HathiTrust Research Center: Improving Scholarly Inquiry

Timothy W. Cole (t-cole3@illinois.edu)
Harriett Green (green19@illinois.edu)

With slides and other contributions from Stephen Downie, Beth Plale, Colleen Fallaw, Megan Senseney, Katrina Fenlon, et al.

CNI Fall 2013 Membership Meeting
Washington, D.C.
9 December 2013
Outline

- The HathiTrust Digital Library (HT)
- The HathiTrust Research Center (HTRC)
- The Workset Creation for Scholarly Analysis (WCSA) Project
- User needs & requirements
- Characterization of bibliographic metadata for corpus
- More about the WCSA RFP & prototyping projects
The HathiTrust Digital Library (hathitrust.org)

- A digital preservation repository coupled with a highly functional access platform
- An international partnership of 80+ research libraries & consortia
- Provides long-term preservation of and access to volumes of member library collections that have been digitized by Google, the Internet Archive, Microsoft & member institutions
- Currently supports ingest of digitized book and journal content, and similar book-like materials
HT DL by the numbers (as of Nov 2013)

- 10,973,063 total volumes
- 6,067,835 distinct bibliographic items:
  - 5,778,450 book (monographic) titles
  - 289,385 serial titles
- 3,803,630,600 pages
- 487 terabytes
- 3,512,404 volumes (~32% of total) digitized from public domain originals
More than just US Libraries

97% of bibliographic records specify resource language

Only 7% specify more than 1
HT DL Searching & Data Availability

- **Web User Interface** ([http://www.hathitrust.org/home](http://www.hathitrust.org/home)):
  - Full text keyword (includes indexed metadata)
  - Bibliographic metadata keyword
  - Advanced (field-specific) bibliographic metadata searching

- **Bibliographic metadata** ([http://www.hathitrust.org/data](http://www.hathitrust.org/data)):
  - OAI-PMH & custom bib API – [http://www.hathitrust.org/bib_api](http://www.hathitrust.org/bib_api)
  - HathiFiles (tab delimited metadata) – [http://www.hathitrust.org/hathifiles](http://www.hathitrust.org/hathifiles)

- **Full-text** ([http://www.hathitrust.org/datasets](http://www.hathitrust.org/datasets)):
  - ~300,000 digitized volumes in the public domain – contact for bulk download or use API for volume-by-volume access
  - ~3,500,000 volumes digitized by Google from public domain – available by arrangement, typically using rsync. Must agree to conditions of use.
How many pages per volume?

For volumes digitized from public domain sources:

- XL - Greater than 550 pages: 13%
- S - Less than 150 pages: 36%
- L - Between 350 and 550 pages: 21%
- M - Between 150 and 350 pages: 30%
The HathiTrust Research Center (1)

HTRC is a collaboration between HT, Indiana University and the University of Illinois at Urbana-Champaign

- Goal is to provide computational access to researchers: initially to all content digitized from public domain eventually to the entire HT DL corpus

- Currently hosts
  - complete copy of HT metadata
  - copy of OCR of all HT volumes digitized from public domain
  - copy of OCR of all public domain volumes in HT

Supported by the Sloan Foundation, the Mellon Foundation, IU, & UIUC
The HathiTrust Research Center (2)

HTRC end-user access (so far)

- **HTRC Portal**
  
  [https://htrc2.pti.indiana.edu/HTRC-UI-Portal2/](https://htrc2.pti.indiana.edu/HTRC-UI-Portal2/)
  
  Must login; pull-down login menu (upper right) to sign up (free)

- **HTRC Workset Builder**
  
  [https://htrc2.pti.indiana.edu/blacklight](https://htrc2.pti.indiana.edu/blacklight)
  
  Must login to this interface also; same credentials used.

- **HTRC Sandbox (contact us)**
  
  [http://sandbox.htrc.illinois.edu:8080](http://sandbox.htrc.illinois.edu:8080)
  
  Clone of Portal, but accessing 250,000 digital public domain volumes
  
  Supports use data api – [http://wiki.htrc.illinois.edu/display/COM](http://wiki.htrc.illinois.edu/display/COM)
  
  As well as HTRC Solr Proxy api – [http://chinkapin.pti.indiana.edu:9994](http://chinkapin.pti.indiana.edu:9994)

More: [http://www.hathitrust.org/htrc/faq](http://www.hathitrust.org/htrc/faq), [http://wiki.htrc.illinois.edu/display/OUT](http://wiki.htrc.illinois.edu/display/OUT)
Welcome to the HathiTrust Research Center!

What Can You Do With HTRC Portal?

