Case Study:
Creating Worksets in HTRC

J. Stephen Downie and Megan Senseney

Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign
Outline

• Introduction
• HTRC Workset Builder
• HTRC Portal
• Other HTRC Tools and Services
• Hands-On Exercises
• Case Study Consultation and Recommendations
Introduction
In this session, we are presenting HTRC to you as an active project case study.

Instructors will give participants a general tour of HTRC Tools and Services.

Participants will:

- Learn about creating and analyzing worksets in HTRC
- Examine the HTRC initiative from the perspective of data curators
Case Study Questions

1. What are the data?
2. Who are the stakeholders?
3. What are HTRC’s curatorial responsibilities?
4. How would a researcher using HTRC plan for data curation at the level of an individual project?
5. What documentation is currently available? How easily can you find it? Is it enough?
Case Study Questions (cont’d)

6. How does the idea of “non-consumptive research” affect decisions for curation?

7. Are worksets shareable?

8. What would you need to know about another scholar’s workset to decide whether its relevant for your research purposes?

9. What would you need to know about another scholar’s workset to make use of it in your own research?
Personal Account Creation

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.

Sign In to Begin

https://htrc2.pti.indiana.edu/HTRC-UI-Portal2/
Registration

User Registration (ALL fields are required)

- User ID
- Password
- Retype-Password
- First Name
- Last Name
- E-mail

Submit
Summary of Steps for Workset Creation and Analysis

• Create a workset
• Select algorithm
  – Provide job name
  – Select workset
  – Adjust parameters
• Run
• View results
Workset Builder
https://htrc2.pti.indiana.edu/blacklight/
Workset Builder

• Based on Blacklight 3.5
• Provides a familiar search search interface to the entire HTRC collection
• Searches based on full text, author, title, or subject
• Allows for creating/updating worksets

NB: You’ll need to log in to portal and workset builder separately
Workset Builder: Search

NB: Click “More options” for advanced search options

https://htrc2.pti.indiana.edu/blacklight/
Workset Builder: Search Results

An author search for “Ruskin, John” filtered using left-hand sidebar, first by author then by subject.

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

Once texts are selected, click “Selected Items”, located in upper right-hand menu.
Workset Builder: Create Workset

When you’re done selecting items, review your choices, and select “Create/Update Workset”
NB: you have the option to add selected items to an existing workset or to create a new workset.

You can also alter or remove existing worksets by choosing “Manage Worksets”.
Example: after creating the Ruskin workset, I went back and removed all duplicate titles, reducing workset from 113 items to 12.
HTRC Portal
https://htrc2.pti.indiana.edu/HTRC-UI-Portal2/
Portal: Workset Menu

Click “List” in the Workset menu to review your workset in the Portal

Don’t forget to sign in!
Portal: Workset Details

Downloadable CSV includes volume ID and title as well as dividing authors by gender.

HTRC-added metadata includes author gender, page count, and word count for each item.
Choose among 10 possible algorithms and set parameters.

Each algorithm includes a basic description of its use, purpose, and any computational limits.
Meandre Data Flow

Flow Parameter → Universal Text Extractor → Search Text → HTRC Page Retriever → Text Cleaner

Sentence Detector → Sentence Tokenizer → Named Entity → Add Tuple Attribute → Add Tuple Attribute → Tuple Aggregator

Tuple To HTML → Stream Delimiter Filter → Push Text → Write To File
Portal: Results

Even a small workset will take a few minutes to process. The larger the workset, the longer it takes.

Refresh your browser to track status updates.
Job Details: Topic Modeling

Topic modeling data is now available as a word cloud or as a structured XML document.
Named entities for person, organization, and location in a tabular format.
How Ruskin compares to 200 other works on painting.
# A Note on Performance

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Number of Volumes</th>
<th>Execution Time (Minutes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meandre-Classification_NaiveBayes</td>
<td>1000</td>
<td>85.28</td>
</tr>
<tr>
<td>Meandre_Dunning_LogLikelihood_to_Tagcloud</td>
<td>1000</td>
<td>34.35</td>
</tr>
<tr>
<td>Meandre_OpenNLP_Date_Entities_to_Simile</td>
<td>100</td>
<td>21.93</td>
</tr>
<tr>
<td>Meandre_OpenNLP_Entities_List</td>
<td>100</td>
<td>26.22</td>
</tr>
<tr>
<td>Meandre_Spellcheck_Report_Per_Volume</td>
<td>100</td>
<td>16.02</td>
</tr>
<tr>
<td>Meandre_Tagcloud</td>
<td>1000</td>
<td>2.00</td>
</tr>
<tr>
<td>Meandre_Tagcloud_with_Cleaning</td>
<td>1000</td>
<td>6.79</td>
</tr>
<tr>
<td>Meandre_Topic_Modeling</td>
<td>1000</td>
<td>84.90</td>
</tr>
</tbody>
</table>
Other HTRC Tools
HTRC Bookwork Instance: Settings

http://sandbox.htrc.illinois.edu/bookworm/
HTRC Bookworm Instance: Search
What are women writing about hysteria in 1888?
HTRC Bookworm Instance: Drilling Down

Links to full text in HathiTrust for all items in the public domain reveal:
- 1 volume by medical doctor
- 4 works of fiction
- 1 likely outlier
Welcome to the HathiTrust Research Center Sandbox!

About Us

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.

What is the Sandbox?

The HTRC Sandbox is distinct from the main production portal of the HTRC. The HTRC Sandbox is meant to be an arena for users to try out experiments and do exploratory work. The dataset available on the sandbox is a much smaller subset of that associated with the HTRC's main production portal. The HTRC Sandbox dataset consists of the non-Google-digitized public domain volumes (approximately 250,000 volumes) from the HathiTrust corpus.

The HTRC Data API is available for experimentation and several additional feature data and tools, such as HTRC-Bookworm, are being connected to this data for exploratory analysis. HTRC users can write their own programs, in programming languages of their choice, accessing the data through the HTRC Data API programmaticallly as

https://sandbox.htrc.illinois.edu/HTRC-UI-Portal2/
Researchers performing text analysis usually process documents into a *machine-readable* set of features.
HTRC Feature Extraction: Features

• Persistent volume IDs throughout HTRC
• Per page:
  – Token counts
  – Tokens
  – Part of speech counts per token
  – Line counts
  – Empty line counts
  – Sentence count
  – Beginning and end of line characters
• Identify header, body, and footer
HTRC Feature Extraction: Benefits

• Provides a path toward sharing non-consumptive versions of documents without detracting from most uses of such data.

• Saves processing and development time for scholars.

• Offers value-added processing (e.g., re-joining hyphenation and identifying headers).

Coming Soon! Workset-specific feature download
Hands-On Exercise
Discussion
Special thanks to Loretta Auvil, Stacy Kowalczyk, and Peter Organisciak.