Ninety-six and Abbeville District wills and bonds, as on file in the Abbeville, South Carolina, courthouse.
8. The Five Laws of Library Science, by S. R. Ranganathan.
9. Godey's Magazine, v.40-41, 1850.
10. Roster of the Confederate soldiers of Georgia, 1861-1865, v.3.

Current Initiatives

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

Photo detail from http://babel.hathitrust.org/cgi/pt?id=mdp.39015087610286;view=1up;seq=14

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

Shared Print Monographs Archive

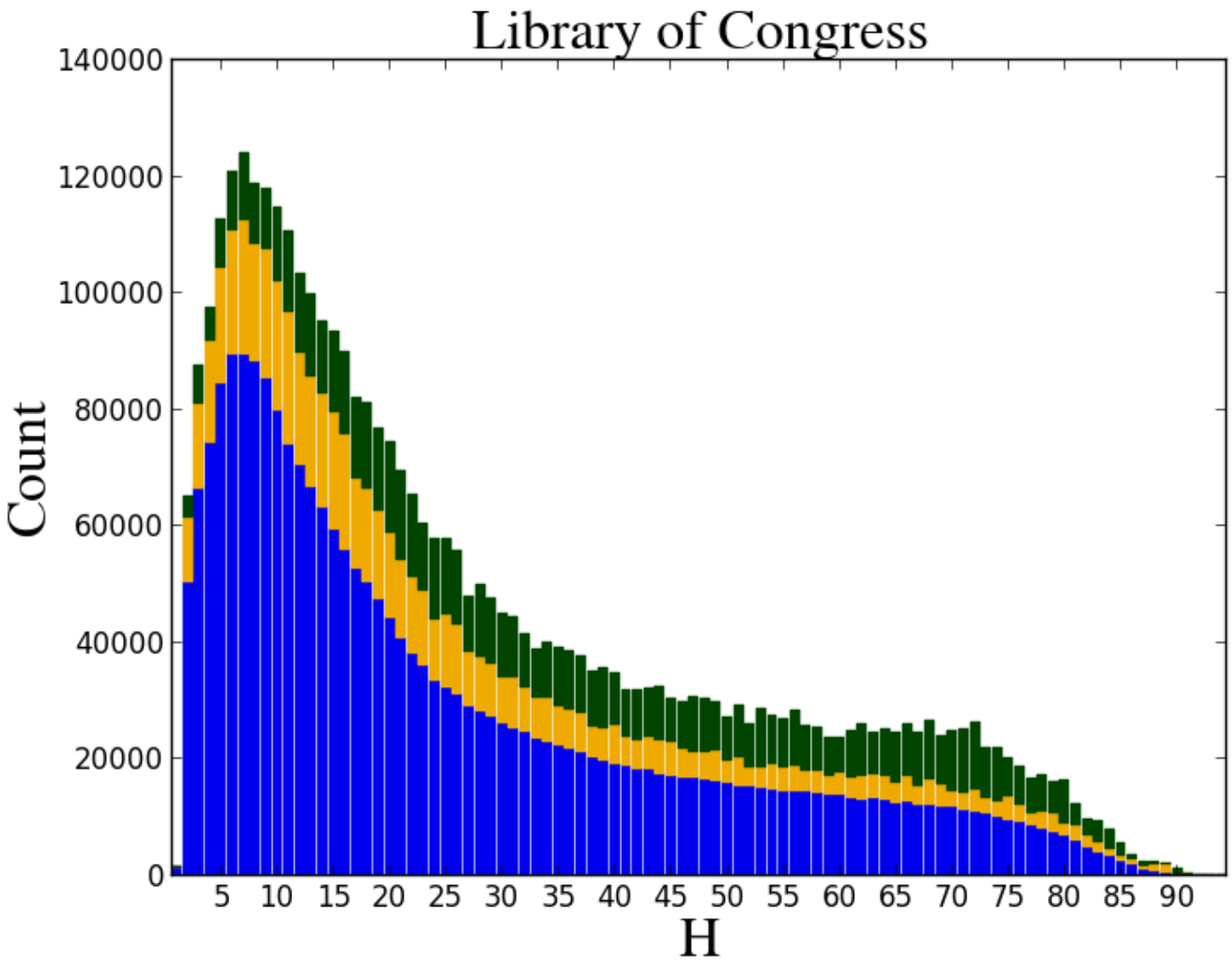
- Ballot Initiative passed at the 2011 HT Constitutional Convention (Con-Con)
  - "To develop a print monographs archive corresponding to volumes represented within the HathiTrust"
- Focus
  - Ensure preservation of print and digital collections
  - Catalyze national/continental collective management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

