HATHITRUST RESEARCH CENTER

What's Next: HathiTrust Research Center

November 10, 2016 | HT Member Meeting

HTRC Executive Management Team

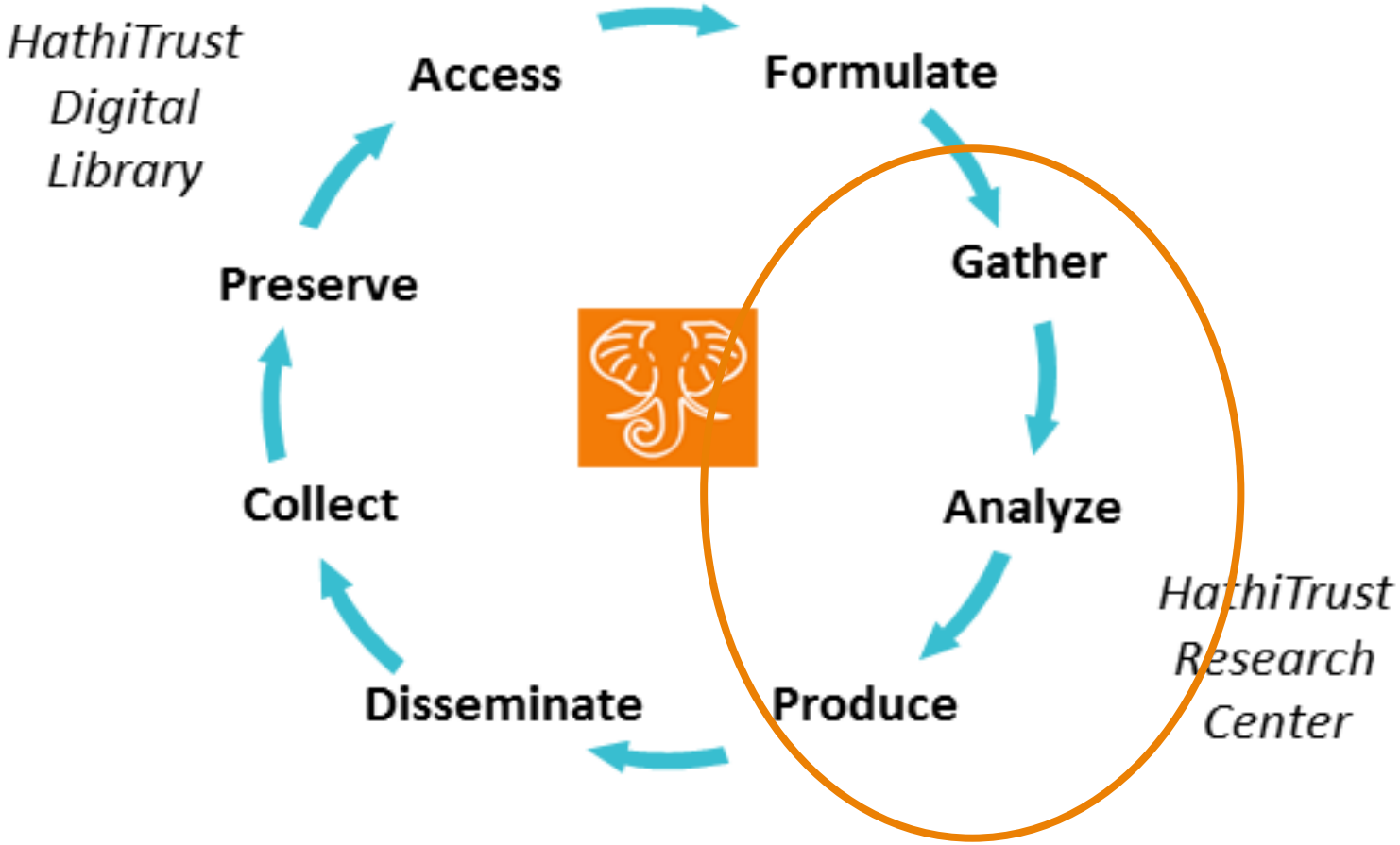
HTRC Overview

About the HathiTrust Research Center

- Facilitates text analysis of HTDL content
  - Large-scale, computational research
- Research & Development
  - Conducting user studies
  - Finding technical solutions
  - Building tools and services
- Collaboration:
  - HathiTrust | University of Illinois Urbana-Champaign | Indiana University

