HATHITRUST

A Shared Digital Repository

Sharing Collections through Shared Stewardship

A HathiTrust Progress Report

TRLN 2014 Annual Meeting
23 July 2014
Mike Furlough
Executive Director, HathiTrust

This Morning's Conversation

- Do you really know HathiTrust?
  - How things work
  - Collections and data
- What are we working on now?
- How has the world changed since we began?
  - And what does that mean for HathiTrust

The Mission and Partnership

Mission

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

The Goals

- To build a reliable and increasingly comprehensive digital archive of library materials converted from print that is co-owned and managed by a number of academic institutions.
- To dramatically improve access to these materials in ways that, first and foremost, meet the needs of the co-owning institutions.
- To help preserve these important human records by creating reliable and accessible electronic representations.
- To enable the digital archive to be accessible to persons who have print disabilities.
- To stimulate redoubled efforts to coordinate shared storage strategies among libraries, thus reducing long-term capital and operating costs of libraries associated with the storage and care of print collections.
- To create and sustain this “public good” in a way that mitigates the problem of free-riders.
- To create a technical framework that is simultaneously responsive to members through the centralized creation of functionality and sufficiently open to the creation of tools and services not created by the central organization.

Highlights and Accomplishments

- Launch (2008)
- TRAC certification (2011)
- Constitutional convention (2011)
- HathiTrust Research Center (2011)
- 10 million volumes (2012)
- New governance established (2012)
- Current bylaws and fee structure (2013)
- 11 million volumes (2014)

Partnership

Allegheny College
Arizona State University
Baylor University
Boston College
Boston University
Brandeis University
Brown University
California Digital Library
Carnegie Mellon University
Colby College
Columbia University
Cornell University
Dartmouth College
Duke University
Emory University
Florida State University
Getty Research Institute
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
Northwestern University
The Ohio State University
The Pennsylvania State University
Princeton University
Purdue University
Rutgers University
Stanford University
Syracuse University
Temple University
Texas A&M 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
The University of Chicago
University of Connecticut
University of Delaware
University of Florida
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
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

How are costs shared?

- Public domain volumes: All partners share in infrastructure costs for each item.
- In copyright volumes: Partners share costs based on their holdings.
- Infrastructure cost per volume: ~$0.155 per volume per year.
- All partners pay an additional amount above costs to fund new programs and investigations.

