HATHI TRUST
RESEARCH CENTER

Text Mining HathiTrust as Big Data Exemplar in PRAGMA

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

INDIANA UNIVERSITY

RESEARCH CENTER

ILLINOIS

Tweet us - @HathiTrust #HTRC

HathiTrust Digital Library

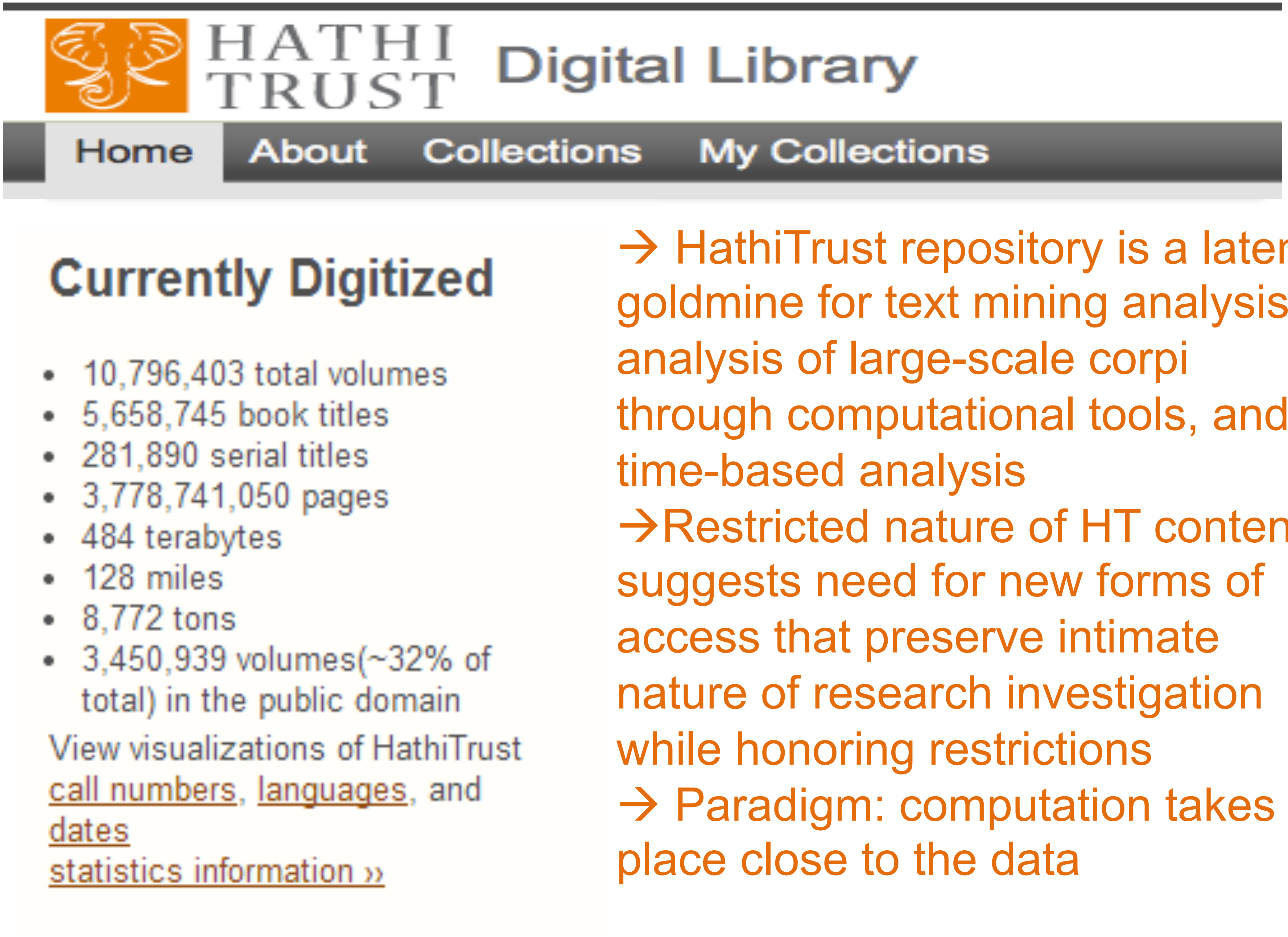
- HathiTrust is a partnership of academic & research institutions, offering a collection of millions of titles digitized from libraries around the world.
    - Founding members of HathiTrust along with University of Michigan are Indiana University, University of California, and University of Virginia

