



HATHITRUST

A Shared Digital Repository

The HathiTrust Research Corpus

University of California Davis
ICIS Seminar
February 25, 2015
Mike Furlough
Executive Director, HathiTrust

Today

- Interrupt me
- Why does HathiTrust exist?
- Major current work
- HathiTrust Research Center
- Next?





HATHITRUST.ORG

[LOG IN](#) ▾

Search HathiTrust's digital library

[FULL-TEXT](#)[CATALOG](#)

All Fields ▾

 [Advanced catalog search](#) | [Search tips](#) Full view only[? Should I search catalog or full-text?](#)

Want to get the most out of HathiTrust?

Log in with your partner institution account to access the largest number of volumes and features.

[Not with a partner institution? »](#)

HathiTrust is a [partnership](#) of academic & research institutions, offering a collection of millions of titles digitized from libraries around the world.

WHAT CAN YOU DO WITH HATHITRUST?



BROWSE COLLECTIONS

Explore user-created [featured collections](#).



READ BOOKS ONLINE

Read millions of titles online — [like this one!](#)



READ BOOKS ON THE GO

Take the library's books anywhere with our [mobile website](#).



DOWNLOAD BOOKS* & CREATE COLLECTIONS

**requires institutional login*

Go to a Section Search: All of Technology

- [Technology Home](#)
- [Circuits](#)
- [Product Reviews](#)
- [How To's](#)
- [Deals](#)

Advertisement

Google Is Adding Major Libraries to Its Database

By JOHN MARKOFF
and EDWARD WYATT

Published: December 14, 2004

Google, the operator of the world's most popular Internet search service, announced today that it had entered into agreements with some of the nation's leading research libraries and Oxford University to begin converting their holdings into digital files that would be freely searchable over the Web.

It may be only a step on a long road toward the long-predicted global virtual library. But the collaboration of Google and research institutions that also include Harvard, the University of Michigan, Stanford and the New York Public Library is a major stride in an ambitious Internet effort by various parties. The goal is to expand the Web beyond its current valuable, if eclectic, body of material and create a digital card catalog and searchable library for the world's books, scholarly papers and special collections.

Google - newly wealthy from its stock offering last summer - has agreed to underwrite the projects while also adding its own technical abilities to the task of scanning and digitizing tens of thousands of pages a day at each library.

[Enlarge This Image](#)



Thor Swift

A book is scanned at Stanford University. Google's plans for digital files include the University of Michigan and the New York Public Library.

ARTICLE TOOLS

SANFORD P. DUMAIN (SD-8712)
MILBERG WEISS BERSHAD
& SCHULMAN LLP
One Pennsylvania Plaza
New York, NY 10119-0165
Telephone: (212) 594-5300
Facsimile: (212) 868-1229

MICHAEL J. BONI
KATE REZNICK
KOHN SWIFT & GRAF, P.C.
One South Broad Street, Suite 2100
Philadelphia, PA 19107
Telephone: (215) 238-1700
Facsimile: (215) 238-1968

Counsel for Plaintiffs

UNITED STATES DISTRICT COURT
SOUTHERN DISTRICT OF NEW YORK

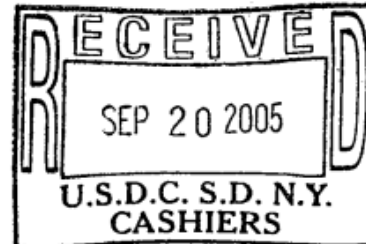
The Author's Guild, Associational Plaintiff, :
Herbert Mitgang, Betty Miles and Daniel Hoffman, :
Individually And On Behalf Of All Others Similarly :
Situated, :
Plaintiffs, :
v. :
Google Inc., :
Defendant. :

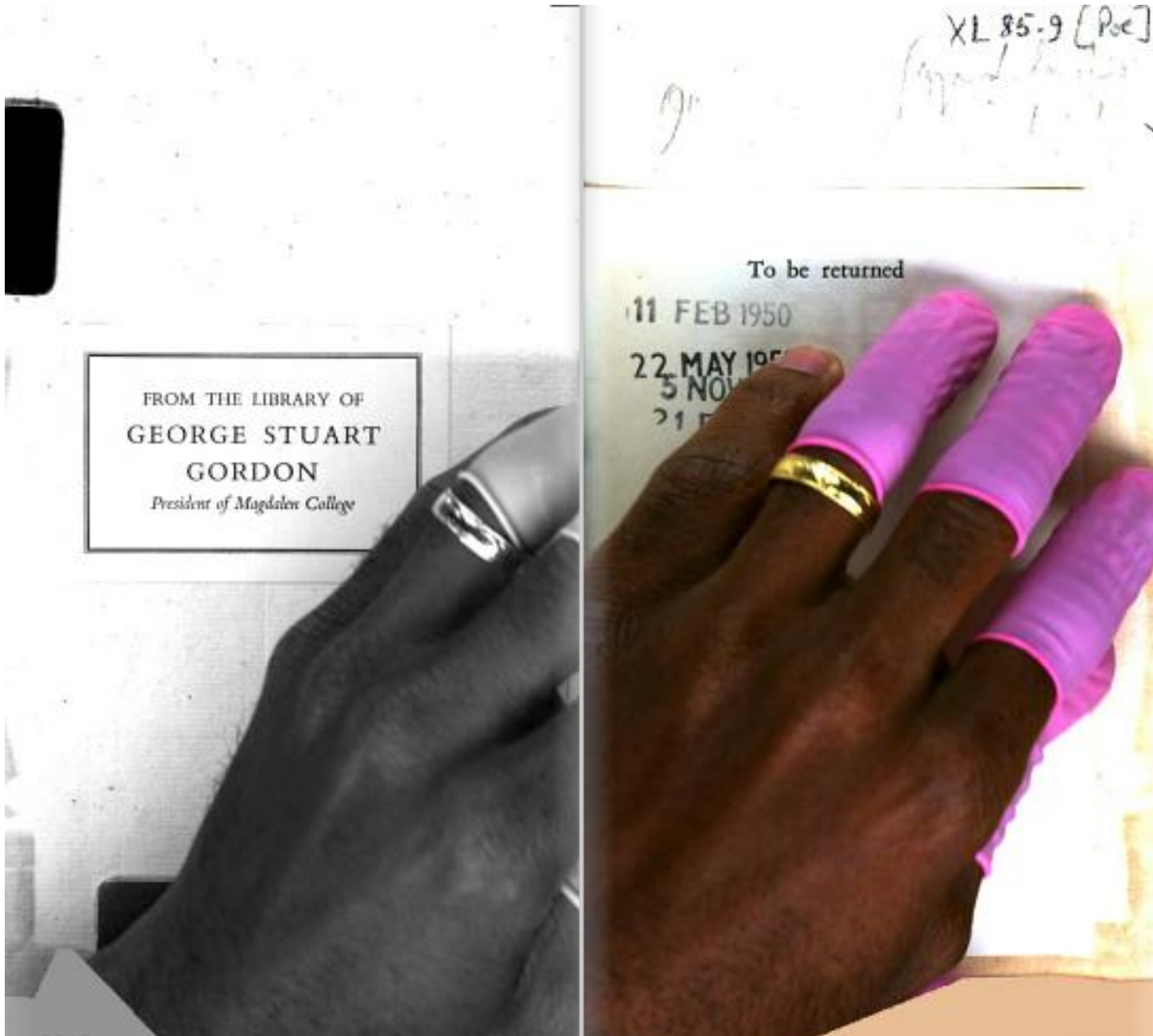
CLASS ACTION COMPLAINT

JURY TRIAL DEMANDED

JUDGE SPRIZZO

05 CV 8136





25 February 2015
From The Art of Google Books: <http://theartofgooglebooks.tumblr.com/post/74936156541/married-employees-hand-over-bookplate-and>

QOTD: Google books

March 6, 2007

Lorcan Dempsey Twitter: [@lorcanD](#)

Categories: [Books, movies and reading ...](#) • [Libraries - organization and services](#) • [ebooks and other e-resources](#)



Peter Brantley, the new Director of the Digital Library Federation writes about the library digitization initiatives with Google:

We poisoned our hand before we played it. We were approached singly, charmed in confidence, the stranger was beguiling, and we embraced. For the love of selfish confidence, we spoke neither our fortune nor our misgivings with our neighbors or our friends. We felt special, invited to loud weddings on far away islands of adventure; in the quiet we may wonder if we were given broken jewelry. [[shimenawa - Google and the books](#)]



Building a digital archive of global content
for universal access

[Home](#)

[About](#)

[Contributors](#)

[FAQ](#)

[Participate](#)

[Press](#)

[Milestone achieved »](#)

Sloan Foundation Grant Awarded

The Sloan Foundation announced today, that it has awarded \$1 million to the Internet Archive to help pay for digital copies of collections owned by the Boston Public Library, the Getty Research Institute, Johns-Hopkins Libraries and the Bancroft Library of the University of California. – 2006 December 19.

