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
HTRC Eco-System

HathiTrust Digital Library

Access
Preserve
Collect
Disseminate

Formulate
Gather
Analyze
Produce

HathiTrust Research Center
HTRC 2014-2018 Org Chart

- Administrative Support
- Core Development
- Advanced Research
- Advanced Collaborative Support
- Scholarly Commons

HTRC Executive Mgmt
HTRC Growth 2014-2016

923 New Users
1127 Registered Users
130 Data Capsule Users
257 Total No of Institutions
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
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
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
HTRC & Libraries
# HT Contributions by Library - Nov 2015

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

![Bar chart showing contributions by library](chart.png)
HT Call Number Distribution

- P - Language and Literature: 23%
- E - History of America: 2%
- D - History General: 12%
- H - Social sciences (General): 12%
- Q - Science (General): 7%
- T - Technology (General): 5%
- M - Printed Music: 2%
- L - Education: 2%
- K - Law: 2%
- J - General legislative and executive papers: 3%
- N - Visual arts: 4%
- R - Medicine (General): 3%
- S - Agriculture (General): 2%
- U - Military Science (General): 1%
- V - Naval science (General): 0%
- Other: 1%
- A - General Works: 1%
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
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
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
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.
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
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
ACS Goals YIII

• Next round of ACS RFP
  – QI 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
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
Welcome to the HathiTrust Research Center!

The HathiTrust Research Center (HTRC) provides research access to the public domain corpus of the HathiTrust Digital Library. The HTRC is a collaborative research center launched jointly by Indiana University and the University of Illinois, along with the HathiTrust Digital Library, to help meet the technical challenges of dealing with massive amounts of digital text that researchers face by developing cutting-edge software tools and cyber-infrastructure to enable advanced computational access to the growing digital record of human knowledge. The HTRC provides an infrastructure to search, collect, analyze, and visualize the full text of nearly 3 million public domain works and is intended for nonprofit and educational researchers.

What would you like to do today?

- Create Workset
  - Create workset using our workset builder.

- Upload Workset
  - Upload a workset by specifying the necessary data about its volumes through a text file.

- Browse Workset
  - Browse through already created worksets.

- Execute Algorithms
  - Select and execute text analysis algorithms for word count to more sophisticated approaches.
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).

HTTPS://ANALYTICS.HATHI TRUST.ORG/DATASETS
HathiTrust Bookworm

Search for trends in 4.6M public domain texts from HathiTrust Digital Library

open
in All texts
access
in All texts

open [all texts (1750–1923); 3.7M texts, 610B words]
access [all texts (1750–1923); 3.7M texts, 610B words]

Year
Number of Words
0 00000
400000
600000
800000
1000000

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
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)
HTRC Non-Consumptive Use Policy
HTRC Useful Links

- **HTRC Portal**
  - [https://analytics.hathitrust.org](https://analytics.hathitrust.org)
- **HTRC Extracted Features Dataset**
  - [https://analytics.hathitrust.org/features](https://analytics.hathitrust.org/features)
- **HTRC FAQ**
- **HTRC+BW**
  - [https://bookworm.htrc.illinois.edu](https://bookworm.htrc.illinois.edu)
- **HTRC-Educause Review**
HTRC@Upcoming Events

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