http://www.hathitrust.org

→ Distinguished from

RESEARCH CENTER

http://www.hathitrust.org/htrc



Currently Digitized

- 10,796,403 total volumes
- 5,658,745 book titles
- 281,890 serial titles
- 3,778,741,050 pages
- 484 terabytes
- 128 miles
- 8,772 tons
- 3,450,939 volumes(~32% of total) in the public domain

View visualizations of HathiTrust call numbers, languages, and dates

statistics information ››

→ HathiTrust repository is a latent goldmine for text mining analysis, analysis of large-scale corpi 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

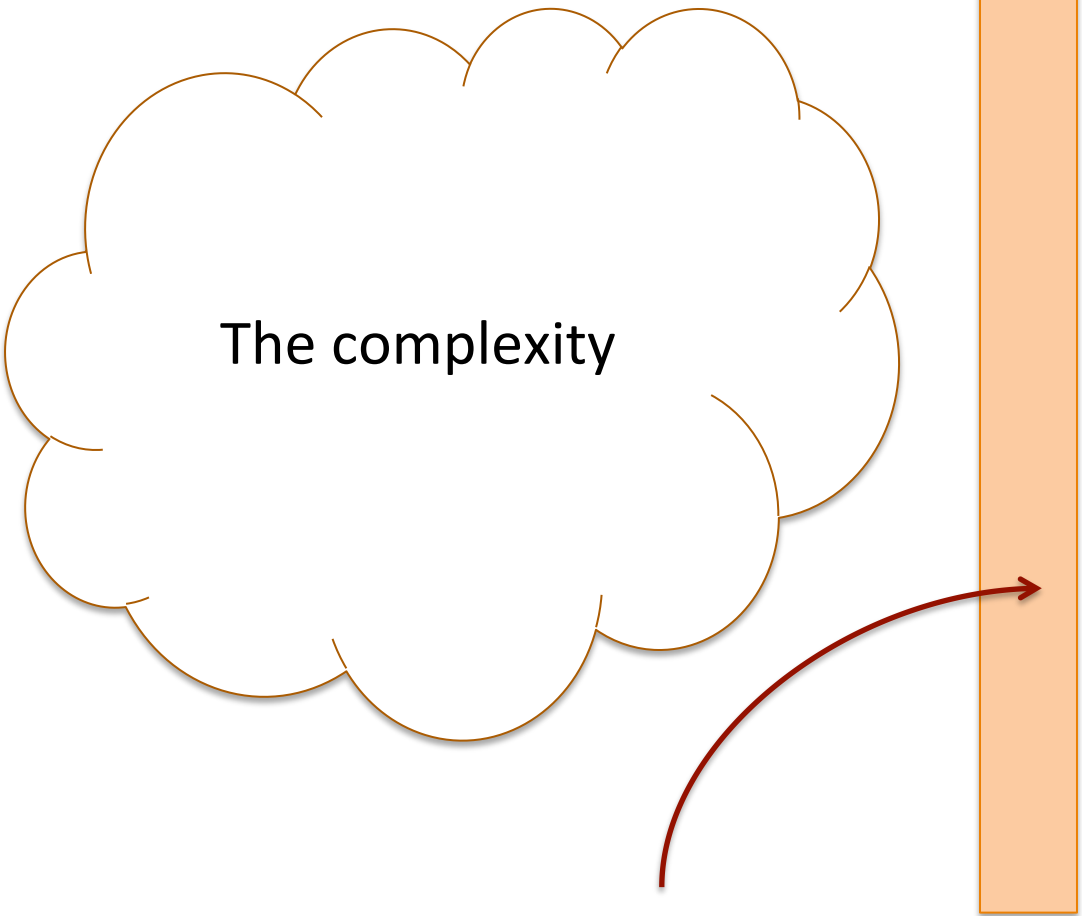
→ Paradigm: computation takes place close to the data