- Create Workset
- Upload Workset
- Browse Workset
- Execute Algorithms

The HathiTrust Research Center (HTRC) provides research access to the public domain text 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 cyberinfrastructure 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.
HTRC Workset Builder (as it is now)

Search tips
- Select "match all" to require all fields.
- Select "match any" to find at least one field.
- Combine keywords and attributes to find specific items.
- Use quotation marks to search as a phrase.
- Use "-" before a term to make it required. (Otherwise results matching only some of your terms may be included).
- Use "~" before a word or phrase to exclude.
- Use "OR", "AND", and "NOT" to create complex boolean logic. You can use parentheses in your complex expressions.
- Truncation and wildcards are not supported - word-stemming is done automatically.

More Search Options

Find items that match all of the fields below:

- Full Text:
- Title:
- Author: mark twain
- Subject:
- Publish Date:

AND have these attributes:

Sort results by relevance

Start over Search

Timothy W. Cole (t-cole3@uiuc.edu) University of Illinois at UC

CNI 2013 Fall Membership Meeting
9 December 2013
Create a small workset / collection

1. A Connecticut Yankee in King Arthur's Court / by Mark Twain.
   - Title: A Connecticut Yankee in King Arthur's Court / by Mark Twain.
   - Author: Twain, Mark, 1835-1910.
   - Language: English
   - Published: 1917

2. A tramp abroad / by Mark Twain [pseud.]
   - Title: A tramp abroad / by Mark Twain [pseud.]
   - Author: Twain, Mark, 1835-1910.
   - Language: English
   - Published: 1907, 1907

3. How to tell a story, and other essays.
   - Title: How to tell a story, and other essays.
   - Author: Twain, Mark, 1835-1910
   - Language: English
   - Published: 1902
Submit a small workset / collection for analysis
Your completed and pending analytical jobs

<table>
<thead>
<tr>
<th>Job Title</th>
<th>Last Updated</th>
<th>Status</th>
<th>Cancel?</th>
</tr>
</thead>
<tbody>
<tr>
<td>tc3Test2</td>
<td>2013-12-07 13:27:38</td>
<td>Staging</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Job Title</th>
<th>Last Updated</th>
<th>Status</th>
<th>Delete/Save?</th>
<th>Saved?</th>
</tr>
</thead>
<tbody>
<tr>
<td>tc3Test1</td>
<td>2013-12-07 13:23:43</td>
<td>Finished</td>
<td></td>
<td>unsaved</td>
</tr>
</tbody>
</table>
Workset Creation for Scholarly Analysis

Premise

The ability to slice through a massive corpus constructed from many different library collections, and out of that to construct the precise workset required for a particular scholarly investigation, is an example of the “game changing” potential of the HathiTrust...
Collections, corpora, worksets, ...:

- Scholars & librarians aggregate items in a variety of contexts:
  - Archival
  - Curatorial
  - Experimental
  - Referential
  - Thematic

These worksets facilitate, sometimes enable certain kinds of scholarly inquiry.

*Analogy: HathiTrust worksets for analysis are as the contents of a scholar’s carrel in a library*

Anecdotal feedback from UnCamp 2013

My workset should contain...

- Volumes pertaining to Japan / in Japanese
- All volumes relevant to the study of Francis Bacon
- Music scores or notation extracted from HT volumes
- Volumes in HT similar to TCP-ECCO novels
- 19th c. English-language novels by female authors
- Representative sample (by pub date & genre) of French language items in HT
What is a Workset (in context of HTRC)?

- A workset is an aggregation brought together for the purpose of analysis, i.e., to facilitate inquiry.
- Worksets are conceptual and need to be expressible in a variety of ways.
- A workset encapsulates the specific materials that share specific attributes or satisfy some set of criteria.
- May be large, e.g., tens of thousands of items.
- Can be constructed by machine as well as human agents.
- Attributes and criteria not always bibliographic.
- Items aggregated may be more granular than a volume.
Why Worksets?

- The result of a first-level, rough filter
- Better scale for intensive analytics
- Provides essential scope for certain analytics
- Some tools (are trained to) work best on a narrow, homogeneous work-set
- Eliminate noise that would otherwise arise by asking questions across whole of HT
HathiTrust Corpus

...Passages, Images, Annotations...

Pieces of Books

Metadata

Books

(Linked) Data / Metadata

Secondary Lit

Media

Workset (Research collection)

<serialized dataset(s)>

Scope
Workset Creation for Scholarly Analysis

Prototyping Project

Collection analysis, data modeling and prototype tools & services to facilitate workset creation

- Principal investigators:
  J. Stephen Downie, Tim Cole, Beth Plale
- Funded by a grant from the Andrew W. Mellon Foundation
- 1 July 2013 - 30 June 2015

- Will feature 4 $40K sub-awards for prototyping/demonstration projects illustrating how worksets from HT DL can be created and used and can be useful for scholarly analysis
  - Methods & tools for metadata enrichment, including with links
  - Analytical services over full text useful for defining worksets
Key research questions for WCSA project

- Can we formalize the notion of collections and worksets in the HTRC context?