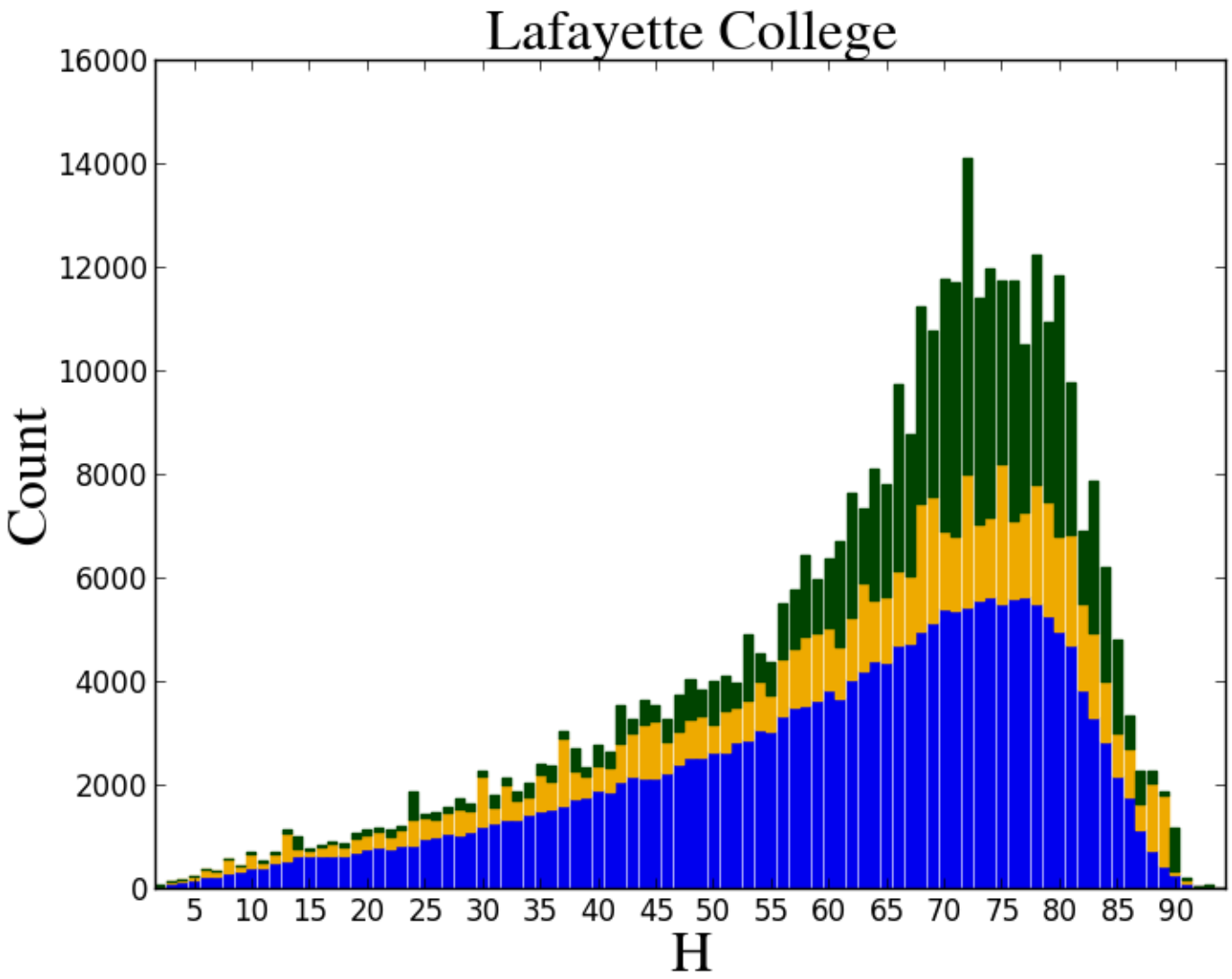
Why A Shared Print Archive Program

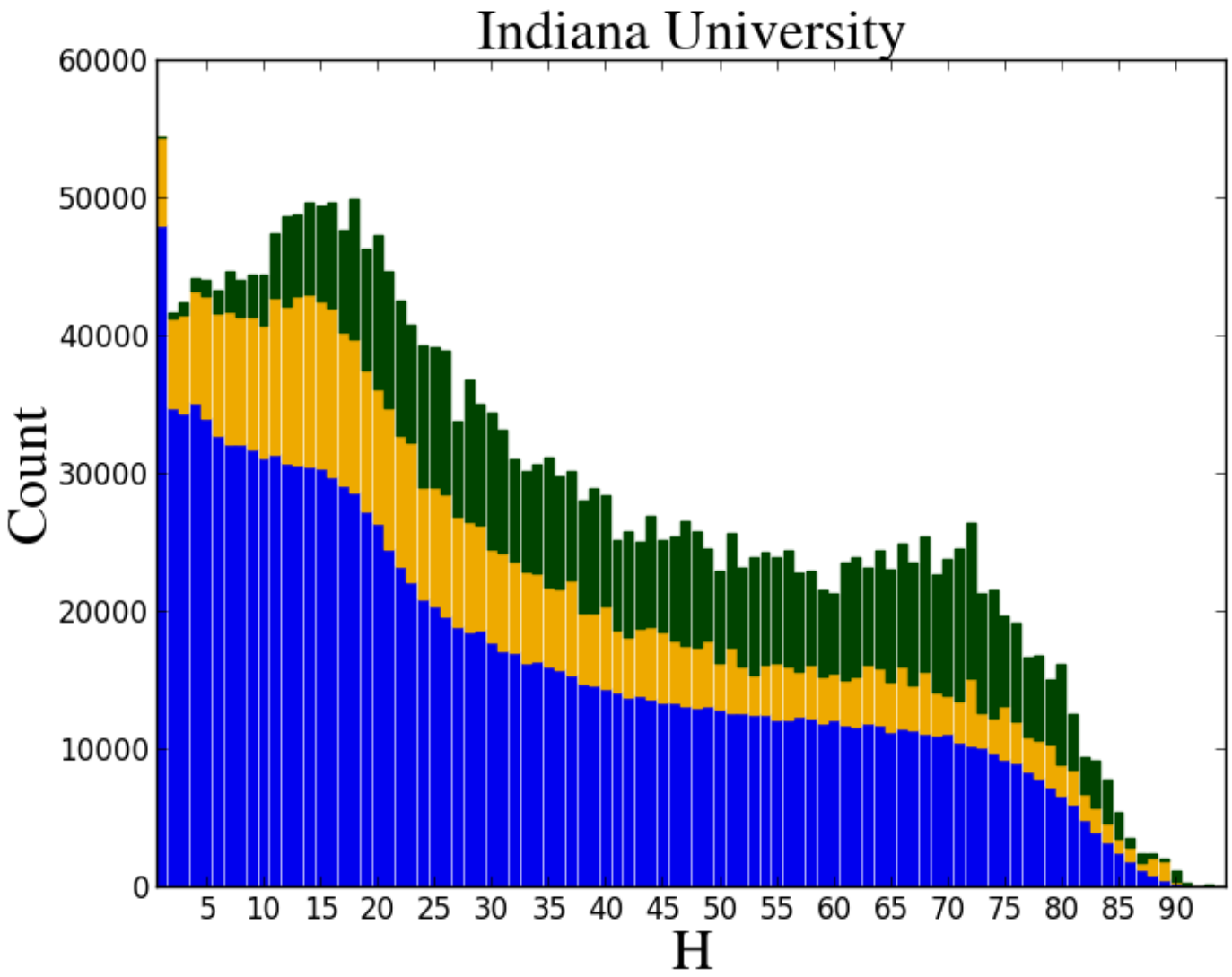
- Creation of the digital corpus provides significant overlap with research collections
- Significant need and desire to reduce costs of collection management and associated footprint
- Many regional efforts, but limited national/ international coordination
- Strengthens preservation commitments
  -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