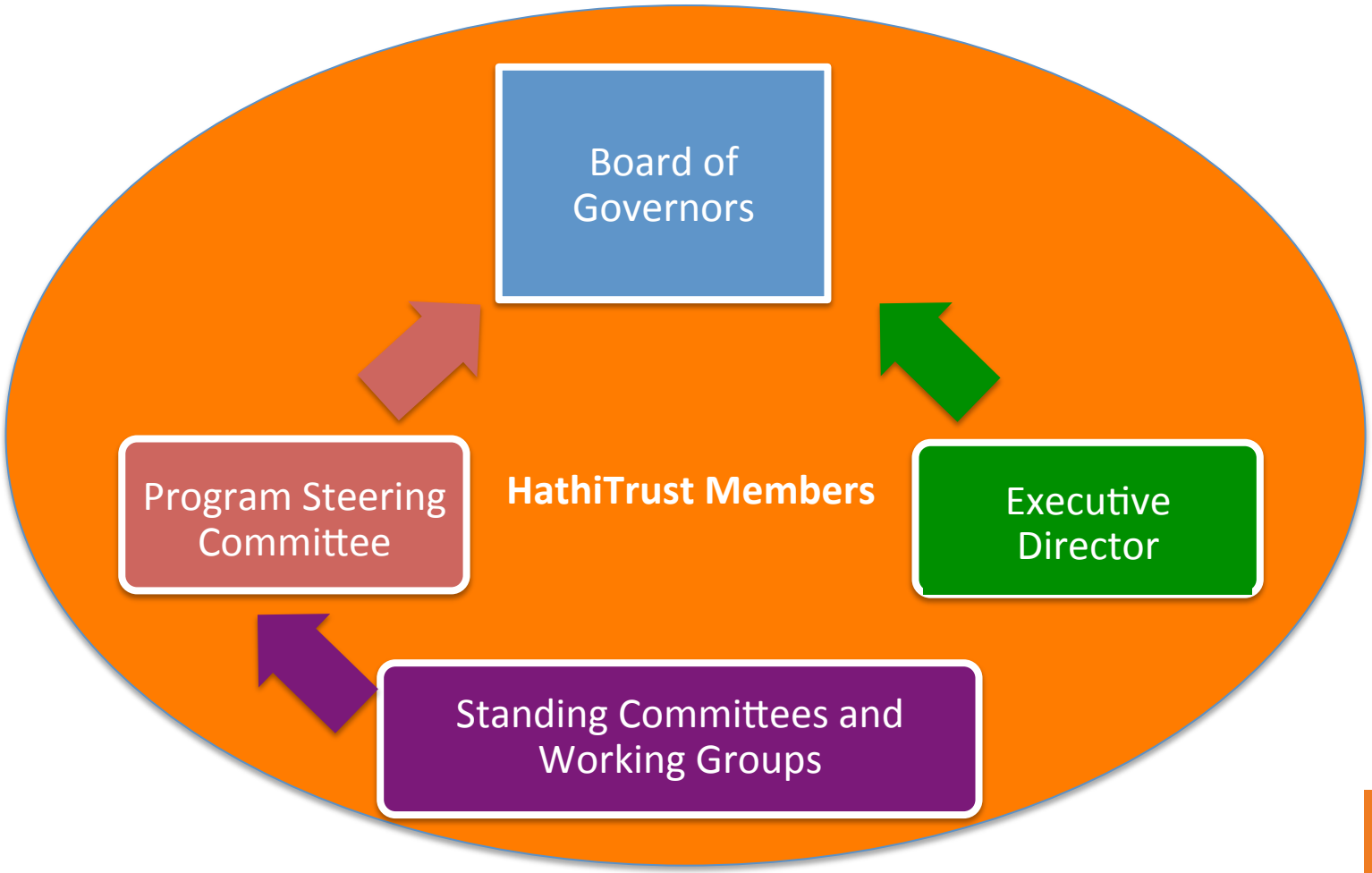
Where does work get done?

- HathiTrust is legally constituted as part of the University of Michigan, but functions are distributed.
  - Preservation repository and access services
    - University of Michigan
    - Mirror site: Indiana University
  - Metadata management services (Zephir)
    - California Digital Library
  - HathiTrust Research Center
    - Indiana University and University of Illinois

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

Governance


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

Five-year terms (beginning Jan 2013)
Betsy Wilson (University of Washington)
Bob Wolven (University of Columbia)

Four year terms
Richard Clement (University of New Mexico)
Patricia Steele (University of Maryland)

Three year terms:
Carol Mandel (New York University)
Sarah Michalak (University of North Carolina-Chapel Hill)

Members appointed by the founding institutions:
James Hilton (University of Michigan)
Carol Diedrichs (Ohio State University)
Laine Farley (California Digital Library)
Wendy Lougee (University of Minnesota)
Brian Schottlaender (UC, San Diego)
Brenda Johnson (Indiana University)

Ex Officio (Board, PSC, Executive Committee):
Mike Furlough, Executive Director

Executive Committee
- Chair
- Past Chair
- Treasurer
- Chair of PSC
- Executive Director

HathiTrust Board of Governors

Program Steering Committee

- Serves at the direction of the Board of Governors to…
  - Reviews HathiTrust's development agenda, shaping initiatives and strategies for Board discussion and decision-making, and considering the implications of those initiatives for the future.
  - Recommends alterations in the development agenda based on such reviews. Based on its reviews, develops position papers for the member community to encourage debate or mobilize discussion with regard to particular issues.
  - Works with the Board of Governors to develop policies for HathiTrust and its members.

