

HATHI TRUST RESEARCH CENTER

HathiTrust Research Center: Challenges and Opportunities in Big Text Data

Digital Library Brown Bag | 5.Mar.14

Presented by Miao Chen
Research Associate, Data To Insight Center, IU

Beth Plale – @bplale
Professor, School of Informatics and Computing
Director, Data To Insight Center
Indiana University



Tweet us –
#dlbb
@HathiTrust #HTRC

Thanks to sponsors



**ALFRED P. SLOAN
FOUNDATION**



INDIANA UNIVERSITY



ILLINOIS



The Andrew W. Mellon Foundation



HathiTrust Digital Library

- HathiTrust is a partnership of academic & research institutions, offering a collection of millions of titles digitized from libraries around the world.
 - Founding members of HathiTrust along with University of Michigan are Indiana University, University of California, and University of Virginia



<http://www.hathitrust.org>

→ Distinguished
from



<http://www.hathitrust.org/htrc>



Currently Digitized (by 2/11/2014)

- 11,014,179 total volumes
- 5,750,943 book titles
- 286,864 serial titles
- 3,854,962,650 pages
- 494 terabytes
- 130 miles
- 8,949 tons
- 3,637,649 volumes (~33% of total) in the public domain

http://www.hathitrust.org/statistics_info

→ HathiTrust repository is a latent goldmine for text mining analysis, analysis of large-scale corpora through computational tools, and time-based analysis

→ Restricted nature of HT content suggests need for new forms of access that preserve intimate nature of research investigation while honoring restrictions

