Citation Key

for more information see: http://open.umich.edu/wiki/CitationPolicy

Use + Share + Adapt

{ Content the copyright holder, author, or law permits you to use, share and adapt. }

Public Domain – Government: Works that are produced by the U.S. Government. (17 USC § 105)

Public Domain – Expired: Works that are no longer protected due to an expired copyright term.

Public Domain – Self Dedicated: Works that a copyright holder has dedicated to the public domain.

Creative Commons – Zero Waiver

Creative Commons – Attribution License

Creative Commons – Attribution Share Alike License

Creative Commons – Attribution Noncommercial License

Creative Commons – Attribution Noncommercial Share Alike License

GNU – Free Documentation License

Make Your Own Assessment

{ Content Open.Michigan believes can be used, shared, and adapted because it is ineligible for copyright. }

Public Domain – Ineligible: Works that are ineligible for copyright protection in the U.S. (17 USC § 102(b)) *laws in your jurisdiction may differ

{ Content Open.Michigan has used under a Fair Use determination. }

Fair Use: Use of works that is determined to be Fair consistent with the U.S. Copyright Act. (17 USC § 107) *laws in your jurisdiction may differ

Our determination **DOES NOT** mean that all uses of this 3rd-party content are Fair Uses and we **DO NOT** guarantee that your use of the content is Fair.

To use this content you should **do your own independent analysis** to determine whether or not your use will be Fair.

HATHI TRUST

A Shared Digital Repository

Building A Future By Preserving Our Past

The Preservation Infrastructure of HathiTrust Digital Library

Jeremy York

IFLA 2010

August 15, 2010

Current Partners

- Columbia University
- New York Public Library
- University of California system
- CIC (Committee on Institutional Cooperation)
  - University of Chicago
  - University of Illinois
  - Indiana University
  - University of Iowa
  - University of Michigan
  - Michigan State University
  - University of Minnesota
  - Northwestern University
  - Ohio State University
  - Pennsylvania State University
  - Purdue University
  - University of Wisconsin-Madison
- University of Virginia
- Yale University

Mission

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

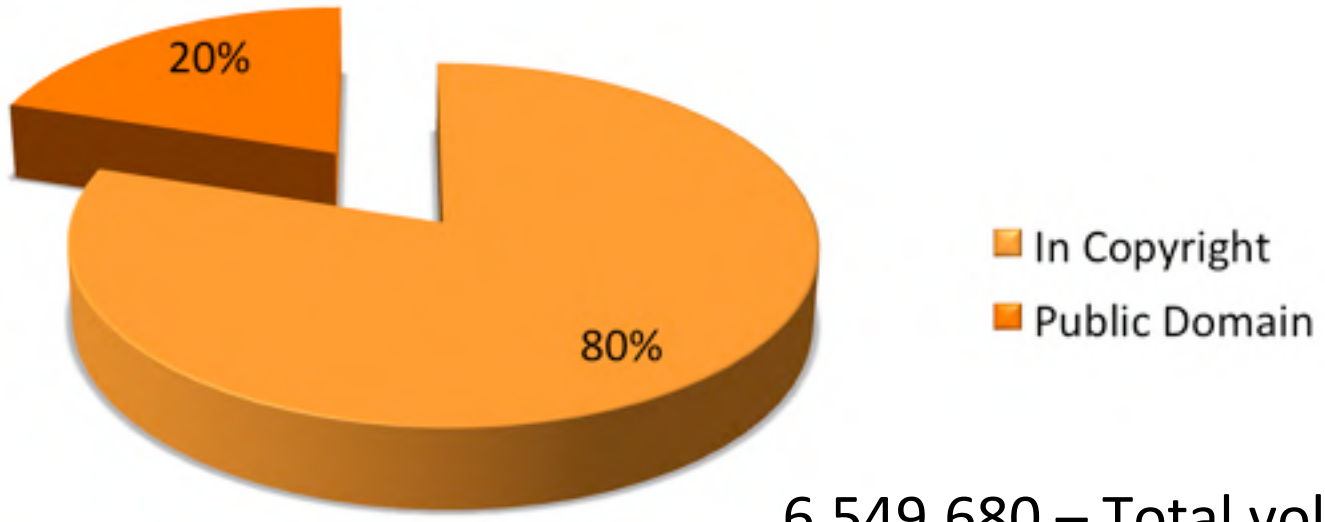
HathiTrust


Universal Library
Common Goal
Single Entity, Many Partners

Goals

- Comprehensive collection
- Preservation…with Access
- Shared strategies
  - Collection management, development
  - Preservation
  - Copyright
  - Efficient user services
- Openness