Program Steering Committee Membership

- Ivy Anderson (CDL)
- John Butler (Minnesota)
- Chris Freeland (Washington University)
- Todd Grappone (UCLA)
- Martha Hruska (UC San Diego)
- Martin Kurth (New York University)
- Erika Linke (Carnegie Mellon University)
- Robert McDonald (Indiana)
- Matthew Sheehy (Harvard)
- Elaine Westbrooks (Michigan)
- Bob Wolven, Chair (Columbia)

Standing Committees and Working Groups

- Collections Committee
- Rights and Access Working Group
- User Support Working Group

- On hiatus, pending review:
  - Communications
  - User Experience

Annual Membership Meeting

- Required by the bylaws.
- 1st Annual Meeting:
- October 10, 2014 in Washington, DC
- Member representatives or a designated substitute.

Collections and Access

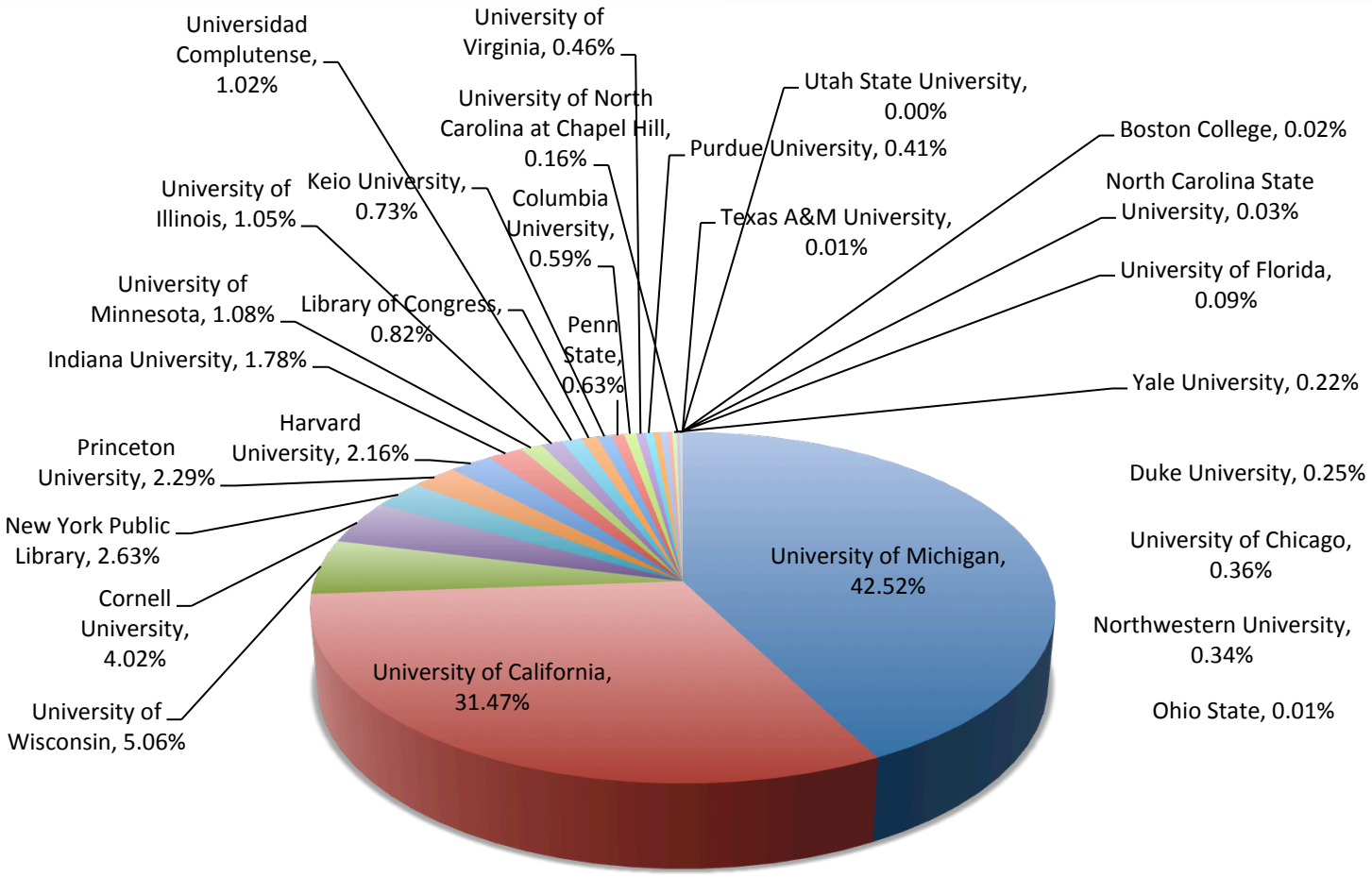
Preservation with Access

- Cost effective preservation and access services
- Preservation
  - TRAC-certified
  - Robust infrastructure
  - Long-term commitments on digital content facilitate planning, decision-making
  - Facilitate activities such as discovery, copyright review, use of materials

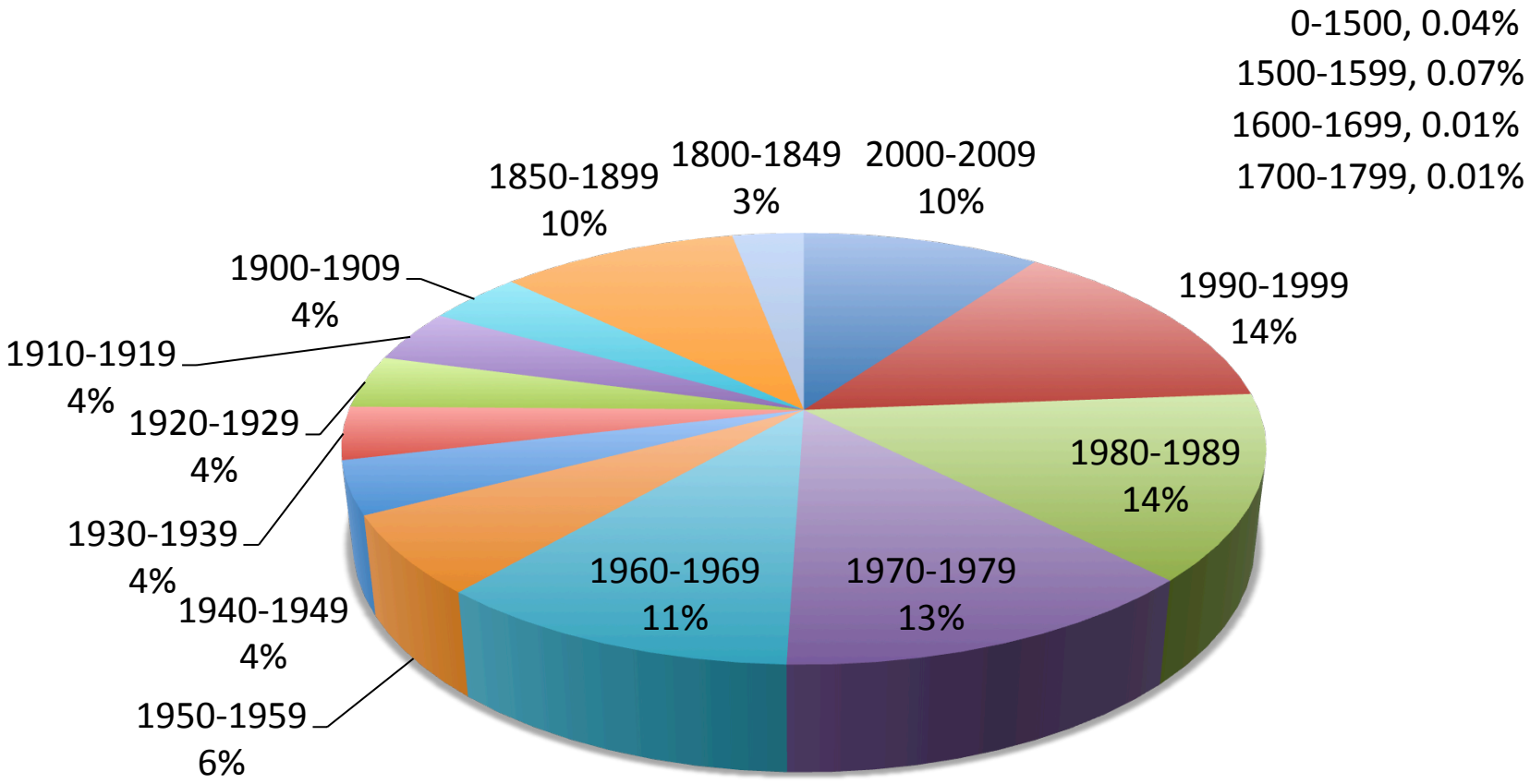
Preservation with Access (2)

