Collection and Data Overview

Jeremy York

Stacy Kowalczyk

RESEARCH CENTER

HATHITRUST

A Shared Digital Repository

HathiTrust Data Overview

September 10, 2012
Jeremy York
Project Librarian, HathiTrust

Content and Metadata

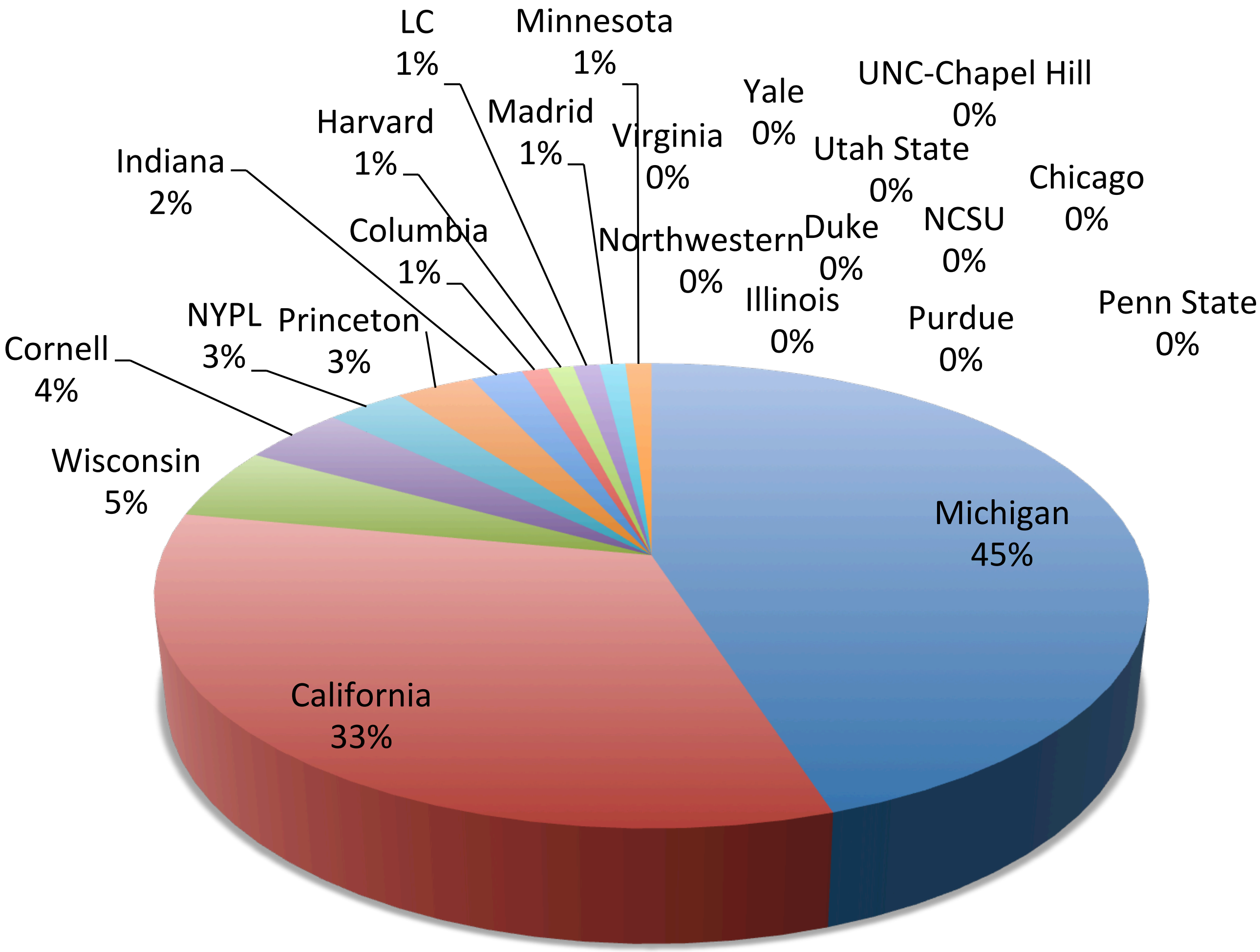
Outline

- Content and Metadata
  - Data formats
- Repository Organization
- Data availability
  - Availability mechanisms
  - Rights and agreements

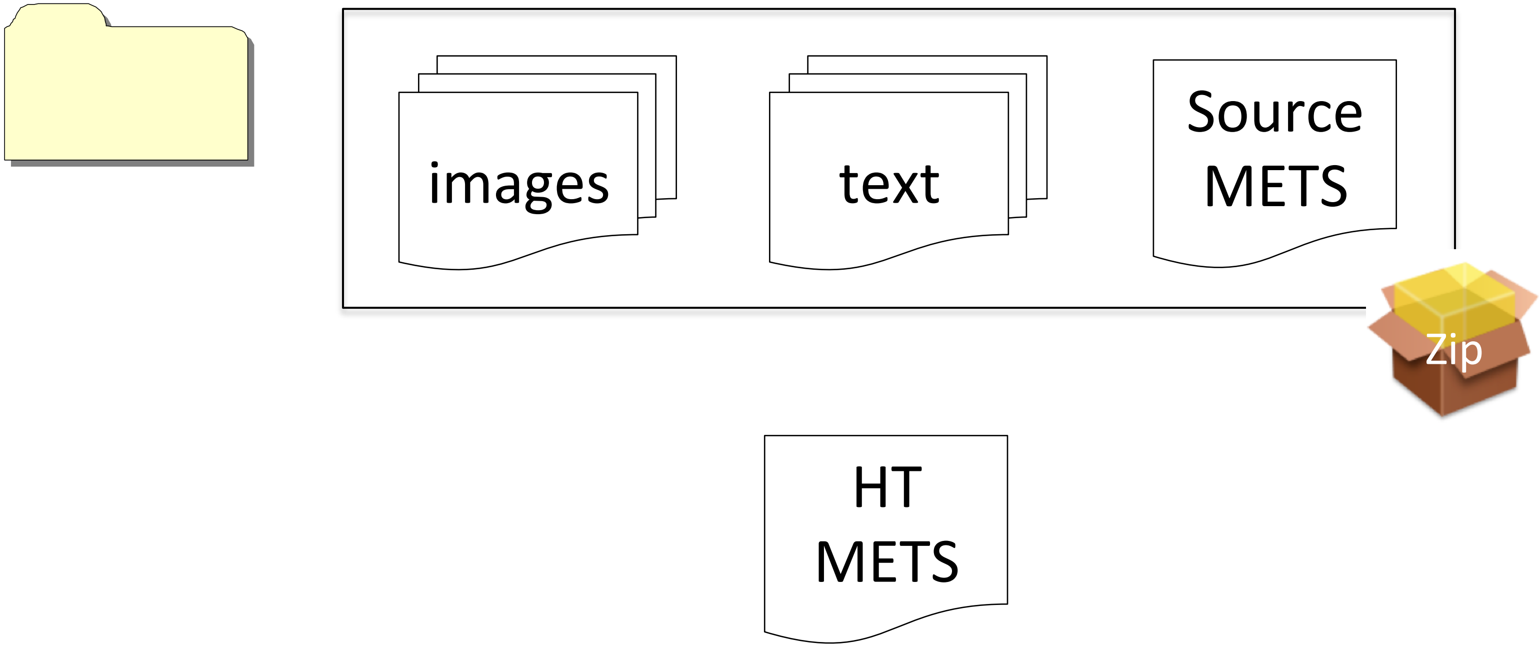
Content

- Books and journals
  - Pilots around images, audio, born-digital
- Digitization sources
  - Google (96.8%, 10,162,104)
  - Internet Archive (2.9%, 301,972)
  - Local (0.3%, 31,840)