Microsoft Launches Live Search Books

by Greg R. Notess



December 11, 2006 — A little over a year ago, Microsoft announced that it was joining the Open Content Alliance (OCA; <http://www.infotoday.com/newsbreaks/nb051031-2.shtml>) to create a database of full-text books. Planning to launch a beta version of its books database sometime in 2006, Microsoft has now met its goal with the launch of Live Search Books (beta) at <http://books.live.com>. All of the books available on Live Search Books are out-of-copyright titles, thus avoiding the copyright controversy at Google Books. The initial load at Live features titles from several libraries' collections, including the University of California, the University of Toronto, and The British Library. Microsoft also announced the addition of The New York Public Library and the American Museum of Veterinary Medicine as future contributors.

Google

≠



25 February 2015

10



Mission

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

Efforts include, but are not limited to

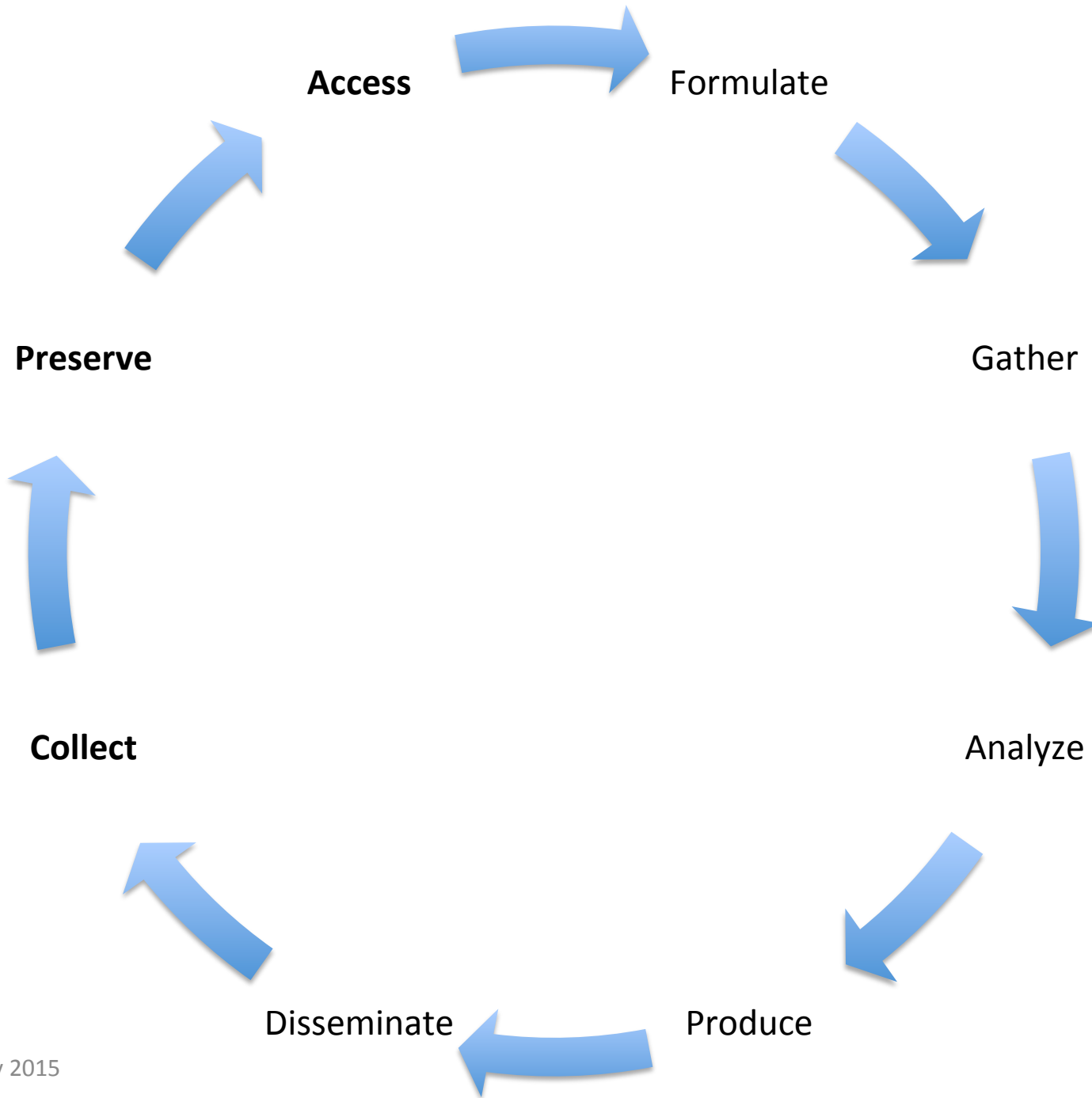
- ...building comprehensive collections co-owned and managed by partners.

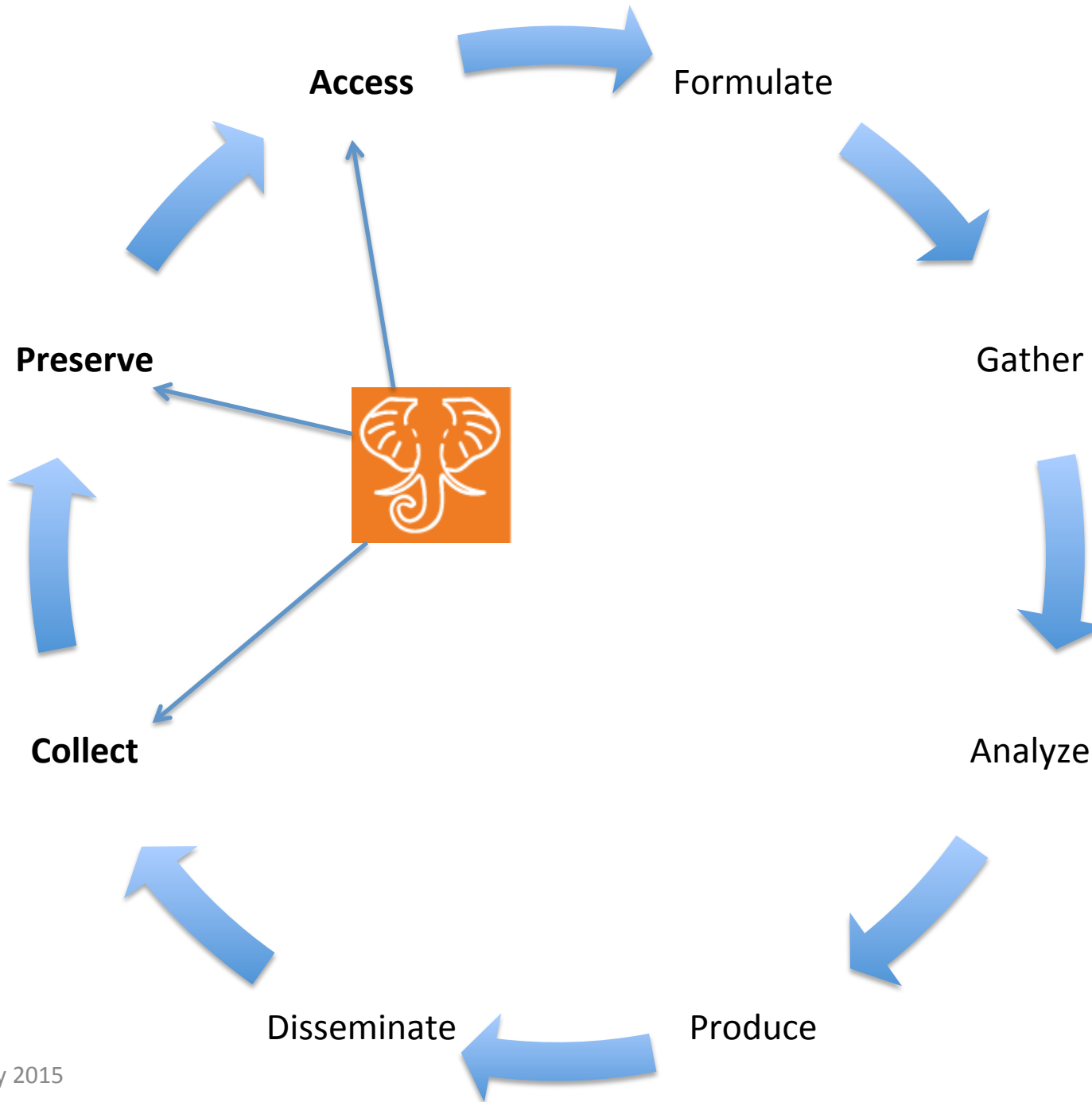
- ...enabling access by users with print disabilities.