→ Paradigm: computation takes place close to the data

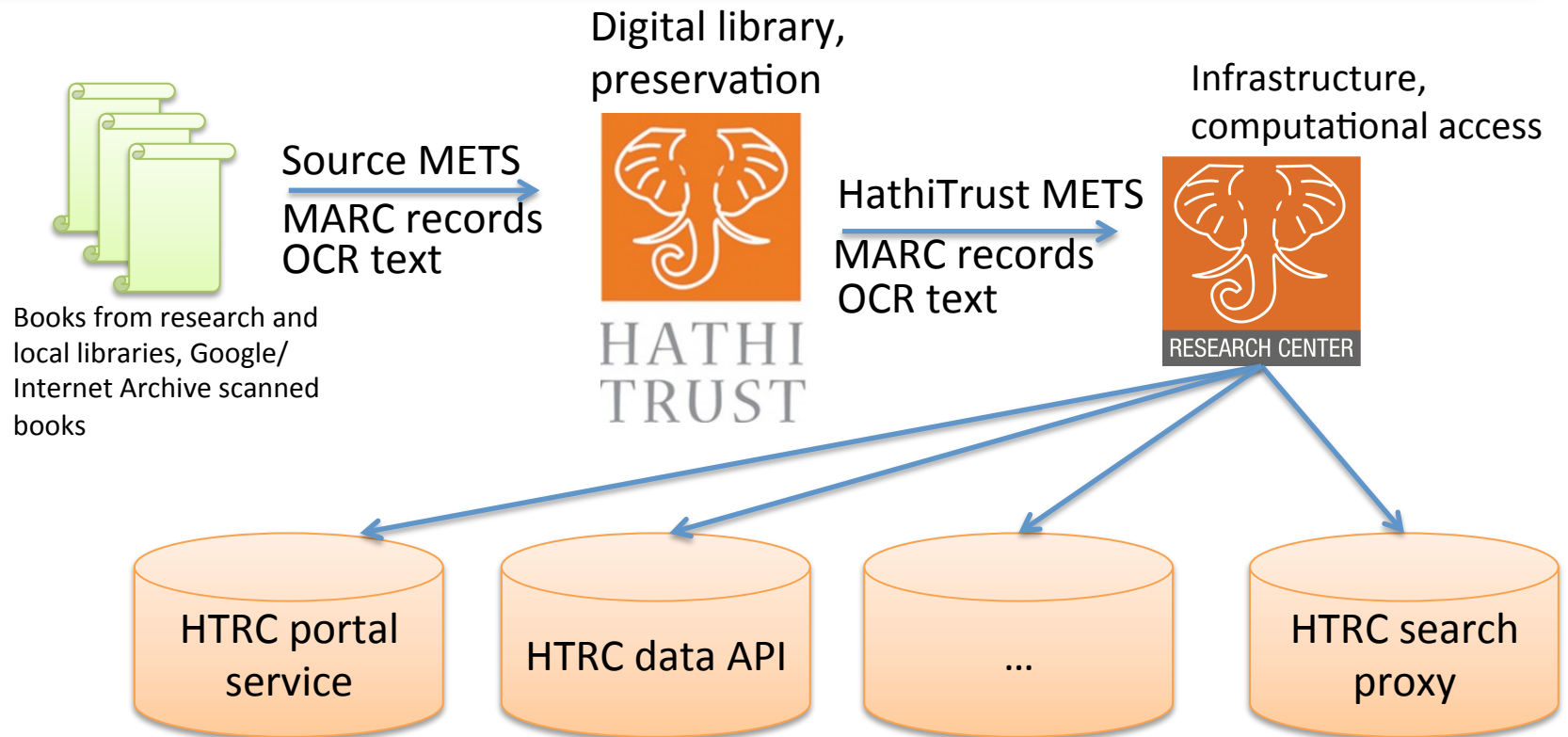


RESEARCH CENTER

Mission of HT Research Center

- Research arm of HathiTrust
- Goal: enable researchers world-wide to carry out computational investigation of HT repository through
 - Develop model for access: the ‘workset’
 - Develop tools that facilitate research by digital humanities and informatics communities
 - Develop secure cyberinfrastructure that allows computational investigation of entire copyrighted and public domain HathiTrust repository
- Established: July, 2011
- Collaborative effort of Indiana University, University of Illinois, and HathiTrust

Books => HT => HTRC



Partitioning the HTRC Collections

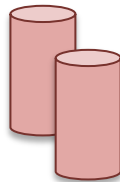
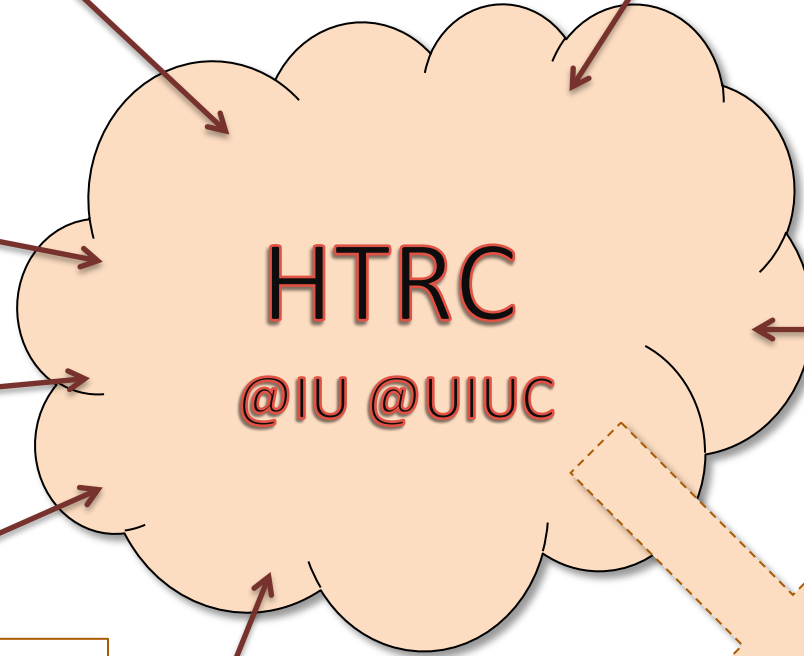
- Public domain corpus (~3.6M)
 - Non-Google digitized (~250K)
 - Also referred to as the “open-open” corpus
 - On sandbox
 - Google digitized (~2-3M)
 - Production stack
- In-copyright corpus (~7.5M)
 - in progress



TEXT MINING
TOOLS

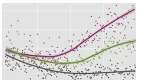
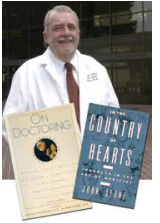


