What is the HathiTrust Research Center?

http://www.hathitrust.org/htrc

• Jointly led by the University of Illinois at Urbana-Champaign and Indiana University Bloomington
• Facilitates text analysis of HTDL content → Focus on large-scale, computational research
• Research & Development
  • Finding technical solutions
  • Building tools and services
  • Conducting user studies
Scholarly Practices with Digital Collections and Tools

How humanities scholars use digital collections: Brockman et al., 2001; Palmer and Neumann, 2002; Babeu 2011; Rutner and Schonfeld, 2011; Green and Courtney, 2015

How humanities scholars use digital tools: Frischer et al., 2006; Warwick 2008; Toms and O’Brien, 2008; Gibbs and Owen, 2012

Tools and resources for textual analysis: ARTFL and Philologic (Argamon et al., 2009; Horton et al., 2009), MONK (Unsworth, 2011), Wordseer (Muralidharan and Hearst, 2013), Voyant and TaPOR (Rockwell et al., 2010), and Lexos (LeBlanc et al., 2013)
Mellon Foundation funded research project, 2013-2014

GOAL: Find out how researchers collect together digital materials and build textual corpora for research purposes.

Findings (Green et al. 2014, Fenlon et al. 2014):

- Need the ability to create and manipulate collections as reusable datasets and research products
- The ability to work at different units of analysis
- Access to highly enriched metadata
HTRC User Requirements Study: Research Goals

- Learn how researchers use digitized textual corpora, apply relevant methods and approaches, and seek needed tools
- Develop illustrative use cases of text analysis research that will help shape the development and expansion of HTRC research services and training curricula for scholars
- Obtain information that can inform development of text analysis data providers and research services
HTRC Users Requirements Study: Methods

- Recruited interviewees from professional conferences and meetings on digital libraries and digital humanities
- Semi-Structured Interviews with 15 scholars
- All interviews transcribed and coded in ATLAS.ti 7 by HTRC Scholarly Commons members
- Qualitative content analysis
Findings

*What are scholars’ needs and practices for conducting textual analysis with large text archives?*

Research Practices and Culture

Textual Data Acquisition and Use

Generating and Sharing Results
Research Practices and Culture

“When I say people have been studying this time period for 300 years—people who are much smarter than me, better writers, have better access to the archives, who can read more than I can—the only way we can say something new is if we get new perspective on old data.”

“I’m not worried about publishing venues, I’m not worried about reproducibility, I’m not worried about statistics. My own knowledge of that is pretty good. But the collaborative work style is really hard.”

“I wish more people were archiving their data and their algorithms from the source code, as you see CS papers that will benchmark results against a dataset…. Then how do you try to replicate or beat those results? It becomes impossible to evaluate your own methods against theirs and really slows down the pace of research, because if one could surpass state of the art, then that’s an application and [a] step forward.”
"We need that corpus, and we need basically the data exposable. I mean we have to be able to view, not only manipulate, but also view the data."

“The newspaper archives that exist... all have their own siloed, siphoned search system and metadata collection mechanisms, and these things do not necessarily always talk to each other, try as I will to get everything into Zotero in some kind of unified form."

"I did work with ProQuest and the New York Times. I had an article that came out in an academic journal and I worked with them to get permission to use an image before, so I know that it is possible. But the process was so long and it was for three images. I’m going to have thousands of files."
## Where Do They Get Data and How Do They Analyze It?

<table>
<thead>
<tr>
<th>Data Sources</th>
<th>Tools</th>
</tr>
</thead>
<tbody>
<tr>
<td>HathiTrust Digital Library</td>
<td>Voyant</td>
</tr>
<tr>
<td>Internet Archive</td>
<td>Juxta</td>
</tr>
<tr>
<td>Self-digitized content</td>
<td>Gephi</td>
</tr>
<tr>
<td>Early English Books Online</td>
<td>Tableau</td>
</tr>
<tr>
<td>JSTOR</td>
<td>MorphAdorner</td>
</tr>
<tr>
<td>ARTFL</td>
<td>D3.js</td>
</tr>
<tr>
<td>Google Scholar</td>
<td>MALLET</td>
</tr>
<tr>
<td>English Short Title Catalog</td>
<td>Zotero</td>
</tr>
<tr>
<td>Project Gutenberg</td>
<td>Weka</td>
</tr>
<tr>
<td>ProQuest Historical Newspapers</td>
<td>Python: NLTK, SciKit Learn</td>
</tr>
</tbody>
</table>
Generating and Sharing Results

“My general like philosophical approach to these things is I like to do things small. I build my corpora. I like to read them myself. I’m a little weary of like big distant reading approaches, especially with stuff as far away from the present as my stuff. So I’m still trying to perfect the stuff that I’m currently doing.”

“Document navigation would be extremely helpful, and that’s the kind of thing that people have to do a lot of: searching, bookmarking, grouping things, and looking at several segments together.”

“In some ways GitHub is an integral part of this. We can try to describe this code, or you can go look at our code, so it’s interesting in that if you read the paper without actually looking at the code, you’ve gotten sort of a broad overview of the method, but you couldn’t replicate it… So it’s a bit of a hybrid publication.”
Findings: User Personas

<table>
<thead>
<tr>
<th>Digital Project Librarian</th>
<th>Faculty Member</th>
<th>Graduate Student</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Wants flexible, transparent tools</td>
<td>• Wants computational resources</td>
<td>• Wants examples</td>
</tr>
<tr>
<td>• Role: Research Support staff</td>
<td>• Role: Experienced Researcher</td>
<td>• Role: New Researcher</td>
</tr>
<tr>
<td>• Challenges: Inaccessible data, matching tool to researcher</td>
<td>• Challenges: Collaboration, Finding texts</td>
<td>• Challenges: Understanding stats, choosing areas of interest</td>
</tr>
</tbody>
</table>

Credits: Alex Kinnaman, Peter Organisciak, Eleanor Dickson
Looking Forward

- Expanding study with interviews with social scientists
- IMLS-funded “Digging Deeper, Reaching Further: Libraries Empowering Users to Mine the HathiTrust Digital Library”: http://teach.htrc.illinois.edu
- HT + Bookworm development (NEH)
- Data Capsule development (Mellon)
- Revision to HTRC Portal and Workset Builder
Interested in working with HTRC?

http://www.hathitrust.org/htrc

HTRC Announcements:

htrc-announce-l @ list.indiana.edu

Questions?

htrc-help@hathitrust.org

HTRC User Group:

htrc-usergroup-l @ list.indiana.edu

Advanced Collaborative Support program:

htrc.acs.awards@gmail.com
HTRC Scholarly Commons Team

University of Illinois:
Beth Sandore Namachchivaya
Stephen Downie
Harriett Green
Eleanor Dickson
Peter Organisciak
Alex Kinnaman (graduate assistant)
Sayan Bhattacharyya (former CLIR postdoc)

Indiana University:
Robert McDonald
Angela Courtney
Nicholae Cline
Leanne Nay
Ewa Zegler-Poleska
Thank you!

Harriett Green
green19@illinois.edu  |  @greenharr

Eleanor Dickson
dicksone@illinois.edu  |  @elliedickson

Angela Courtney
ancourtn@Indiana.edu  |  @englishlitlib