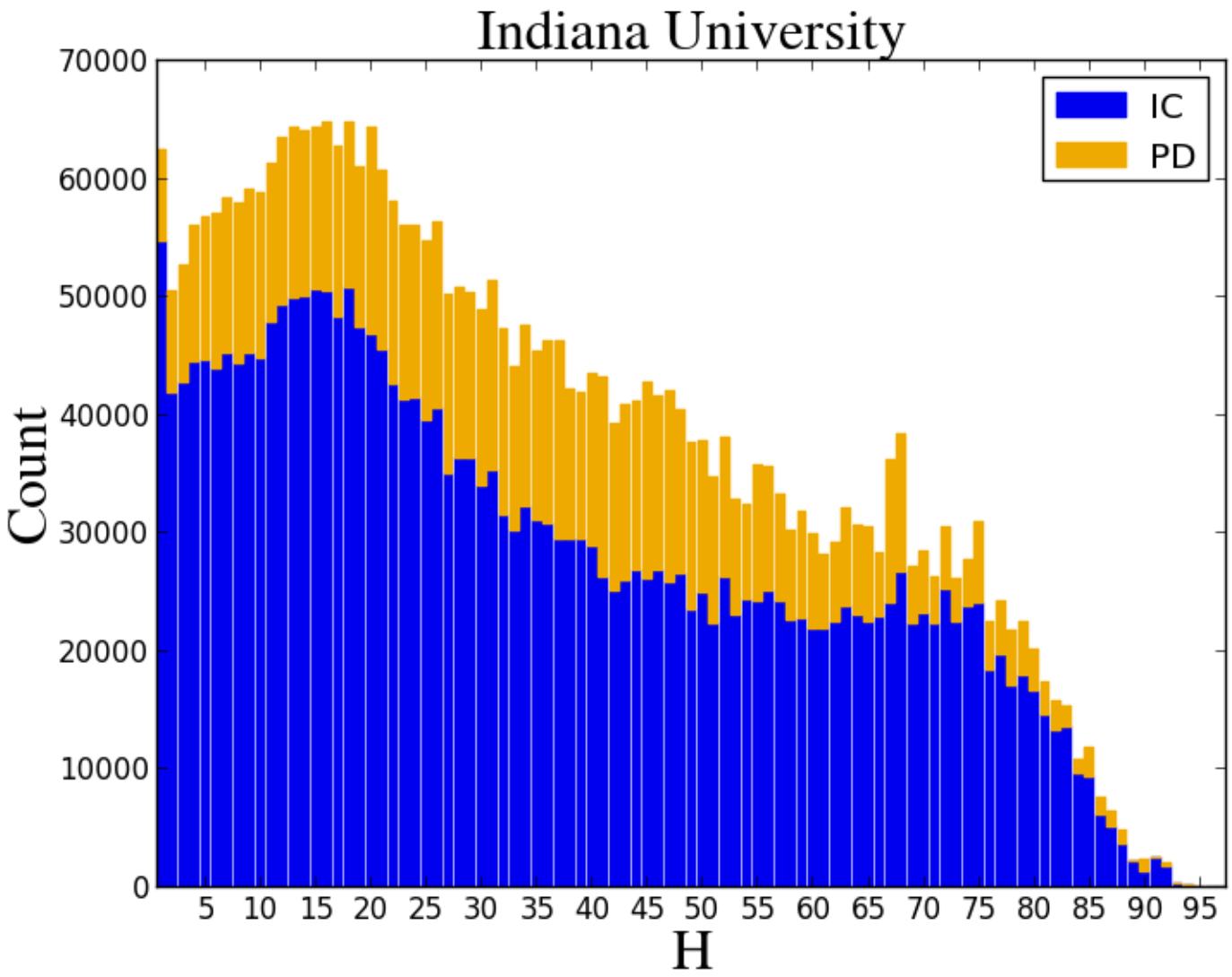
Task Force proposal…

- Defines the 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


Library of Congress
Count
H
0
20000
40000
60000
80000
100000
120000
140000
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90


Lafayette College
Count
16000
14000
12000
10000
8000
6000
4000
2000
0
5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90
H


Indiana University
Count
H
60000
50000
40000
30000
20000
10000
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90


Indiana University
IC
PD
Count
H
70000
60000
50000
40000
30000
20000
10000
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95