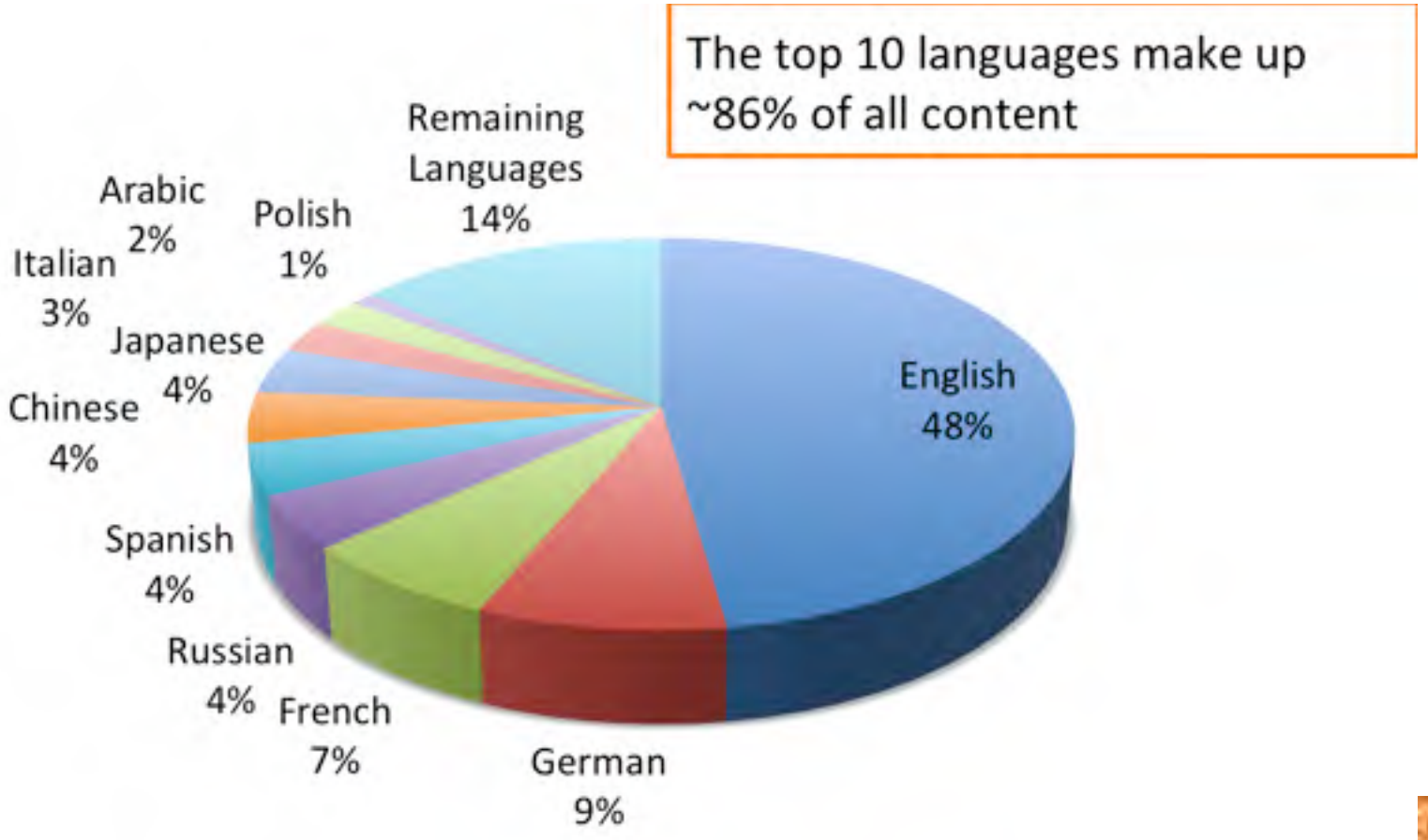
Content Distribution



6,549,680 – Total volumes
1,300,896 – Public Domain
3,798,116 Book titles
153,311 Serial titles

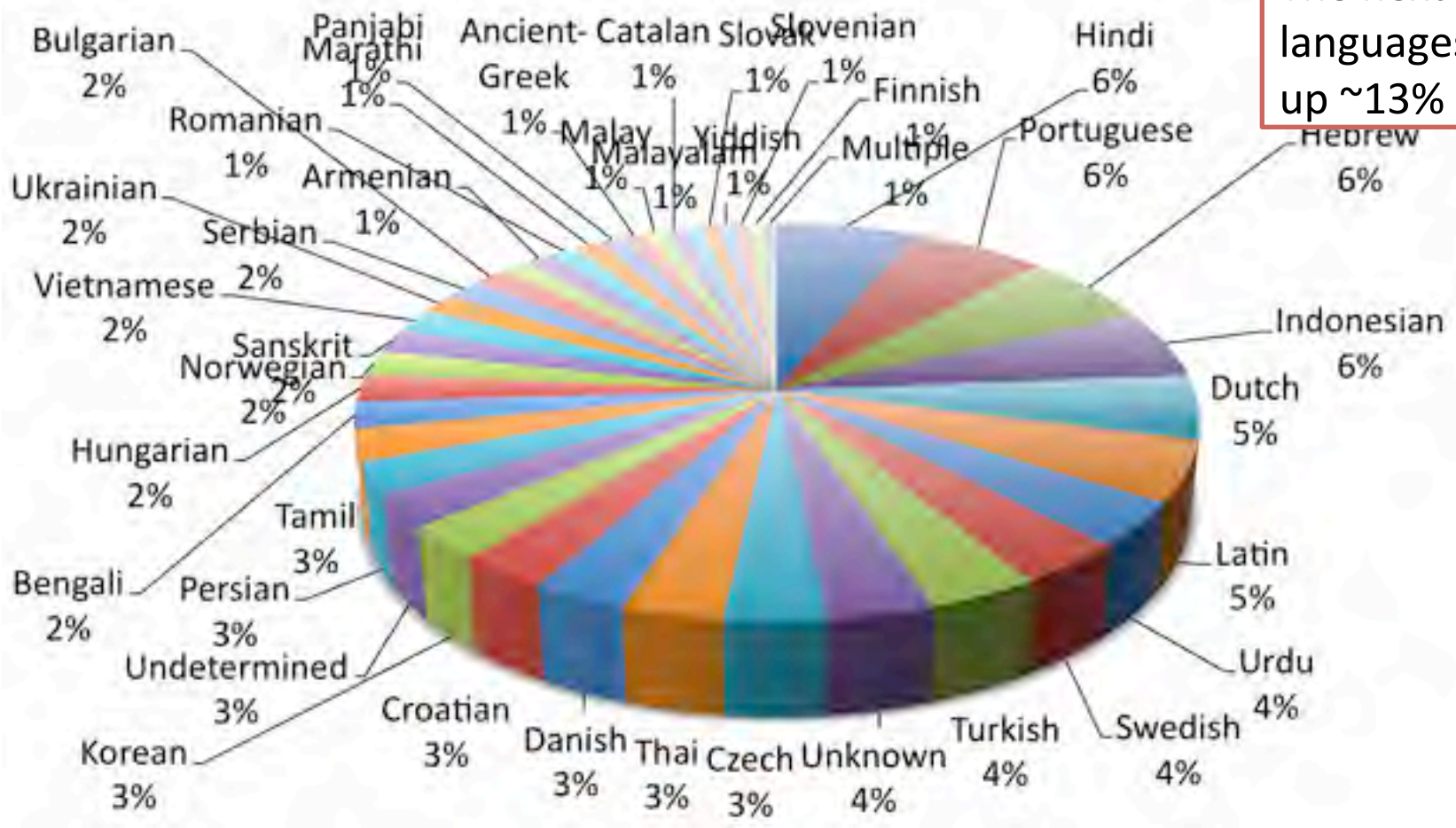
* As of August 13, 2010

Language Distribution (1)



* As of August 13, 2010

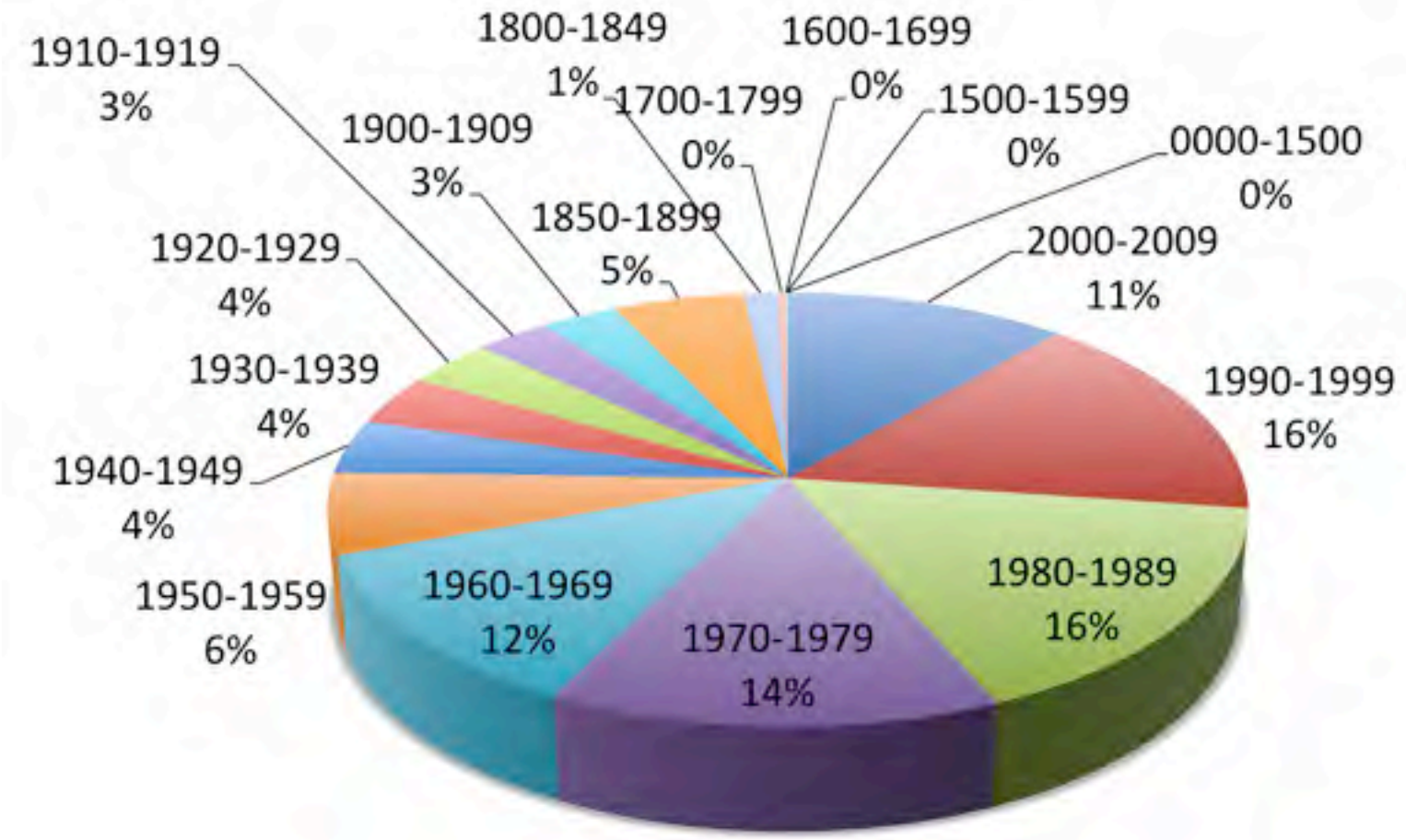
Language Distribution (2)