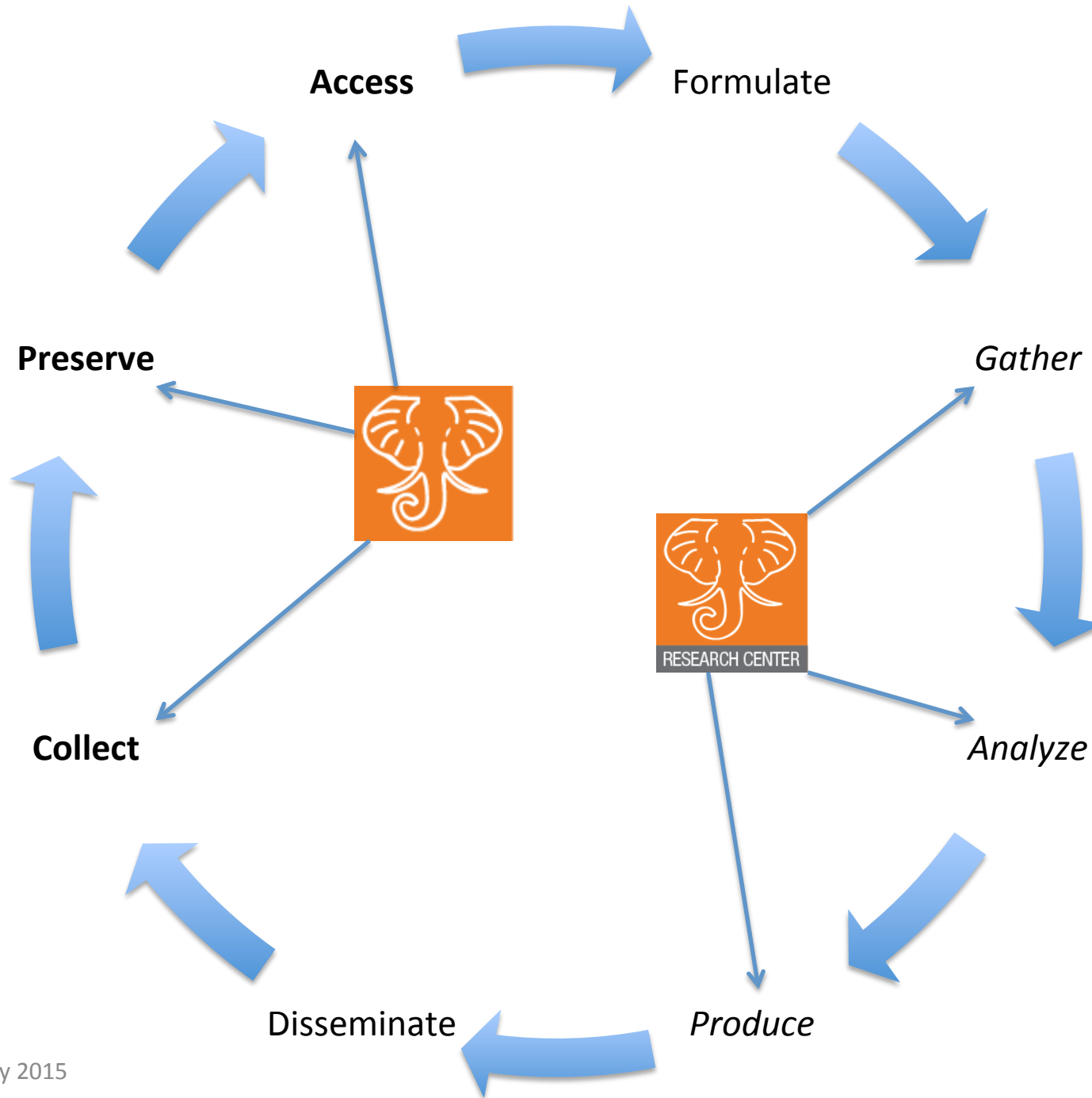
- ...supporting computational research with the collections.

- ...stimulating shared collection storage strategies among libraries.











DIGITAL PUBLIC LIBRARY
OF AMERICA

Discovery

Dissemination



Preservation and Access

Dark preservation



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
California Digital Library
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
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



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)



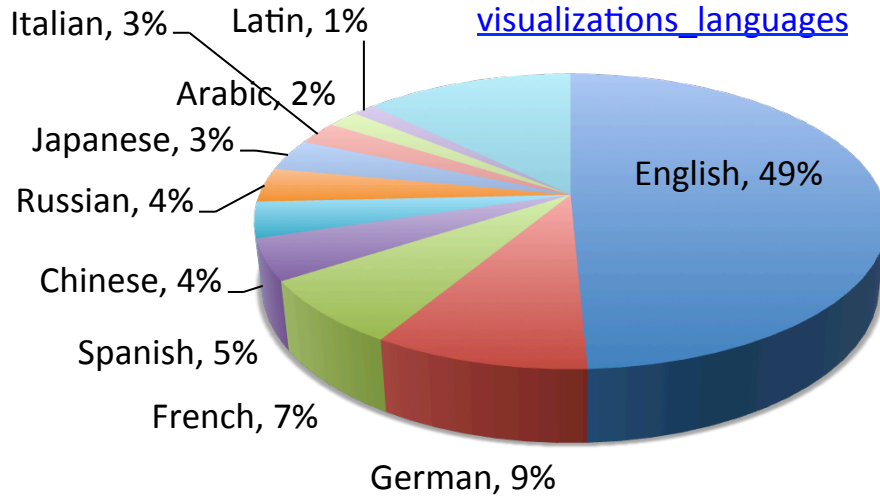
HathiTrust Collections in 2015

- 13 million total items
 - 6.7 million book titles
 - 345,000 serial titles
 - 604,000 US federal government documents
 - 4.9 million items open (public domain & CC-licenses)
 - A handful of images and thimbleful of audio files

