HATHITRUST

A Shared Digital Repository

HathiTrust: Putting Research in Context

HTRC UnCamp
September 10, 2012
John Wilkin, Executive Director, HathiTrust

Introduction

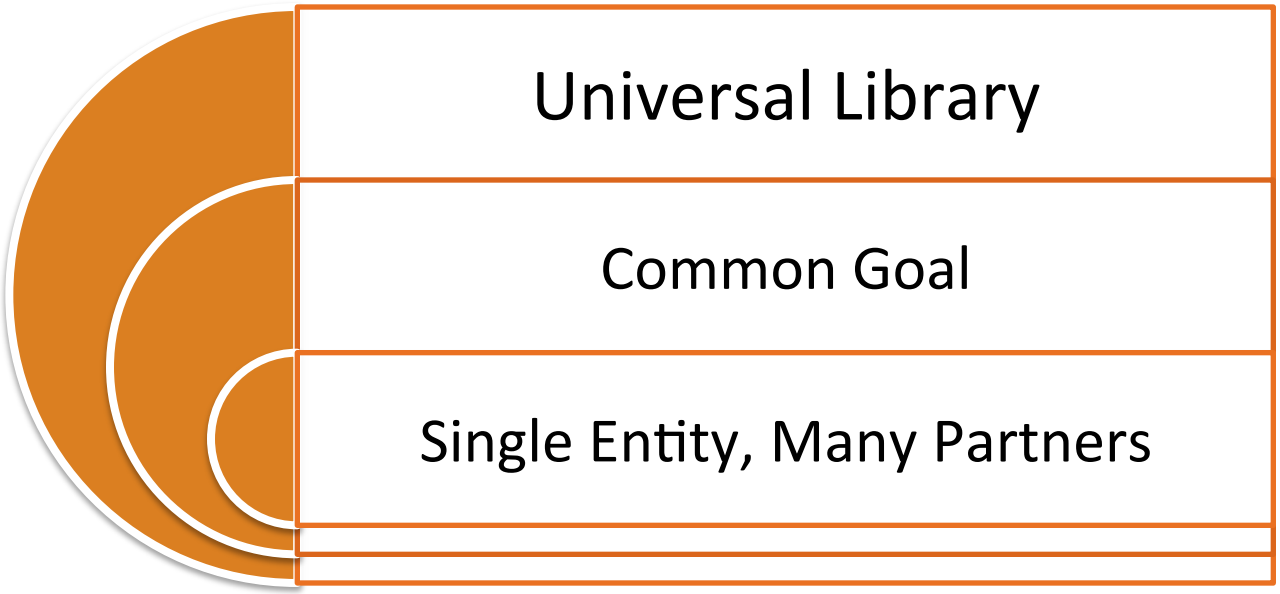
Partnership

Arizona State University
Baylor University
Boston College
Boston University
California Digital Library
Columbia University
Cornell University
Dartmouth College
Duke University
Emory University
Florida State University
Getty Research Institute
Harvard University Library
Indiana University
Johns Hopkins University
Lafayette College
Library of Congress
Massachusetts Institute of Technology
McGill University`
Michigan State University
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
Stanford University
Texas A&M University
Universidad Complutense de Madrid
University of Arizona
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 Illinois
University of Illinois at Chicago
The University of Iowa
University of Maryland
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 Pennsylvania
University of Pittsburgh
University of Utah
University of Virginia
University of Washington
University of Wisconsin-Madison
Utah State University
Washington University
Yale University Library