EXTRACTED
FEATURE
SETS



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

Complexity hiding interface



BLUE WATERS

HTRC architecture



- Philosophy: computation moves to data
- Web services (REST) architecture and protocols
- WS02 Registry for worksets and results
- Solr Indexes: full text, MARC, and new metadata
- noSQL (Cassandra) store as volume store
- Authentication using WSO2 Identity Server
- Portal front-end, programmatic access
- Mining tools: currently SEASR

ssh client



Portal

Blacklight



Secure Capsule Service

Secure Capsule Instance Manager

User session management

Sigiri job deployment

SEASR analytics service

Meandre Orchestration

HTRC Data API v0.1

Registry Services, worksets

Solr index

Volume store (Cassandra)

Page/volume tree (file system)

Identity Server

IU compute resources

Secure Capsule Cluster

Hadoop Cluster (MapReduce/HDFS)

SEASR service execution

rsync

HathiTrust corpus

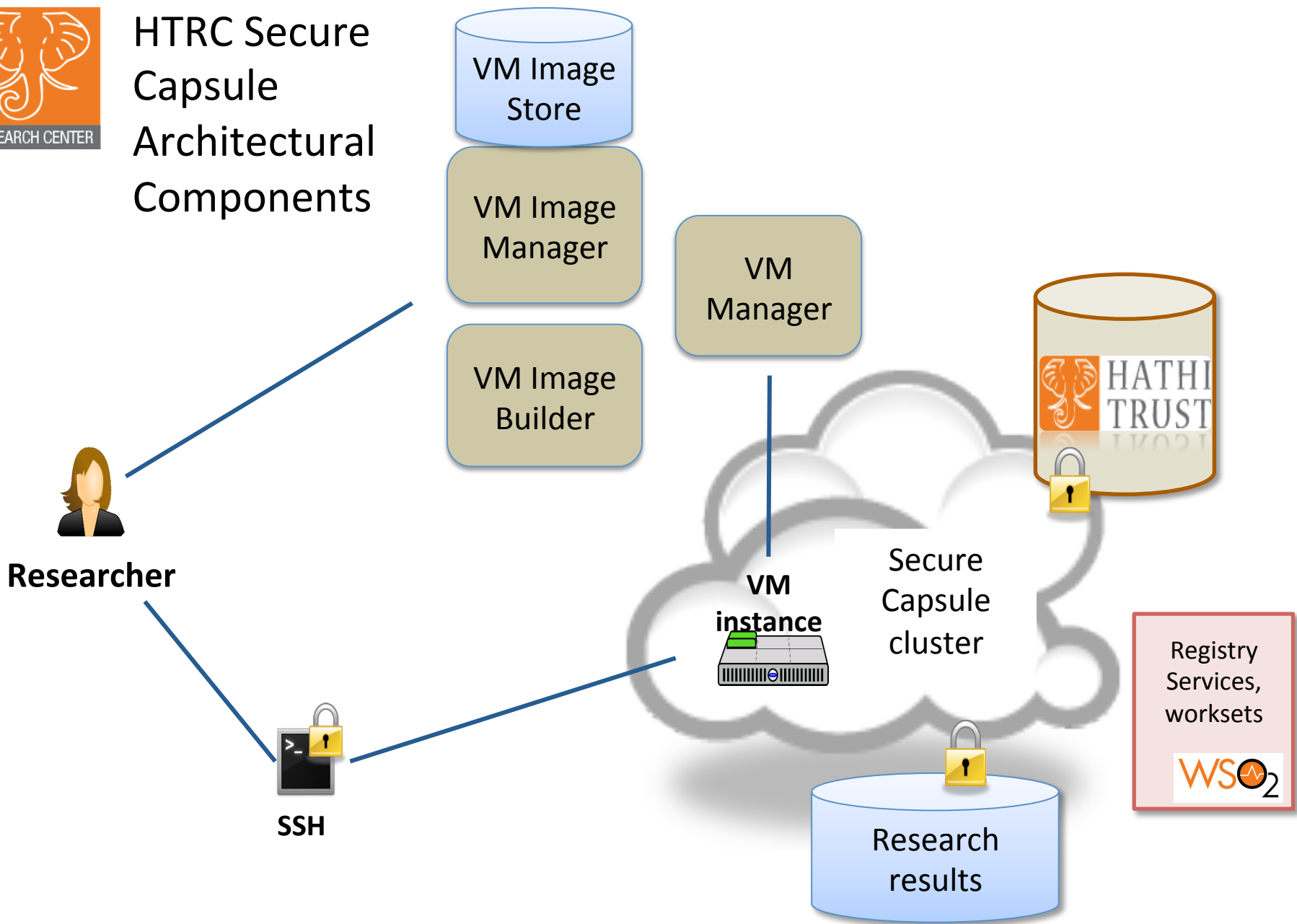
University of Michigan

HTRC's guiding principle to computational access

- *No computational 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 the HT repository to reassemble pages from collection for reading*
- Definition disallows collusion between users, or accumulation of material over time.
- Defining “sufficient information”: research has shown need to interact directly with select texts. How much of a text to show? Google withholds from showing to reader every 10th page of a book (Int'l NYTimes Nov 16-17, 2013)