- What are the attributes that define and describe a workset in the context of HTRC?

- How can we balance rigor with extensibility & flexibility?

- What roles do data, metadata, annotations, tags, feature sets, and so on, play in the conception, creation, use and reuse of collections and worksets?

*Can we demonstrate the utility & practicality of worksets for HTRC?*
WCSA Timeline

- July 2013: Project Start
- Q1: User needs assessments / focus groups
- Q2: HT Corpus characterization
  Request For Prototype Proposals (RFP)
- Q3: RFP Finalist Workshop (Chicago)
  Prototype experiment funding awarded
- Q4-6: Prototype experiments done
  Metadata workflow & workset modeling
- Q7-8: Planning for prototype to production
  Report out
- June 2015: Project ends
Harriett Green, English and Digital Humanities Librarian

- Preliminary results
- An early deliverable of WCSA Project
Who Are Our Researchers?

- Humanities scholars? Computer programmers and technologists? Digital humanities research teams?
- Previous research in scholarly use of digital resources (Duff and Cherry 2000; Brockman et al. 2001; Warwick et al., 2008; Sukovic, 2008 and 2011; RIN 2011)
- Identify use cases for HTRC and large-scale, digitized text corpora
Report prepared for the HTRC in 2011 by UIUC researchers at GSLIS’s Center for Informatics Research in Science and Scholarship (CIRSS)

Interviewed researchers who were awarded Google Digital Humanities Research Awards on research needs

Findings for scholarly requirements included improved metadata, accurate OCR, data curation

Report available to download at http://www.hathitrust.org/htrc
Feedback from UnCamp 2013

My work-set should contain...
- Volumes pertaining to Japan / in Japanese
- Music scores or notation extracted from HT volumes
- Volumes in HT similar to TCP-ECCO novels

General Needs:
19th c. English-language novels by female authors
- User-friendly interfaces
- Documentation on the portal
- Avenues for community input in HTRC portal development
Scholarly Requirements

We are interested in understanding how scholars and researchers that use digital book and serials collections decide which texts (or parts of texts) to include in collections used for analysis. This includes:

- How researchers identify, select and obtain access to texts to include in their analysis
- Understanding the specific fields/disciplines that work with these sources along with the types of research questions and analysis applied.
- Desired units of analysis (works, manifestations, pages, n-grams OCR, images, etc.)
- Transformation and preprocessing steps;
- Understanding sources and criteria used for identifying texts
- Specific methods of selection
- Methods of analysis
- Challenges to working with these digital collections (e.g., OCR quality, duplication)
Focus Groups and Interviews

- Conducted at DH 2013, JCDL 2013, and HTRC Uncamp conferences in summer and fall 2013

- **Goal:** To understand practices of humanities researchers using digital collections, especially in the context of large-scale text corpora

- Survey instrument queried users about their experiential practices of organizing datasets
Participant Demographics

- **Positions:**
  - Junior and senior faculty at liberal arts colleges and universities
  - Computer programmers
  - Librarians
  - Data scientists
  - Academic technologists
  - Graduate students

- **Domains:**
  - English literature, classics, linguistics, library and information science, history

- **Institutions:**
  - Academic institutions in Great Britain, Singapore, Germany, France, and United States
1. General types of data, materials, or collections
2. Purposes of collections
3. Selection or inclusion/exclusion criteria
4. Sources, acquisition, and access
5. Pre-processing and analysis
6. Post-analysis
7. Challenges
Analysis

**Methodology:** Qualitative content analysis of user responses

- A “directed” approach based on inductive reasoning to condense raw data (transcriptions of audiorecordings of interviews and focus groups) into categories and themes

**Goal:** To identify common themes and patterns in users’ responses
Coding (still ongoing)

Coding manual consisting of category names, rules for assigning codes, and examples:

- Challenges — access rights
- Challenges — OCR quality
- Collections — comprehensiveness
- Objects — data
- Sources — Google Books
- Sources — Selection Criteria — Language etc.
Selected examples for categories

- **Category:** Challenges—Access Rights
  - User: “I check to see if a volume has substantial copyrighted text included in it already as quotes or extracts”

- **Category:** Objects — Temporal
  - User: “Classic materials”
  - User: “single-authored books of poetry between 1840 and 1900”

Etc.
Early Findings

- Roles of collections
- Need to implement granular, actionable units of analysis
- Importance of expert-enriched, shareable metadata
“collection-building is scholarly activity... we also need to think about how to document not just the status of different versions but also the labor that goes into and the kinds of knowledge that go into the decisions in making a collection, and the knowledge that’s gained from that process.”