Mission

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

HathiTrust


Universal Library
Common Goal
Single Entity, Many Partners

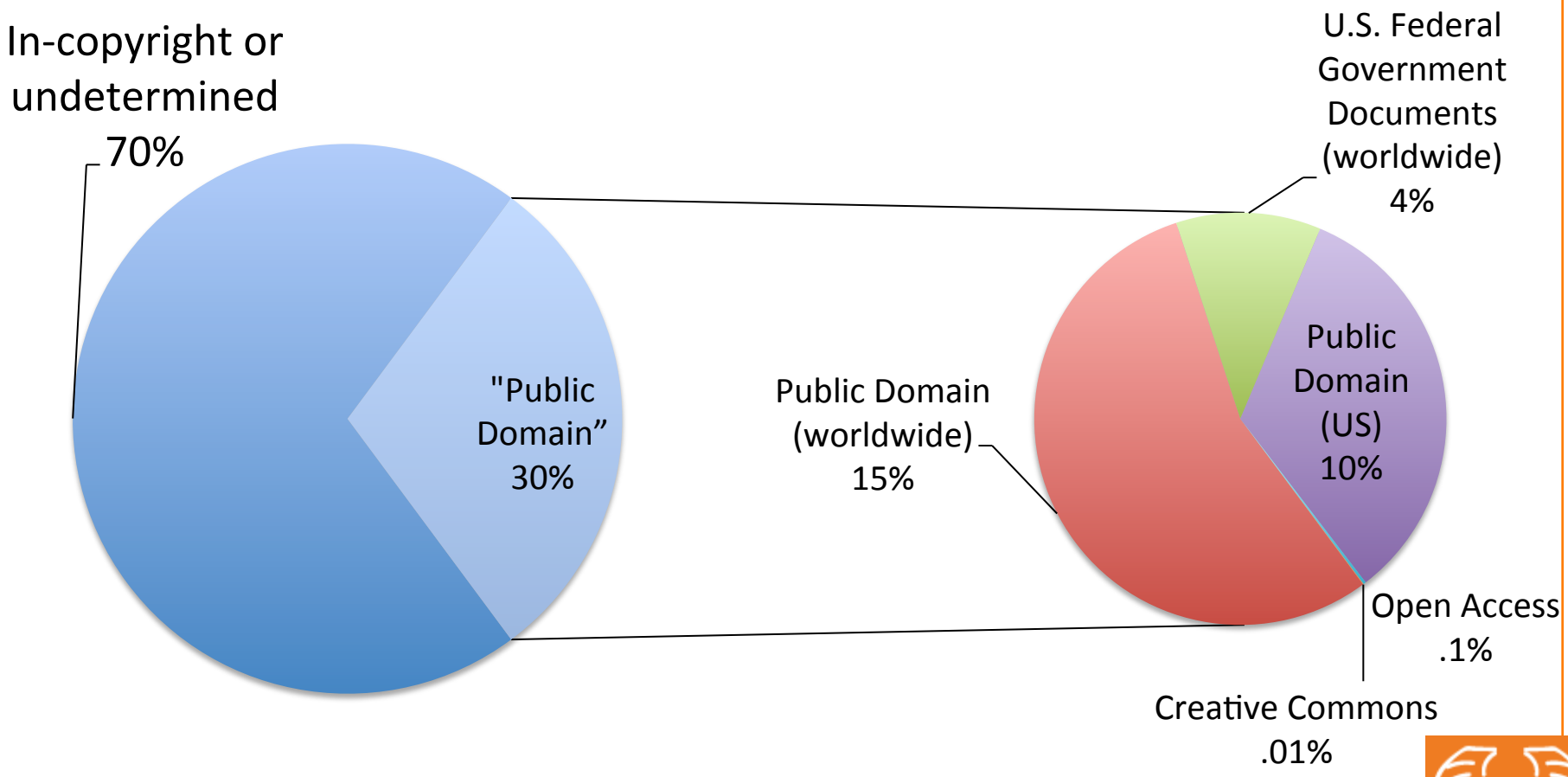
Digital Repository

- Launched 2008
- Initial focus on digitized book and journal content
  - 10.5 million total volumes
  - 5.5 million book titles
  - 270,000 serial titles
  - 3.2 million public domain (~30%)