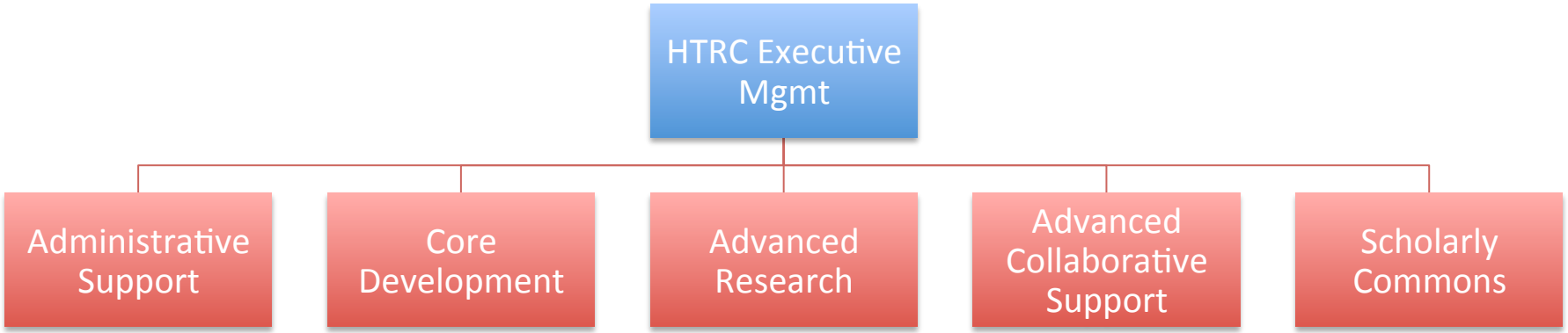
RESEARCH CENTER

HTRC Eco-System



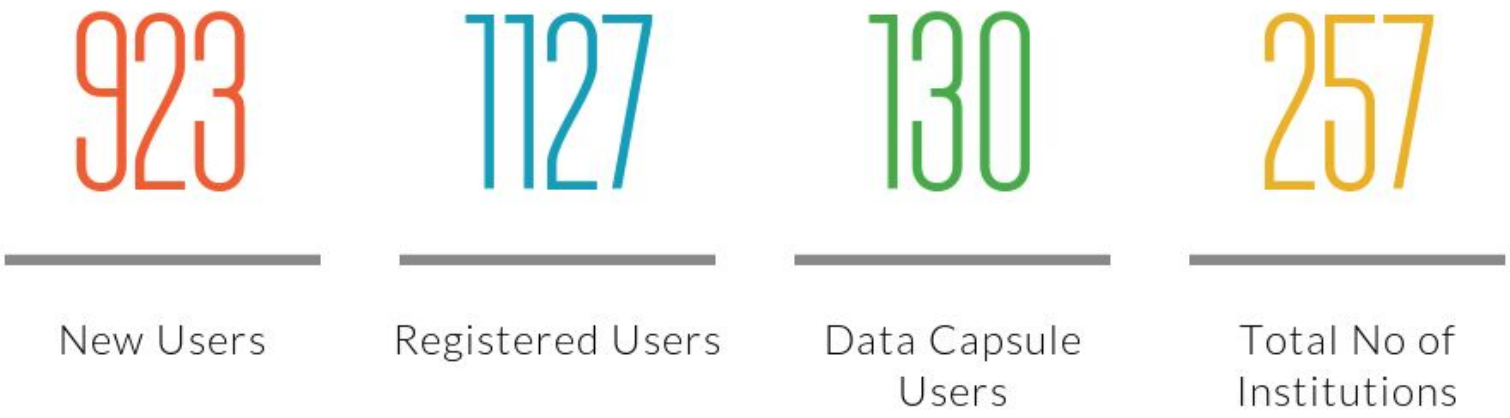
RESEARCH CENTER

HTRC 2014-2018 Org Chart



RESEARCH CENTER

HTRC Growth 2014-2016



RESEARCH CENTER

New Advisory Board (Pt. 1)

- *Wolfram Horstmann*, University Librarian, Göttingen Library & Project Lead, TextGrid
- *Nancy Ide*, Professor, Department of Computer Science, Vassar
- *Allan Lu*, Vice President of Research Tools, Services, and Platform, ProQuest
- *Greg Raschke*, HathiTrust Program Steering Committee member, Associate Director for Collections and Scholarly Communication, North Carolina State University
- *Matthew Sag*, Professor of Law, Loyola University, Chicago

RESEARCH CENTER

New Advisory Board (Pt. 2)

- *Claire Stewart*, Associate University Librarian for Research and Learning, University of Minnesota Libraries
- *Craig Stewart*, Executive Director, Pervasive Technology Institute, Indiana University
- *Stefan Sinclair*, Associate Professor, Department of Languages, Literatures, and Cultures, McGill University & Project Lead, Voyant Tools
- *John Towns*, Executive Director for Science and Technology, National Center for Supercomputing Applications (NCSA)
- *Jennifer Vinopal*, Librarian for Digital Scholarship Initiatives, New York University