The next 40 languages make up ~13% of total

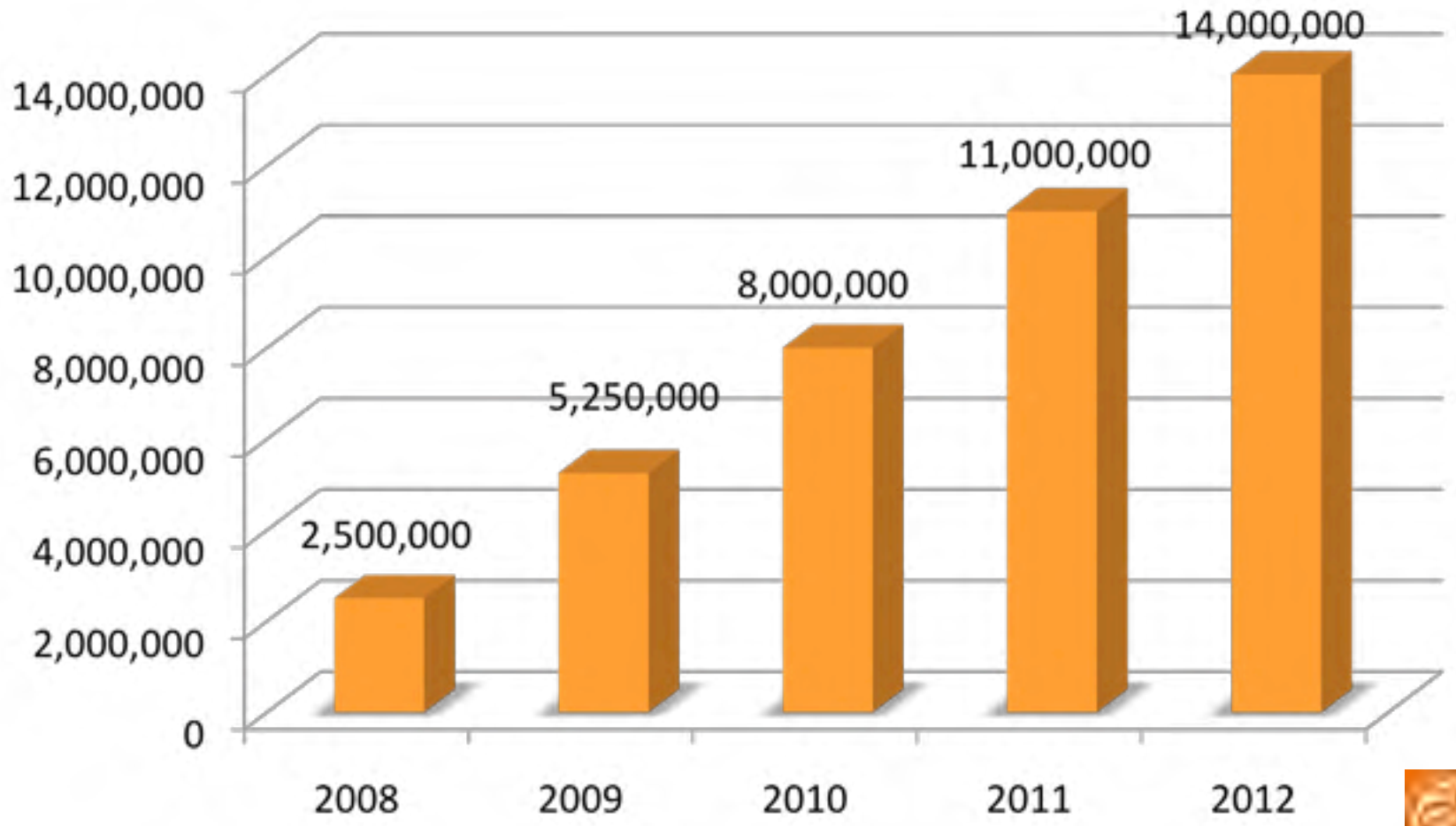
* As of August 13, 2010

Dates



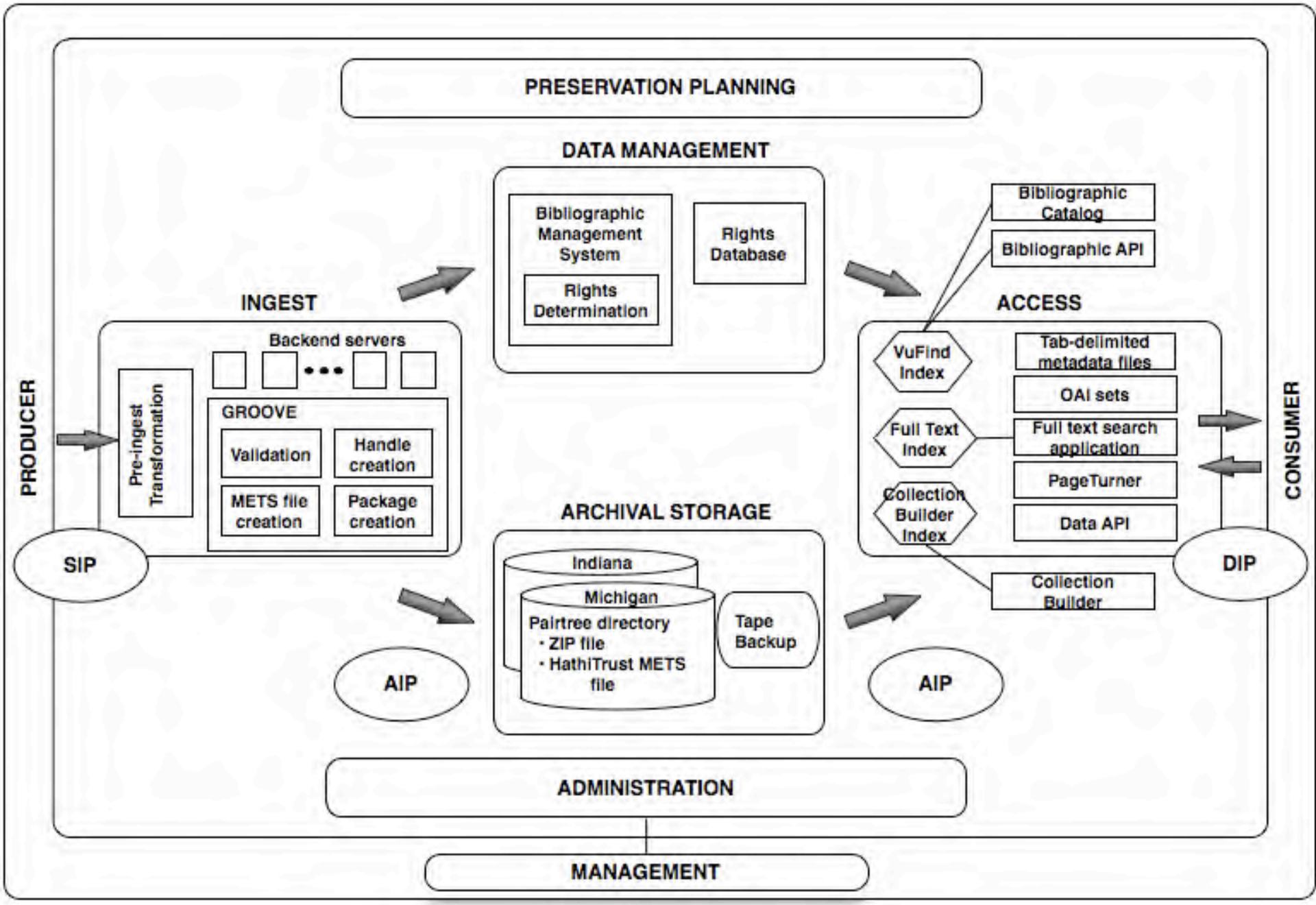
* As of August 13, 2010

Content Growth



Repository Philosophy/Design

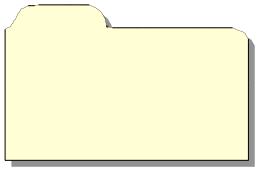
- OAIS/TRAC
- Consistency
- Standardization
- Simplicity (in design, not function)
- Practicality
- Sustainability

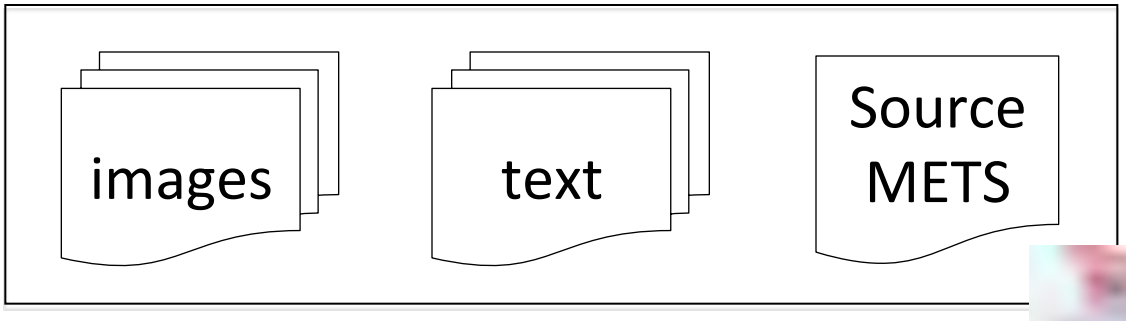

PRESERVATION PLANNING
DATA MANAGEMENT
Bibliographic Management System
Rights Database
Rights Determination
Bibliographic Catalog
Bibliographic API
INGEST
ACCESS
Backend servers
PRODUCER
CONSUMER
Pre-ingest Transformation
GROOVE
Validation
Handle creation
METS file creation
Package creation
VuFind Index
Tab-delimited metadata files
OAI sets
Full Text Index
Full text search application
PageTurner
Collection Builder Index
Data API
SIP
