The HathiTrust Research Center: An Overview of Advanced Computational Services

Robert H. McDonald | Indiana University | @mcdonald
Beth Sandore Namachchivaya | University of Illinois at Urbana-Champaign

Tweet us @hathitrust | #HTRC | #EDU15
Outline

• HTRC Update from #EDU13
• Scholars Commons and Outreach
• Current Services and Tools
• Data and Infrastructure
• Discussion Questions
Update: Progress Since #EDU13

- Core development of data mining services
- WCSA Projects completed (Mellon)
- Bookworm+ HTRC available (NEH)
- First round of ACS projects completed
- Multi-year operational plan with HT support

About the HathiTrust Digital Library

• Repository
  – 13.7+ million volumes | 3+ billion pages
  – 50% of volumes are in English
  – Material from the 15th C. on | 20th C. concentration
  – 70% in copyright or undetermined | 30% open

• Interface
  – Search and read books in the public domain
HTRC Engaging the HT Community

- Engage with tools via Portal
- Download and work with derived datasets
- Engage with capability through Scholars Commons experts and metadata
- Give feedback via user group meetings and UnCamp
- Work directly with HTRC staff (large compute and ACS)
- Advisory Board

HTRC
HT Points of Access

- HathiTrust Digital Library
  - HT data request
  - HathiFiles
  - Bib and Data APIs
- HTRC datasets
  - Extracted Features dataset
    - [https://sharc.hathitrust.org/features](https://sharc.hathitrust.org/features)
  - Genre in English Language Literature (Underwood)
- HTRC Data Capsule
- HTRC Portal and Workset Builder
Scholarly Commons and Outreach

- Design, Development and Delivery
- Outreach Activities
- Education Activities
Scholarly Commons: Education

• “Digging Deeper, Reaching Further: Libraries Empowering Users to Mine the HT DL Resources”
• IMLS Laura Bush 21st Century Librarian Program
• Led by Harriett Green (Illinois)
• Collaborating institutions:
  • University of Illinois
  • Indiana University
  • University of North Carolina (Chapel Hill)
  • Lafayette College
  • Northwestern University
Current Services & Tools

- Portal & Workset builder
- Data Capsule
- Bookworm
- WCSA (Workset Creation for Scholarly Analysis)
- Advance Collaborative Support (ACS)
Portal/Workset Builder & Data Capsule

A set of tools for assembling collections of digitized text and performing text analysis on them
A workset is a user-created collection of text from the HathiTrust Digital Library.

- Textual datasets
- Built through searching OCR data and/or metadata
  - In full text and titles
  - By author, time period, subject classification, etc.
- Currently contains non-copyrighted material NOT digitized by Google
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.

https://sharc.hathitrust.org
Workset Builder: Search Results

An author search for “Austen, Jane”.

Select all or some of the returned search items for your workset.

Filtering results by “Austen, Jane, 1775-1817” reduces results from 3,148 to a more manageable 390.
Data Capsule: Non-consumptive research environment (Private Virtual Machine)
Researcher is assigned a virtual machine (VM)

Can configure the VM
- as if one’s own computer
- download custom tools/algorithms to it

Can load VM with HTRC textual data
- Only when it’s in "secure mode"
  - Network and other data channels are restricted

After analysis of the textual data,
- results emailed to the user, text not...
How it works?
HTRC Infrastructure
# Data & Infrastructure for HTRC

## HathiTrust Data Protection Levels

<table>
<thead>
<tr>
<th>Level</th>
<th>Content</th>
<th>Login</th>
<th>Download</th>
<th>Provenance</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Derived Factual Data</td>
<td></td>
<td>Bulk</td>
<td></td>
</tr>
</tbody>
</table>
| 1     | In public domain  
No 3\textsuperscript{rd} party restrictions | X     | Volume   |            |
| 2     | In public domain  
With 3\textsuperscript{rd} party restrictions | X     | None     |            |
| 3     | In Copyright      | X     | None     | Required   |
Extracted Feature Set and Distant Reading

Distant Reading – “understanding literature not by studying particular texts, but by aggregating and analyzing massive amounts of data.” – Moretti (Stanford Literary Lab)

— HTRC Distant Readers:

• Ted Underwood (University of Illinois)
• Andrew Goldstone (Rutgers University)
• Jonathan Goodwin (University of Louisiana)
• Chris Forster (Syracuse University)
HathiTrust + Bookworm

A tool for visualizing and analyzing word usage trends in the HathiTrust Digital Library

http://bookworm.htrc.illinois.edu
HT+BW Components

The HathiTrust Digital Library (HT)

- Text data (OCR-ed)
- Metadata (classification, genre, dates, format, etc.)