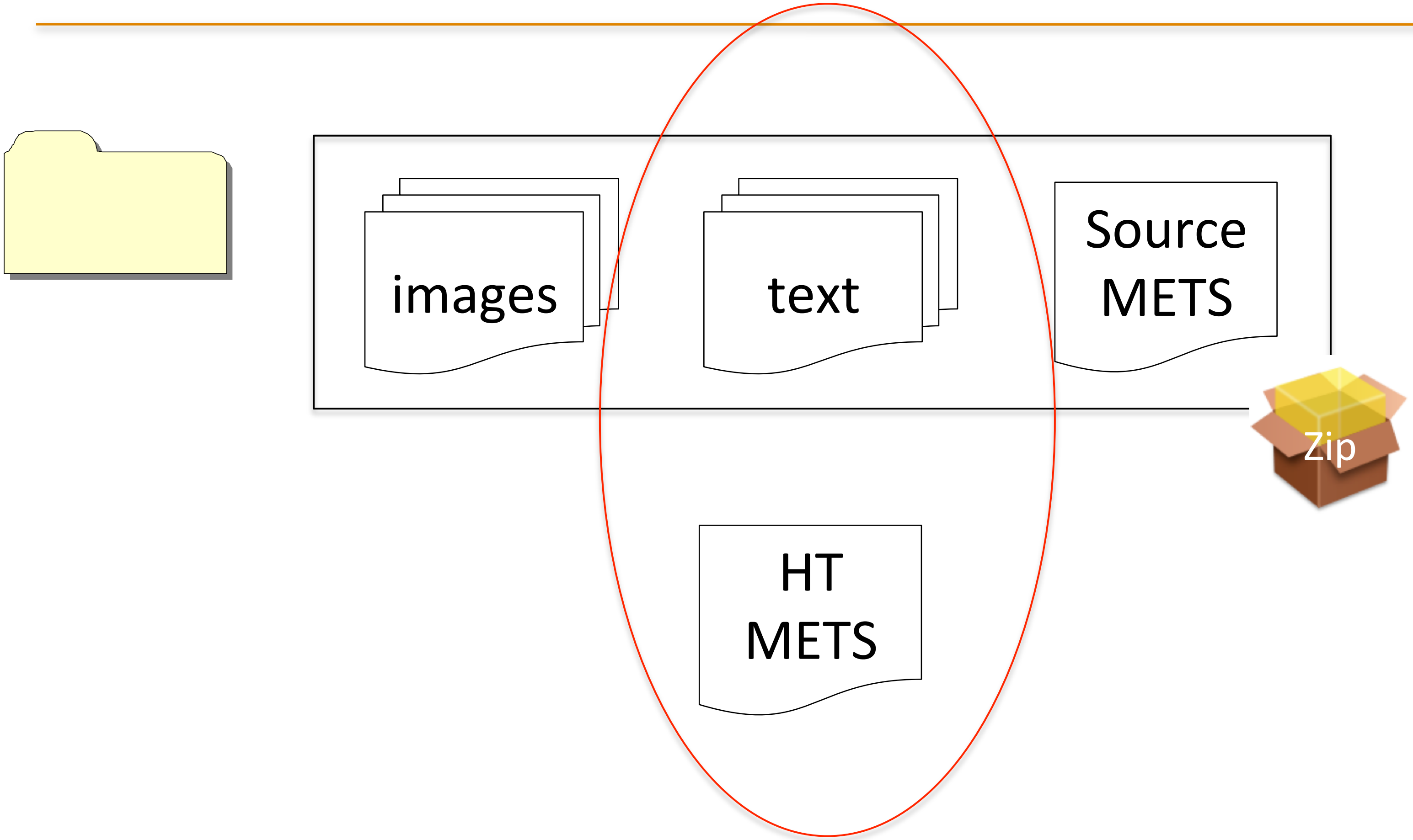
Content Sources



Content Package



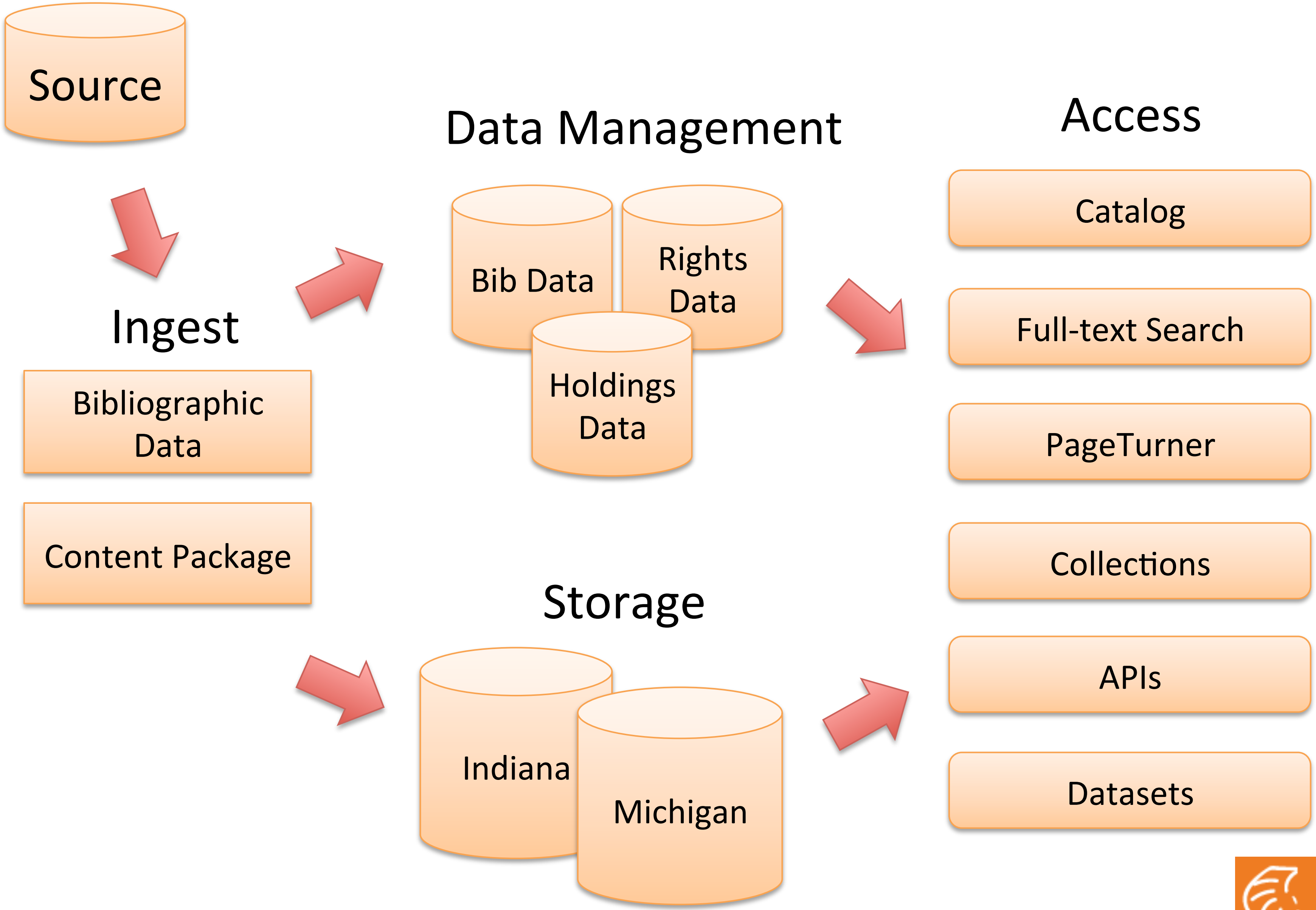
Content Package



Metadata

- Bibliographic
- Structural
- Rights
- Administrative (preservation)
- Holdings

Repository Organization



Source

Ingest

Bibliographic Data

Content Package

Data Management

Bib Data

Rights Data

Holdings Data

Storage

Indiana

Michigan

Access

Catalog

Full-text Search

PageTurner

Collections

APIs

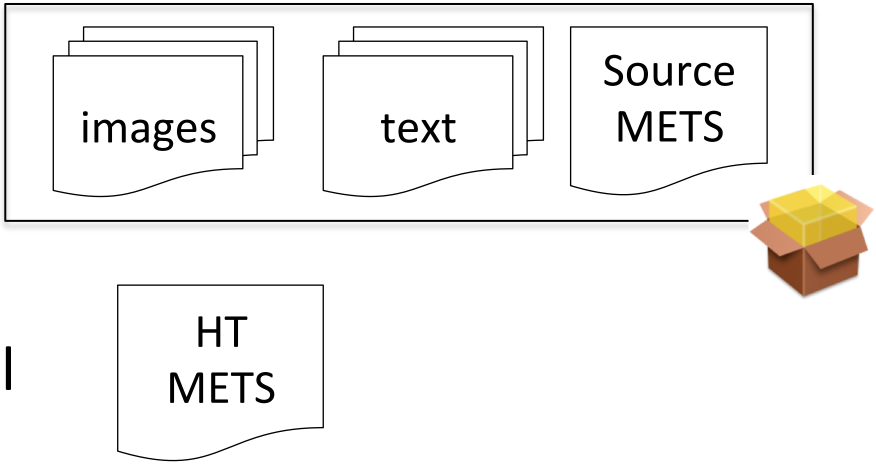
Datasets

File System

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

b34543486.zip

b34543486.mets.xml



Example ids:

wu.89094366434
mdp.39015037375253

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

Data Availability

APIs

- Data API
  - Zip package
  - Single page images or OCR
  - Volume and rights metadata (XML)
- Bib API (JSON)
  - Volume and rights metadata
  - MarcXML