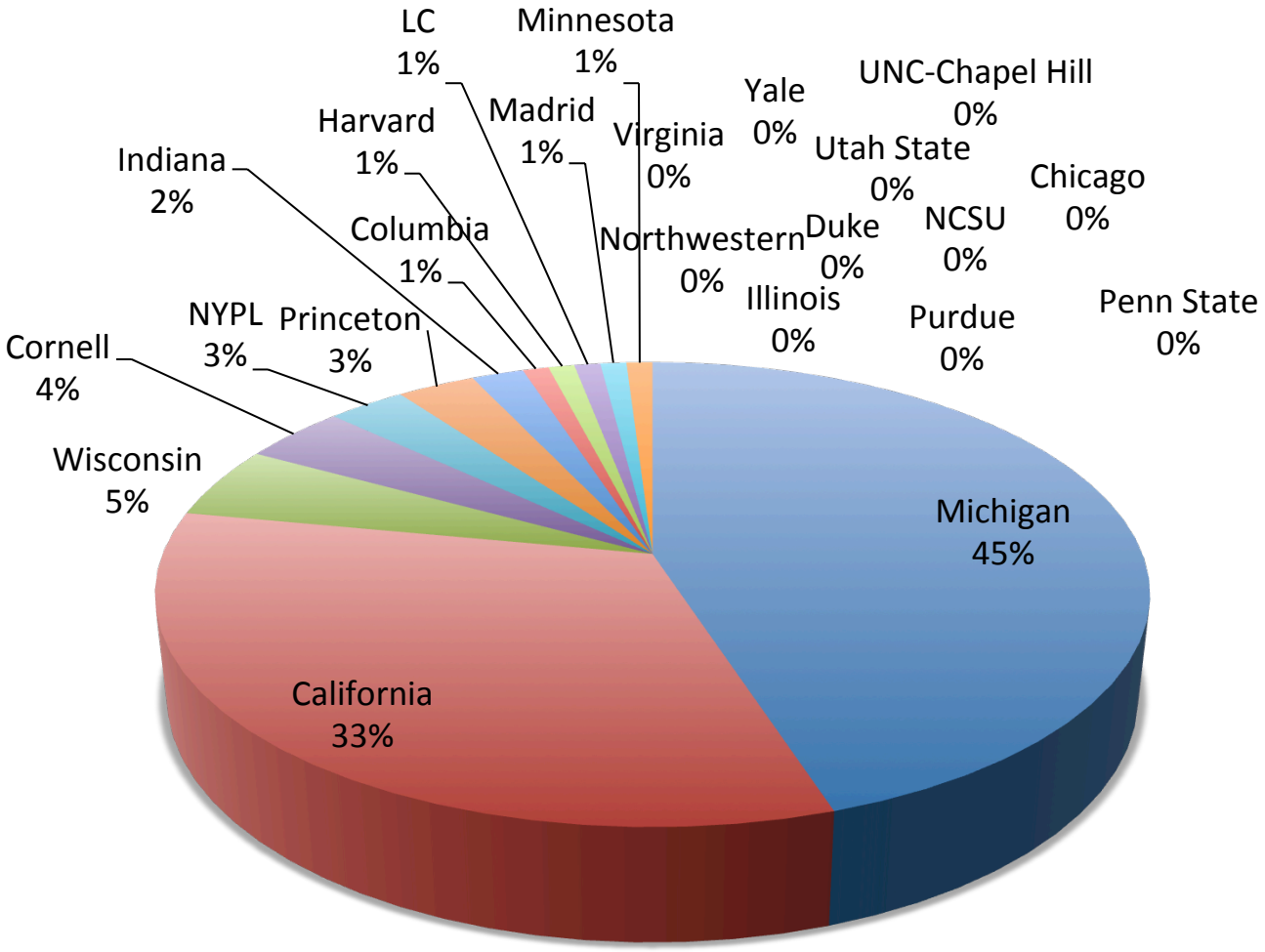
Goals

- Reliable and comprehensive archive of materials converted from print…co-owned
- Improve access …to meet the needs of the co-owning institutions
- Ensure the long-term preservation of content
- Coordinate shared storage strategies
- “public good” …sustaining the historical record
- Simultaneously …centralized …open