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 and University of Illinois

HTRC

RESEARCH CENTER

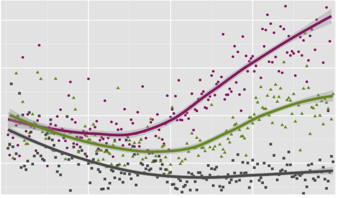


The complexity

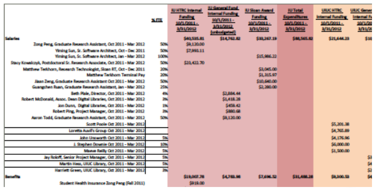
Complexity hiding interface

Request

Spatial plots

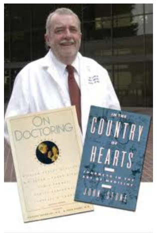


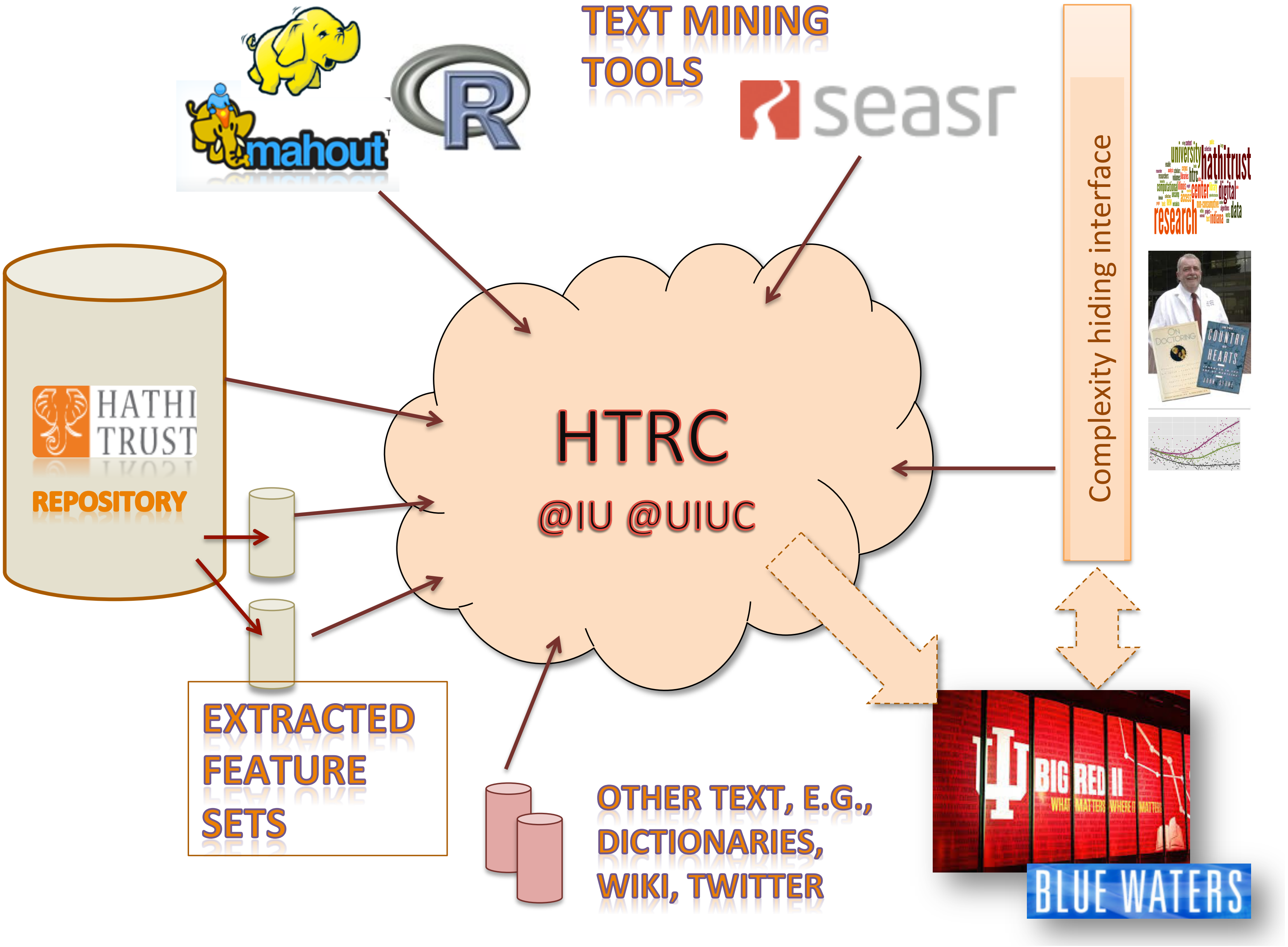
Statistical plots



	% FTE	IU HTRC Internal Funding 10/1/2011 - 3/31/2012	IU General Fund Internal Funding 10/1/2011 - 3/31/2012 (unbudgeted)	IU Sloan Award Funding 10/1/2011 - 3/31/2012	IU Total Expenditures 10/1/2011 - 3/31/2012	UIUC HTRC Internal Funding 10/1/2011 - 3/31/2012	UIUC Gene Internal F 10/1/20 3/31/2
Salaries		$40,535.81	$14,762.82	$33,267.19	$88,565.82	$21,644.23	$10
Zong Peng, Graduate Research Assistant, Oct 2011 - Mar 2012	50%	$9,120.00					
Yiming Sun, Sr. Software Architect, Oct - Dec 2011	50%	$7,993.11					
Yiming Sun, Sr. Software Architect, Jan - Mar 2012	100%			$15,986.22			
Stacy Kowalczyk, Postdoctoral Sr. Research Associate, Oct 2011 - Mar 2012	50%	$23,422.70					
Matthew Terkhorn, Research Technologist, Sloan RT, Oct - Dec 2011	20%			$3,045.00			