DIP
ARCHIVAL STORAGE
Indiana
Michigan
Pairtree directory
• ZIP file
• HathiTrust METS file
Tape Backup
Collection Builder
AIP
AIP
ADMINISTRATION
MANAGEMENT

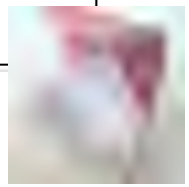
Content

- Largely uniform in technical characteristics
- 4 formats
  - ITU G4 TIFF
  - JPEG2000
  - JPEG
  - Unicode (with and without coordinates)

Object Package







malachus,
Flickr.com

HT
METS

Metadata

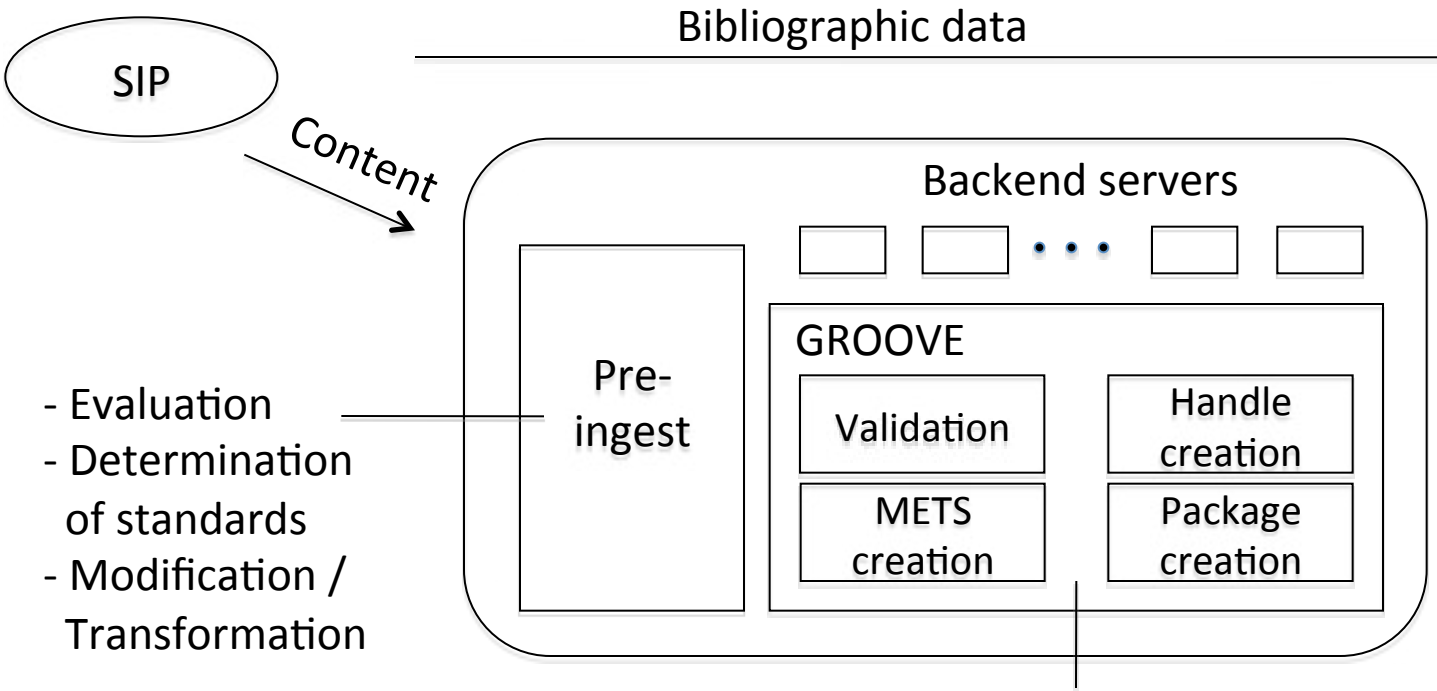
- Details and specifications at repository level
  - Object specifications / Validation criteria
  - Page-tagging
- Variations at object level
  - Files missing
  - Non-valid files
  - Incorrect file checksums

http://www.hathitrust.org/digital_object_specifications

Ingest

- Bibliographic Data
  - Must be present prior to content ingest
  - MARCXML, as complete as possible
- Content
  - Pre-ingest
  - Ingest