“Today it is viewed as something very technical to prepare a corpus. But I think it’s getting more and more... interesting to do. And one day, it will be unrelated to technical stuff, and it will get closer to something of value.”

“the valorization of corpus-building...The recognition at the scientific level”

“I’m learning a lot through this organizing of my material and it’s informing what will be the main argument of my research”

“[If] I have a corpus and nobody is allowed to see it but wonderful things come out of it... That’s not really research... We are tying to get accountability for the kind of work we are doing. And it’s important for us to show the basis of our work.”

Figure 1. Selected focus group and interview excerpts on collection- and workset-building.
“...we need ways to slice this book. So we need to slice it by page...We need to slice it by poem, which doesn’t conveniently overlap or match the page boundaries. We potentially need to slice it by sections within a poem...”

“They use a lot of corpus configurations, like subcorpora. Subcorpus building... And partitions-building. Partition is to slice the corpus in parts, the sum of which is the whole. So this is for contrastive analysis”

“Books are often not interesting without knowledge of the logical works or units within...”

“That’s a whole different dicing intellectually ... Being able to support the huge variety of those kinds of ways of thinking about [texts] at that logical level is a bit challenging. But I think it’s one that somehow has to be approached...”

“We have words, text units, and intermediate structure. Those three levels hold different types of properties”

Figure 2. Selected focus group and interview excerpts on divisibility and objects of analysis.
“The book is not a unit of great interest – you want all the poems that aren’t listed in the metadata. The metadata from the library is very coarse, especially in respect to the goal you have. There’s no opportunity for the experts to provide the deep metadata to share in the broad infrastructure that librarians do very well.”

“Collaborative curation... You could create the data collaboratively, and then explore them collaboratively”

“one thing is getting the data out. But then the next step is, you’ve done all this work, and you then have the authoritative metadata. You have the best metadata in the world, and no one will take that from you. Because it has not been blessed.”

“it would be very important to have the ability to say [of the metadata], this is wrong ...having a workflow which supports that would be important. So the whole idea of social addition comes really into play here.
Use Case 1: Gender

- Scholar wants to compare works by gender, based on the Library of Congress headings.
- This information is in the metadata, but hard to text mine.
- Questions:
  - How can I track gender of authored texts across time?
  - What correlations are there between gender of the author and sentiment analysis of the text?
  - How people and characters of different genders are treated in books over time?
Use Case 2: Serials

A scholar wants to find a series of an author’s works that were originally serialized across several issues or volumes of a periodical.

- Serials vs. volumes as manifestations of works
- Map the pages for content
- Might be able to investigate questions as:
  - What was the original instantiation of the work in serialized form?
  - How can I text mine for sentiment and themes across the serialized texts?
Use Case 3: Images

Scholar wants to find texts of Victorian travel narratives and the images depicted in them. Investigate questions such as:

- What are patterns/themes of images depicted in Victorian England travel narratives?
- What is the frequency of images in travel narratives?
Scholar wants to identify conversational dialogue between characters in novels.

- Requires OCR that detects boundaries: can we detect quote marks and signal words for dialogue?
- Create a training set of curated texts (i.e., TCP texts) matched with HTRC texts, apply detection algorithm
- Enable questions such as:
  - How are characters connected across the narrative—who interacts most frequently?
  - What would sentiment analysis or topic modeling reveal about the dialogue in comparative novels of the genre?
User Needs for Worksets

Comments from interviews/focus groups:

“How do I gather works similar to those I currently have in hand? Can I define different kinds of similarity?”

“How do I merge a HathiTrust collection of works and metadata with my set of works and tags and my colleague’s annotations?”
How useful is existing metadata for creating worksets?