Matthew Terkhorn Terminal Pay	20%			$1,315.97			
Jiaan Zeng, Graduate Research Assistant Oct 2011 - Mar 2012	50%			$10,640.00			
Guangchen Ruan, Graduate Research Assistant, Jan - Mar 2012	25%			$2,280.00			
Beth Plale, Director, Oct 2011 - Mar 2012	4%		$2,884.44				
Robert McDonald, Assoc. Dean Digital Libraries, Oct 2011 - Mar 2012	3%		$1,418.28				
Jon Dunn, Digital Libraries, Oct 2011 - Mar 2012	1%		$459.42				
Robert Ping, Project Manager, Oct 2011 - Mar 2012	3%		$880.68				
Aaron Todd, Graduate Research Assistant, Oct 2011 - Mar 2012	50%		$9,120.00				
Scott Poole Oct 2011 - Mar 2012						$5,201.38	
Loretta Auvil's Group Oct 2011 - Mar 2012						$4,765.89	
John Unsworth Oct 2011 - Mar 2012	5%					$4,176.96	
J. Stephen Downie Oct 2011 - Mar 2012	10%					$6,000.00	
Maeve Reilly Oct 2011 - Mar 2012	5%					$1,500.00	
Jay Roloff, Senior Project Manager, Oct 2011 - Mar 2012	5%						$3
Martin Hesz, UIUC Library, Oct 2011 - Mar 2012	5%						$4
Harriett Green, UIUC Library, Oct 2011 - Mar 2012	3%						$2
Benefits		$19,007.78	$4,783.98	$7,696.52	$31,488.28	$9,300.53	$4
Student Health Insurance Zong Peng (Fall 2011)		$919.00					