Ingest (2)

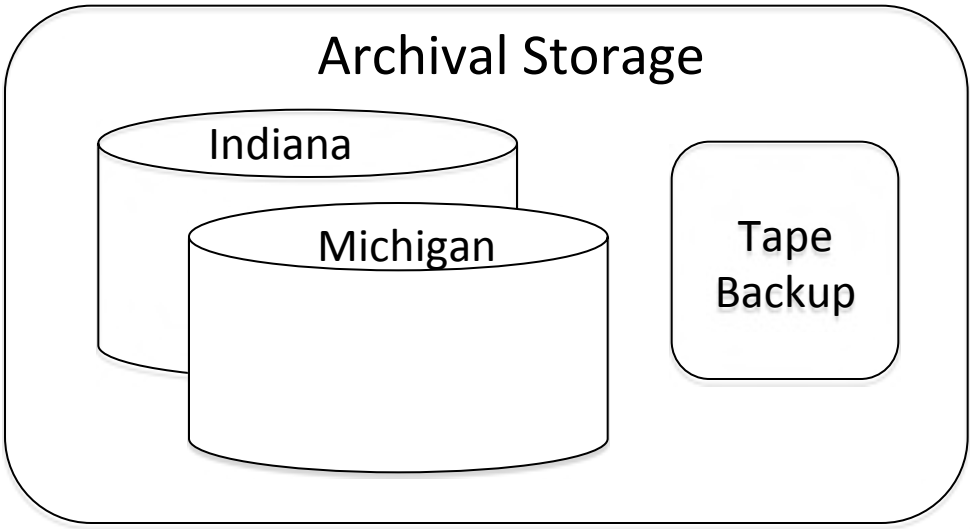

SIP
Bibliographic data
Content
Backend servers
Pre-ingest
GROOVE
Validation
Handle creation
METS creation
Package creation
- Evaluation
- Determination of standards
- Modification / Transformation
- Ensure conformance
- Barcode
- Fixity
- Consistency
- Well-formedness
- Prepare archival package

Archival Storage

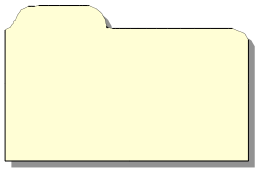
- Reliability – ensure integrity
- Redundancy – in single and multiple sites
- Scalability – including ease of management
- Accessibility – for repository processes and services
- Platform-independence – for data/object management

Media & Architecture

- Isilon Systems
- Load balancing and failover
- Ingest at Michigan, replicated to Indiana
- Replacement on 3-4 year cycle

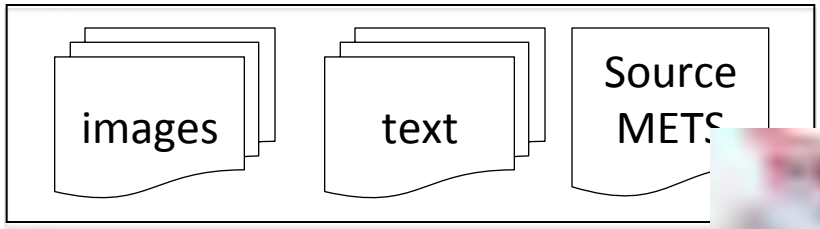


Architecture & Management



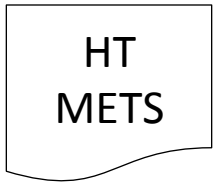
../**uc1**/pairtree_root/b3/54/34/86/b34543486

b34543486.zip



malachus, Flickr.com

b34543486.mets.xml



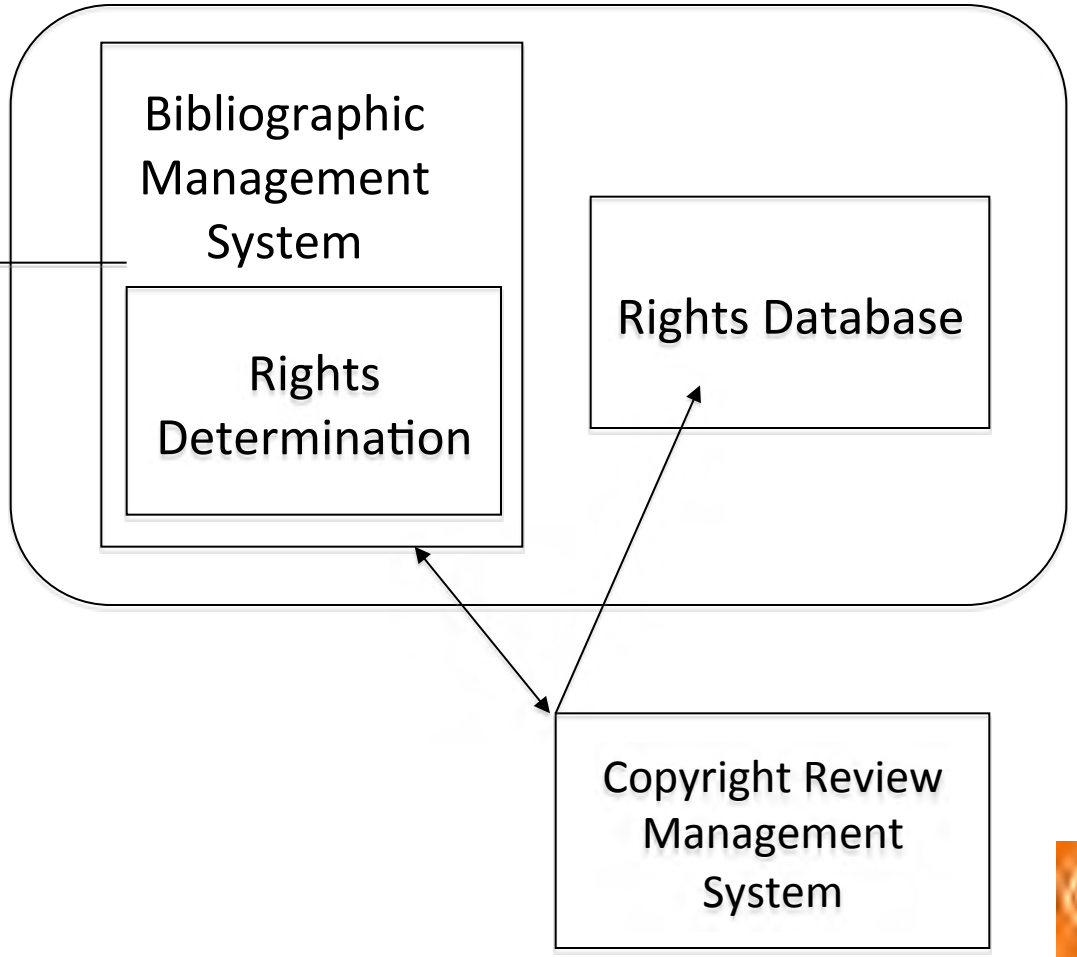
Example ids:

wu.89094366434
mdp.39015037375253

uc2.ark:/1390/t26973133
miua.aaj0523.1950.001

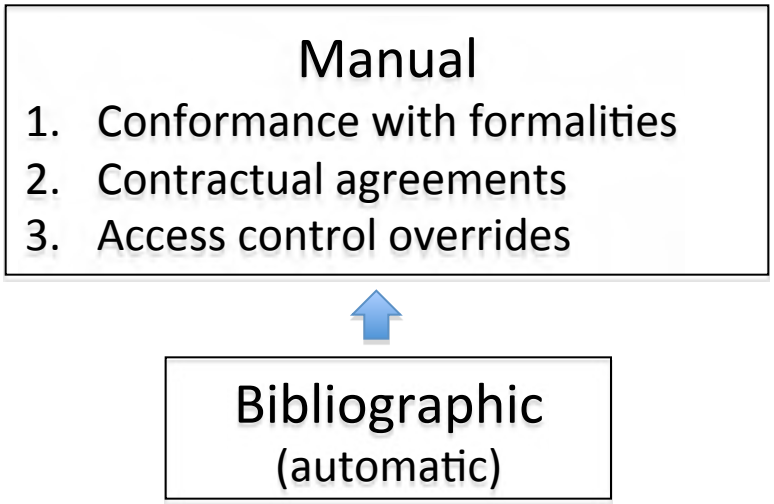
Data Management

- Inventory
- Loading and updating records
- Duplicate detection and collation
- Solr indexes behind VuFind catalog
- Source of information for Access services
- Rights determination (automated and support for manual review)