RESEARCH CENTER

HTRC Access

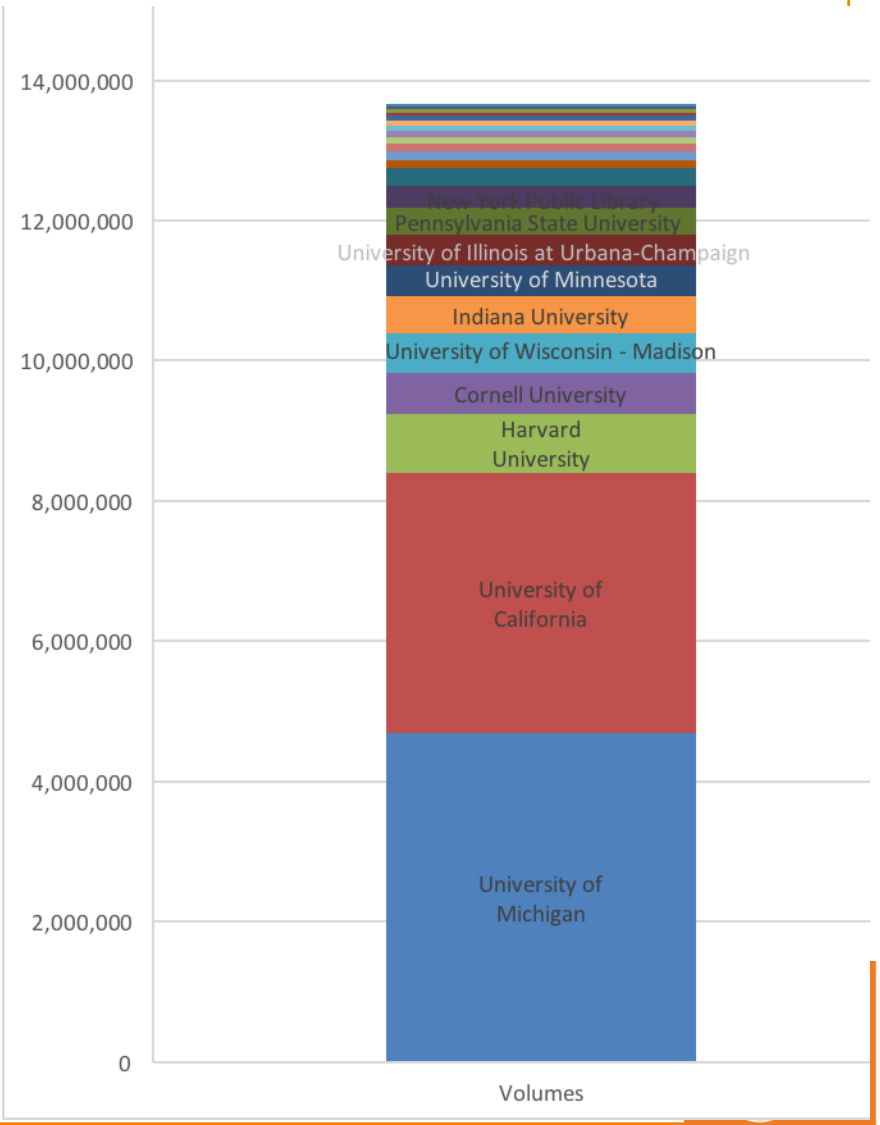
- HTRC Portal
  - Workset Builder – Predefined Algorithms (Inspired by Monk)
  - Access to Data Capsule| Bookworm | Extracted Features
- HTRC Data Capsule
  - Run your own algorithm/program in secure environment
- HTRC Extracted Features Workset
  - Currently 13.7M set available Nov 2016