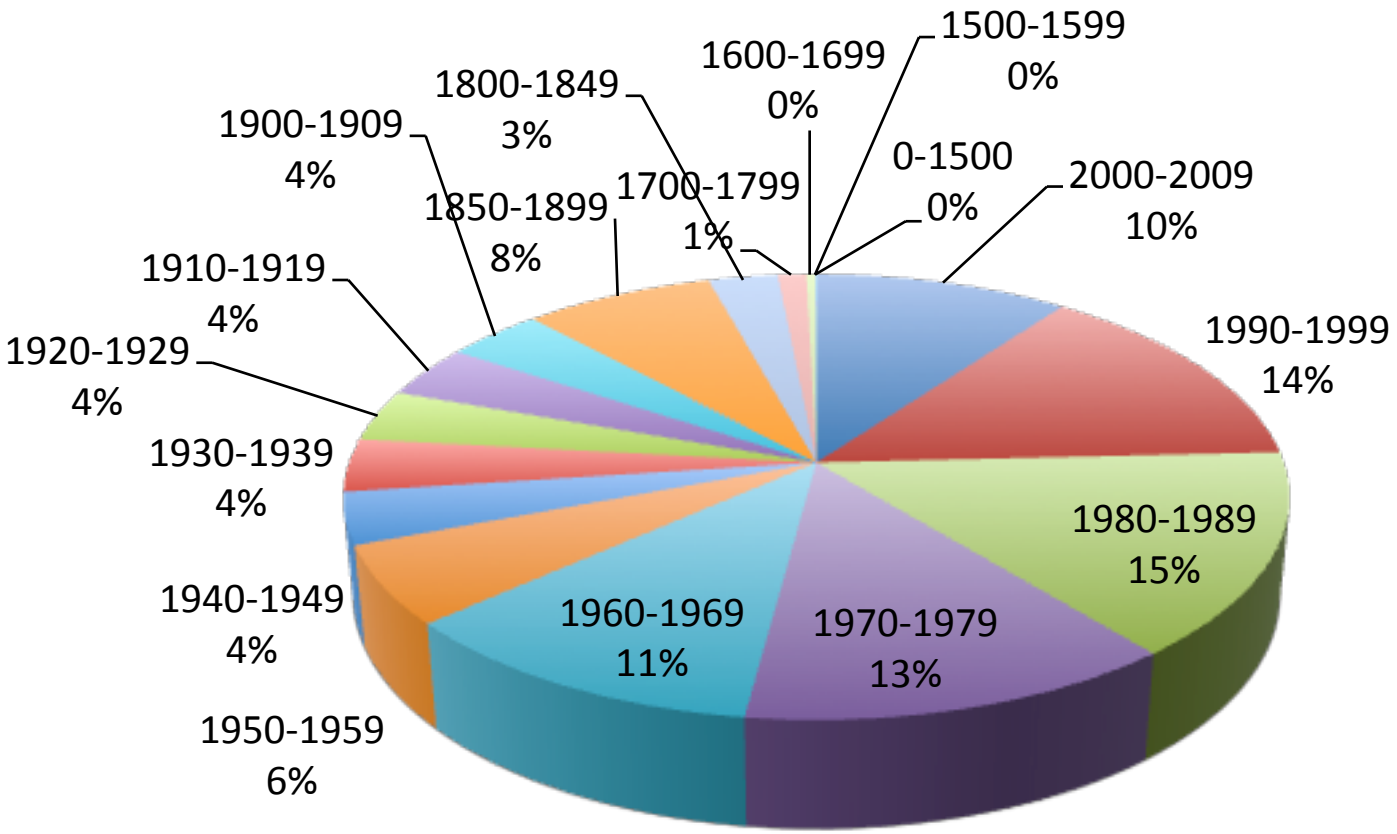
Content Distribution



Content Sources

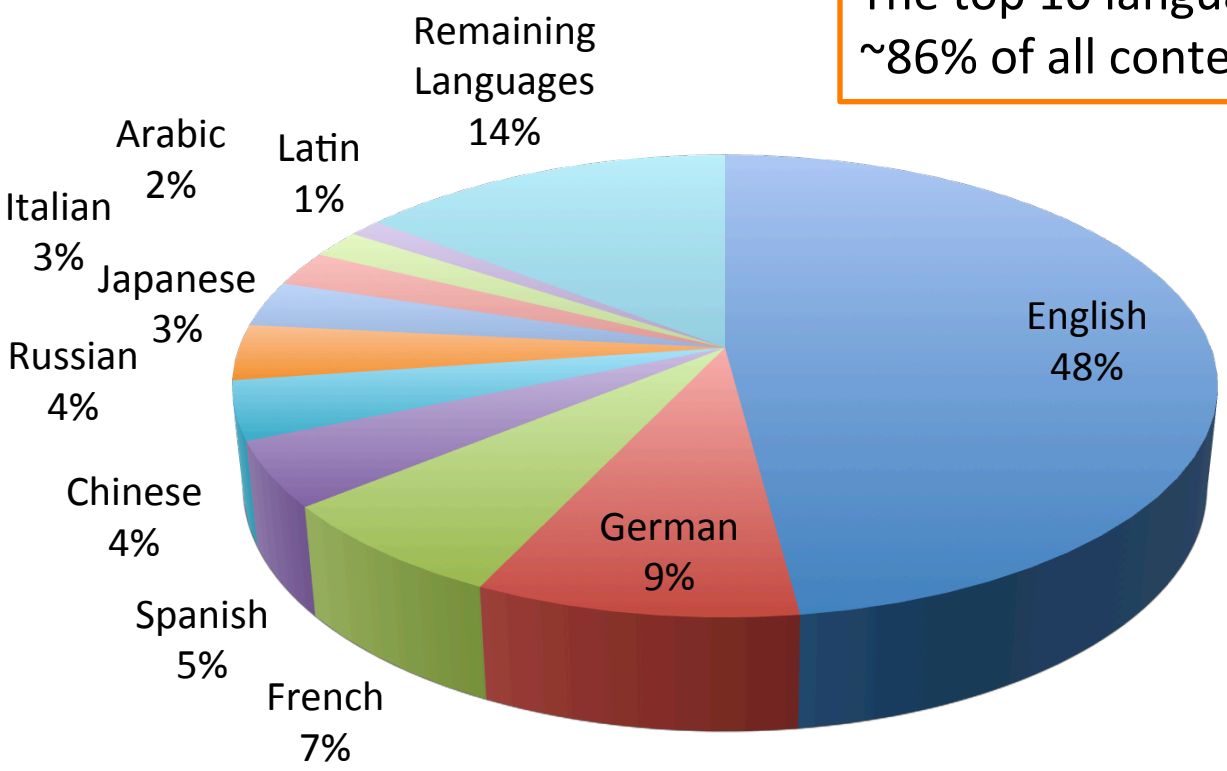


Dates

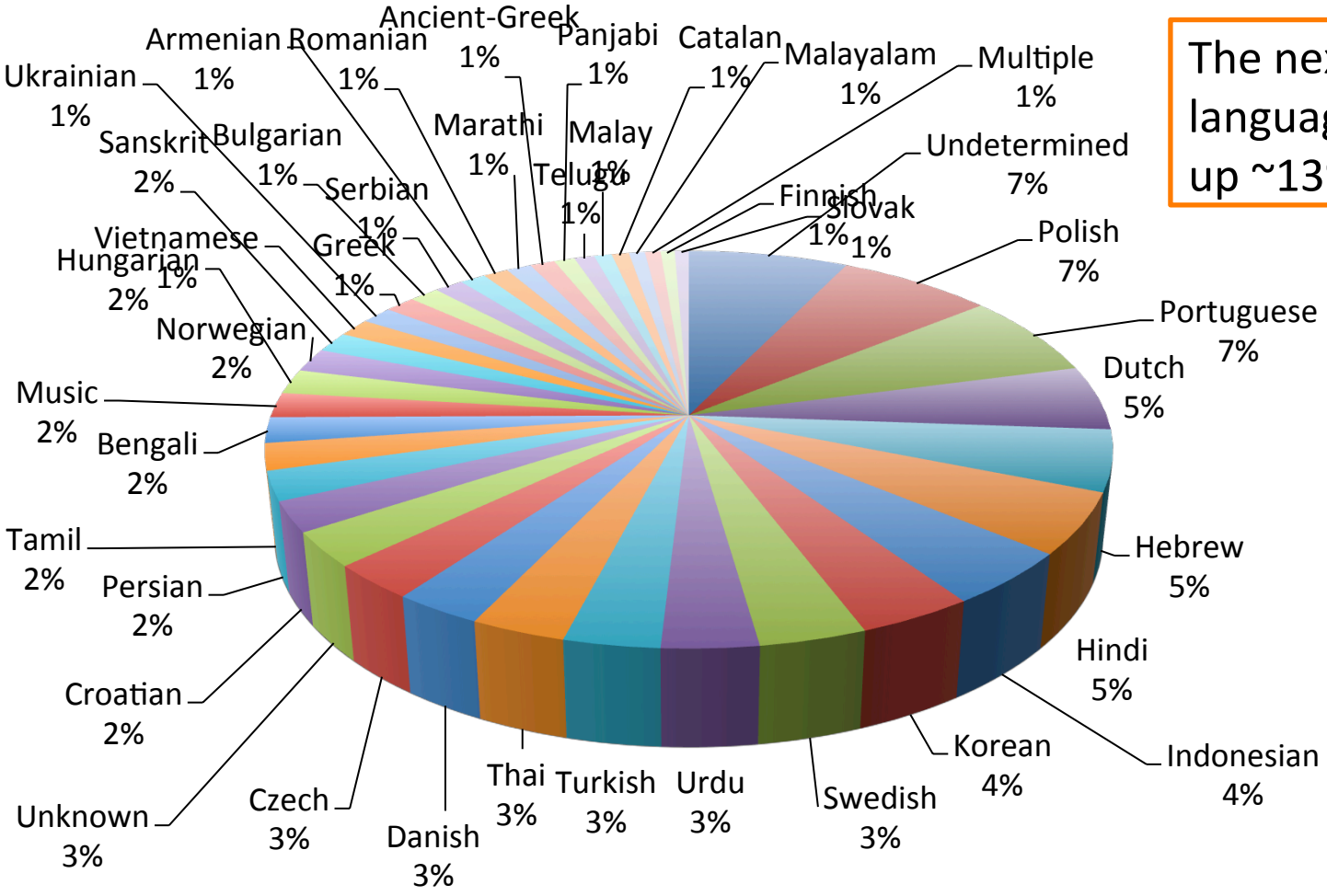

