Digging Deeper, Reaching Further

Module 2.1: Gathering Textual Data

Finding and Curating Text Data
In this module we’ll…

- Explore the concept of a text data and where to find it
  - *Provide data reference for researchers*

- Build a HathiTrust workset
  - *Gain experience in building a textual dataset*

- Learn how Sam built a *Creativity Corpus* of HathiTrust volumes
  - *Understand real-world data collection strategies*
Where we’ll end up

Create a collection of volumes from the HathiTrust Digital Library and prepare it for analysis in HTRC Analytics as a workset
“Text analysis projects share in common 3 challenges. **First**, data of interest must be found. **Second**, data must be gettable. **Third**, if it’s not already formed according to wildest dreams, ways must be known of getting data into a state that they are readily usable with desired methods and tools.”

Kludging: Web to TXT (Padilla, 2015)
http://www.thomaspadilla.org/2015/08/03/kludge/
Finding text

- Not always easy
  - copyright restrictions
  - licensing restrictions
  - format limitations
  - hard-to-navigate systems

** issues more pronounced at scale**
Vendor databases

- Be aware of licensing restrictions

- Strategies
  - Addendums to libraries’ contracts
  - Vendor-provided services
  - Asking for special permission case-by-case

- Example: JSTOR Data for Research
Library/archives digital collections

- Wealth of material, but:
  - Often siloed
  - Access not formulated for research at scale

- Things to look for:
  - Plain text
  - Bulk download

- Example: UNC’s DocSouth Data
Social media

- Popular with social science researchers

- To access:
  - Some provide systems to access text
  - Or there are 3rd-party tools on the market

- Example: Twitter API (Application Programming Interface)
Hands-on activity

Building a corpus for political history, what are the strengths and weaknesses of each of these broad sources for textual data?

<table>
<thead>
<tr>
<th>Source</th>
<th>Strengths</th>
<th>Weaknesses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vendor database</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Library/archives digital collections</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Social media</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Evaluating sources of text data

<table>
<thead>
<tr>
<th>Does the researcher already have a data source in mind?</th>
<th>Is the text they want to use already digitized?</th>
<th>Are there copyright and licensing concerns?</th>
<th>How technically experienced is the researcher?</th>
</tr>
</thead>
<tbody>
<tr>
<td>What is the period, place, person of interest?</td>
<td>How much flexibility is needed for working with the data?</td>
<td>Does the researcher have funding?</td>
<td>What format does the researcher expect the data in?</td>
</tr>
</tbody>
</table>
Building corpora

- Identify texts through full text search
  - Use a key term or phrase

- Identify texts through metadata
  - Search by certain author(s)
  - Search within a date range
  - Search for a specific genre

- Or some combination of the two!
Building corpora

- Process usually involves deduplication

- What to keep/discard is project dependent

- Examples of deduplication:
  - OCR quality
  - Earliest edition
  - Editions without forewords or afterwords
HTRC Worksets

- User-created collections of text from the HathiTrust Digital Library
  - think of them as textual datasets
- Can be shared and cited
- Suited for non-consumptive access
HTRC Worksets

Workset viewed on the web

Workset manifest
Building worksets

- Stored in HTRC
  - Require account with university email address

- Ways to build:
  - Import from HT Collection Builder
  - Compile volume IDs elsewhere
Sample Reference Question

I’m a student in history who would like to incorporate digital methods into my research. I study American politics, and in particular I’d like to examine how concepts such as liberty change over time.

Approach:

- Create a textual dataset of volumes related to political speech in America with the HT Collection Builder, and upload it to HTRC Analytics as a workset for analysis.
Hands-on activity

In this activity, you will log in to HTDL and create a collection containing volumes of the public papers of the presidents of the United States, and import it into HTRC Analytics as a workset. Follow the instructions on the handout to build your workset.

Websites:

- HTDL: https://www.hathitrust.org
- HTRC Analytics: https://analytics.hathitrust.org
Go to HTDL interface

https://www.hathitrust.org
Log in
Log in

Find your partner institution:

University of Illinois at Urbana-Champaign

Why isn't my institution listed?