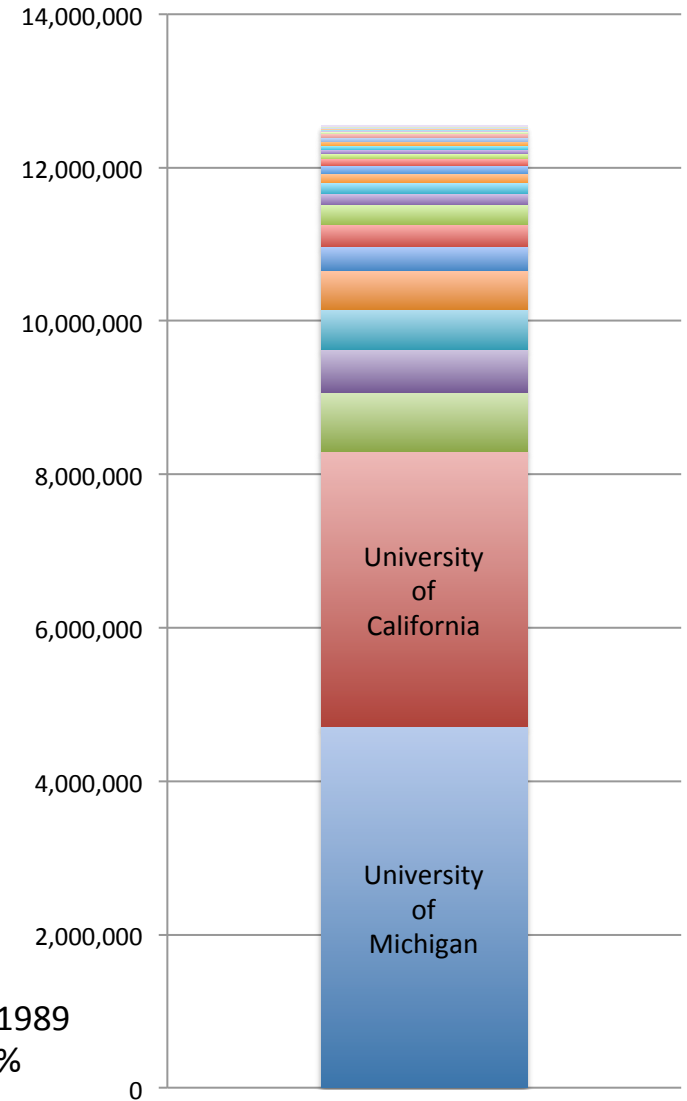
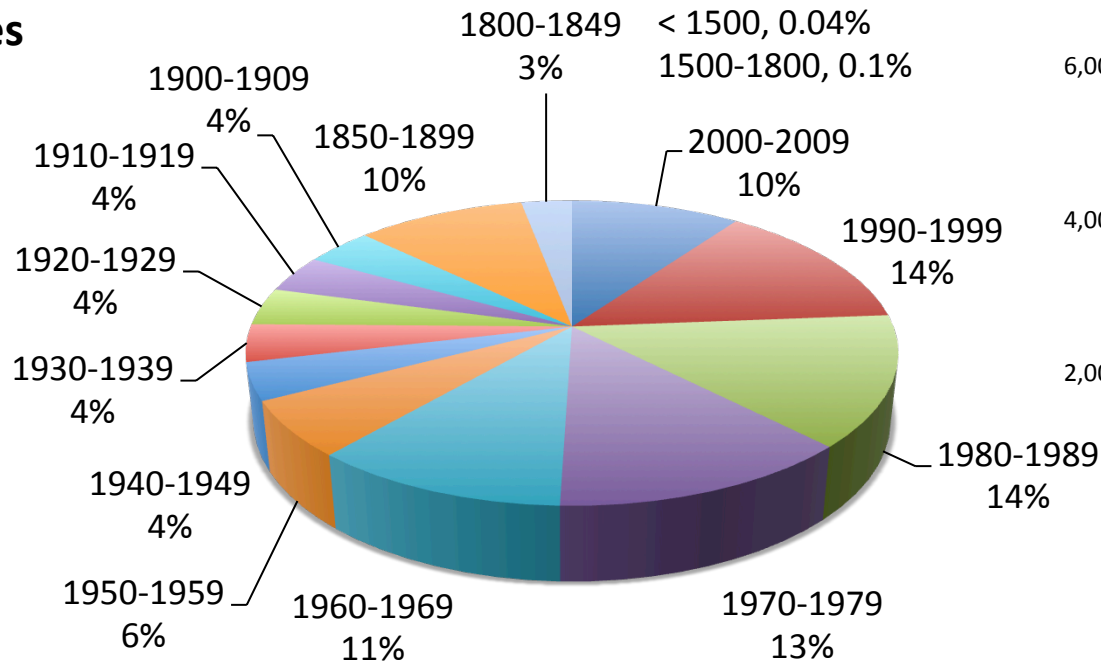


Top 10 Languages

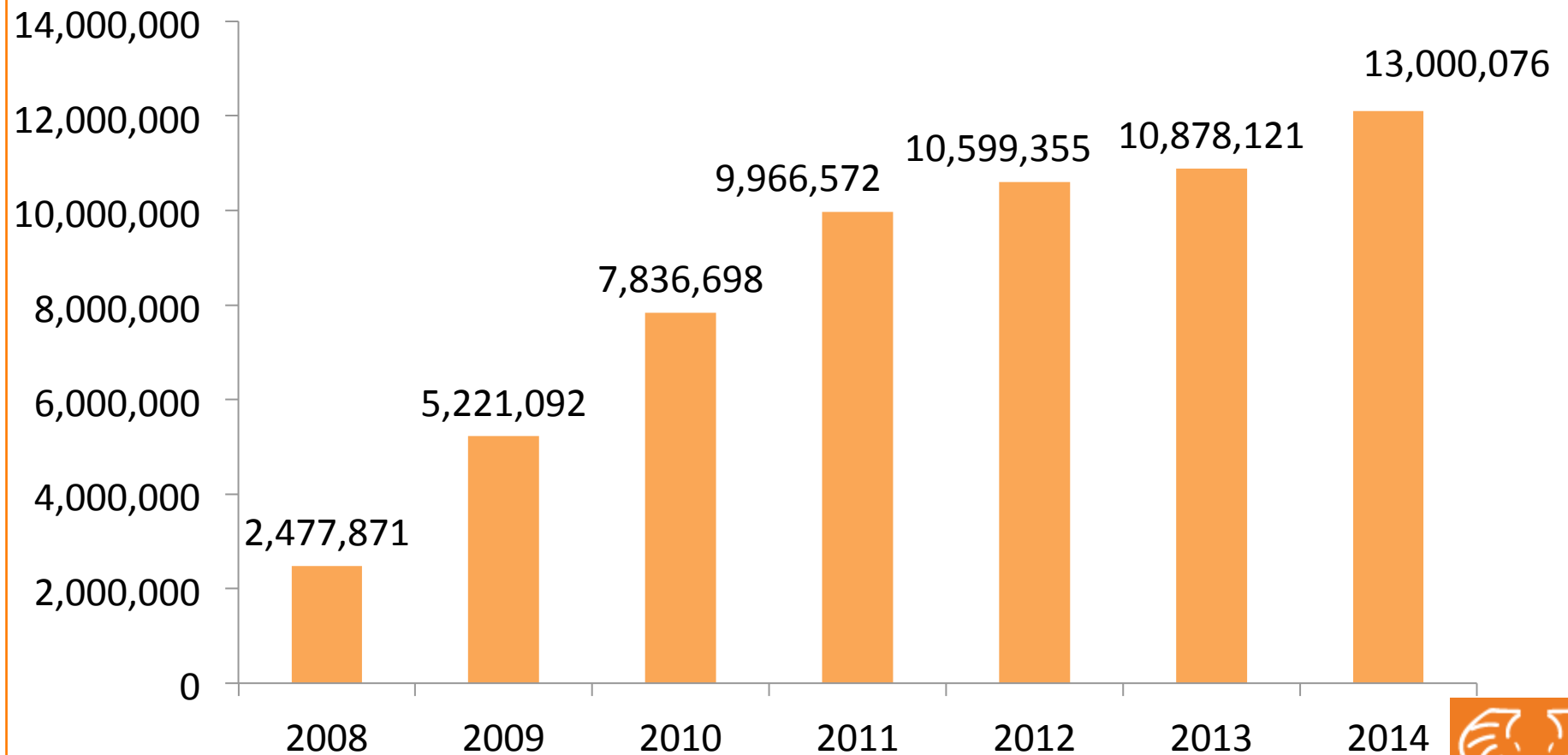
http://www.hathitrust.org/visualizations_languages



Dates



Growth of Collection



Largest Contributors (end 2014)

Volumes Added	Total Volumes
University of Michigan	4,712,752
University of California*	3,612,596
Harvard University	838,110
University of Wisconsin	560,775
Indiana University	528,811
Cornell University	510,065
Penn State University	387,717
University of Illinois	318,131
New York Public Library	294,835
Princeton University	252,808

Davis:

Total: 39,823

Google: 36,194 (all public domain)

IA: 3629

NRLF:

Total: 2,600,753

Google: 2,530,463 - including
323,738 public domain

IA: 70,290

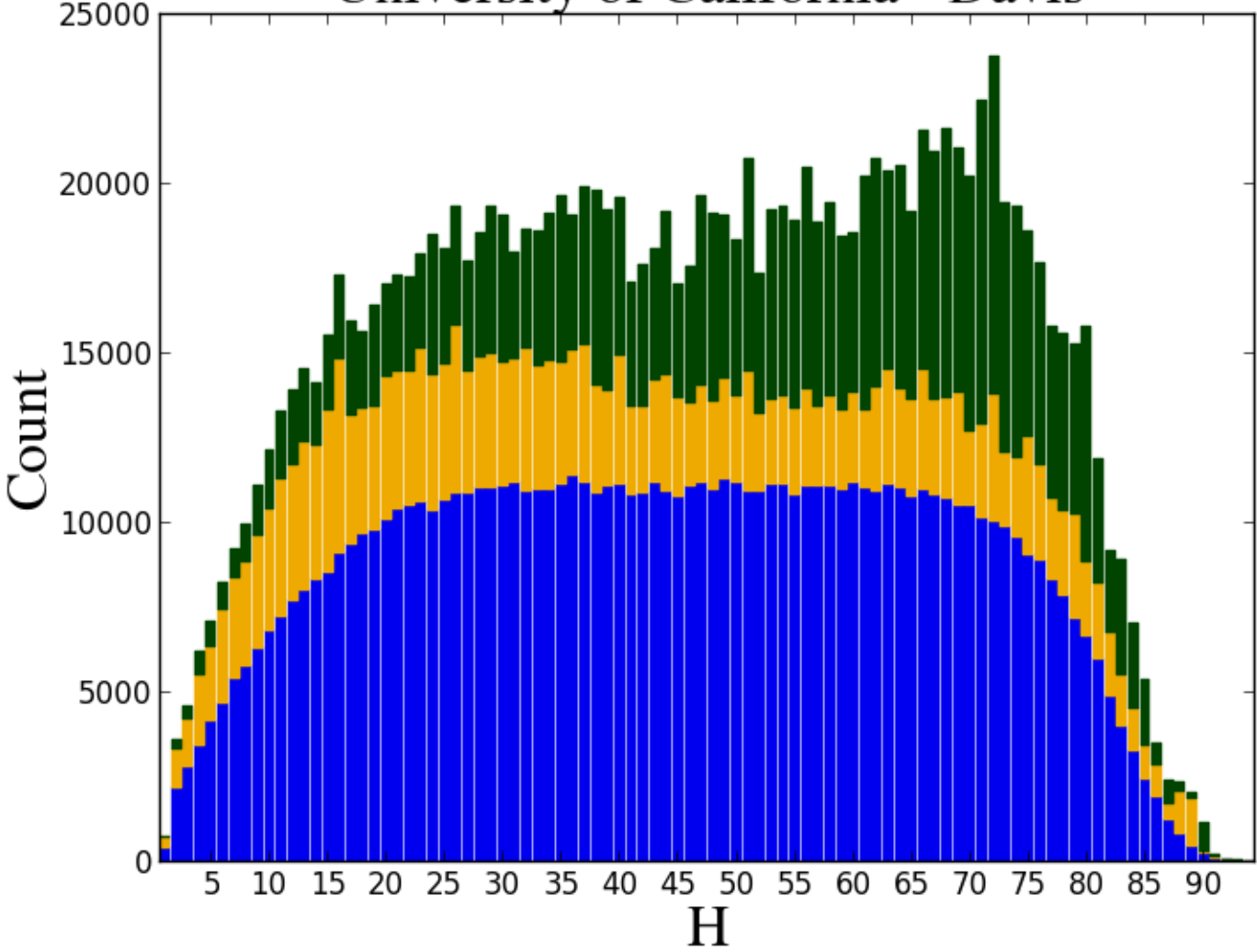


UC Collection Overlap (by titles Feb 2015)

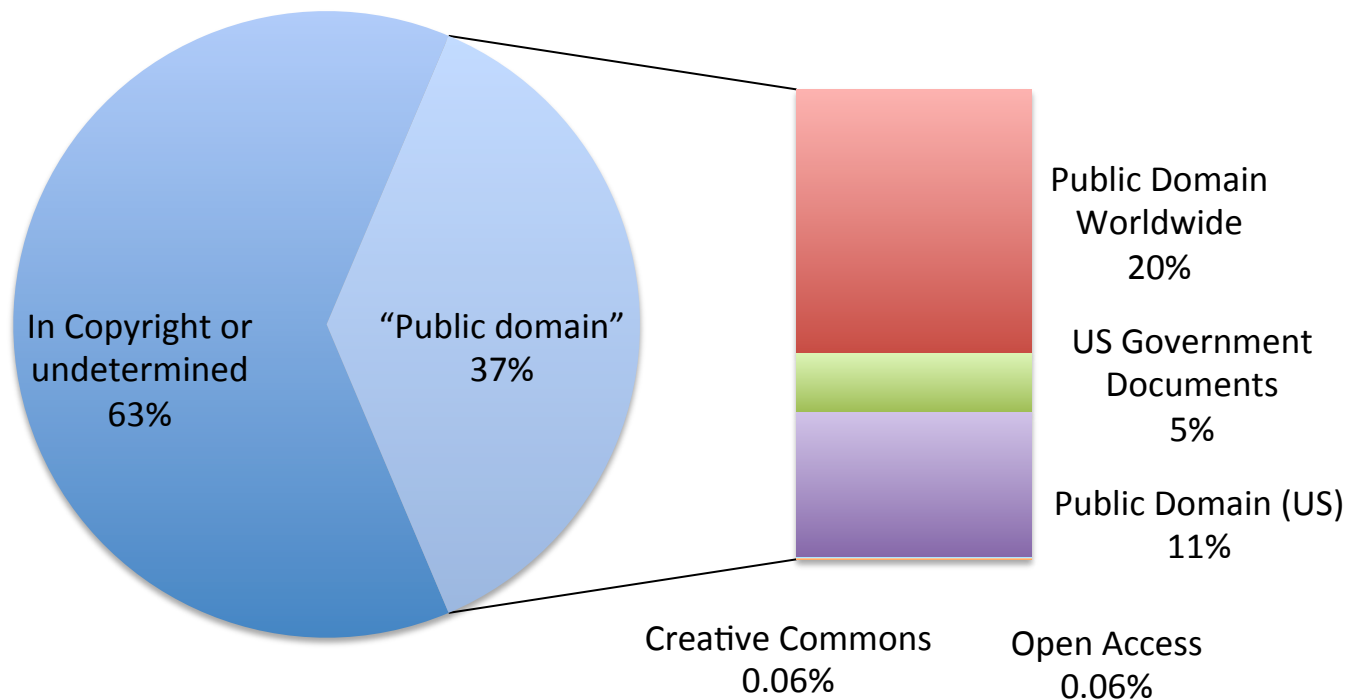
	Submitted	In Hathi	Percent
single-part monographs	9,766,951	3,332,926	34.1%
multi-part monographs	985,087	467,318	47.4%
serials	349,422	89,545	25.6%
TOTAL	11,101,460	3,889,789	35.0%



University of California - Davis



Copyright Distribution



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



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**
4. Expanding services for users with print disabilities
5. Maturing operations



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

- Many regional efforts, but limited national/international coordination
- Strengthens preservation commitments
 - Connects both print and digital preservation
- Significant need and desire to reduce costs of collection management and associated footprint