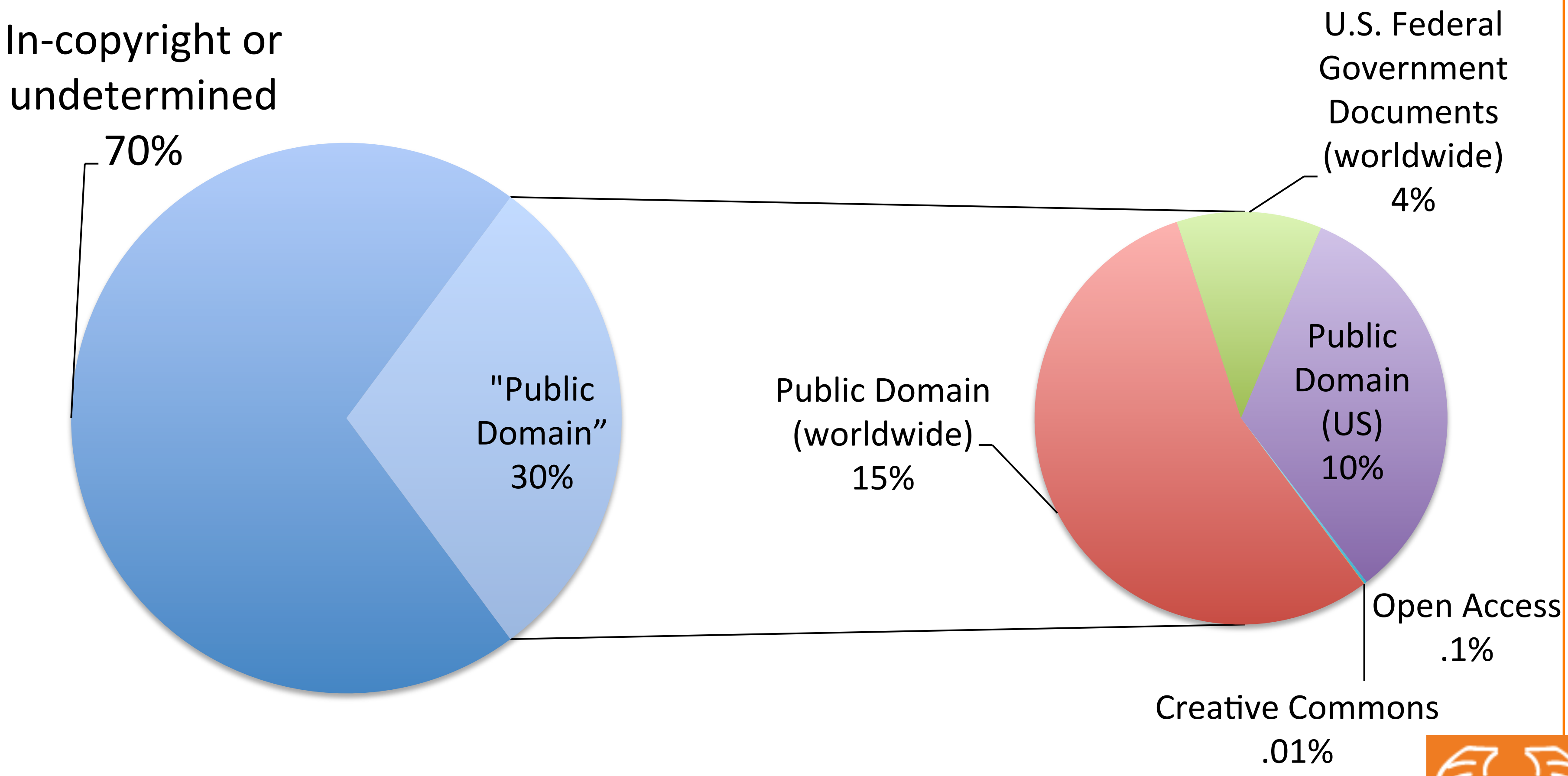
Data Feeds

- OAI
  - MarcXML
  - Dublin Core
- Hathifiles
  - Tab-delimited inventory files
  - Contain
    - Identifiers
    - Limited bibliographic information
    - Rights, language, gov docs status information

Datasets

- Content
- Bibliographic data
- Content organization

Content Distribution



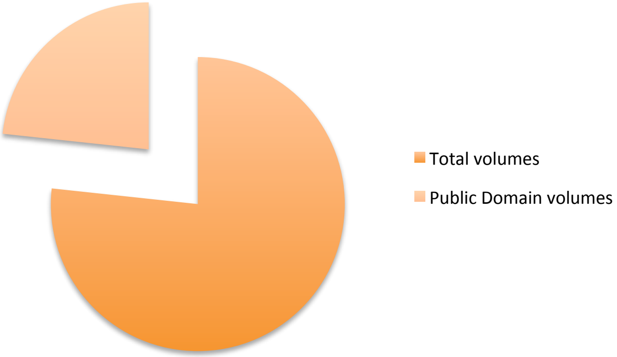
HathiTrust Research Collection Overview

Stacy Kowalczyk

RESEARCH CENTER

The HTRC Collection

- Public Domain Materials of the HatihTrust
  - 2,592,097 Volumes
  - Gigabytes
    - 2.3 TB in raw OCR’d text
    - 3.7 TB of managed OCR’d text
    - 1.85 TB solr Index
  - Monthly Updates
    - And irregular data ‘take down’ requests

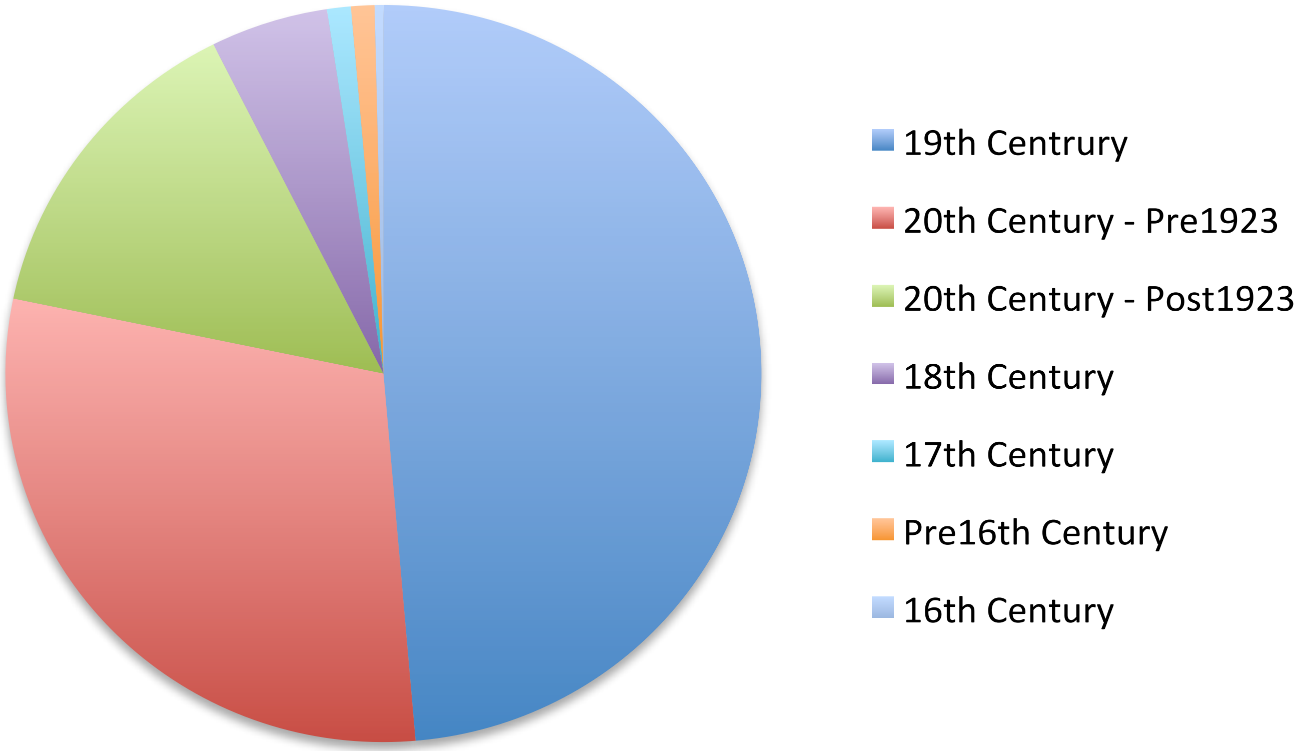


Exploring the Collection

- Publication Data
  - Date of publication
  - Country
  - Publisher
- Language
- Topical Coverage
- Authors