1500-1599
0%
1600-1699
0%
1800-1849
3%
1900-1909
4%
0-1500
0%
2000-2009
10%
1700-1799
1%
1850-1899
8%
1910-1919
4%
1990-1999
14%
1920-1929
4%
1930-1939
4%
1980-1989
15%
1940-1949
4%
1960-1969
11%
1970-1979
13%
1950-1959
6%

Language Distribution (1)

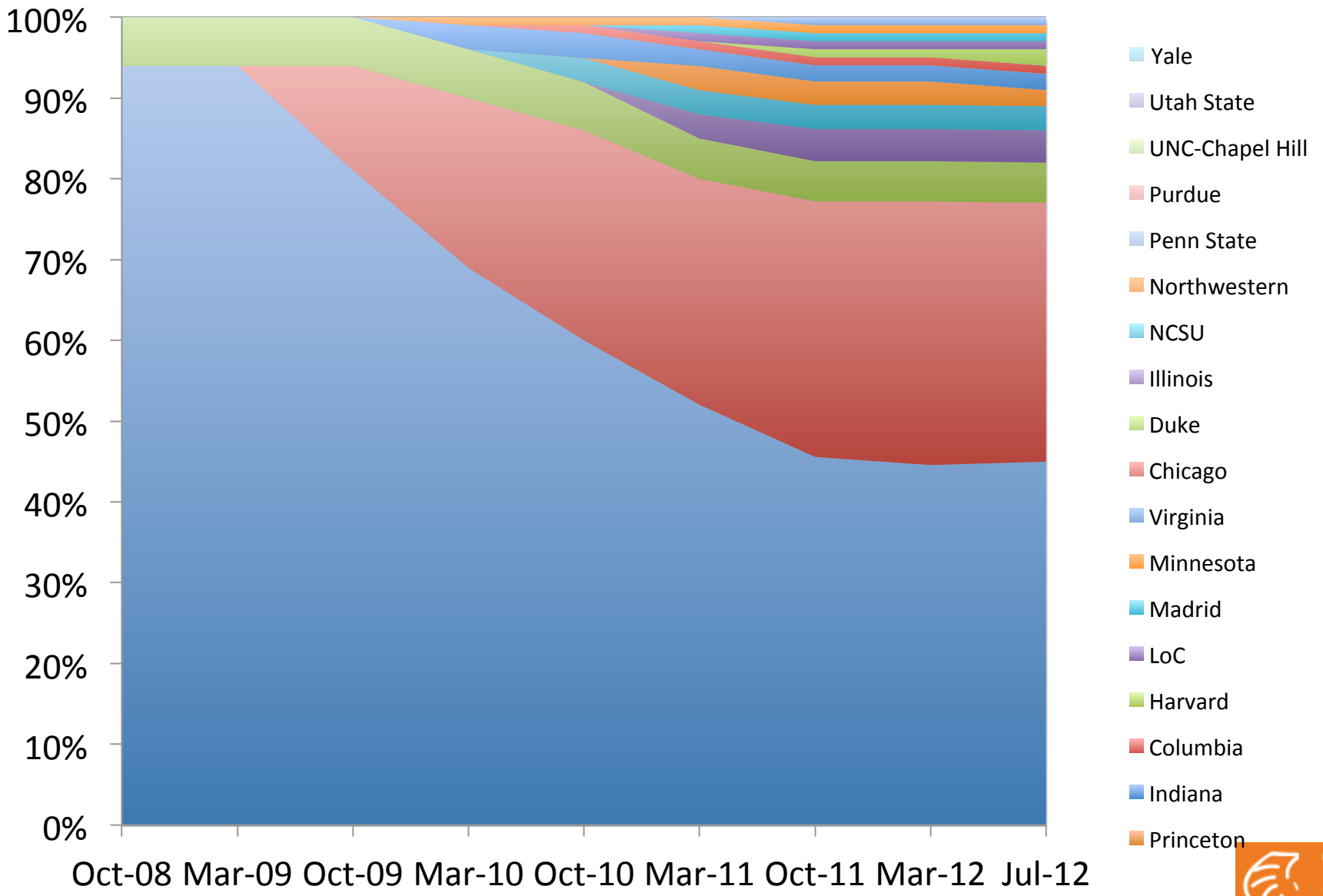
The top 10 languages make up ~86% of all content