RESEARCH CENTER

HTRC & Libraries

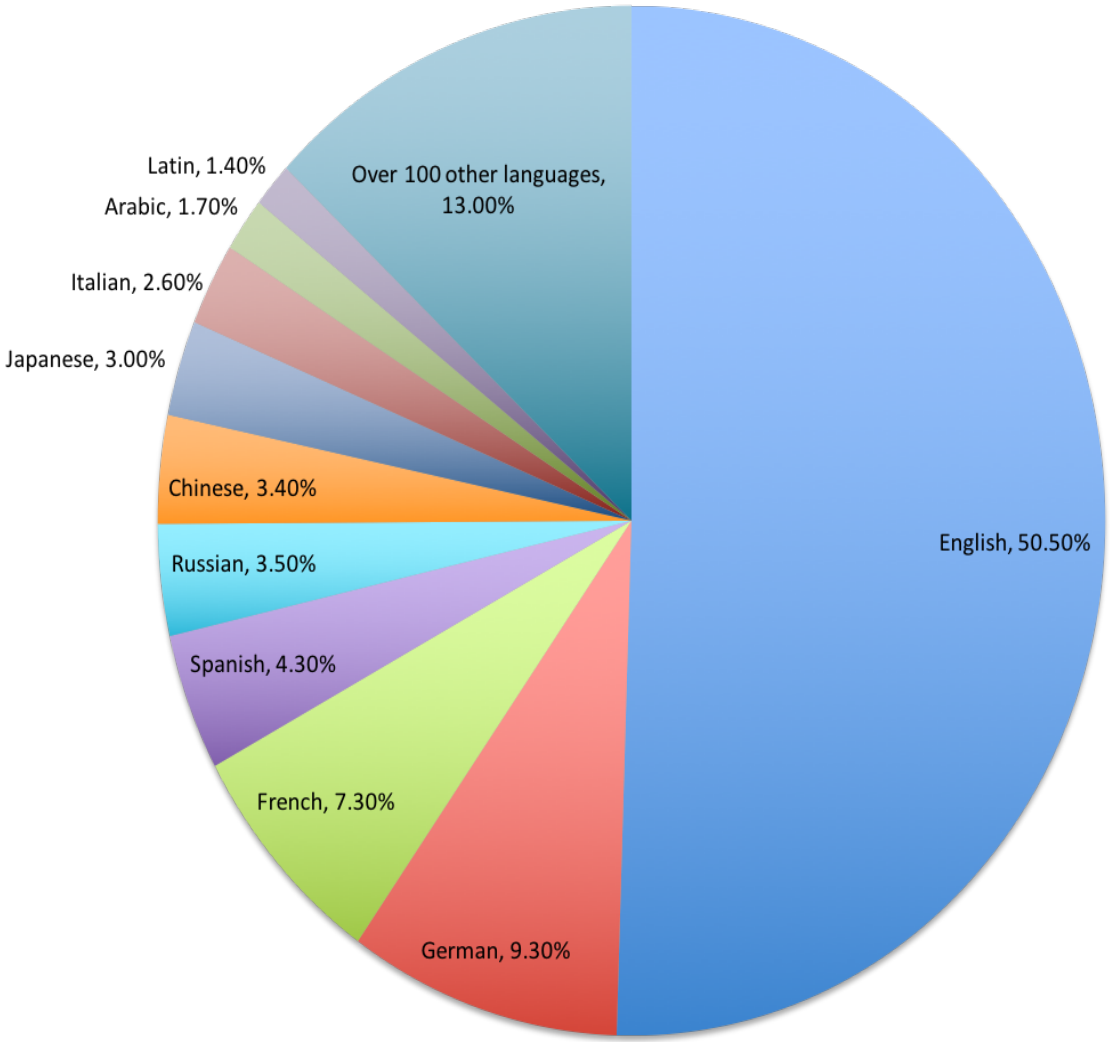
HT Contributions by Library-Nov 2015

Institution	Volumes
University of Michigan	4,696,618
University of California	3,707,214
Harvard University	838,344
Cornell University	584,875
University of Wisconsin - Madison	561,700
Indiana University	530,588
University of Minnesota	438,134
University of Illinois at Urbana-Champaign	437,288
Pennsylvania State University	390,087
New York Public Library	310,737
Princeton University	252,885
The Ohio State University	118,513
Universidad Complutense de Madrid	117,508
Library of Congress	108,892
University of Chicago	99,181
Keio University	90,126
University of Alberta	76,114
Columbia University	74,514
Northwestern University	57,142
University of Virginia	51,220
Purdue University	47,490
University of Iowa	40,622
Technical Report Archive & Image Library	35,923



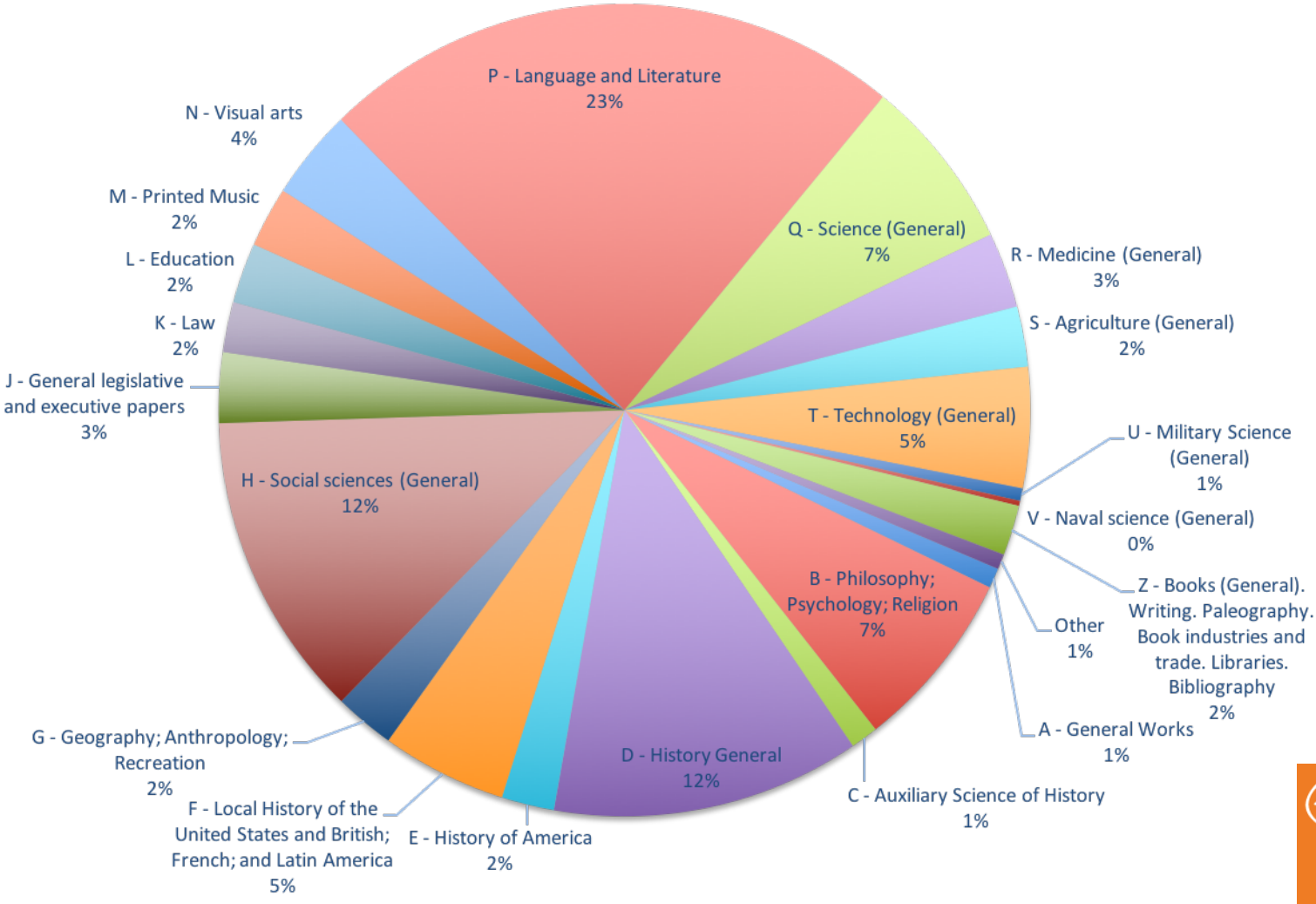
RESEARCH CENTER

Language Distribution of All Works in Hathitrust, Nov 2015



RESEARCH CENTER

HT Call Number Distribution



RESEARCH CENTER

HTRC: Scholars Commons

- Focus on pedagogy and support for librarians and beginning researchers.
- Startup: Scholars Commons programs at Indiana University and the University of Illinois libraries
- IMLS "Digging Deeper Reaching Further" Grant developing librarian training workshops with:
  - University of North Carolina
  - Northwestern University
  - Lafayette College