Publication Dates

- 2,562,283 Bib records with pub dates

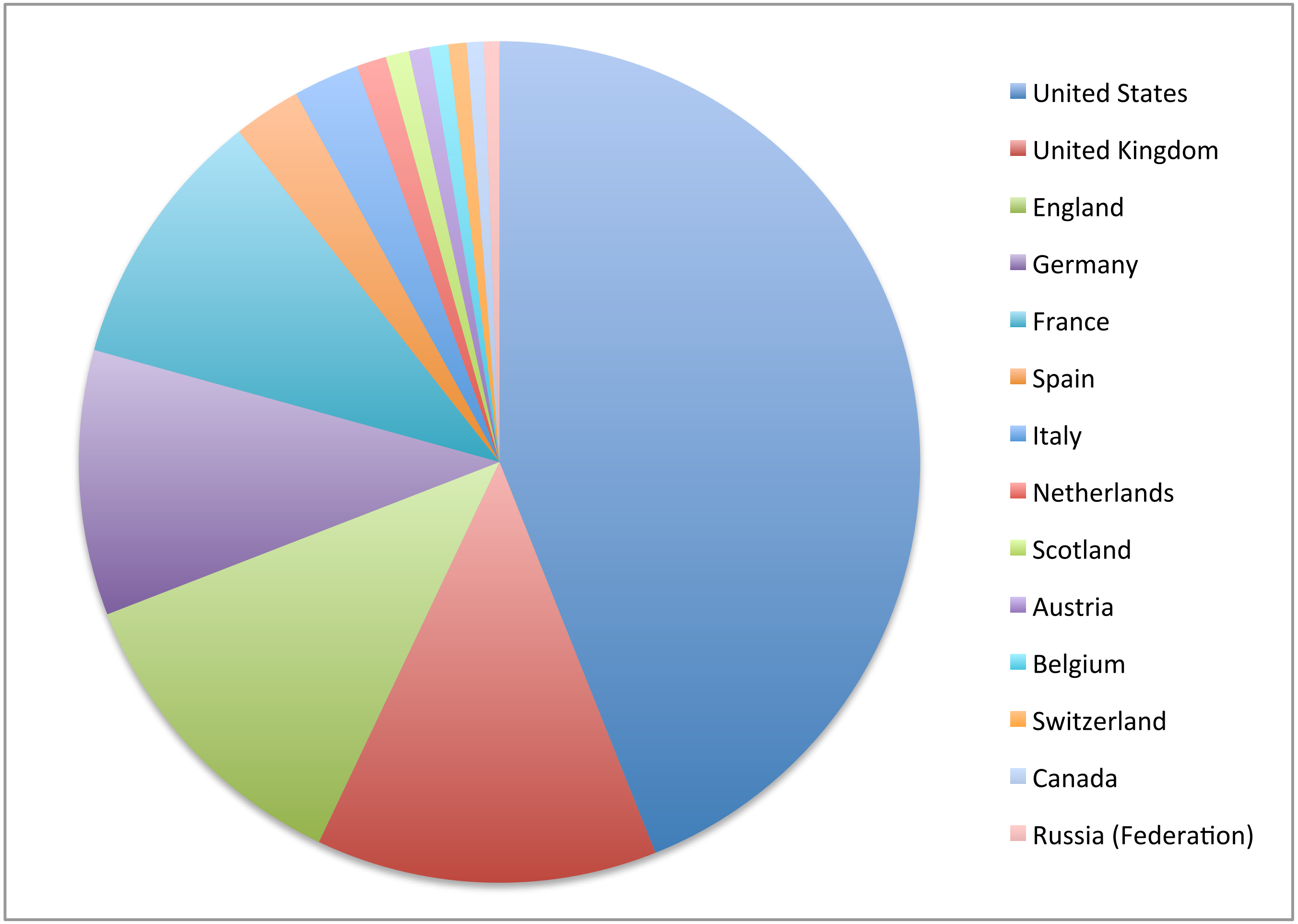


Country of Publication

Country of Publication

- 244 different countries of publication
- 2,578,341 bib records
- 400,000 records have more than one country of publication
- The top 11 countries accounted for nearly 90%
- 229 counties accounted for 6%
- Unknown country indicated 5%

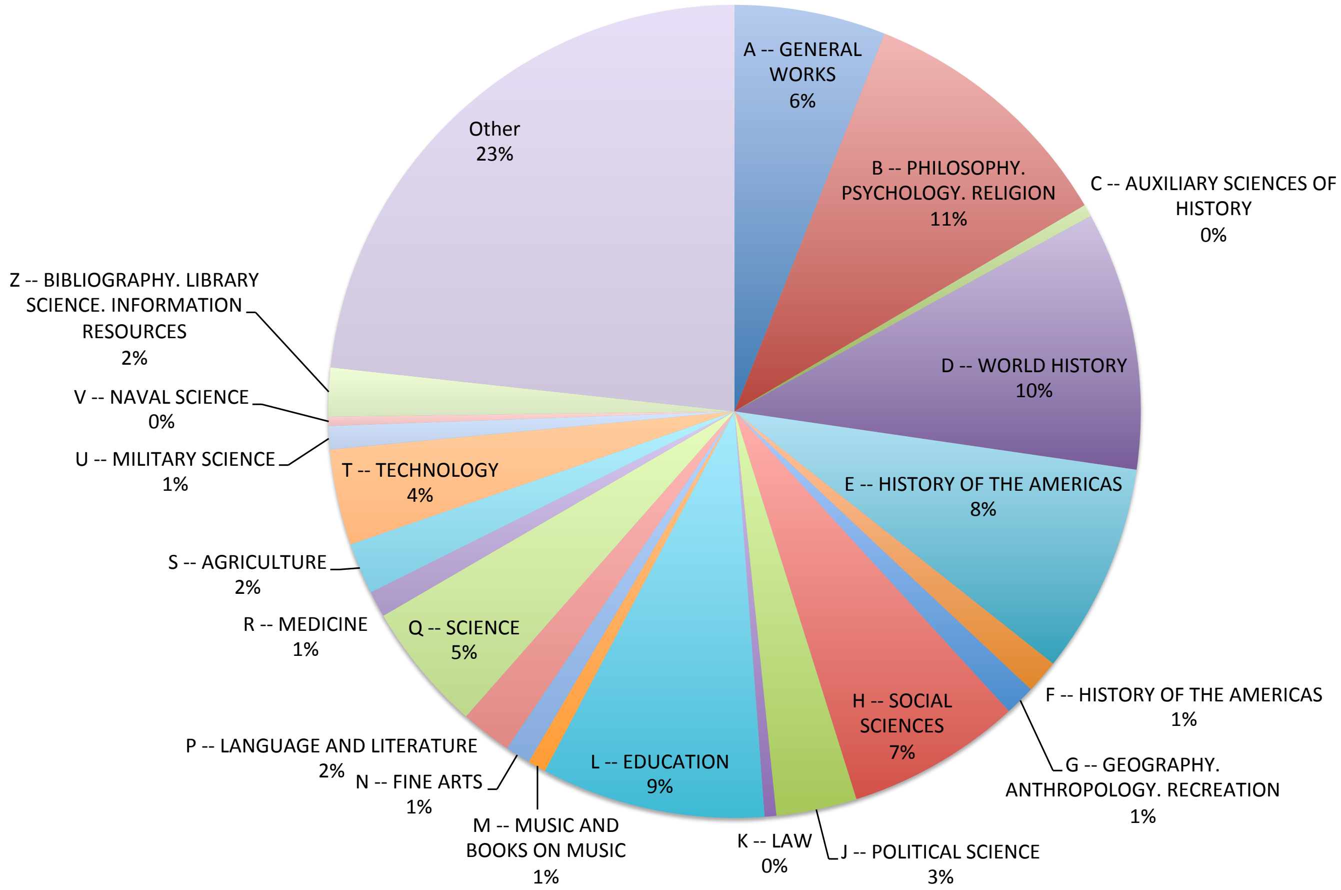
Country of Publication



Topical Coverage

- Call numbers
  - 335,446 unique call numbers
  - 691,131 bib records
- Topic Strings
  - 589,428 unique subject headings
  - 1,948,999 bib records
  - 2,315,070 occurrences

Call Number Distribution



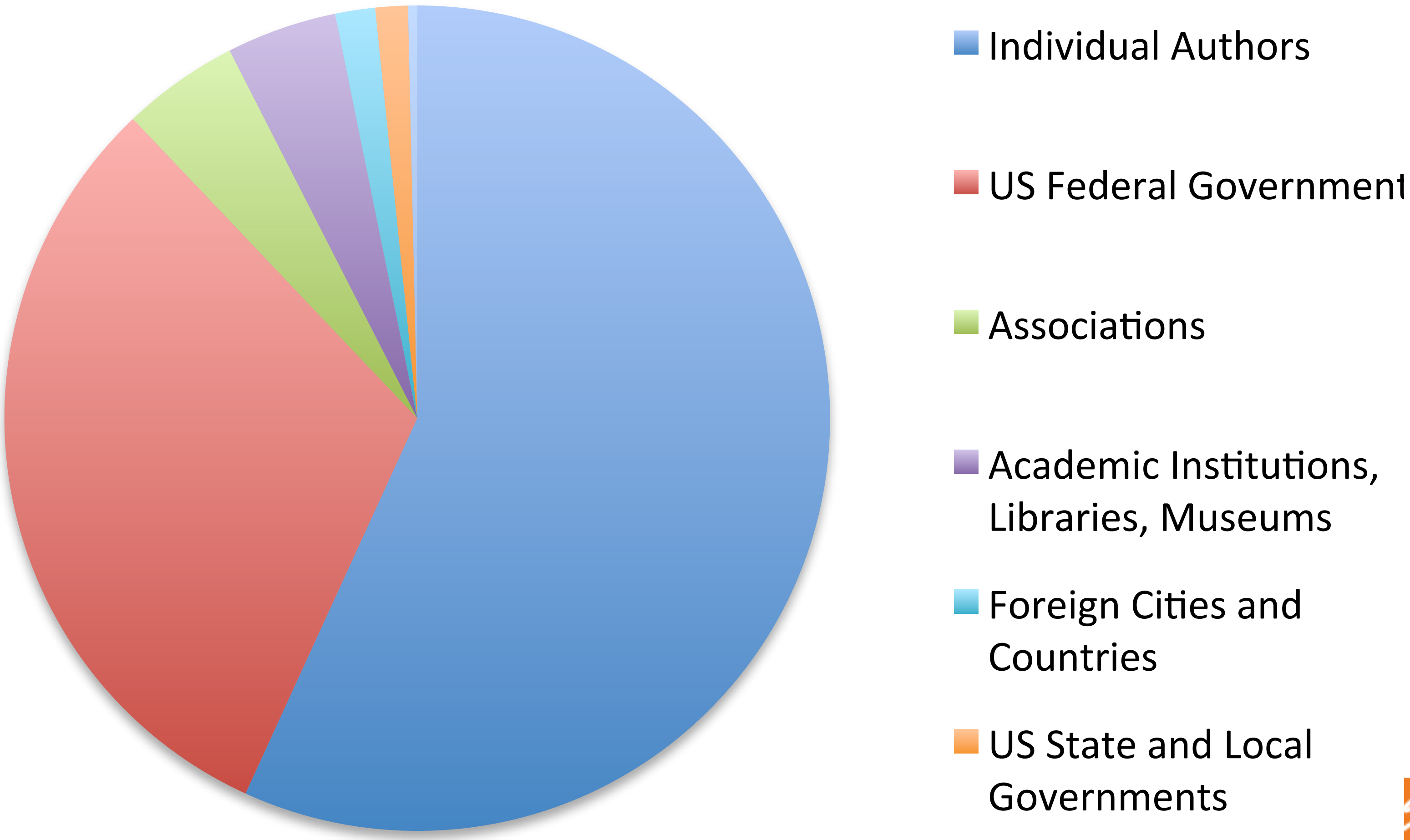
Standard Numbers

- SuDocs
  - 117,095 unique SuDoc numbers
  - 259,718 bib records
- ISBN
  - 23,765 ISBN numbers
  - 34,855 bib records
- ISSN
  - 8,658 unique ISSN numbers
  - 234,554 bib records
- OCLC numbers
  - 434,589 unique OCLC number
  - 1,112,499 bib records
- LCCN
  - 432,563 unique LCCN
  - 1,104,696 bib records