Language Distribution (2)



The next 40 languages make up ~13% of total

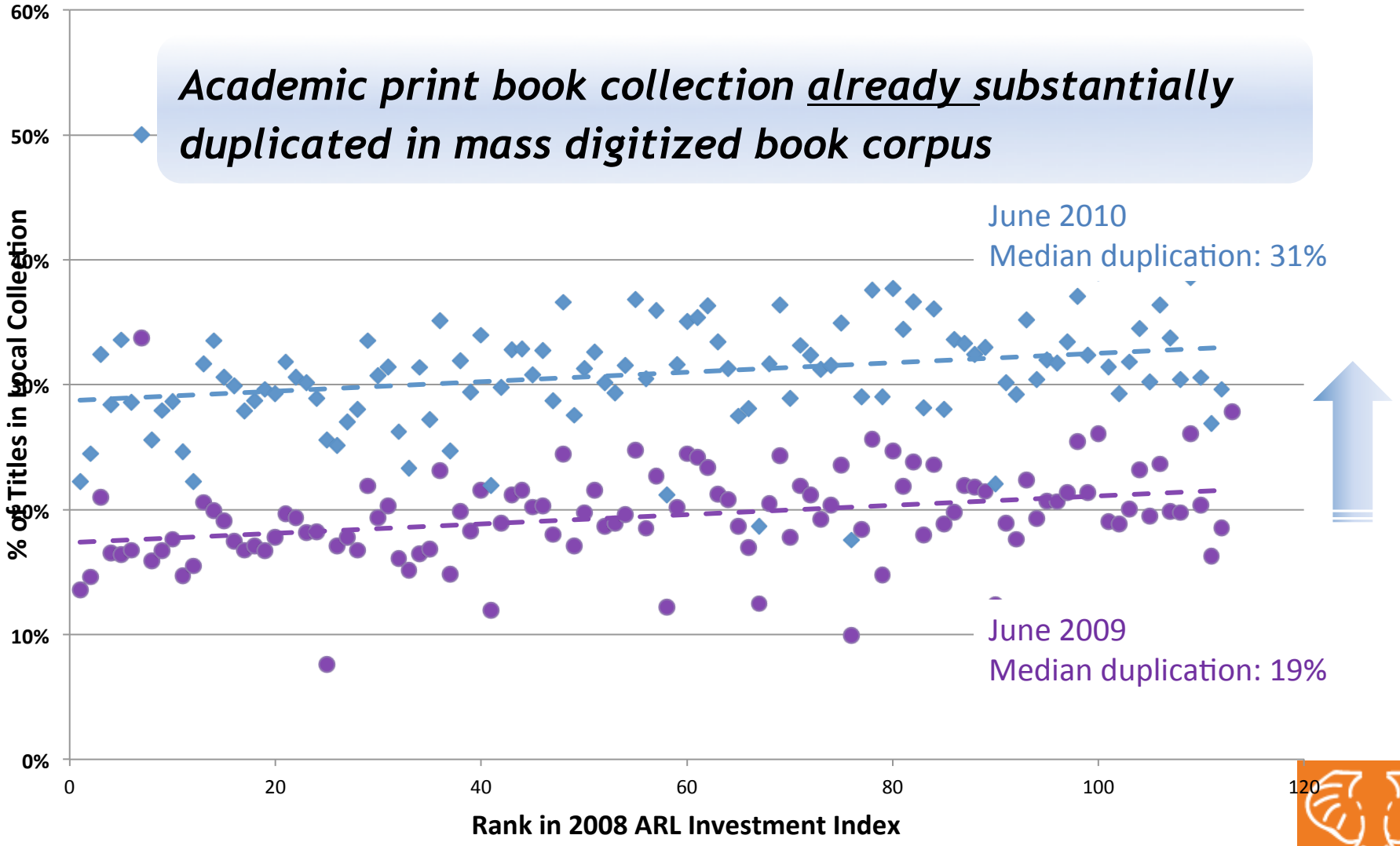

100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Oct-08
Mar-09
Oct-09
Mar-10
Oct-10
Mar-11
Oct-11
Mar-12
Jul-12
Yale
Utah State
UNC-Chapel Hill
Purdue
Penn State
Northwestern
NCSU
Illinois
Duke
Chicago
Virginia
Minnesota
Madrid
LoC
Harvard
Columbia
Indiana
Princeton

Services

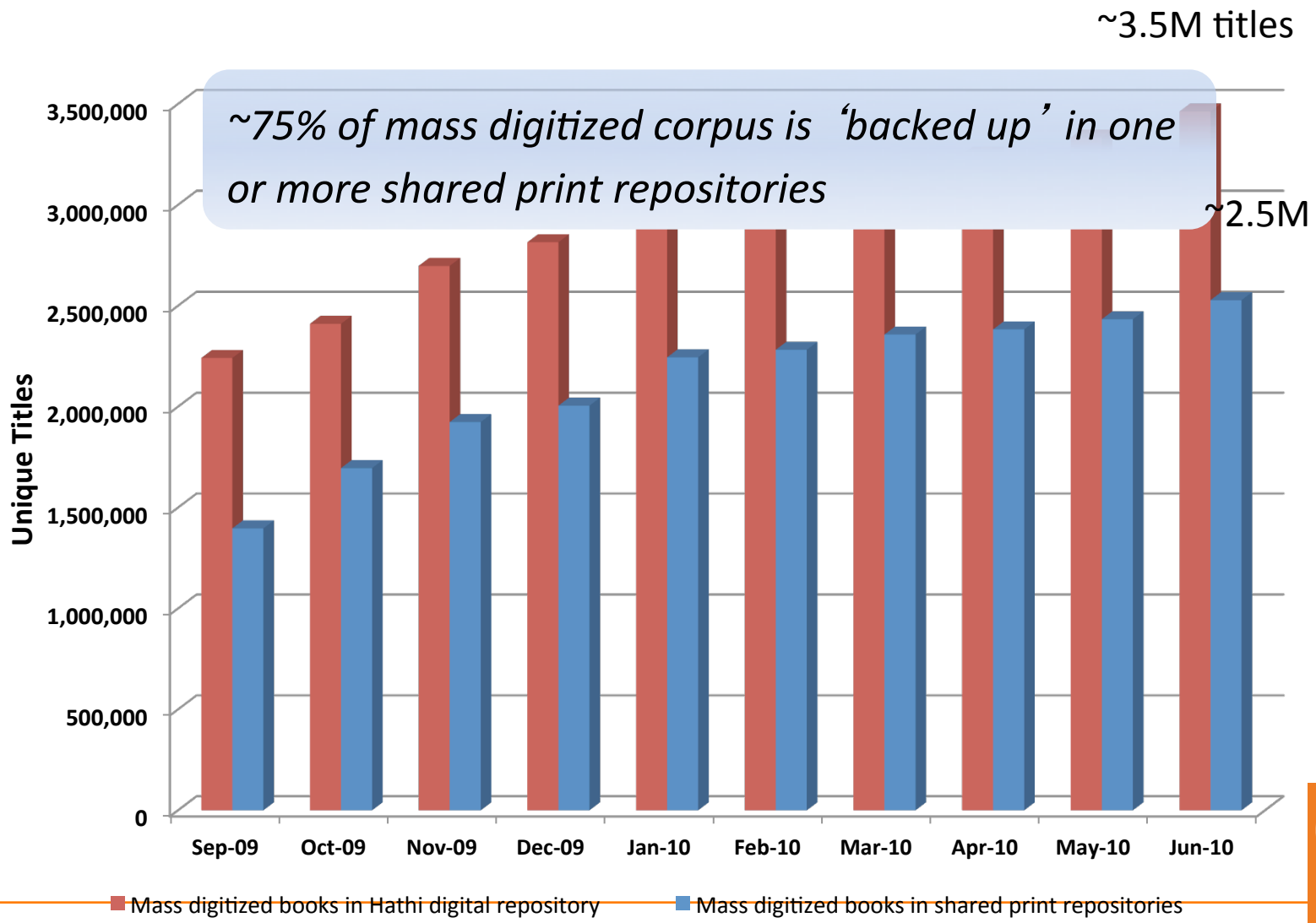
- Long-term preservation
  - Bit-level and migration
- Bibliographic search
- Full-text search
- Reading and download capabilities
- Print on demand
- Collections
- Datasets, Research Center