RESEARCH CENTER

SC Accomplishments (Pt. 1)

What do users need?

- *Phase 1*: Interviewed humanities scholars on use of text analysis and mining tools (2015-16)
- *Phase 2*: Interview social science scholars (2016-17)
- Results inform development of analysis tools, services, training, support.

How do we train librarians?

- Developed training (in-person and online) for the Portal and Workset Builder, Bookworm, and Data Capsule. "Beginner" and "advanced" workshops meet needs of diverse user community.
- Assessment workshop outcomes

RESEARCH CENTER

SC Accomplishments (Pt. 2)

Communication & training in action:

- DH2016 Krakow, Poland (June 2016)
- Digital Humanities Summer Institute Workshop (June 2016)
- Berkeley DH Institute (August 2016)
- Digital Frontiers (September 2016)
- University of Wisconsin HTRC Workshop (October 2016)
  - Showcasing current beginning curricular materials for train the trainer
- Charleston Conference (November 2016)
  - Showcasing research methods studies
  - Showcasing extracted features worksets
- DLF Forum (November 2016)
  - Showcasing text-mining pedagogy

RESEARCH CENTER

HTRC Working With Scholars: Advanced Collaborative Support

Benefits of ACS Program

- Enables HTRC to embed a tools expert within the research group of established researchers.
- Maps the researchers questions directly to the HT corpus via HTRC tool set.
- Enables new concepts and tools to develop within HTRC to support ongoing work with the HT corpus.

RESEARCH CENTER

2015 ACS Projects
Round 1

- Detecting Literary Plagiarisms: The Case of Oliver Goldsmith (Doug Duhaime) – ***Notre Dame***
- Literary Geography at Scale (Matthew Wilkins) - ***Notre Dame***
- Taxonomizing the Texts: Towards Cultural-Scale Models of Full Text (Colin Allen) – ***Indiana University***
- Trace of Theory (Geoffrey Rockwell, Laura Mandell, Stefan Sinclair, Matthew Wilkens, Susan Brown) – ***University of Alberta, Texas A&M, Notre Dame***
- Tracking Technology Diffusion Over Time (Michelle Alexopolous) – ***University of Toronto***

RESEARCH CENTER

2016 ACS Projects
Round 2

- Fighting Fever in the Caribbean: Medicine and Empire, 1650-1902 – ***University of Iowa***
- Inside the Creativity Boom – ***Brown University***
- The Chicago School: Wikification as the First Step in Text Mining in Architectural History – ***Illinois Institute of Technology***
- Signal and Noise and Pride and Prejudice: Toward an Information History of Romantic Fiction – ***Augsburg College***

RESEARCH CENTER

ACS Goals YIII

- Next round of ACS RFP
  - Q1 2017
  - Special emphasis on in-copyright materials
  - Special emphasis on Data Capsule use
- Showcase R I & II ACS projects at, for example, user group meetings and outreach and instructional sessions, to assist future submissions to ACS
- Expand use of Worksets, tools and EF data

HTRC: Future Forward

YIII Targets

- WCSA+DC
- Portal Access to Full HT Collection Q3 2017
- Extracted Features: Research Dataset
- Bookworm + HT
- Release New Curricular Materials (DDRF)
- Reduce Barriers from Research to Results
- New Communities: Social Science
- Modeling New Partnerships