Authors

- 849,753 unique author strings
- 2,41,0,788 bibliographic records
- Organized into subcategories
  - US governmental agencies
  - US state and local governments
  - Foreign country and city governments
  - Companies
  - Associations/societies
  - Academic Institutions, Libraries, Museums
  - Individual Authors

Authors


Individual Authors
US Federal Government
Associations
Academic Institutions, Libraries, Museums
Foreign Cities and Countries
US State and Local Governments

Collection Access

- Known item
  - Title
  - Author
  - Standard number
- Key word access
  - All words in OCR’d text
  - All words in bibliographic data
- Sparsely populated data

To Learn More

Sessions tomorrow

- **Data in Detail** – Jeremy York and J. Stephen Downey
  - – 9:30 am Main Lobby/Atrium
  - – 1 pm Main Lobby/Atrium
- **Building Collections and Analyzing Data**
  - – 1 pm Flex Lab 005

HATHITRUST

A Shared Digital Repository

HathiTrust Research Center Architecture Overview

Robert H. McDonald | @mcdonald

Executive Committee-HathiTrust Research Center (HTRC)

Deputy Director-Data to Insight Center

Associate Dean-University Libraries

Indiana University

INDIANA UNIVERSITY

RESEARCH CENTER

ILLINOIS

Follow Along

http://slidesha.re/U4z1gW

HTRC Architecture Group

Indiana University

- Beth Plale, Lead
- Yiming Sun
- Stacy Kowalczyk
- Aaron Todd
- Jiaan Zeng
- Guangchen Ruan
- Zong Peng
- Swati Nagde

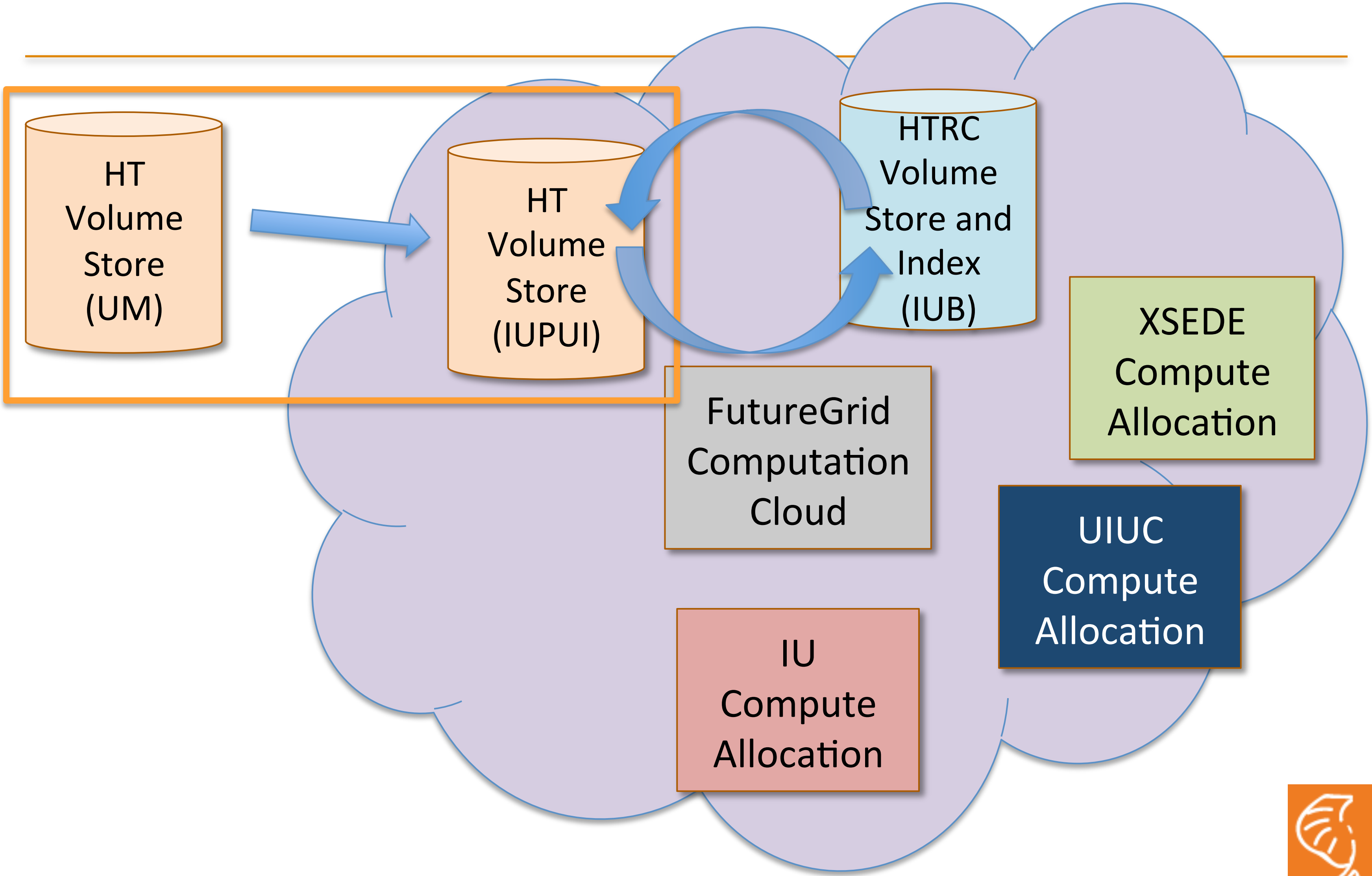
University of Illinois

- J. Stephen Downie
- Loretta Auvil
- Boris Capitanu
- Kirk Hess
- Harriett Green

Presentation Overview

- Considerations for Current Architecture
- Architecture - Use Case Methodology
- Technical Overview
- UnCamp Sessions for Further Review

Main Case – Data Near Computation



HT Volume Store (UM)

HT Volume Store (IUPUI)

HTRC Volume Store and Index (IUB)

FutureGrid Computation Cloud

XSEDE Compute Allocation

UIUC Compute Allocation

IU Compute Allocation

Non-Consumptive Research Paradigm

- *No action or set of actions on part of users, either acting alone or in cooperation with other users over duration of one or multiple sessions can result in sufficient information gathered from collection of copyrighted works to reassemble pages from collection.*
- Definition disallows collusion between users, or accumulation of material over time. Differentiates human researcher from proxy which is not a user. Users are human beings.

Amicus Brief and NCR

- Jockers, Sag, Schultz –
- http://tinyurl.com/cy34hhr

Use Cases for Phase 1 Architecture

- Use Case #1 - Previously registered user submitted algorithm retrieved and run with results set
- Use Case #2 - HTRC applications/portal access (SEASR)
- Use Case #3 – Blacklight Lucene/Solr faceted access
- Use Case #4 - Direct programmatic access through Secure Data API (right now only for UnCamp and open content)