Impact

A global change in the library environment



Digitized Books in Shared Repositories



Collection Management, Development

- Overlap
  - More than 50% median overlap with ARL institutions; higher for small liberal arts colleges
- Pricing model based on Print holdings
  - Requires print holdings database
  - Also support expansion of legal uses, efforts in de-duplication
  - Facilitate individual and collaborative collection development and management operations
- Print monographs archiving

Discovery and Use

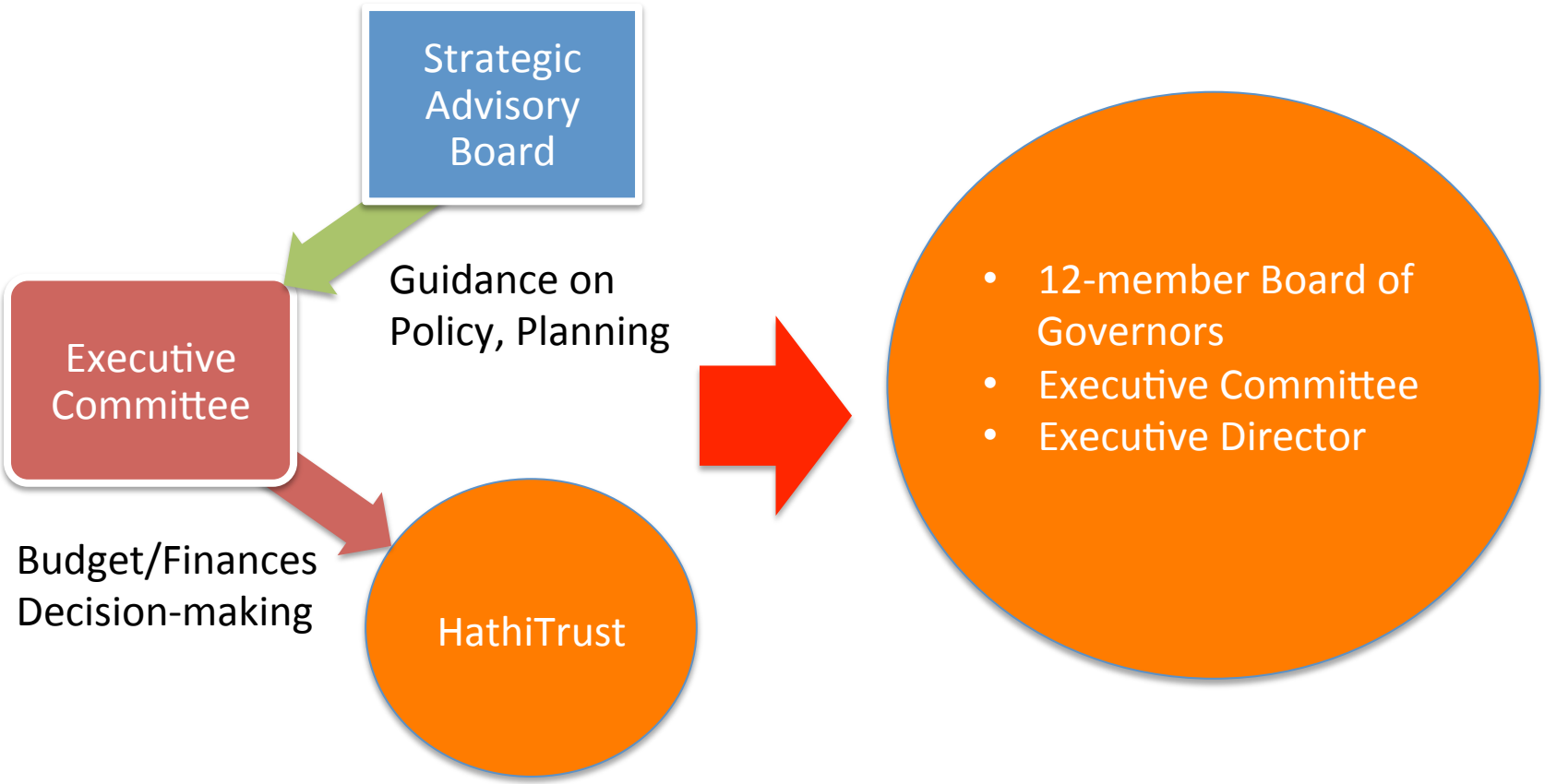
- Search, collections, online access
- APIs and data feeds
  - Data API
  - Bibliographic API
  - “Hathifiles” inventory files
  - OAI
- Computational Research
  - Distribution of datasets
  - Protocol-based access
  - Research Center

Research Center in Context

Institutional Support / Sustainability

Constitutional Convention

- October 2011
- 52 partners
- 3-year review overseen by SAB
- Ballot Proposals
  - Print monograph storage
  - Approval Process for development initiatives
  - U.S. Government Documents
  - Fee-for-service content deposit
  - Governance


Strategic Advisory Board
Guidance on Policy, Planning
Executive Committee
Budget/Finances Decision-making
HathiTrust
• 12-member Board of Governors
• Executive Committee
• Executive Director

Collaborative Support

- New pricing model
- Base infrastructure costs
  - Public domain
  - In-copyright/undetermined
- Funds for programmatic initiatives

The Future

Concluding thoughts

Thank you!