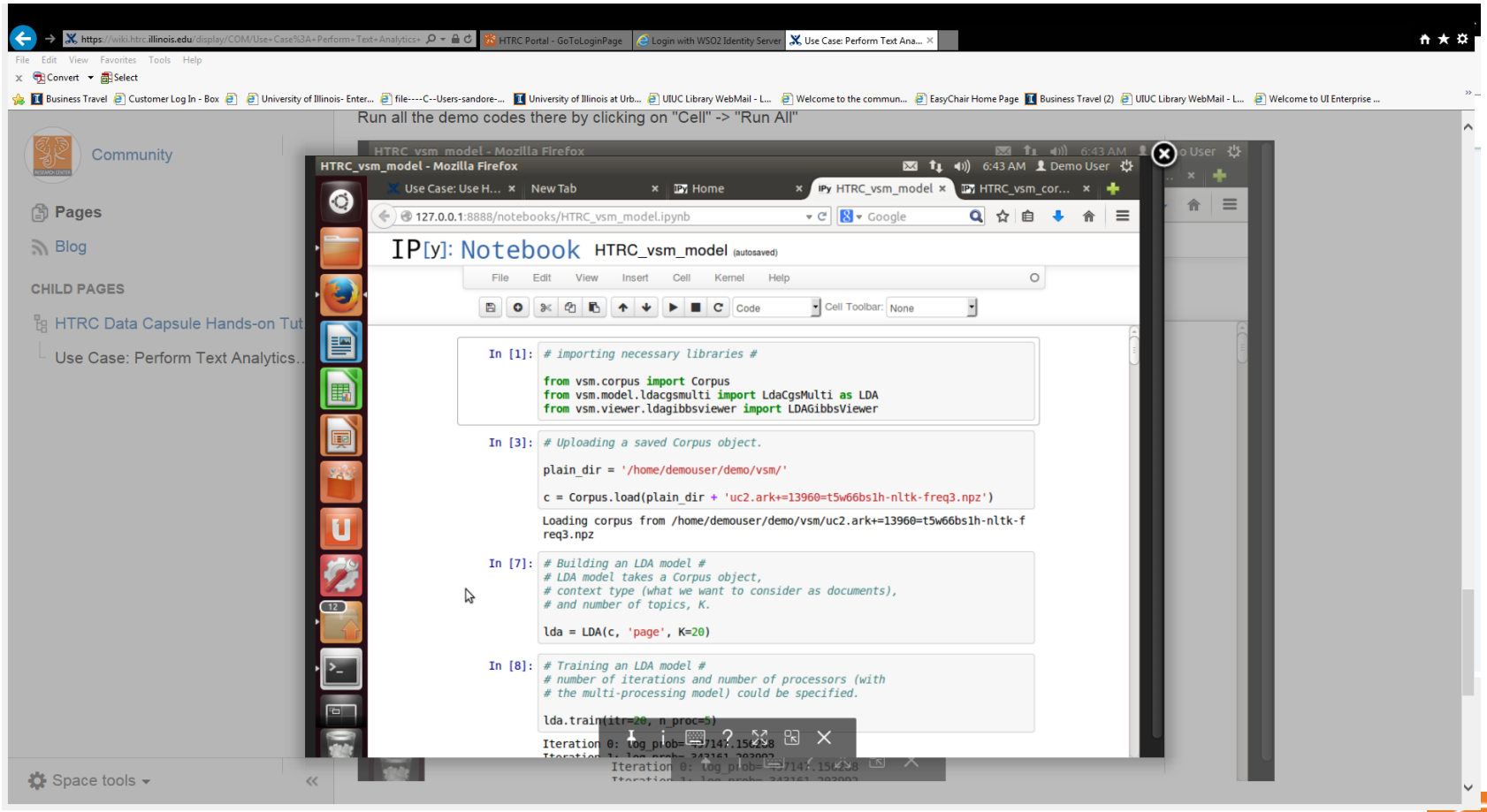
Computational Access Initiatives

- 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

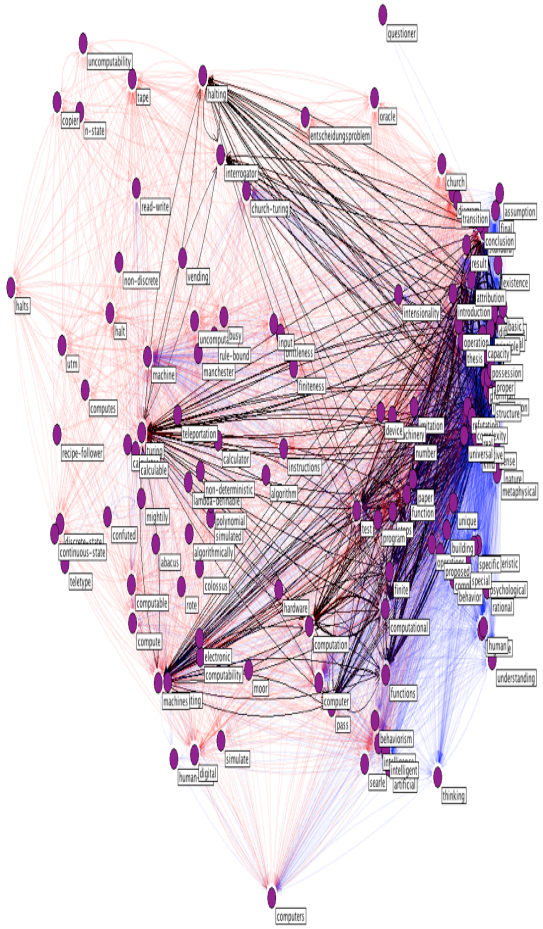
- Research arm of HathiTrust
- Provide a persistent and sustainable structure to enable original and cutting edge research.
  - Leverage data storage and computational infrastructure at Indiana & Illinois
  - Stimulate community development of new functionality and tools
  - Use tools to enable discoveries 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 corpus.

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.

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 (Gov't Docs)
  - Research Center data transfers
  - Print monograph archiving.
- 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
  - Deeper engagement among members
  - Engagement with researchers and libraries
  - 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: feedback@issues.hathitrust.org
- Blogs: http://www.hathitrust.org/blogs
  - Large-scale Search
  - Perspectives from HathiTrust

Thank you!

furlough@hathitrust.org
@MikeFurlough