- Discovery
  - Bibliographic and full-text search of all materials
  - Extended discovery (ProQuest, EBSCO, OCLC, Ex Libris)
  - Mechanisms for local loading of records
- Access and Use
  - Full text search
  - Public domain and open access works
    - Full download of materials where possible
    - Print on demand
  - Collections and APIs
  - Research Center
  - Lawful uses of in-copyright works

Content Sources



Dates



* As of February 17, 2014

Lawful uses of in copyright works

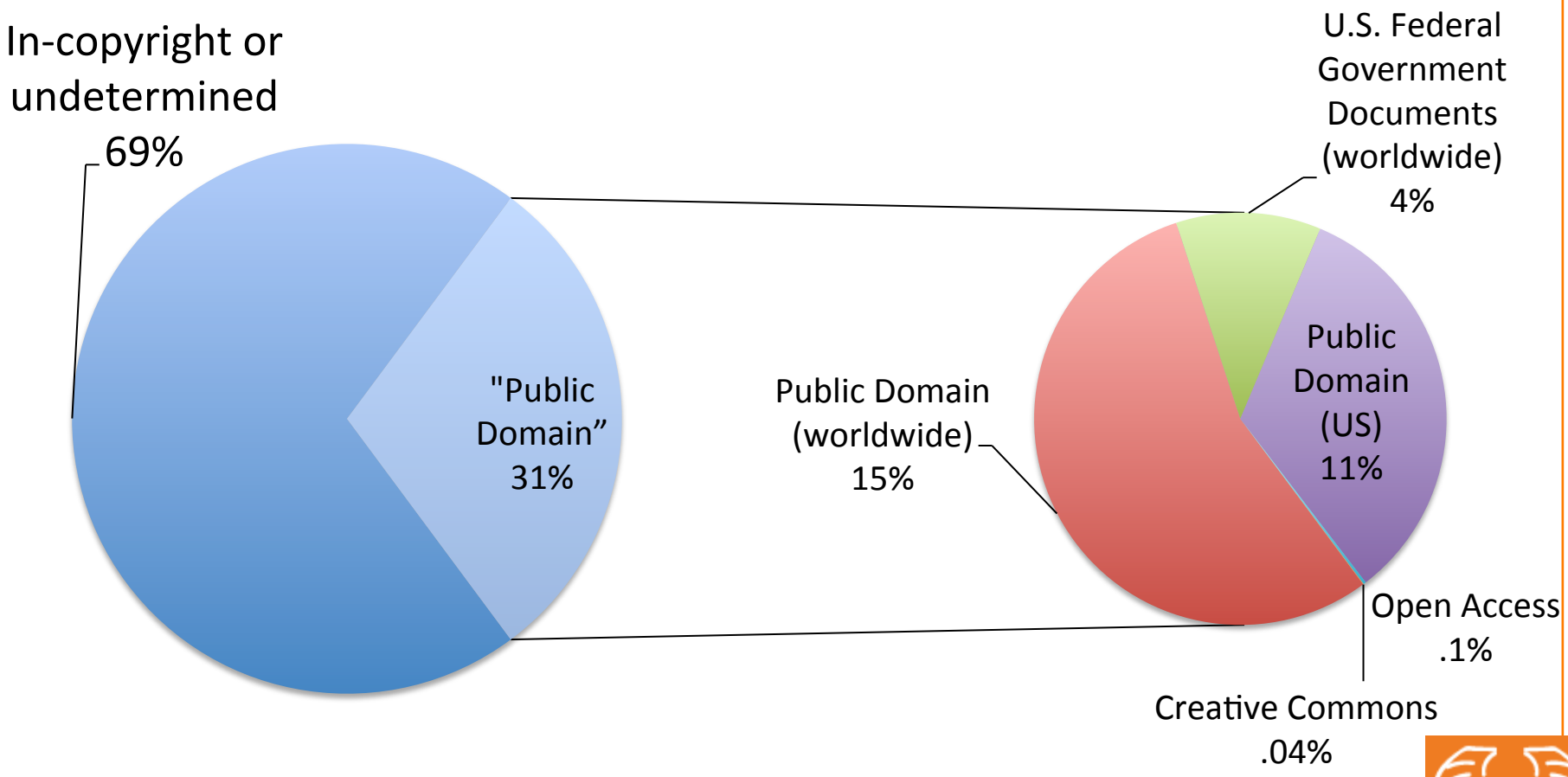
- Sensitive to multiple legal regimes
  - Full-text search (everywhere)
  - Access to users who have print disabilities (US, and where law permits)**
  - Access works that are damaged or missing and also out of print and unavailable (US only)