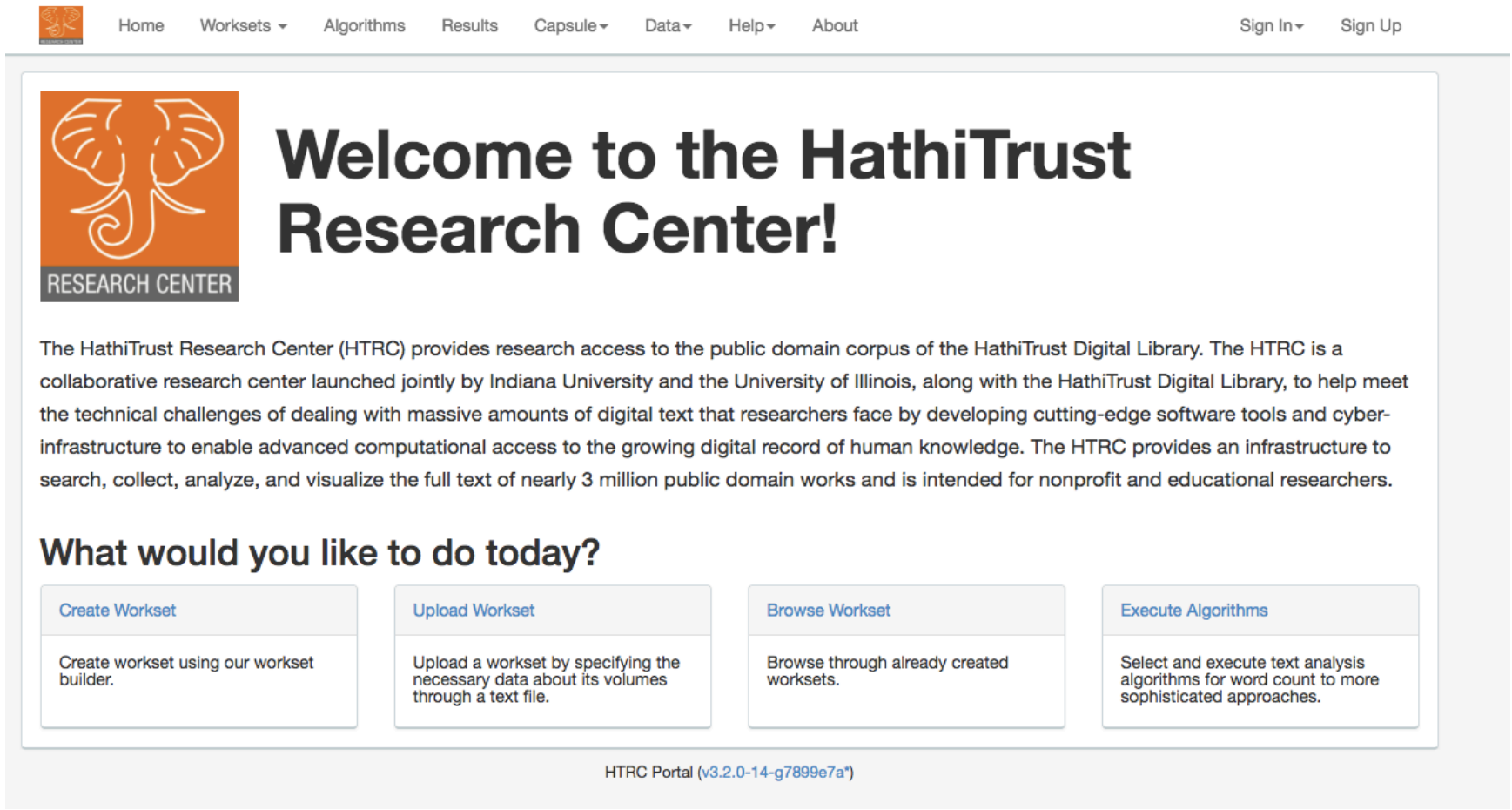
RESEARCH CENTER

WCSA+DC

- Mellon-funded: $1.17 Million, 2 years
- Roll out enhanced Workset Builder
  - New interface
  - Linked data metadata
  - Test page-level search
  - Connecting linked data + SOLR
- Roll out enhanced Data Capsules
  - Handle larger worksets
    - From 10K to 1M Use Cases
  - Incorporate new linguistic tools
  - In-copyright content

RESEARCH CENTER

Portal Access



RESEARCH CENTER

Home Worksets Algorithms Results Capsule Data Help About

Research Datasets

Downloadable, non-consumptive book data.

HTRC Extracted Features Dataset

Page-level features from 13.7 million volumes [v.1.0]

Description

The HTRC Extracted Features Dataset v.1.0 is comprised of page-level features for 13.7 volumes in the HathiTrust Digital Library. This version contains non-consumptive features for both public-domain and in-copyright books.

Features include part-of-speech tagged term token counts, header/footer identification, marginal character counts, and much more.

A full explanation of the dataset's features, motivation, and creation is available at the EF Dataset documentation page

Download the data

All 13.7 million files as well as custom subsets of the EF data are accessible using `rsync`, as described in the documentation.

A sample is available for download through your browser – sample.zip – as well as thematic collections: DocSouth (87 volumes), EEBO (355 volumes), ECCO (505 volumes).