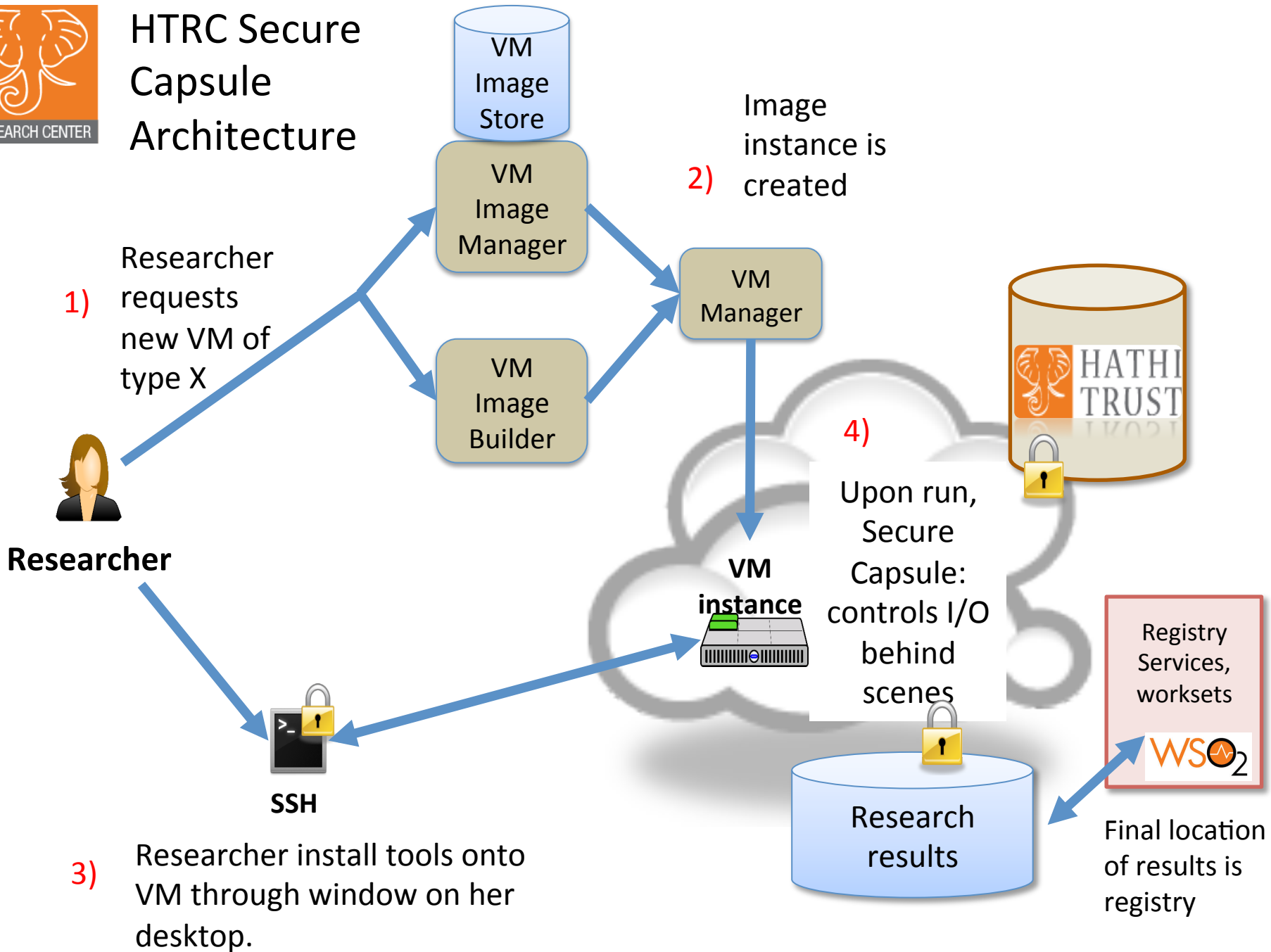


HTRC Secure Capsule Architectural Components



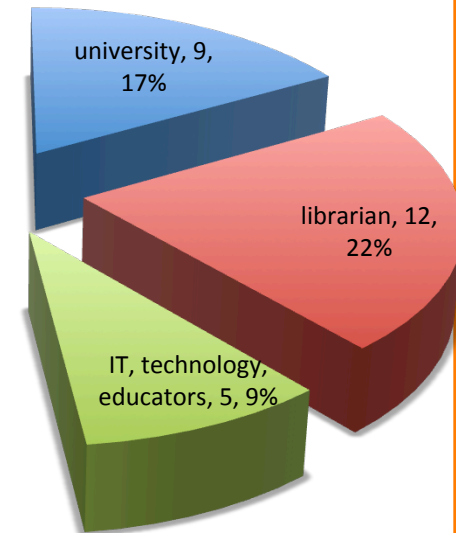
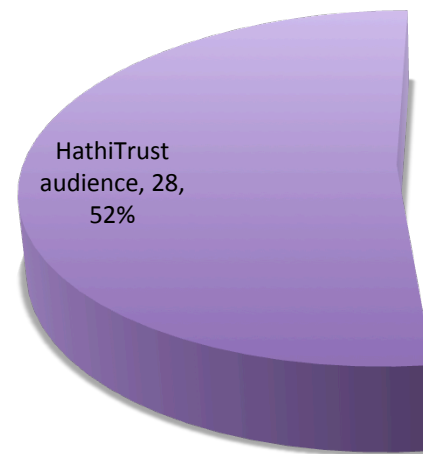
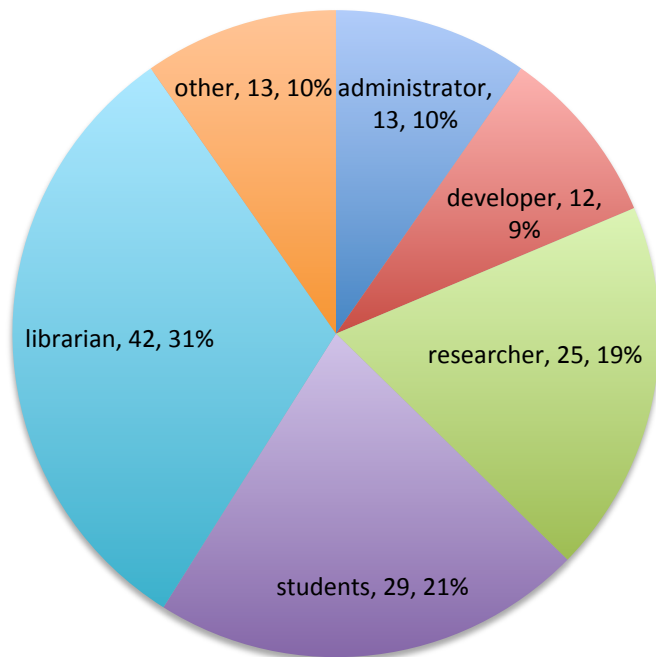


HTRC Secure Capsule Architecture



HTRC and library

- Out of the 134 UnCamp participants, 42 were librarians (31%)
- Target audience of HTRC news (22% to librarians)



Metadata!

HTRC metadata

- Volume
 - The central entity for metadata description
 - Books, journal, serials, government docs...
- MARC records (xml uploaded from libraries)
 - Specs: http://www.hathitrust.org/bib_specifications
 - Contains OCLC number (master record number)
- METS(metadata encoding and transmission standard)
 - Source METS files, for preservation (from Google, Internet Archive)
 - HathiTrust METS files, for both preservation and access
- Plus HTRC specific metadata fields

HTRC metadata

You searched for:

information representation x

Showing item 1 of 869,789 from your search.

Health and Human Services : update on Hispanic representation in HHS Region VIII : fact sheet for the Honorable Timothy E. Wirth, U.S. Senate / United States General Accounting Office.