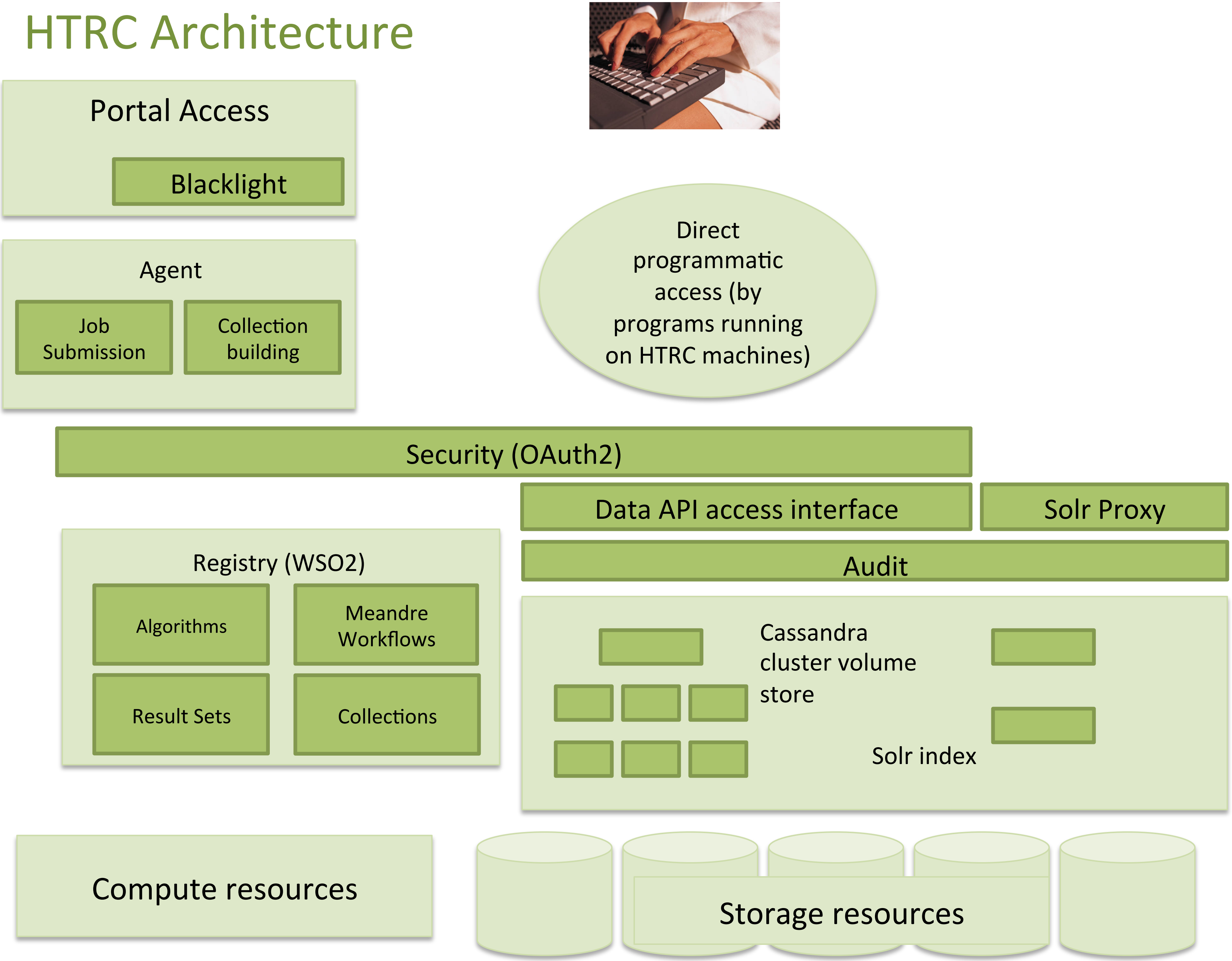
HTRC Current Infrastructure

- Servers
  - – 14 production-level quad-core servers
    - 16 – 32GB of memory
    - 250 – 500GB of local disk each
  - – 6-node Cassandra cluster for volume store
  - – Ingest service and secure Data API access point
- Storage (IU University Infrastructure)
  - – 13TB of 15,000 RPM SAS disk storage
  - – Increase up to 17TB by end of 2012
  - – 500TB available in late year 2-year 3

Key Components of Architecture

- Portal Access
- Blacklight Access
- Agent
- Registry
- Secured Data API Access
- Solr Proxy

HTRC Architecture



Portal Access

Blacklight

Agent

Job Submission

Collection building

Direct programmatic access (by programs running on HTRC machines)

Security (OAuth2)

Data API access interface

Solr Proxy

Registry (WSO2)

Algorithms

Meandre Workflows

Result Sets

Collections

Audit

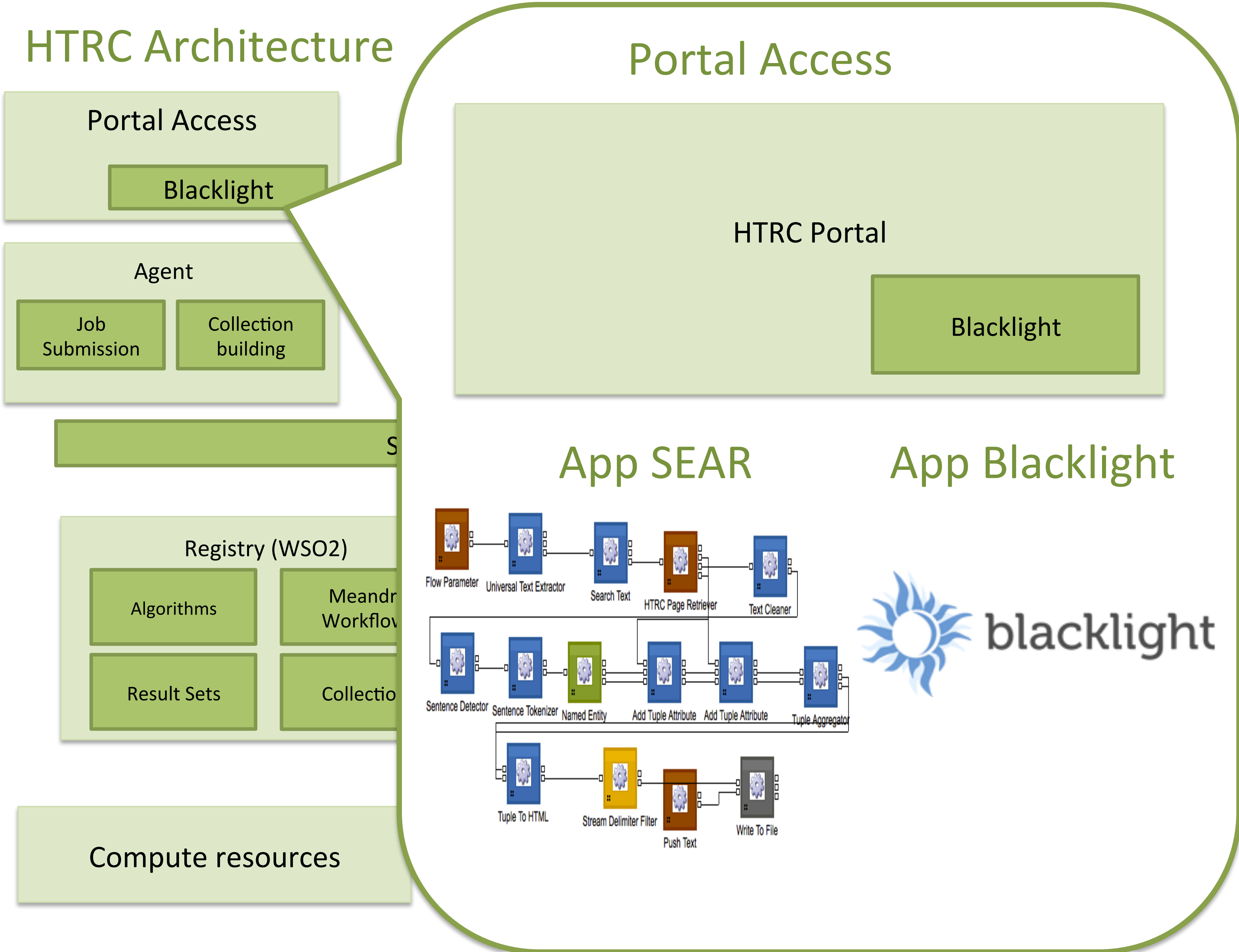
Cassandra cluster volume store

Solr index

Compute resources

Storage resources

HTRC Architecture



Portal Access

Blacklight

Agent

Job Submission

Collection building

S

Registry (WSO2)

Algorithms

Meandr Workflov

Result Sets

Collectio

Compute resources

Portal Access

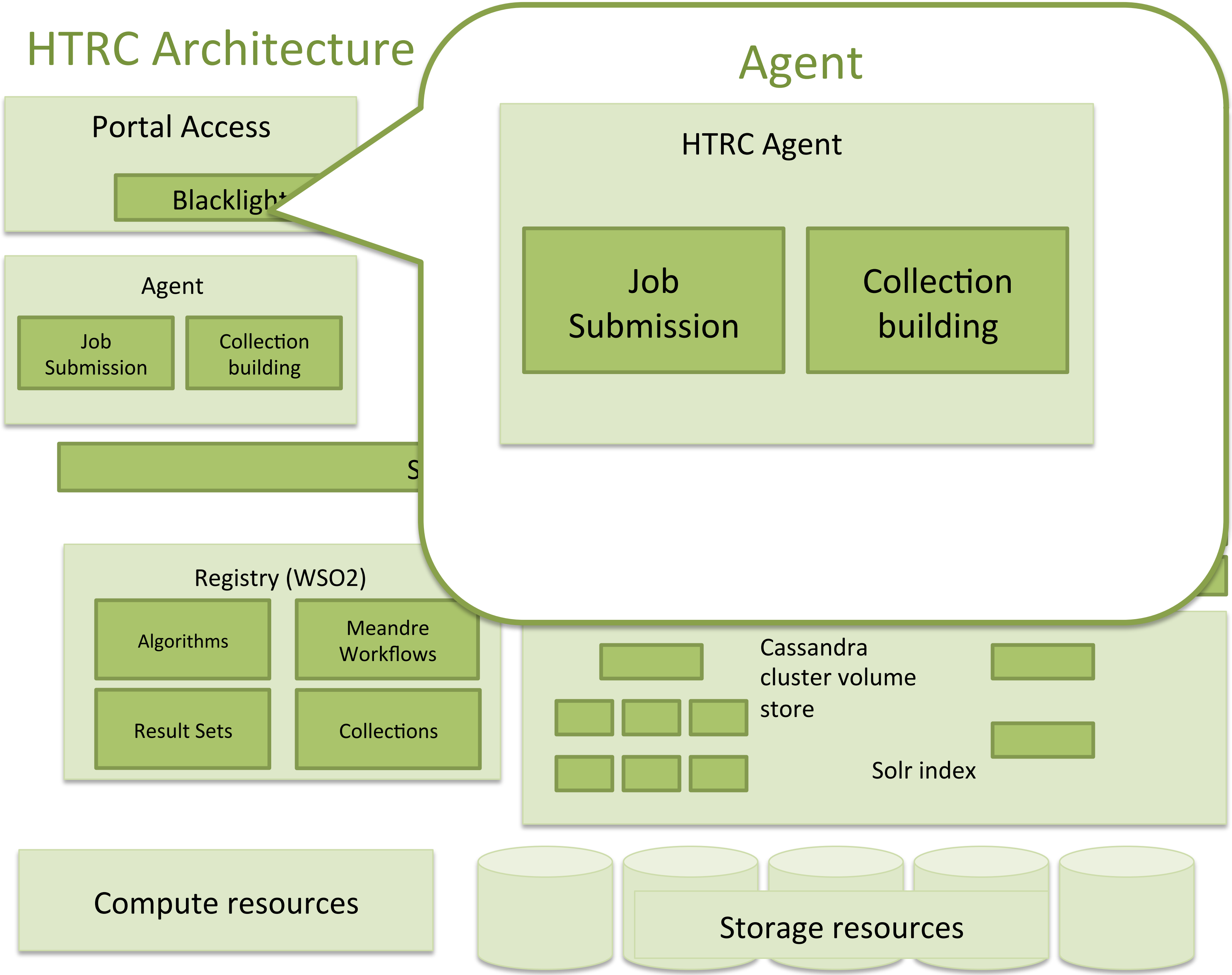
HTRC Portal

Blacklight

App SEAR

App Blacklight

HTRC Architecture



Portal Access

Blacklight

Agent

Job Submission

Collection building

S

Registry (WSO2)