Co

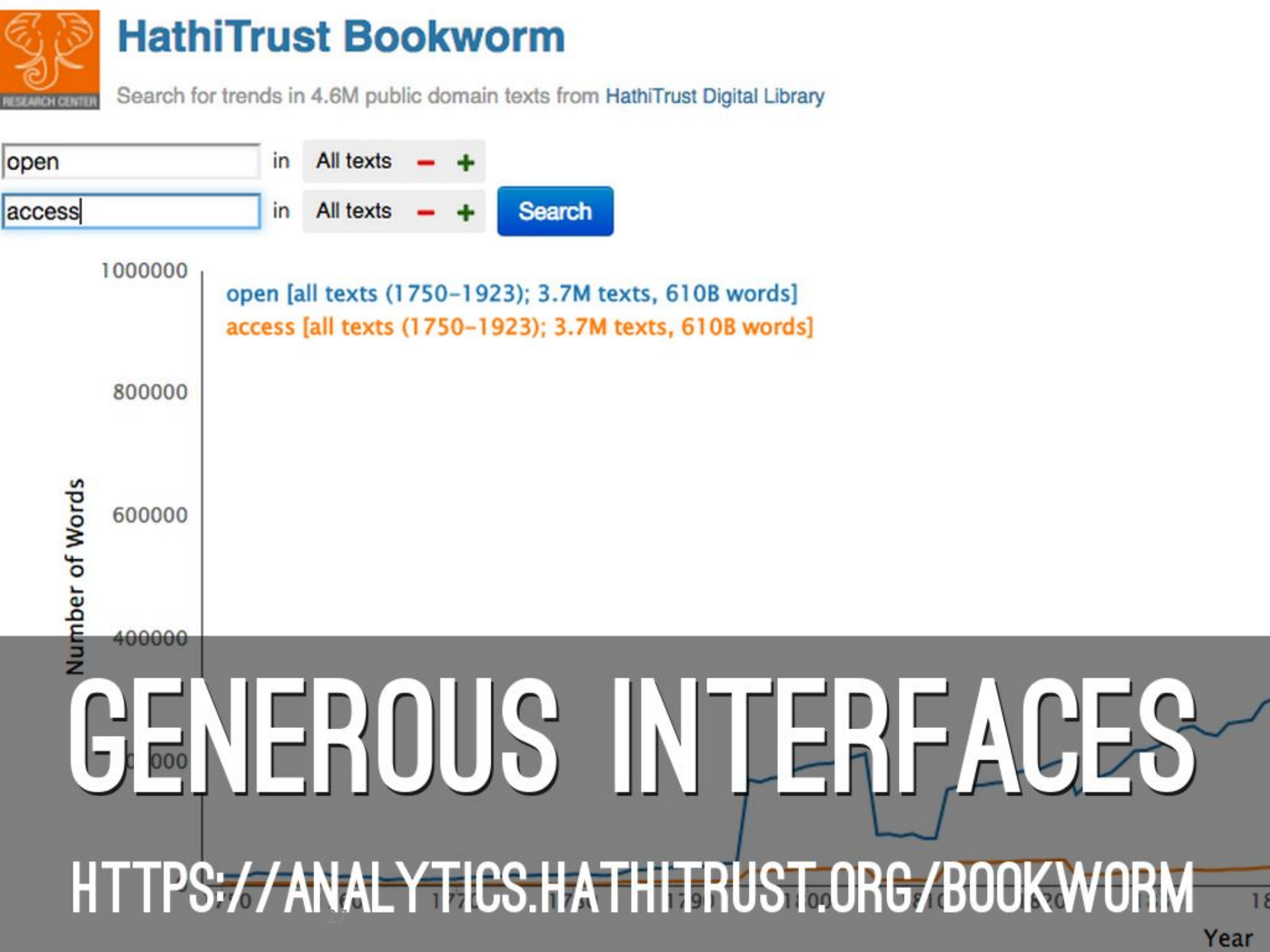
Attribution

HTTPS://ANALYTICS.HATHITRUST.ORG/DATASETS

Boris Capitanu, Ted Underwood, Peter Organisciak, Timothy Cole, Maria Janina Sarol, Stephen Downie, J. Stanton. *The HathiTrust Research Center Extracted Feature Dataset* (1.0) [Dataset]. HathiTrust Research Center, http://dx.doi.org/10.13012/J8X63JT3.