- HT metadata is bibliographic
- Built from MARC records provided by members & OCLC
- Good, consistent quality for author / title / pub info
- Subject less extensive, less consistent
- Genre more hit and miss
- Author gender not present in MARC bib records is present in some MARC authority records
- MARC records provided for serials are about the serial not about the contents of the serial
- No visibility over internal elements (e.g., images, embedded language / genre, dialog, ...) of digitized volumes

Timothy W. Cole (t-cole3@uiuc.edu)
University of Illinois at UC

CNI 2013 Fall Membership Meeting
9 December 2013
<table>
<thead>
<tr>
<th>Category</th>
<th>Count</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Records</td>
<td>6.1 million</td>
<td></td>
</tr>
<tr>
<td>Having at least 1 genre</td>
<td>5.2 million</td>
<td>85%</td>
</tr>
<tr>
<td>Having no genre</td>
<td>0.9 million</td>
<td>15%</td>
</tr>
</tbody>
</table>

- **Conference publication**: 3%
- **Statistics dictionary**: 1%
- **Catalog**: 1%
- **Periodical**: 2%
- **Other**: 3%
- **Non-fiction, not further specified**: 70%
- **Biography**: 5%
- **Fiction, not further specified**: 5%
- **Government publication**: 8%
2.6 million (43%) of bib records include LC Class no.
Not all genres equally described

<table>
<thead>
<tr>
<th>Description</th>
<th>Count</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Volumes identified as fiction</td>
<td>270837</td>
<td></td>
</tr>
<tr>
<td>at least one subject</td>
<td>70706</td>
<td>26%</td>
</tr>
<tr>
<td>no subject</td>
<td>200131</td>
<td>74%</td>
</tr>
<tr>
<td>subjectGeographic</td>
<td>25491</td>
<td>9%</td>
</tr>
<tr>
<td>subjectTemporal</td>
<td>12536</td>
<td>5%</td>
</tr>
<tr>
<td>subjectTopic</td>
<td>61788</td>
<td>23%</td>
</tr>
<tr>
<td>subjectName</td>
<td>13412</td>
<td>5%</td>
</tr>
</tbody>
</table>
Top genres by country of publication

<table>
<thead>
<tr>
<th>bib records</th>
<th>6067835</th>
</tr>
</thead>
<tbody>
<tr>
<td>country specified</td>
<td>5361473</td>
</tr>
<tr>
<td>and genre specified</td>
<td>4776973</td>
</tr>
</tbody>
</table>

United States
- Conference publication: navy blue
- Fiction, not further specified: light blue
- Biography: yellow
- Government publication: red
- Non-fiction, not further specified: dark blue

Germany
- Conference publication: navy blue
- Fiction, not further specified: light blue
- Biography: yellow
- Government publication: red
- Non-fiction, not further specified: dark blue

France
- Conference publication: navy blue
- Fiction, not further specified: light blue
- Biography: yellow
- Government publication: red
- Non-fiction, not further specified: dark blue

Timothy W. Cole (t-cole3@uiuc.edu)
University of Illinois at UC

CNI 2013 Fall Membership Meeting
9 December 2013
Top subjects by country of publication

<table>
<thead>
<tr>
<th>bib records</th>
<th>6067835</th>
</tr>
</thead>
<tbody>
<tr>
<td>country specified</td>
<td>5361473</td>
</tr>
<tr>
<td>and LC class no. specified</td>
<td>2354094</td>
</tr>
</tbody>
</table>

- United States
- Germany
- France

- Philosophy, Psychology, and Religion
- Technology
- Science
- Social Sciences
- Language and Literature
Fiction as proportion of publications by decade

- **Unknown**
- **Fiction**
- **Non-fiction**

Timothy W. Cole  (t-cole3@uiuc.edu) 
University of Illinois at UC
Opportunities

- Computed attributes
  - Author age at time of publication
  - FRBR relationships

- Add attributes not included in bibliographic records
  - Author gender
  - Author nationality

- Improve completeness & accuracy of bib records

- Describe internal components of volumes
More about RFP

- 4 awards to teams of scholars, librarians & developers
  - $40,000 each
  - Period of performance 16 April 2014 – 15 Jan 2015
  - UIUC will supply a testbed of ~250,000 representative volumes; additional volumes (digitized from public domain) available
  - UIUC will collaborate, provide access to HTRC cluster, ...
  - Deliverables: final report; open source software

- Schedule:
  - Letters of Intent Due (preferred): 16 December 2013
  - Final Proposals Due: 13 January 2014
  - Shortlist Meeting Invitations Issued: 20 January 2014
  - Shortlist Meeting: 20 February 2014
  - Award Notification: No later than 15 March 2014
Questions?

Timothy Cole
Mathematics and Digital Services Librarian, UIUC

t-cole3@illinois.edu

Harriett Green
English and Digital Humanities Librarian, UIUC
green19@illinois.edu
Twitter: @greenharr
Discussion Questions

- Key questions to look for in the data
- Alternative approaches and methodologies
- Knowing what we know about user needs to date, what are the implications for formalize the notion of workset
- How does this translate across domains? (e.g., Workset-like objects in science and elsewhere...)
- What are the re-usability and re-producibility implications for such highly individualized and complex digital objects