Algorithms

Meandre Workflows

Result Sets

Collections

Cassandra cluster volume store

Solr index

Compute resources

Storage resources

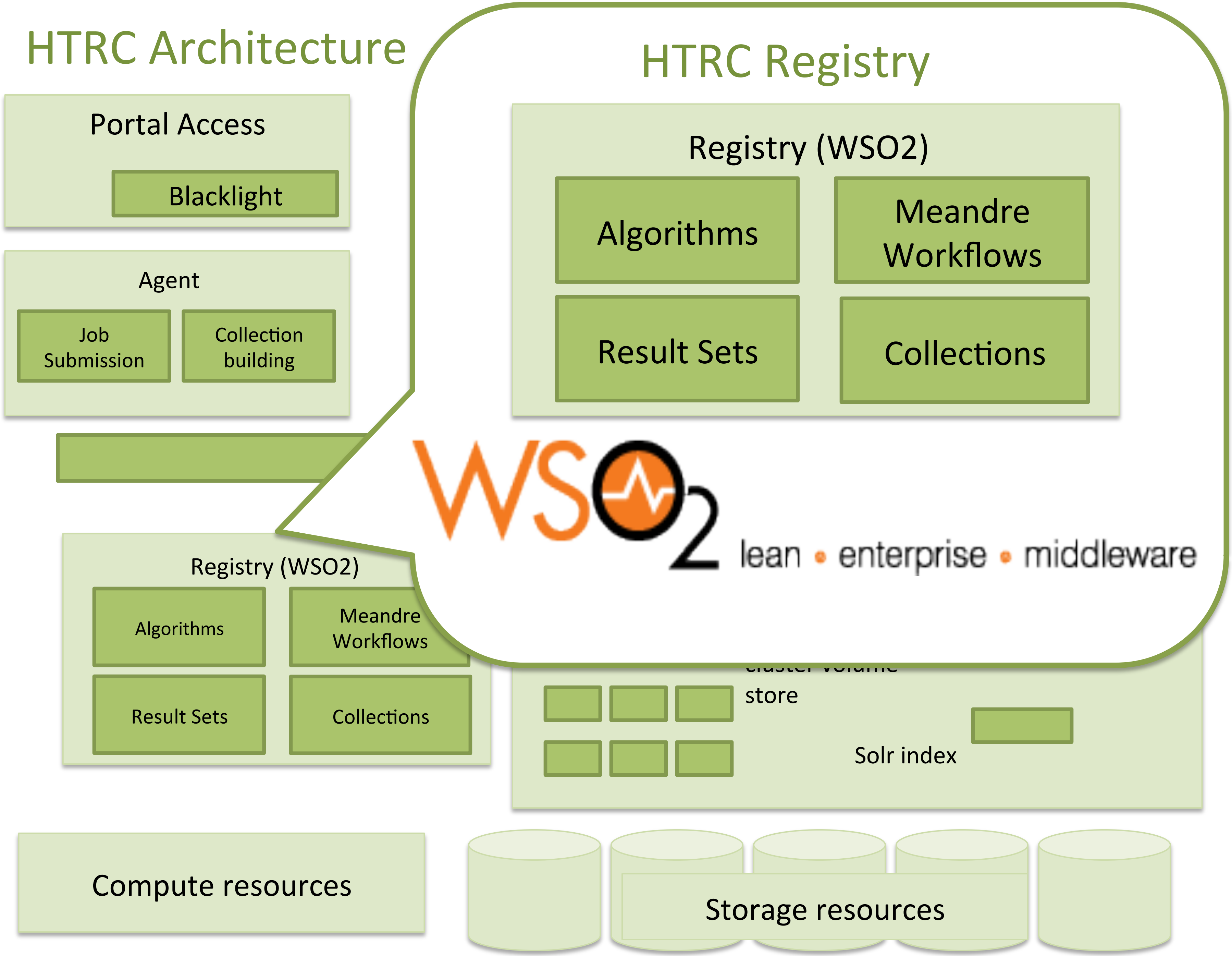
Agent

HTRC Agent

Job Submission

Collection building

HTRC Architecture

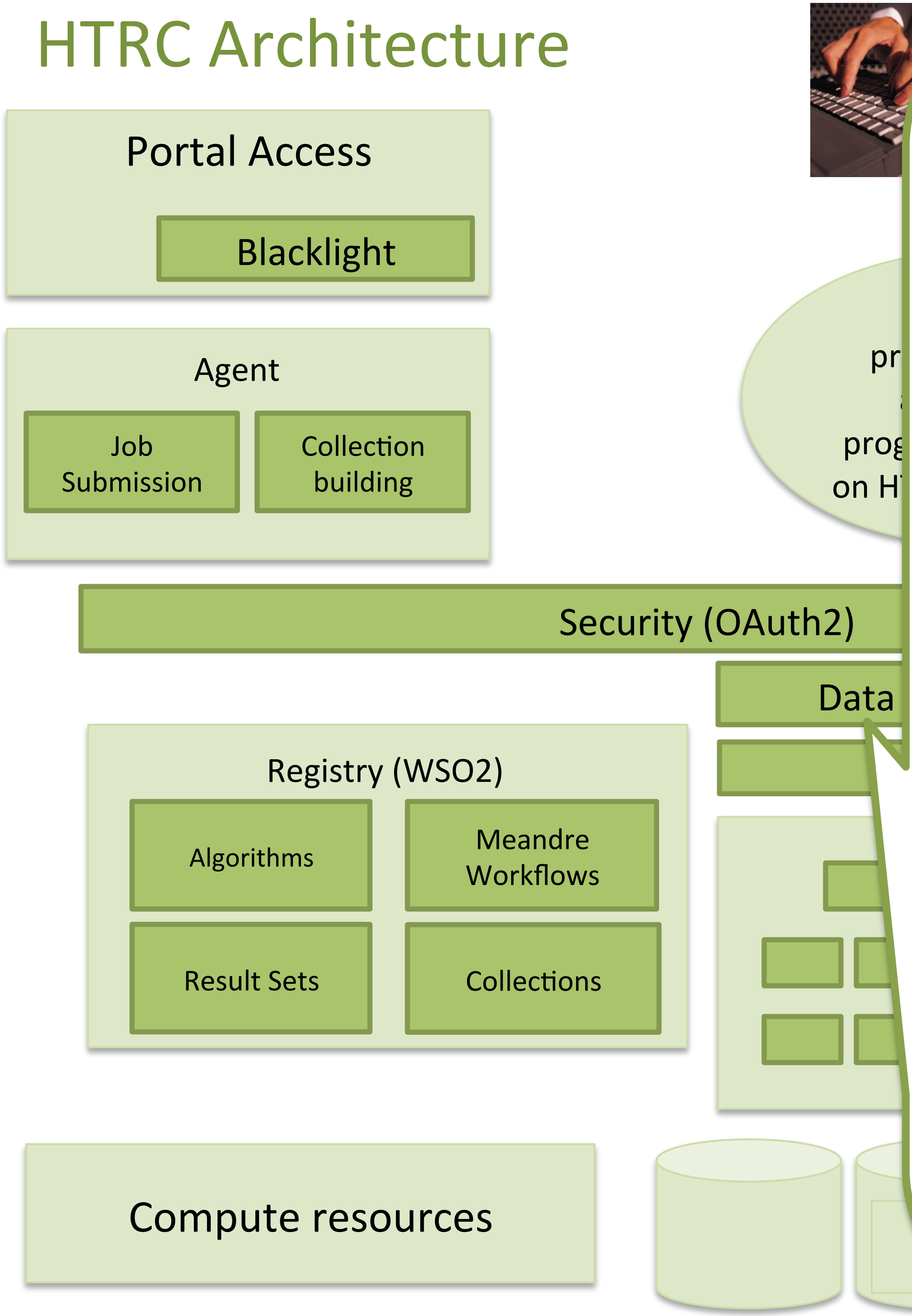


HTRC Registry

Registry (WSO2)