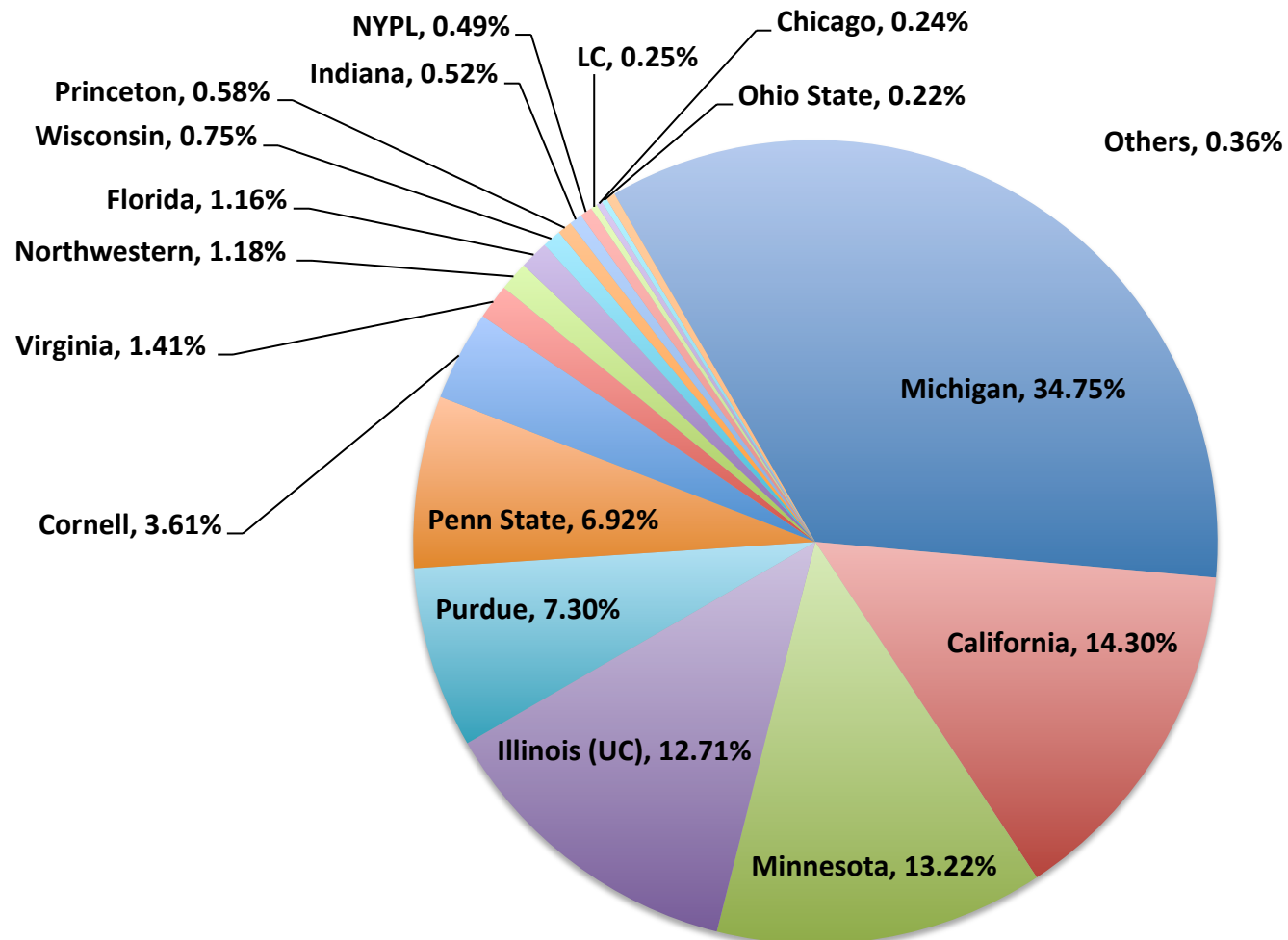


Government Documents Initiative

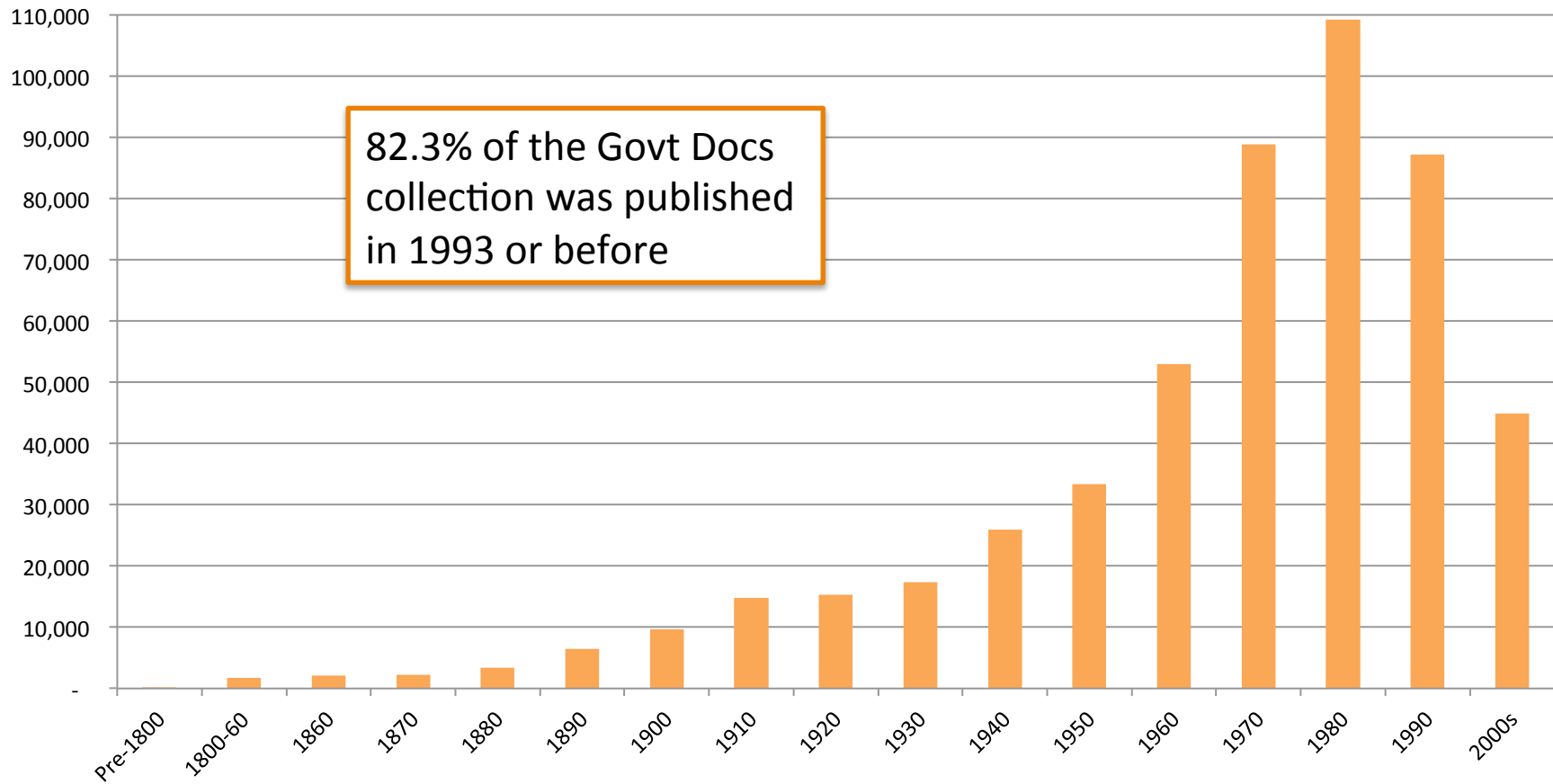
- Ballot Initiative: provide “expanded coverage & enhanced access to U.S. Government Documents.”
- Activities:
 - Developing a registry of US Federal Government Documents
 - Locate materials for inclusion in the collections
 - Improve search and discovery



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

- 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

- 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)



Datasets

- Non-Google-digitized Dataset (400,000+)
 - PD, PDUS, Open Access
 - Signed researcher statement
- Google-digitized (3.2 million+)
 - PD, PDUS, Open Access
 - Agreement between institution and Google
 - Brief proposal
 - Characterize texts
 - Provide ids (custom sets possible)
 - Research, results, use of results
 - Signed researcher statement