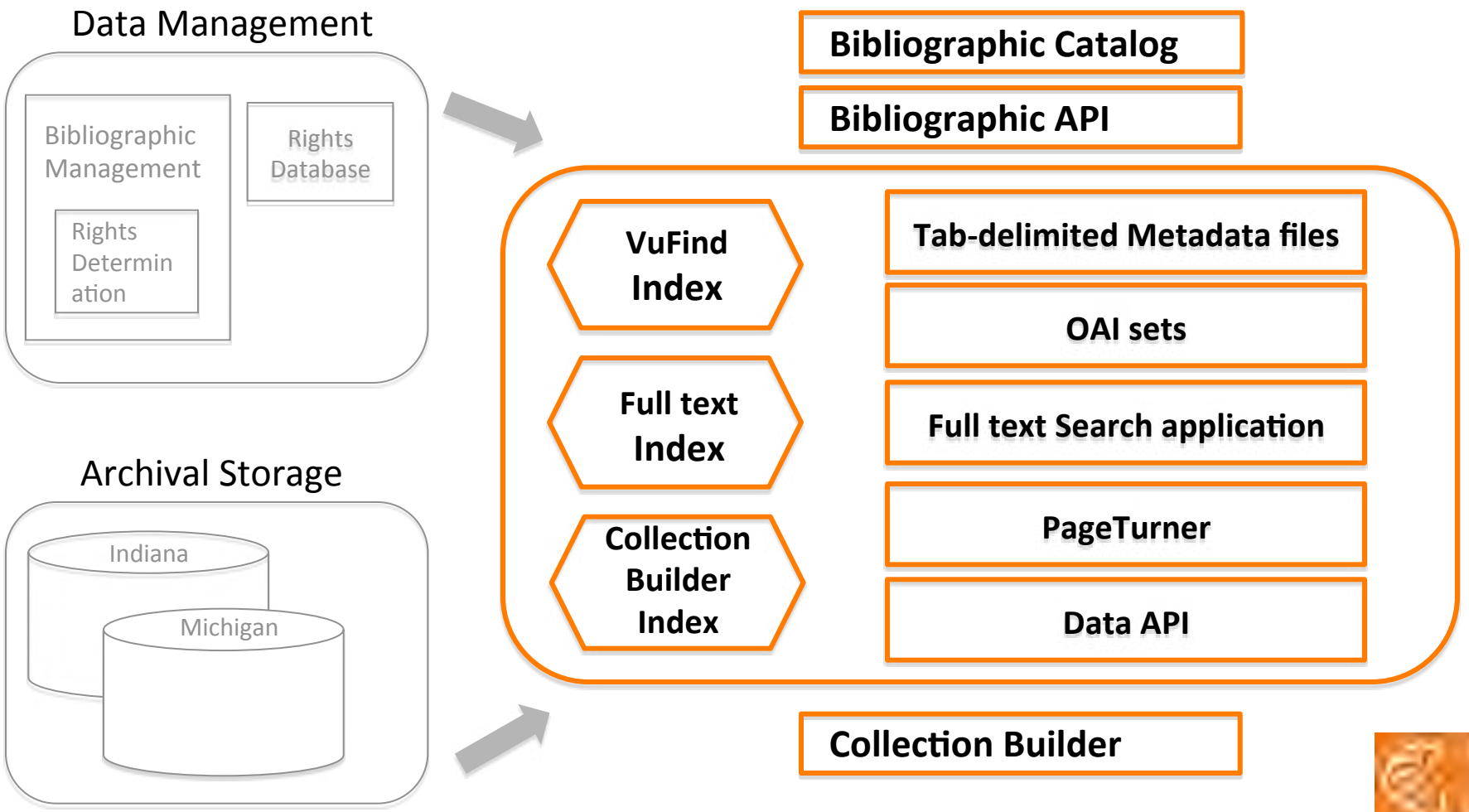
Rights Database

- System of precedence

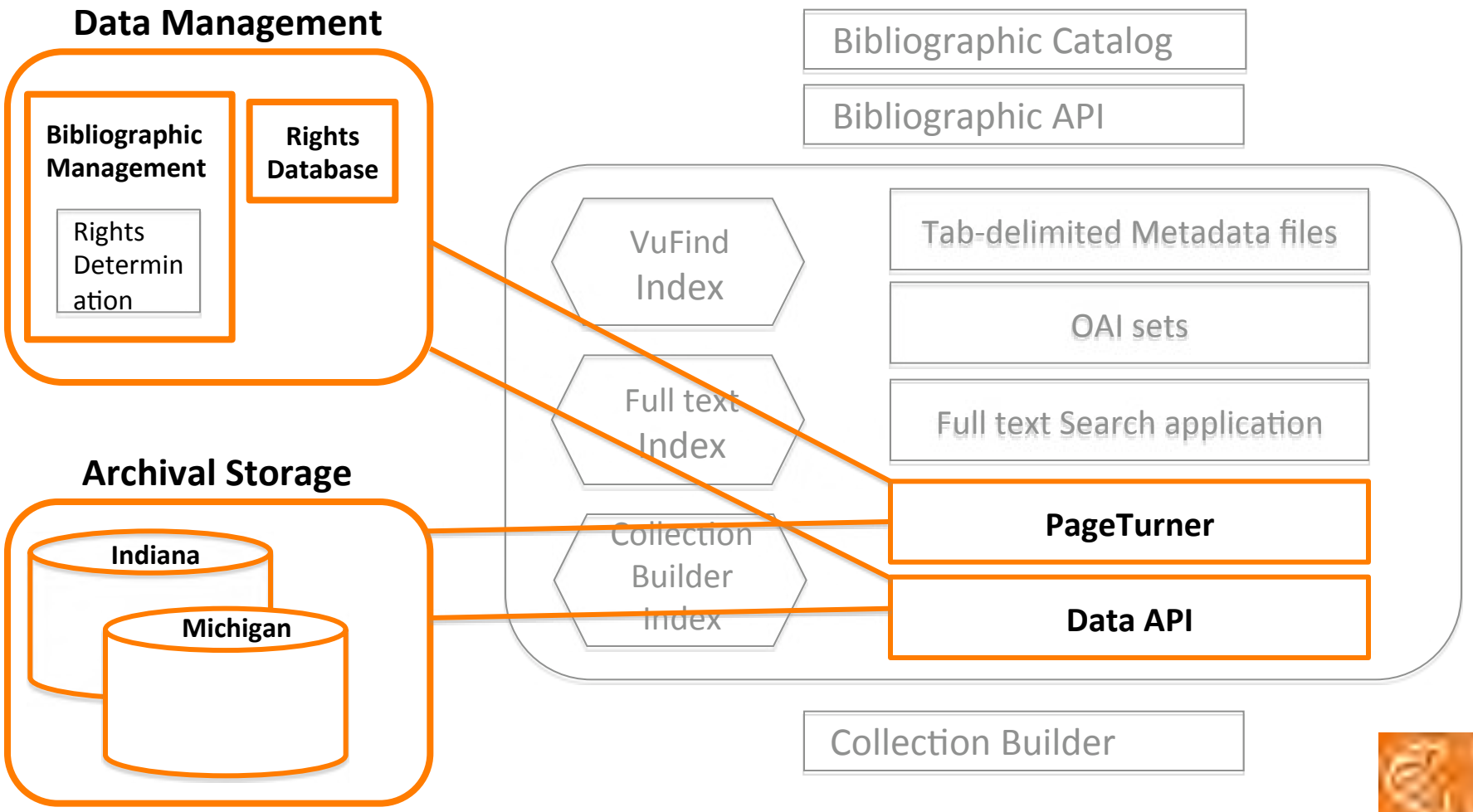


- 9 attributes
- 11 reason codes

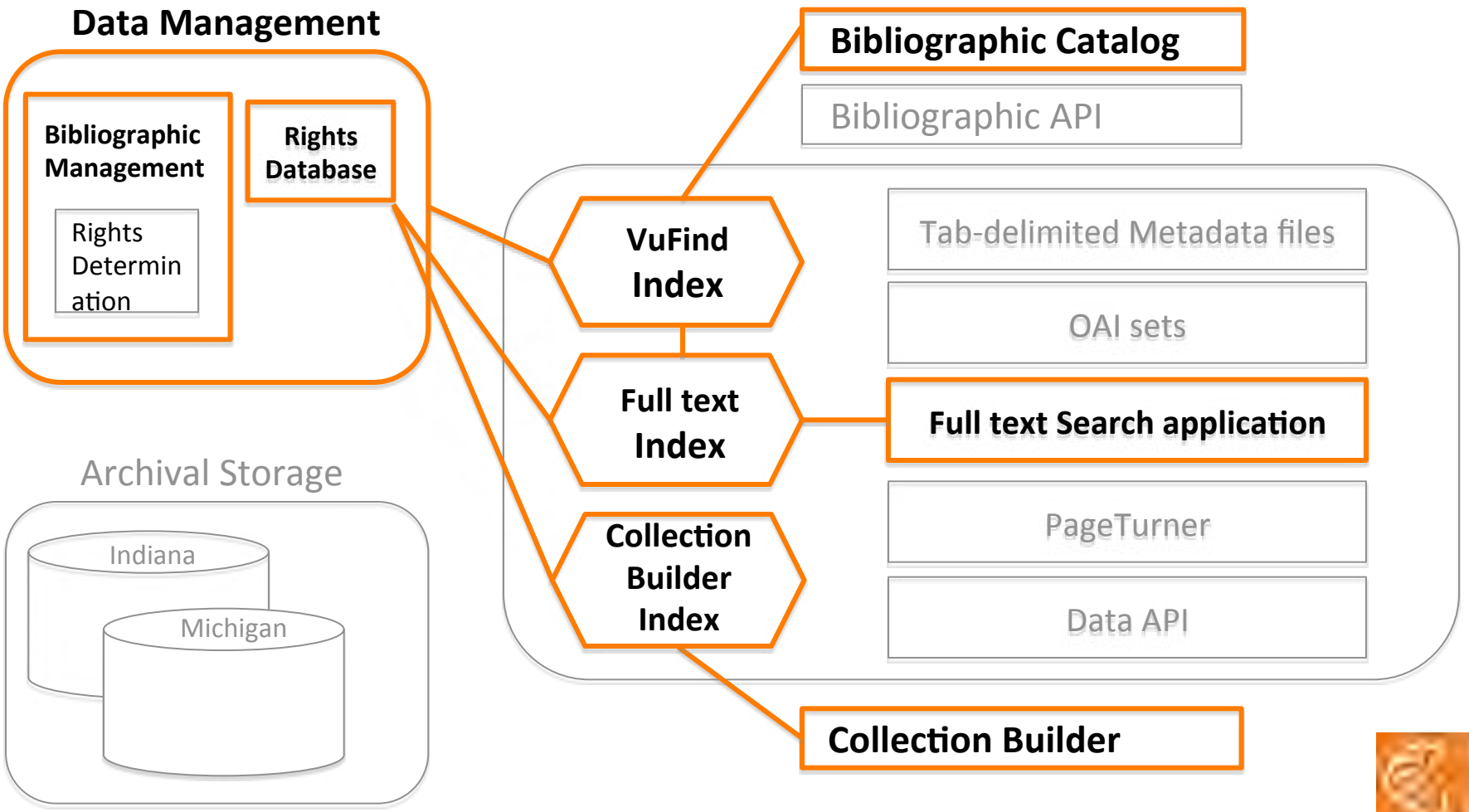
Access


Data Management
Bibliographic Management
Rights Determination
Rights Database
Archival Storage
Indiana
Michigan
Bibliographic Catalog
Bibliographic API
VuFind Index
Full text Index
Collection Builder Index
Tab-delimited Metadata files
OAI sets
Full text Search application
PageTurner
Data API
Collection Builder