Tabular info





TEXT MINING TOOLS

HTRC

@IU @UIUC

REPOSITORY

EXTRACTED FEATURE SETS

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

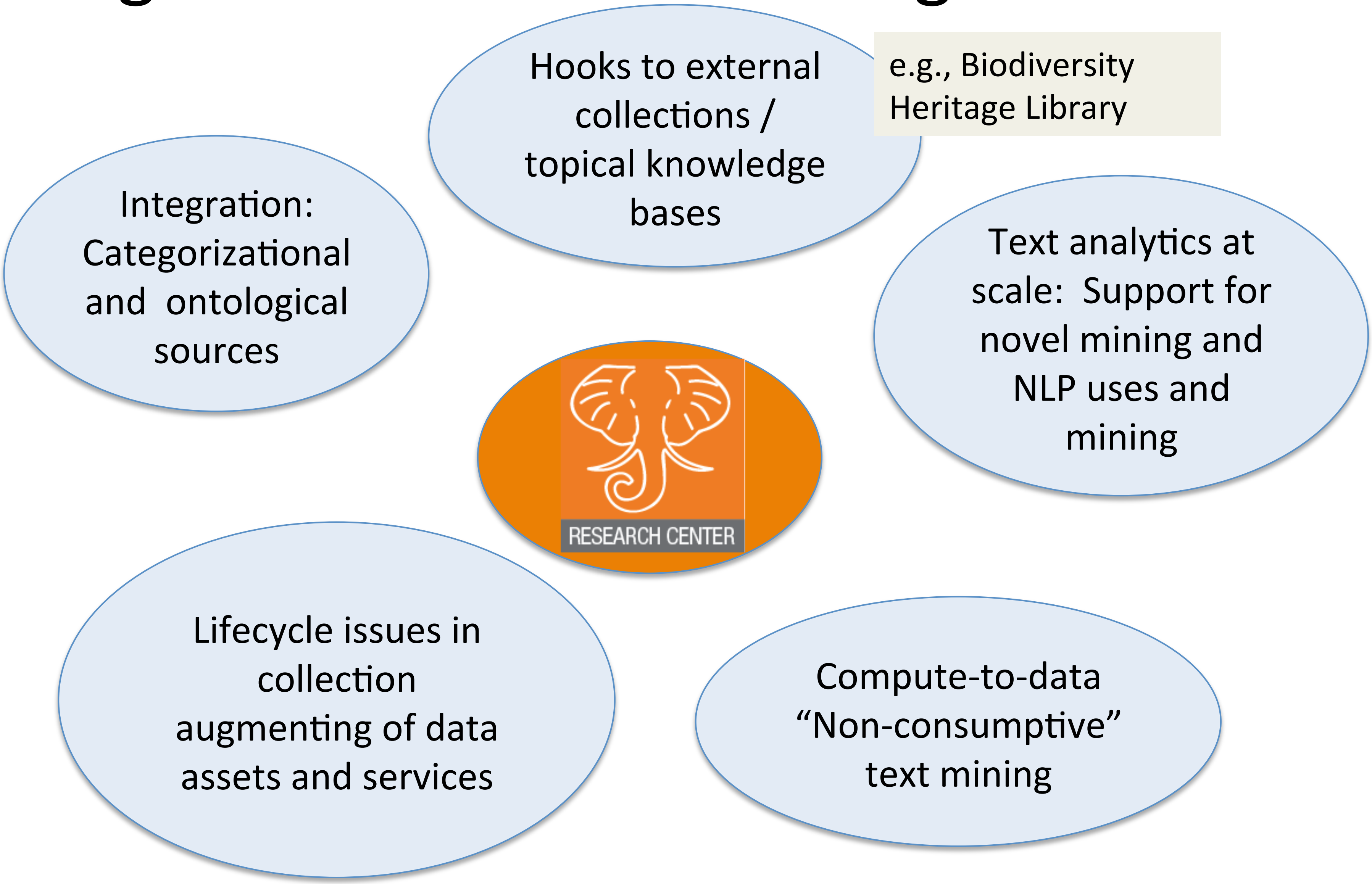
Complexity hiding interface

HTRC architecture

RESEARCH CENTER

- 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 WS02 Identity Server
- Portal front-end, programmatic access
- Mining tools: currently SEASR

Big Data Research Challenges in HTRC



Hooks to external collections / topical knowledge bases

e.g., Biodiversity Heritage Library

Integration: Categorizational and ontological sources

Text analytics at scale: Support for novel mining and NLP uses and mining

Lifecycle issues in collection augmenting of data assets and services

Compute-to-data “Non-consumptive” text mining

Big Data and PRAGMA

- Use PRAGMA member collections as anchor collections around which PRAGMA testbed could exercise experiments in data discovery and interoperability
- PRAGMA uses its testbed to exercise relevant activity of Research Data Alliance. RDA WG on Type Registries is about to deliver; assess usefulness in PRAGMA use cases
- Data on program for PRAGMA 27: devote time to address/expose policy and cultural issues in data sharing; are there tech solutions to policy issues that PRAGMA would be interested in?