- Algorithms
- Meandre Workflows
- Result Sets
- Collections

HTRC Architecture

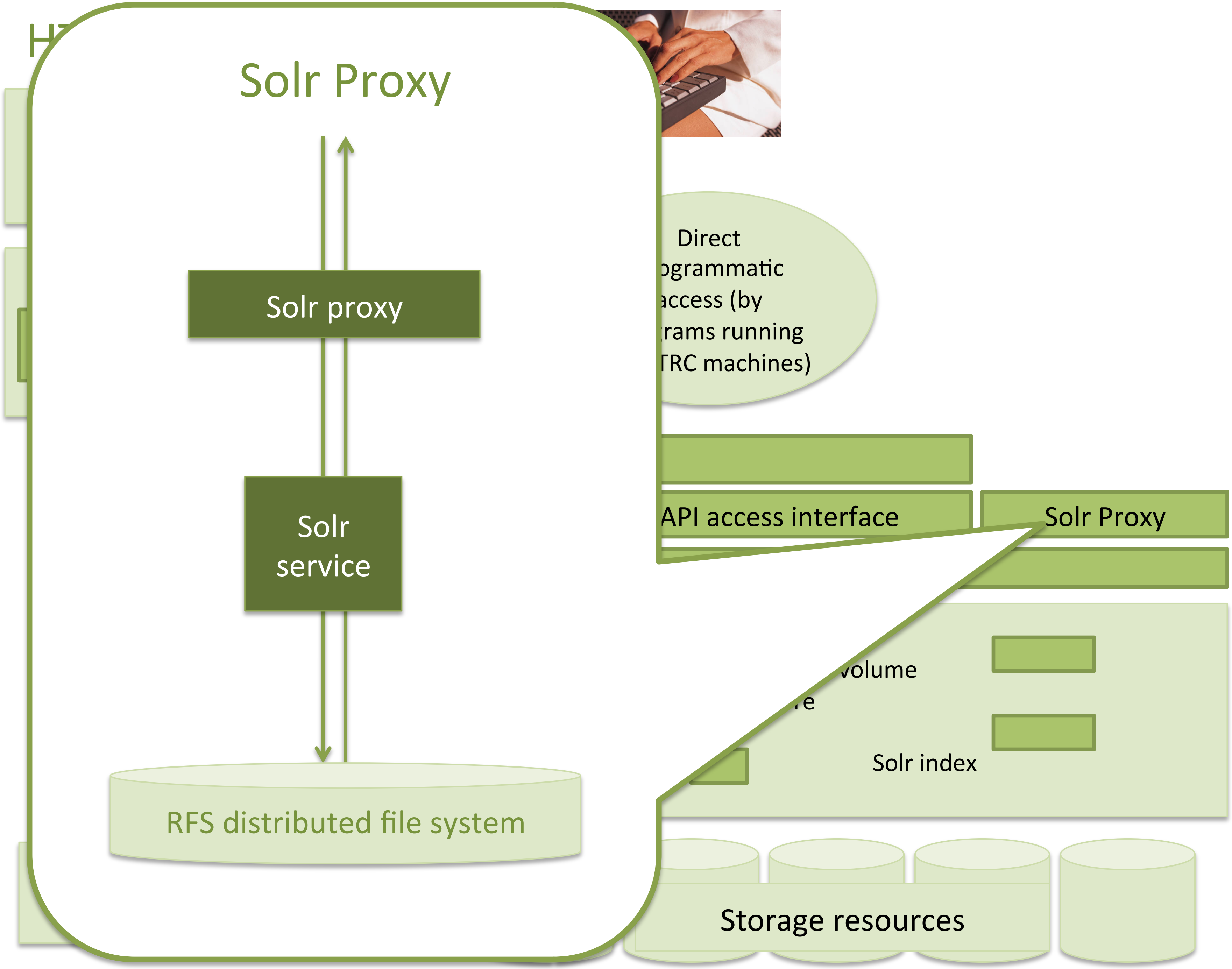


Secure Data API

- RESTful Web Service
  - Language agnostic
  - Clients don’t have to deal with Cassandra
- Simple OAuth2 authentication
- HTTP over SSL
- Audits client access
- Protected behind firewall, accessible only to authorized IPs

HTRC

Solr Proxy



HT

Solr proxy

Solr service

RFS distributed file system

Direct
ogrammatic
access (by
rams running
TRC machines)

API access interface

Solr Proxy

volume

e

Solr index

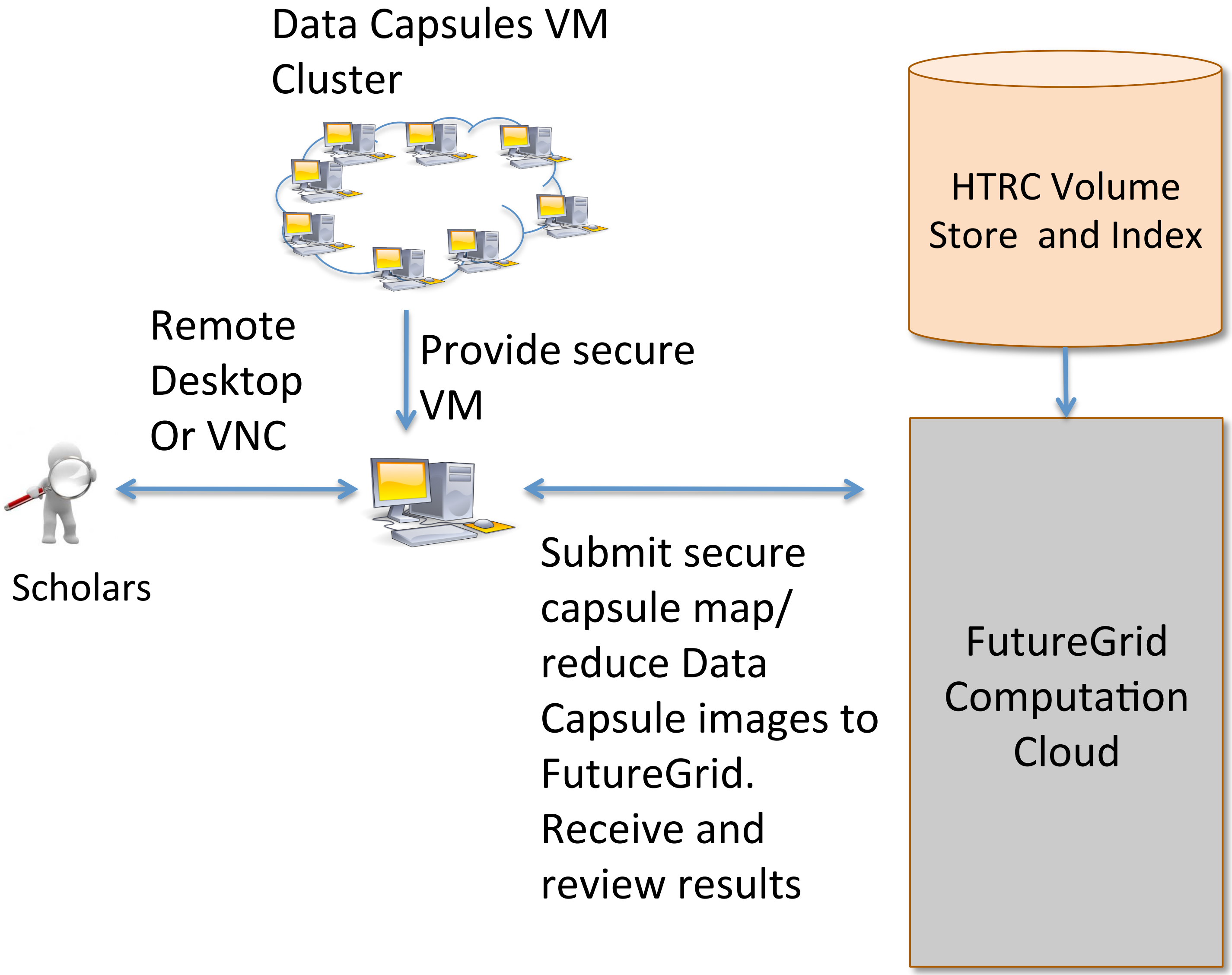
Storage resources

NoSQL Methodology

- Currently HT content is stored in a pair-tree file system convention (CDL)
- Moving these files into a NoSQL store like Cassandra enabled HTRC to aggregate them into larger sets of files for use in retrieval
- Use of Cassandra enabled HTRC to share content over a commodity based Cassandra cluster of virtual machines
- Originally investigated use of MongoDB, CouchDB, Hbase and Cassandra

HTRC Solr Proxy + Solr Service

- Preserves all query syntax of original Solr
- Prevents user from modification
- Hides the host machine and port number HTRC Solr is actually running on
- Creates audit log of requests
- Provides filtered term vector for words starting with user-specified letter
- Filters out “dangerous” requests to Solr
- Adds additional features to Solr
  - E.g. Term Vectors



Data Capsules VM Cluster

Remote Desktop Or VNC

Provide secure VM

Scholars

Submit secure capsule map/ reduce Data Capsule images to FutureGrid. Receive and review results

HTRC Volume Store and Index

FutureGrid Computation Cloud

Non-Consumptive Research-Secure Data Capsule

Sessions for Further Review

- For more on Secure Data API – Tues Topic I/II (Yiming Sun)
- For more on Portal/SEASR – Tues Topic II (Loretta Auvil)
- For more on Portal/Blacklight – Tues Topic III (Stacy Kowalczyk)

Contact Information

- Robert H. McDonald
  - Email – robert@indiana.edu
  - Chat – rhmcdonald on googletalk | skype
  - Twitter - @mcdonald
  - Blog – http://www.rmcdonald.net
  - Twitter Hashtag: #HTRC12