[Full View](#)

MARC standard

Title: Health and Human Services : update on Hispanic representation in HHS Region VIII : fact sheet for the Honorable Timothy E. Wirth, U.S. Senate / United States General Accounting Office.
Subtitle: Update on Hispanic representation in HHS Region VIII.
Author: United States. General Accounting Office.
Language: English
Published: 1992
Country: United States
Call number: HD4903.5.U6 A3523
OCLC: (OCoLC)ocm27370022
Source: University of Michigan
Volume ID: mdp.39015048857141

HTRC specific

Character count: 21,008
Page count: 24
Volume size by page: Small
Volume size by word: Small
Word count: 3,261

Search result page

HTRC specific metadata

Work set display page

The screenshot shows the HTRC Portal interface. The top navigation bar includes 'Home', 'About', 'Worksets', 'Algorithms', 'Results', and 'Help'. The user is signed in as 'miao'. The main content area is divided into 'Available Worksets' and 'Workset Details'. The 'Available Worksets' list includes 'AncientGreek', 'biglaw_sep6upload', 'nlp', 'nlpontology_classlabel', '1darwin-test', '2darwin-english', '2vesalius', 'Anarchism', 'Austen_Dickens_Labels', 'Author_Twain', 'BestCoreComplex', 'BigLaw', 'Bleakhouse', 'Cicero_Orations_Letters', 'Coffee_Books', 'Dickens_as_Authors', 'Dickens_yo', 'Diderot-test', 'DocSouthMatch', 'EAPoeAsAuthor', 'ECCOmatch', 'EEBOmatch', 'Edgar_Allan_Poe', and 'Elizabeth_Gaskell_Works'. The 'Workset Details' section for 'AncientGreek' shows the following information: Name: AncientGreek, Description: query=ancient greek, Author: miao, Last Modified By: miao, Last Modified Time: 2013-09-03T15:37:29-04:00, and Number of Volumes: 20. There are buttons for 'Edit Workset' and 'Download CSV File'. Below this is a pagination control with 'First', 'Prev', 'Next', and 'Last' buttons. The main content area displays two numbered items:

1. Books which we recommend to our members who intend to visit Greece.

Volume Id: yale.39002089540257

Author:

Page Count[HTRC]: 4

Word Count[HTRC]: 806

2. The claim of antiquity, with an annotated list of books for those who know neither Latin nor Greek; issued by the councils of the Societies for the promotion of Hellenic and Roman studies and of the Classical association.

Volume Id: mdp.39015033434559

Author:

Page Count[HTRC]: 36

Word Count[HTRC]: 7989

Metadata Enhancement

- Big data needs good metadata
- Current metadata fields are MARC-based
 - E.g. publication date, authors, title, subject
- MARC fields are fundamental
- Needed more fields of users' interest for granular analytics (Metadata Enhancement)
- Solicit user requirements and prioritize for implementation
 - Mainly digital humanities uses now

Top Metadata Enhancement Items

- 1st round user requirement collection, top 3 items were metadata related:
 - Word frequency count and document length for a volume
 - Metadata de-duplication
 - Author Gender Analysis
- We have added word count and gender fields to HTRC metadata, and more are being planned and investigated.

HTRC specific metadata: Gender

Available Worksets



Workset Details

AncientGreek
biglaw_sep6upload
nlp
nlpontology_classlabel
ldarwin-test
ldarwin-english
2vesalius
Anarchism
Austen_Dickens_Labels
Author_Twain
BestCoreComplex
BigLaw
Bleakhouse
Cicero_Orations_Letters
Coffee_Books
Dickens_as_Authors
Dickens_yo
Diderot-test
DocSouthMatch
EAPoeAsAuthor
ECCOmatch
EBOmatch
Edgar_Allan_Poe
Elizabeth_Gaskell_Works

Name: Austen_Dickens_Labels
Description:
Author: lauvil
Last Modified By: drhtrc
Last Modified Time: 2013-09-03T14:46:32-04:00
Number of Volumes: 12

[Edit Workset](#) [Download CSV File](#)

← First ← Prev

Next → Last →

1. Emma / by Jane Austen.

Volume Id: nyp.33433074943568
Author: Austen, Jane, 1775-1817 **F**
Page Count[HTRC]: 330
Word Count[HTRC]: 82135