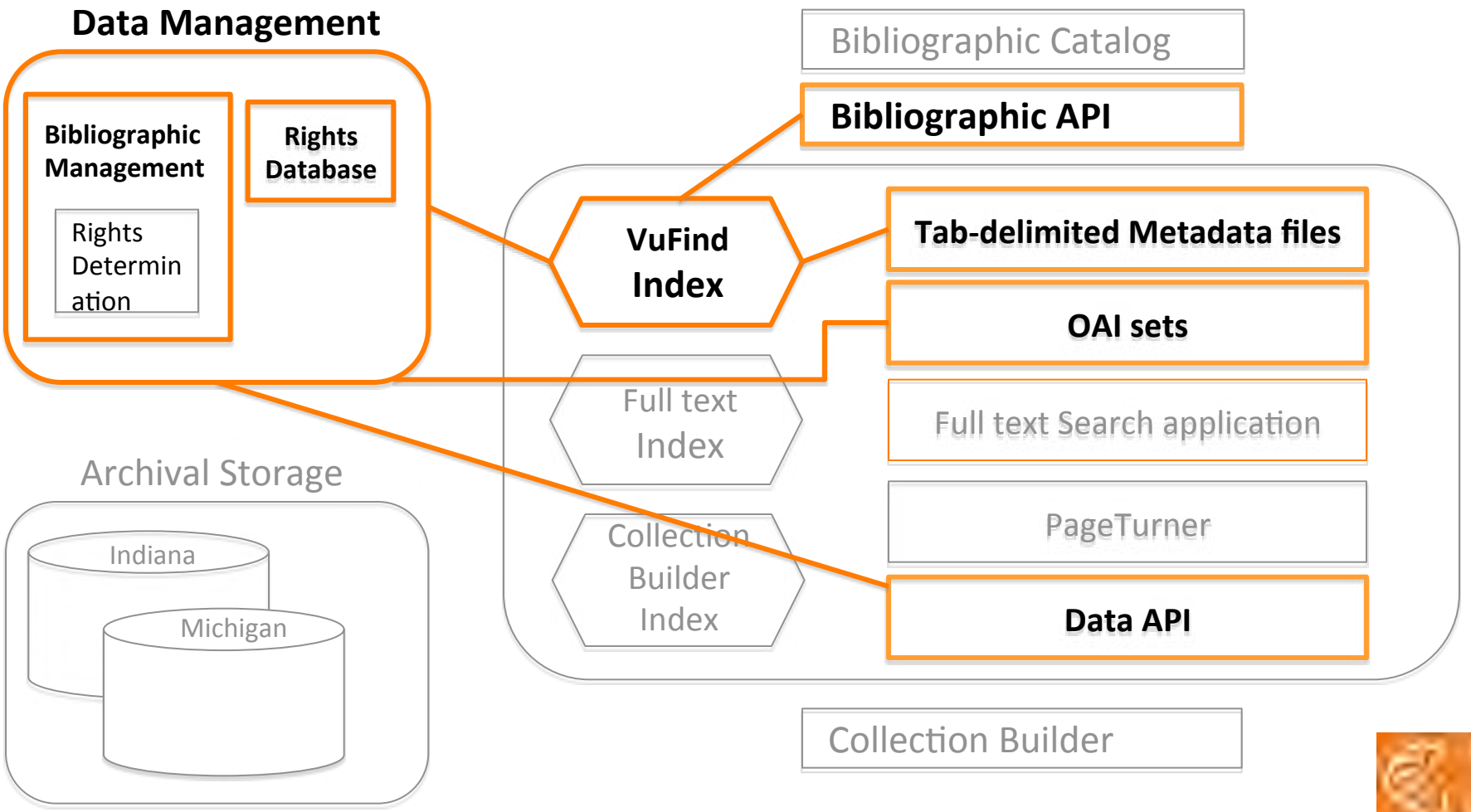
Content Access

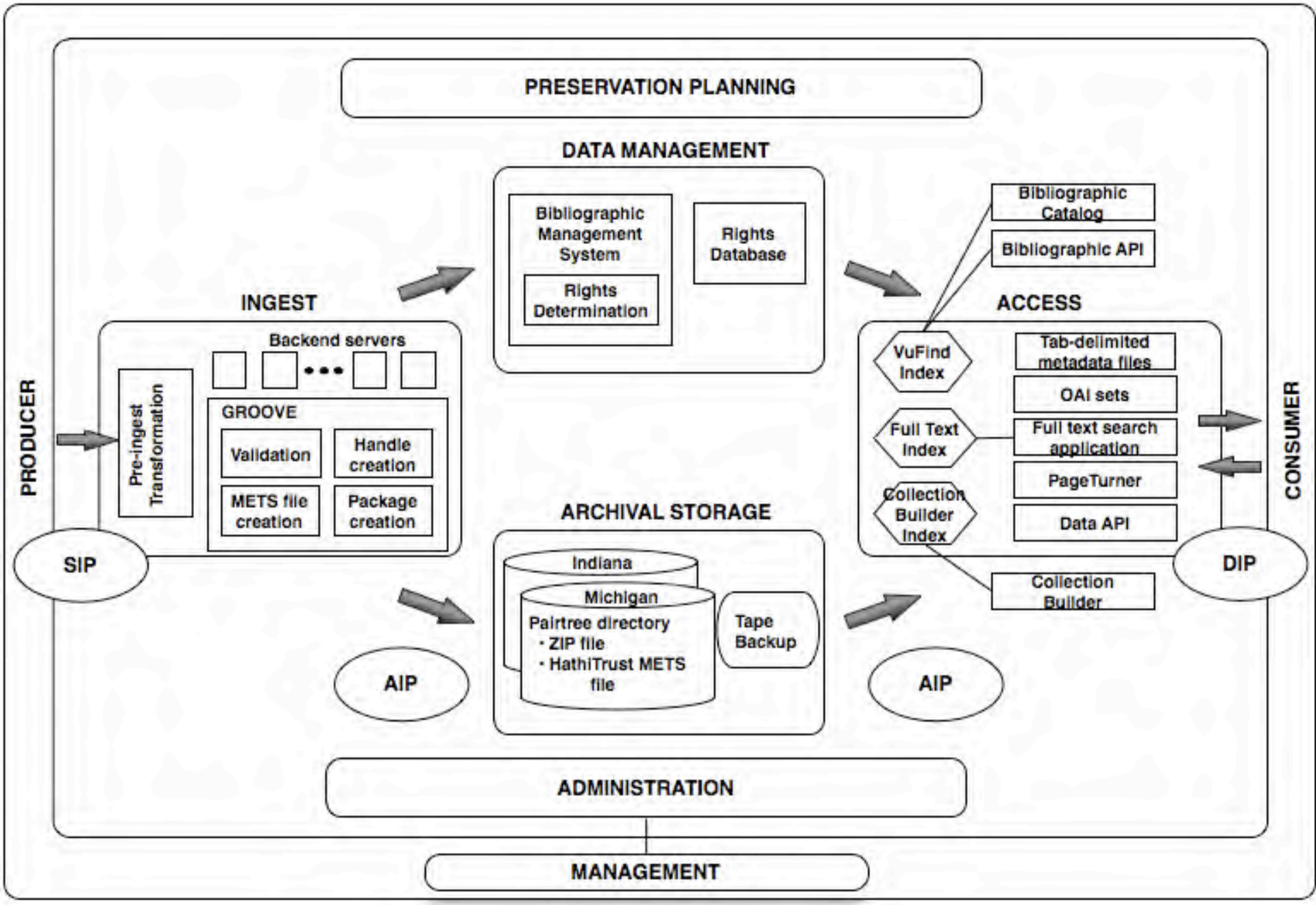


Search and Aggregation Access



Metadata Access





Source Undetermined

Thank you!

hathitrust-info@umich.edu

Additional Source Information

for more information see: http://open.umich.edu/wiki/CitationPolicy

Slide 16, Image 11: malachus, Flickr.com

Slide 22, Image 11: malachus, Flickr.com, http://www.flickr.com/photos/malachus/5152200478/

Slide 29, Image 0: Source Undetermined