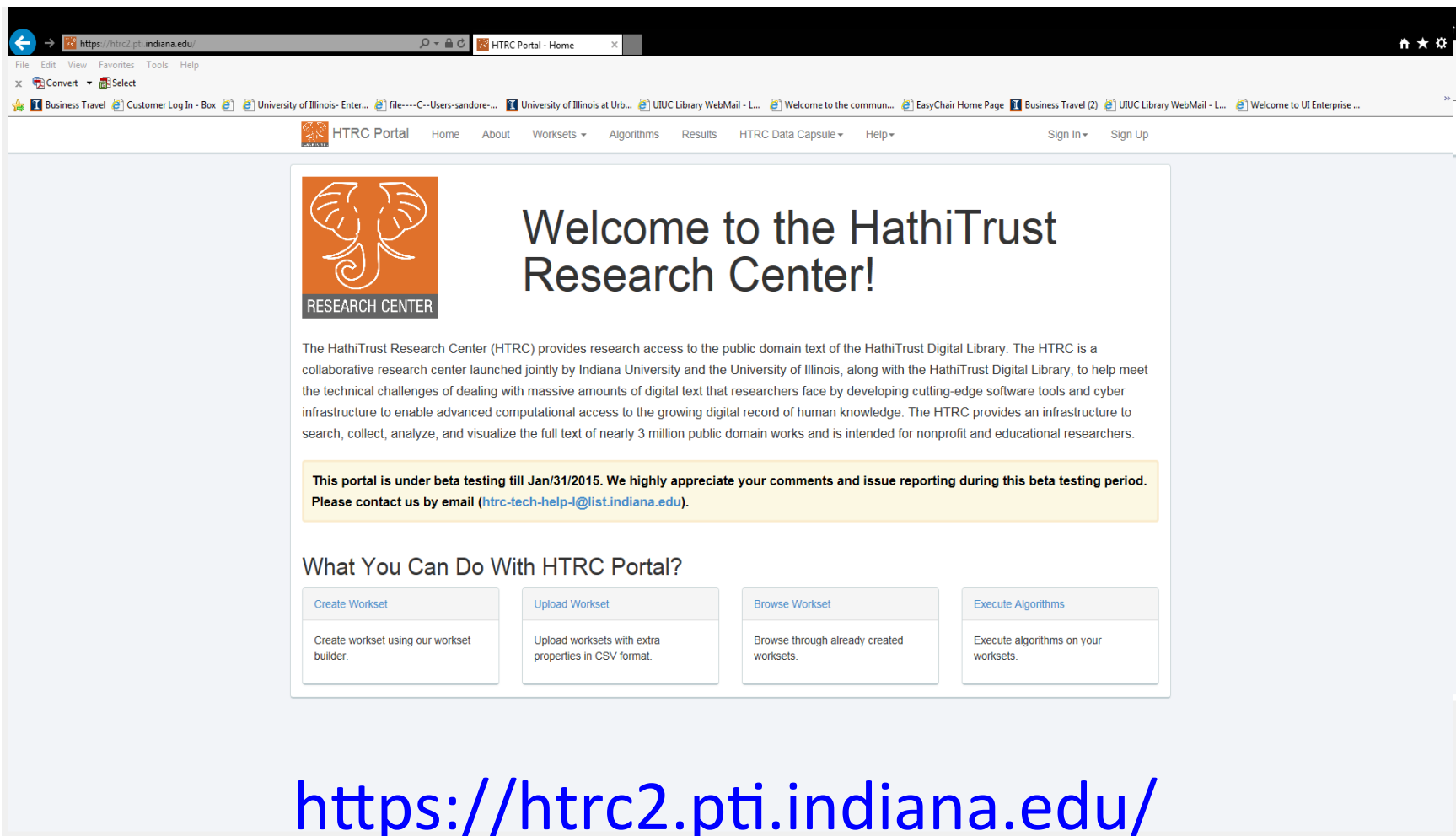
RESEARCH CENTER

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 Portal



The screenshot shows a web browser window with the URL <https://htrc2.pti.indiana.edu>. The page features a navigation bar with links for Home, About, Worksets, Algorithms, Results, HTRC Data Capsule, and Help. There are also links for Sign In and Sign Up. The main content area includes the HTRC logo (an orange elephant head) and the text "Welcome to the HathiTrust Research Center!". Below this is a paragraph describing the center's mission. A yellow box contains a message about beta testing until Jan/31/2015 and provides an email contact: htrc-tech-help-l@list.indiana.edu. At the bottom, there is a section titled "What You Can Do With HTRC Portal?" with four buttons: "Create Workset", "Upload Workset", "Browse Workset", and "Execute Algorithms".

<https://htrc2.pti.indiana.edu/>



HTRC system



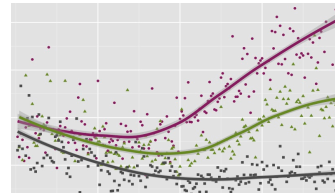
Complexity hiding interface



Request



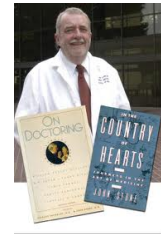
Spatial plots



Statistical plots

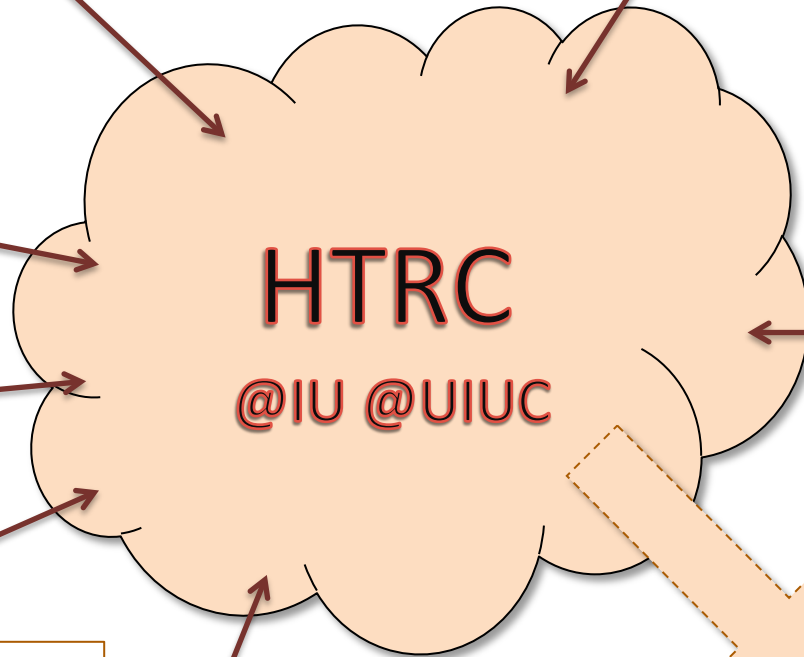
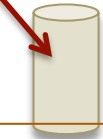
	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
Request	1,234,567	1,345,678	1,456,789	1,567,890	1,678,901	1,789,012	1,890,123	1,901,234	2,012,345	2,123,456
...
Total	12,345,678	13,456,789	14,567,890	15,678,901	16,789,012	17,890,123	18,901,234	19,012,345	20,123,456	21,234,567

Tabular info

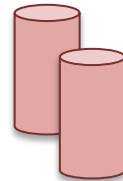




TEXT MINING TOOLS

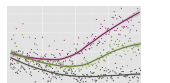
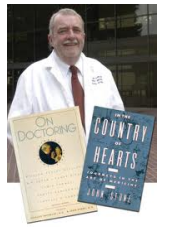


EXTRACTED FEATURE SETS



OTHER TEXT, E.G., DICTIONARIES, WIKI, TWITTER

Complexity hiding interface



BLUE WATERS



Workset builder

The screenshot shows the HTRC Workset Builder web application. The browser address bar displays the URL: <https://htrc2.pti.indiana.edu/blacklight/catalog?&author=alain-fournier&defType=...>. The page header includes a navigation bar with links for "Log Out [sandore]", "Selected Items (0)", "Manage Worksets", and "Portal". The main content area is divided into two columns. The left column, titled "Limit your search", lists various search filters such as "Subject" and "Author", with the "Author" filter expanded to show a list of authors and their associated item counts. The right column features a search interface with a search box, a dropdown menu set to "Full Text", and a "Search" button. Below the search box, there are "More options" and a list of search criteria: "Title > le grand meaulnes" and "Author > alain-fournier". The search results section displays "Displaying items 1 - 10 of 485" and includes a "start over" button. The results are sorted by "relevance" and shown in groups of 10 per page. The first result is "1. Le grand Meaulnes.", which is selected. The details for this result are: Title: Le grand Meaulnes., Author: Alain-Fournier, 1886-1914., Format: Book, Language: French, and Published: 1913. A "25 February 2015" watermark is visible in the bottom left corner of the page.