Res

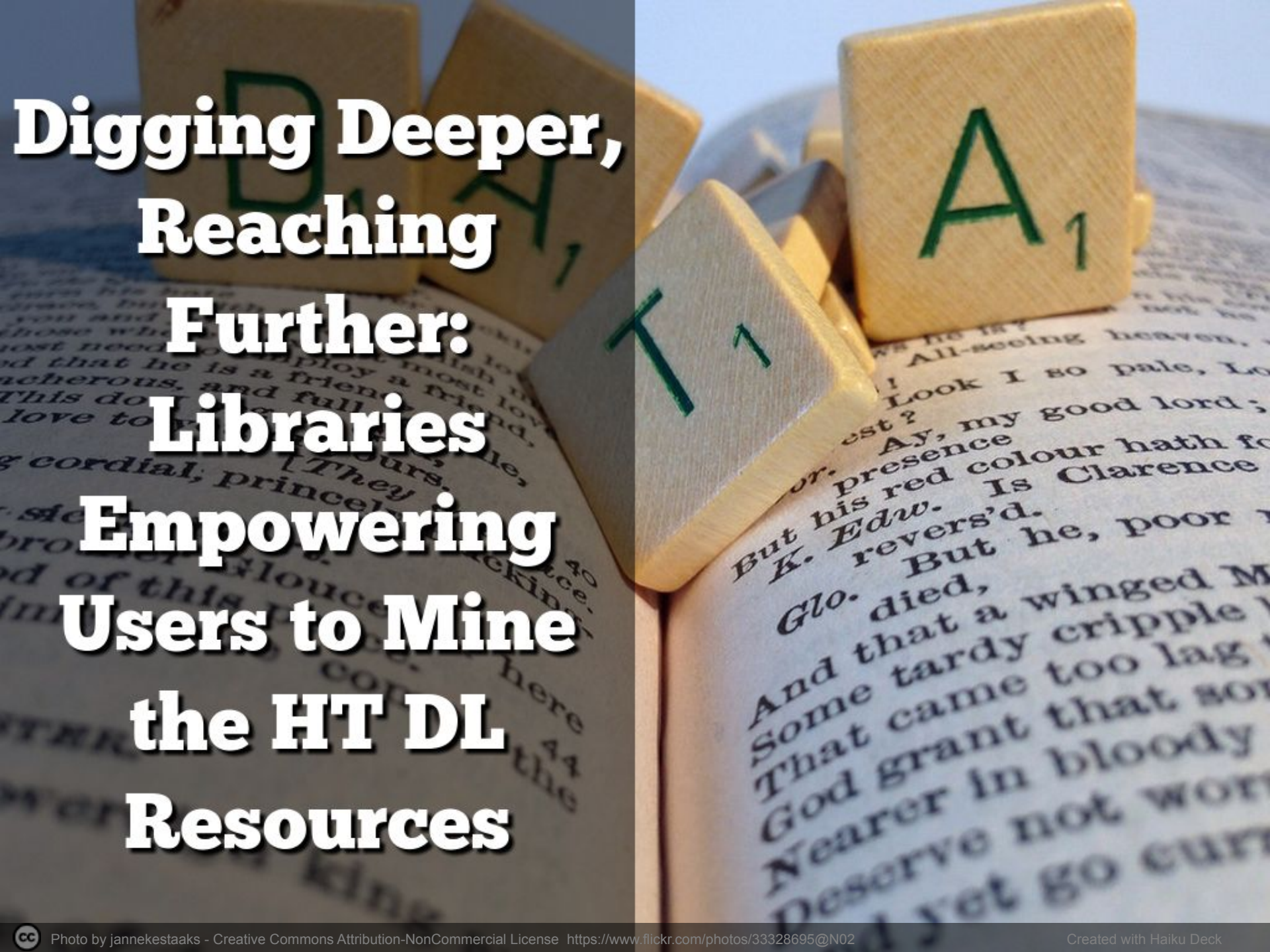
See Data



GENEROUS INTERFACES

HTTPS://ANALYTICS.HATHITRUST.ORG/BOOKWORM

Digging Deeper, Reaching Further: Libraries Empowering Users to Mine the HT DL Resources



New Communities: Social Sciences

- Move beyond traditional Digital Humanities community
- Intuition that the HT corpus is prime for social science scholarship
- Need your input to better understand the needs and uses of social science scholars
- Help us connect with this important community

RESEARCH CENTER

Modeling New Partnerships

- Data and Text-Mining partnerships with other organizations
  - Grow demand for analytical use of HathiTrust
  - Drive down costs through shared resources
  - Develop new resource streams
  - Create sustainability through community involvement
- Cost model for customized solutions
- Current partnership discussions – (Ex. Voyant, Oxford, Ithaka)

RESEARCH CENTER

HTRC Non-Consumptive Use Policy

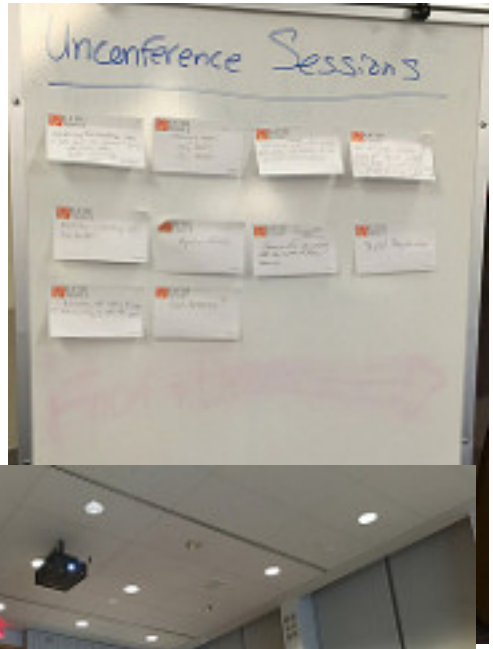
HTRC Useful Links

- HTRC Portal
  - https://analytics.hathitrust.org
- HTRC Extracted Features Dataset
  - https://analytics.hathitrust.org/features
- HTRC FAQ
  - http://bit.ly/HTRCFAQ
- HTRC+BW
  - https://bookworm.htrc.illinois.edu
- HTRC-*Educause Review*
  - http://bit.ly/2e0fkt7

RESEARCH CENTER

HTRC@Upcoming Events

- DLF Forum – Nov 7-9
- CNI Fall Meeting – Dec 12-13
- Planned DPLAFest Chicago
- Planned HTRC UnCamp Fall 2017-Bloomington



RESEARCH CENTER

HTRC Team

HTRC @ Indiana:

- Beth Plale-Co-PI
- Robert McDonald
- Marie Ma
- Samitha Liyanage
- Leena Unnikrishnan
- Jaimie Murdock
- Zong Peng
- Milinda Pathirage
- Inna Kouper
- Angela Courtney
- Nicholae Cline
- Leanne Nay
- Ewa Zegler-Poleska
- Semyon Khokhlov

HTRC @ Illinois:

- J. Stephen Downie-Co-PI
- Beth Namachichivaya
- Tim Cole
- Jacob Jett
- Boris Capitanu
- Eleanor Dickson
- Ryan Dubnicek
- Harriett Green
- Peter Organisciak
- Robert Manaster
- Michael Haberman
- Megan Senseney

RESEARCH CENTER

Funders

- HathiTrust Board of Governors
- Indiana University
- University of Illinois
- Andrew W. Mellon Foundation
- National Endowment for the Humanities
- Social Science and Humanities Research Council
- Institute for Museum and Library Services
- Alfred P. Sloan Foundation

RESEARCH CENTER