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

Copyright Review / Permissions

- CRMS US (since 2008)
  - Published in US, 1923-1963
  - 314,270 determinations
  - 165,340 opened (~53%)
- CRMS-World (since 2012)
  - Published non-US (UK, Canada, Australia, Spain)
  - 117,369 determinations
  - 59,652 opened (~51%)
- Permissions
  - Open access – 6,982
  - Additional Creative Commons – 6,835

Copyright Distribution



Where Do HathiTrust Users Come From?

- Since September 2013:
  - ~43% of all traffic comes from HathiTrust directly (searches or other referrals)
  - The top five non-Hathi sources of traffic:
    1. onlinebooks.library.upenn.edu
    2. worldcat.org
    3. dp.la
    4. clio.columbia.edu
    5. en.wikipedia.org

Current Initiatives

Current Initiatives

1. Developing a shared print monographs archive
2. Expanding coverage and access to US government publications
3. Expanding support for computational (non-consumptive) research

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"
- HathiTrust Board of Governors recently approved appointment of a PSC-designed task force to begin process

Task Force Members

- Tom Teper, Chair (Illinois)
- Clem Guthro (Colby)
- Robert Kieft (Occidental)
- Erik Mitchell (UC Berkeley)
- Jake Nadal (ReCAP)
- Matthew Sheehey (Harvard)
- Emily Stambaugh (CDL)
- Karla Strieb (Ohio State)

Issues to examine…

- Exploration of the model needed to identify and preserve print resources
- Qualifications of participating repositories
- Analysis and identification of appropriate content for inclusion in the archive
- Additional criteria for participation, such as geography, repository type, breadth of contribution, institutional commitment…
- Retention periods
- Discovery, access policies, and service models
- Business and financial models
- Roles and relationships among HT and other libraries and organizations engaged in collaborative management of print collections.

Government Documents Initiative

- Ballot Initiative: provide "expanded coverage & enhanced access to U.S. Government Documents."
- Activities:
  - Developing a registry of US Federal Government Documents
  - HathiTrust Board of Governors recently approved appointment of a PSC-designed Advisory Group to begin process

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.

Advisory Group Membership

- Prue Adler, Association of Research Libraries
- Ivy Anderson, California Digital Library
- Joni Blake, Greater Western Library Alliance
- Kirsten Clark, University of Minnesota
- Richard Clement, Utah State University
- Elizabeth Cowell, University of California, Santa Cruz
- Michael Norman, University of Illinois
- Mark Phillips, University of North Texas
- Mark Sandler, Committee on Institutional Cooperation **(chair)**
- Jonathan Rothman, University of Michigan
- Judith Russell, University of Florida
- Barbara Selby, Univ of Virginia
- Jeremy York, HathiTrust

Issues to Examine

- How to engage existing and potential government documents digitization projects?
- What are the areas of greatest need for access or for collection management?
- The Registry will not be perfect. How will we improve it over time?