2. Mansfield Park / by Jane Austen.

Volume Id: nyp.33433074943618
Author: Austen, Jane, 1775-1817 **F**
Page Count[HTRC]: 328
Word Count[HTRC]: 83658

3. David Copperfield ...

Volume Id: nyp.33433074954060
Author: Dickens, Charles, 1812-1870 **M**
Page Count[HTRC]: 1252
Word Count[HTRC]: 363656

4. Northanger abbey / by Jane Austen.

Volume Id: uc1.31158008377706
Author: Austen, Jane, 1775-1817 **F**
Page Count[HTRC]: 252

Gender field

Other Metadata Enhancement Items

- Stats analysis: tf-idf
- Readability score
- Language
- Topic modeling (e.g. LDA probability)
- Genre
- Era of compilation
- Book length (e.g. short or long)
- Concordance index (indexing with context)

Portal and Work Set Builder

- <https://htrc2.pti.indiana.edu/HTRC-UI-Portal2/>

The screenshot displays the HTRC Workset Builder interface. At the top left is the HTRC Research Center logo. The top right navigation bar includes links for "Log Out [Beth Plale]", "Selected Items (580)", "Manage Worksets", and "Portal". The main header reads "HTRC Workset Builder".

A yellow banner at the top of the main content area states: "All items in this search were successfully selected". Below this is a search bar with a "Full Text" dropdown and a "Search" button. Underneath the search bar are "More options" including filters for "Language > English" and "Subject > World War, 1914-1918".

The main content area displays "Displaying items 1 - 10 of 580" with a "start over" button. Below this is a "Sort by" dropdown set to "relevance" and a "Show 10 per page" dropdown. Navigation links for "Previous" and "Next" are present, along with a page number sequence: "1 2 3 4 5 ... 57 58".

At the bottom of the main content area, there are buttons for "Select items on page", "Deselect items on page", "Select all search items", and "Deselect all search items".

On the left side, there is a "Limit your search" section with a "Subject" filter. The list includes:

- World War, 1914-1918 (580) [remove]
- World War, 1914-1918 Poetry (580) [remove]
- English poetry (19)
- American Field Service (12)
- World War, 1914-1918 Personal narratives, American (12)
- American poetry (11)
- Ambulances (9)
- Ambulances history (9)
- English poetry 20th century History and criticism (9)
- Transportation of Patients (9)
- World War I (9)
- World War I personal narratives (9)
- Great Britain (6)
- War poetry (6)
- Poets, English (5)
- Patriotic poetry (4)
- War (4)
- Canadian poetry (3)
- Poets, Canadian (3)
- American literature (2)

A "more »" link is located at the bottom of this list.

The first item in the list is:

1. "All's well!" : some helpful verse for these dark days of war / by John Oxenham. Select

Metadata for this item:

- Title: "All's well!" : some helpful verse for these dark days of war / by John Oxenham.
- Author: Oxenham, John.
- Language: English
- Published: 1915

Data API

- Data retrieval
 - <http://wiki.htrc.illinois.edu/display/COM/Python+client+for+accessing+volumes+in+bulk+through+HTRC+Data+API>
 - Demo code available in Python and Java
 - Page-level and volume-level word count
 - Option: Concatenate pages
 - Option: return METS metadata, of which MARC record is a part

Some Challenges

- Internationalization
 - Non-western text, especially for text processing
- Users contributed code
 - How to create a mechanism to describe and archive such contributions from the community?
- OCR errors
 - An experiment shows avg number of errors per page is 0.57
 - Crowdsourcing? Automatic correction?
- Implications
 - HTRC as a service or resource to library?

Going beyond the volume level?

- Work set level
 - Should be richer than a list of volume ids
 - The resources that scholars work with
 - Can be within HTRC corpus, or not
 - How to formalize it?
 - What's the metadata for work set?

Going beyond the volume level?

- Page level
 - Emerges as an important unit of analysis in the recent UnCamp in early Sep
 - Important to scholars in some circumstances
 - What are the important metadata fields?
 - E.g. Word frequency count, image/illustration

Recent Updates

- “The HTRC drafted documents covering system architecture, workflows, security measures, and data use cases in preparation for offering “non-consumptive” access to in-copyright volumes in the HathiTrust repository. “

Cited from http://www.hathitrust.org/updates_january2014