+ Bookworm (BW)

(A tool that can visualize language usage trends within *any* repository of digitized text)
HT+Bookworm helps answers questions like: How does language change over time?

• Bookworm originally developed by Erez Lieberman Aiden
  – Co-creator of Google Books Ngram Viewer
  – Partner in the HathiTrust+Bookworm project

• Erez was interested in how irregular verbs tend to become regular over time
  – Originally studied by hand (Nature, 2007)
  – Later, computationally (Science, 2010)
Change of past tense in verbs: an example

‘burned’ (blue graph) and ‘burnt’ (orange graph)
Future Development of HT+BW

- Integration of HTRC worksets
  - Plot trends \textit{within} custom worksets
  - Create worksets \textit{from} plotted trends

- Hierarchical metadata \rightarrow Hierarchical faceting

- APIs
  - Will likely make integration of a bookworm with many digitized collections possible
Advanced Collaborative Support (ACS)

- Computational research by external researchers on material in the HathiTrust, supported by HTRC
- Annual call for proposals—next in 2016
- Three project funded in 2015, receiving dedicated compute cycles and advanced technical support from HTRC
HTRC teams began providing support in Spring 2015
Projects required access at scale, using half a million to a million volumes each
3 ACS projects selected:
  – Literary Geography at Scale
    • Matthew Wilkens, University of Notre Dame
  – Off the Books
    • Michelle Alexopoulos, University of Toronto
  – Taxonomizing the Texts: Towards Cultural-scale models of full text
    • Colin Allen & Jaimie Murdock, Indiana University
Jaimie Murdock and Colin Allen’s

*Topic Explorer* (interface)

This visualization shows the similarity of books in the HTRC-1315 collection (generated by searching the HathiTrust Research Collection in 2013 for the terms "anthropomorphism", "comparative psychology", and "Darwin") to *The principal works of George Meredith: a brief bibliography*.
Get Involved!

HTRC Announcements:

htrc-announce-l @ list.indiana.edu

HTRC User Group:

htrc-usergroup-l @ list.indiana.edu
Discussion Question 1

• How to get involved with HathiTrust Research Center?
  – Join our user group (htrc-usergroup-l @ list.indiana.edu)
    • Monthly Adobe Connect Meeting
  – Advanced Collaborative Support Grants
  – Development of new interfaces using Extracted Feature sets
Discussion Question 2

• How will we develop data use policies for 3\textsuperscript{rd} party data within the research ecosystem?
  – How does this fit with data governance models currently in place?
  – What platform will be used for commercial information content?
Discussion Question 3

• I am working with a researcher at my institution on a data mining problem and we want to use the HTRC tools or data, how do we get access?
  • Published Extracted Feature Set
  • Data Portal
  • ACS Grants
  • Work with us on research grant
    – Developing cost model for ACS that can be utilized in grant budgets
Help Us Improve and Grow

Thank you for participating in today’s session.

We’re very interested in your feedback. Please take a minute to fill out the session evaluation found within the conference mobile app, or the online agenda.

Acknowledgments: HTRC Team

HTRC @ Illinois
(GSLIS and the University Library):
• Stephen Downie
• Beth Sandore Namachichivaya
• Tim Cole
• Loretta Auvil
• Sayan Bhattacharyya
• Boris Capitanu
• Eleanor Dickson

HTRC @ IU
• Beth Plale
• Robert McDonald
• Angela Courtney
• Nicholae Cline
• Miao Chen
• Samitha Liyanage
• Leanne Mobley
• Milinda Pathirage
• Guangchen Ruan
• Leena Unnikrishnan
• Jiaan Zeng

Core Slide Set: Eleanor Dickson, Sayan Bhattacharyya
Questions?

Robert H. McDonald
Associate Dean for Library Technologies & Deputy Director Data to Insight Center
rhmcdona@indiana.edu

Beth Sandore Namachchivaya
Associate University Librarian for Research & Associate Dean of Libraries
sandore@illinois.edu