Computational Access

- Distribution of public domain datasets
- HathiTrust Research Center
  - Developed collaboratively by Indiana University and University of Illinois; launched July 2011
  - Enables computational access to public domain and open access materials; working to support in-copyright materials as well
  - Funding from the Sloan Foundation, Andrew W. Mellon Foundation, and NEH Office of Digital Humanities.
  - Led by Beth Plale (Indiana) and Stephen Downie (Illinois)

Using the HathiTrust Research Center

- http://www.hathitrust.org/htrc
- Portal: sign up, browse volume lists and algorithms, execute algorithms, view results
  - https://htrc2.pti.indiana.edu/HTRC-UI-Portal2/
- Workset Builder
  - https://htrc2.pti.indiana.edu/blacklight
- Sandbox: run own algorithms
- Extracted Features Dataset (alpha)
  - https://sandbox.htrc.illinois.edu/HTRC-UI-Portal2/Features

Example Projects Supported by HTRC

- Muñoz, Trevor, University of Maryland. "Distributed Metadata Correction and Annotation."
  - Correction, annotation and enhancement of HT records and export as linked data
- Page, Kevin, Oxford University. "ElEPHãT: Early English Print in HathiTrust, a Linked Semantic Workset Prototype"
  - Development of secondary worksets based on both HT and the Early English Books Online Text Creation Partnership (EEBO-TCP).
- Burton, Vernon. "The South as 'Other,' the Southerner as 'Stranger.'"
  - Explore how attitudes expressed in print about slavery, southerners, and non-southerners have changed over both time and space.
- Ted Underwood, Associate Professor of English at the University of Illinois, Urbana-Champaign.
  - Using public domain texts received from HathiTrust to explore changing relationships in literary genres from 1700-1899.

Where are we now?

What's Good

- Our mission, collection, and the repository operations are all strong.
- Our brand reputation is outstanding.
- The partnership provides a solid base for action.

What changed since 2008/2011?

- Legal and public policy environment
- Our community's digitization and access strategies
- Development of national digital library infrastructure

Legal and Policy Frameworks

- This was true in 2011, but we have greater certainty today:
  - Digitization for the purposes of full-text search is a fair use of in copyright work.
  - Digitization for the purposes of serving users with print disabilities is a fair use of in copyright work.
- The US Copyright Office is working to update the Copyright Act.
  - Congressional hearings are underway.

Questions for HathiTrust

- Do court rulings help us think differently about our digitization strategies?
- How rapidly, and in what ways, can we expand services for users with print disabilities?
- How can we help the community proactively advocate for educational and research uses of in copyright works that protect user and rightsholder interests?

Our collection digitization strategies

- Mass digitization of books has slowed down but we are nowhere near finished.
- Archives have barely been touched.
- Huge amounts of at-risk media are held in our collections.

Questions for HathiTrust

- How can we collectively define reformatting strategies for the future?
- How will our community fund large scale digitization in the coming years?
- How do we collect newly-created and published work?
- How do we support non-text formats?

National Digital Library Infrastructure

- Since 2011
  - DPLA has launched
  - DPN has been formed
  - APTrust has begun development
  - SHARE is underway
  - Research data management is (more or less) an accepted part of the library portfolio

Questions for HathiTrust

- Are we moving towards consistent discovery layers through federation?
  - "All roads lead to…"
- Does our community know how to talk about digital preservation on the scale of decades?
- How do we stay focused and how can our focus better define the system?
- How do we move to an international infrastructure?

Answers

Assumptions

- The key to our work is alignment in 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 Stay Informed

- The Newsletter
  - Monthly, mid-year, year-end
  - http://www.hathitrust.org/news_publications
- The (irregular) blogs
  - Perspectives from HathiTrust
  - Large Scale Search
  - http://www.hathitrust.org/blogs
- Twitter: @hathitrust

Thank you!

furlough@hathitrust.org
@MikeFurlough