Log Out [sandore] | Selected Items (0) | Manage Worksets | Portal

HTRC Workset Builder

RESEARCH CENTER

Limit your search

Subject

Author

- [Le Sage, Alain René, 1668-1747 \(336\)](#)
- [Fournier, Edouard, 1819-1880 \(36\)](#)
- [Smollett, T. 1721-1771 \(29\)](#)
- [Isla, José Francisco de, 1703-1781 \(25\)](#)
- [Omeval, d', -1766 \(21\)](#)
- [Van Laun, Henri, 1820-1896 \(18\)](#)
- [Lalauze, Adolphe \(14\)](#)
- [Carolet, d. 1739 \(13\)](#)
- [Audiffret, Pierre Hyacinthe Jacques J. B., b. 1773 \(12\)](#)
- [Fournier, François, \(Paris\) \(12\)](#)
- [Alemán, Mateo, 1547-1614 \(11\)](#)
- [Universidad Complutense \(Alcalá de Henares\) \(11\)](#)
- [Francia \(10\)](#)
- [Fuzelier, M. 1672-1752 \(10\)](#)
- [Peña y Marín, Evaristo \(10\)](#)
- [Camacho, Juan Francisco, 1817-1896 \(9\)](#)
- [Combe, William, 1742-1823 \(9\)](#)
- [Brady, John Henry \(8\)](#)
- [Roscoe, Thomas, 1791-1874 \(8\)](#)
- [Defauconpret, Auguste Jean Baptiste \(7\)](#)

in Full Text

More options

Title > le grand meaulnes x Author > alain-fournier x

Displaying items 1 - 10 of 485

Sort by relevance

« Previous 1 2 3 4 5 ... 48 49 Next »

Select items on page Deselect items on page Select all search items Deselect all search items

1. **Le grand Meaulnes.** Select

Title: Le grand Meaulnes.
Author: Alain-Fournier, 1886-1914.
Format: Book
Language: French
Published: 1913

25 February 2015

47

HTRC Data Capsule:

Secure access to copyrighted materials

Secure computing framework that:

- Trusts that researcher will not deliberately leak repository data, but
- Prevents malware acting on user's behalf from leaking data.

Secure support for:

- **Non-consumptive use:** framework for safe handling of large volumes of protected data
- **Openness:** supports user-contributed analysis tools
- **Efficiency:** supports user-contributed analysis tools without lengthy prior review
- **Large-scale and low cost:** protections extend to large-scale national (public) supercomputers



HTRC DataCapsule: Secure Access

Run all the demo codes there by clicking on "Cell" -> "Run All"

```
In [1]: # importing necessary libraries #  
from vsm.corpus import Corpus  
from vsm.model.ldacgsmulti import LdaCgsMulti as LDA  
from vsm.viewer.ldagibbsviewer import LDAGibbsViewer  
  
In [3]: # Uploading a saved Corpus object.  
plain_dir = '/home/demouser/demo/vsm/'  
c = Corpus.load(plain_dir + 'uc2.ark+=13960=t5w66bs1h-nltk-freq3.npz')  
Loading corpus from /home/demouser/demo/vsm/uc2.ark+=13960=t5w66bs1h-nltk-freq3.npz  
  
In [7]: # Building an LDA model #  
# LDA model takes a Corpus object,  
# context type (what we want to consider as documents),  
# and number of topics, K.  
lda = LDA(c, 'page', K=20)  
  
In [8]: # Training an LDA model #  
# number of iterations and number of processors (with  
# the multi-processing model) could be specified.  
lda.train(itr=20, n_proc=5)  
Iteration 0: log prob=-1147.156238  
Iteration 1: log prob=-243161.382092  
Iteration 0: log prob=-1147.156238  
Iteration 1: log prob=-243161.382092
```



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. “EIEPHãT: 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).
- 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.



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?

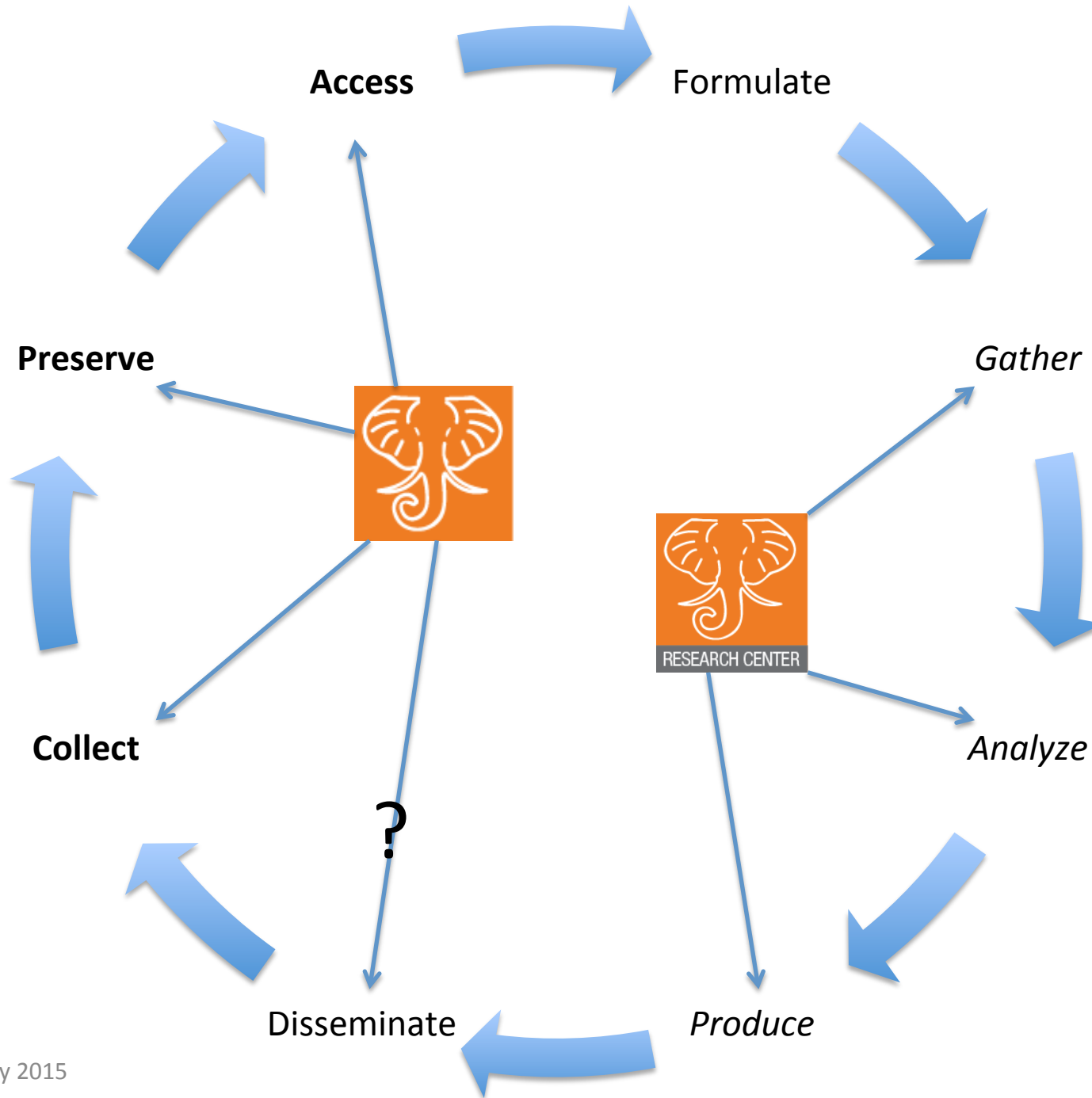
- Our mission, collection, and the repository operations are all strong.
- Our brand reputation is outstanding.
- Our work is solidly supported by the law.
- We have expanded access in unprecedented ways.
- The partnership provides a solid base for action.
- We have very important programs underway.



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: feedback@issues.hathitrust.org
- Blogs: <http://www.hathitrust.org/blogs>
 - Large-scale Search
 - Perspectives from HathiTrust



Thank you!

furlough@hathitrust.org
@MikeFurlough