Not with a partner institution?
See options to log in as a guest
Search for volumes

• Click on “Advance full-text search”
Search for volumes

Advanced Full-text Search:

Search information within or about an item

**Search Tips**

- this exact phrase
- AND
- this exact phrase

United States

Title

Title

+ Add a pair of search fields

Limit to:

Full view only

Year of publication: During or after

Language

Limit to Original Format

All

All
Filter results and select volumes

Filter results on the left sidebar

An advanced search for volumes that contain all the words/phrases below in the title field: “public papers” and “United States”

Select all or some of the returned search items for your collection.
Add volumes to collection

Once texts are selected, click “Select Collection” ➔ choose “[CREATE NEW COLLECTION]” ➔ click “Add Selected”
Add collection metadata

Collection Name: PoliticalSpeech
Description: A collection of volumes of public speeches by the presidents of the United States
View your collection
View your collection

1930s_political_speeches_DDRF
1930s political speeches collection for DDRF workshop
Owner: Ruohua Han (University of Illinois at Urbana-Champaign)
5 items
last updated: 08/31/17
Public: Make Private Delete Collection

1970s_political_speeches_DDRF
1970s political speeches collection for DDRF workshop
Owner: Ruohua Han (University of Illinois at Urbana-Champaign)
16 items
last updated: 08/31/17
Public: Make Private Delete Collection

PoliticalSpeech
A collection of volumes on political speeches by the presidents of the United States
Owner: Ruohua Han (University of Illinois at Urbana-Champaign)
2 items
last updated: 10/05/17
Private: Make Public Delete Collection
Grab the collection URL

https://babel.hathitrust.org/cgi/mb?a=list,c=1848885365
Go to HTRC Analytics

https://analytics.hathitrust.org
Welcome! Returning users signing into the new HTRC Analytics interface for the first time must reset their password using the “Forgot Password” link below.

Sign In to HathiTrust Research Center

Username

Password

Remember me on this computer

SIGN IN

Forgot Password? | Forgot Username? | Create Account
Go to Worksets page
Choose to create a workset
Choose creation method

How would you like to create your workset?

Upload File
Create a workset from a file of HathiTrust volume IDs

Import From HathiTrust
Create a workset from an existing, public HathiTrust collection
Input workset information

Create A Workset

Import a collection from HathiTrust using the collection's URL. While HathiTrust grows daily, HTRC syncs data periodically from the HathiTrust Digital Library. Some volumes you would like to include in your workset may not be available. Any volumes in your workset not available through HTRC will be skipped by the algorithm.

Find collection URL

When viewing your collection on HathiTrust, simply copy the URL from your browser, or copy the "Link to this collection" found on the left sidebar, and paste the URL below.

Import

Hit "Fetch Collection" and your collection will be transformed into an HTRC workset. You may need to edit the default name in order to meet HTRC requirements.

HathiTrust Collection URL

https://babel.hathitrust.org/cgi/mb?id=xilistc.1849865365

Add collection URL here

Create Workset
View created workset

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Volume Count</th>
<th>Last Modified Date</th>
<th>Availability</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1970s_political_speeches_DDRF</td>
<td>1970s political speeches collection for DDRF workshop</td>
<td>16</td>
<td>October 5, 2017</td>
<td>public</td>
<td></td>
</tr>
<tr>
<td>Darwin_Test_2</td>
<td></td>
<td>3</td>
<td>October 6, 2016</td>
<td>private</td>
<td></td>
</tr>
<tr>
<td>1930s_political_speeches_DDRF</td>
<td>1930s political speeches collection for DDRF workshop</td>
<td>5</td>
<td>October 5, 2017</td>
<td>public</td>
<td></td>
</tr>
<tr>
<td>poli_science_DDRF</td>
<td>Political science collection for DDRF workshop</td>
<td>16</td>
<td>October 5, 2017</td>
<td>public</td>
<td></td>
</tr>
<tr>
<td>Charles_Darwin_Test</td>
<td></td>
<td>32</td>
<td>August 24, 2016</td>
<td>private</td>
<td></td>
</tr>
</tbody>
</table>
Workset review

- How did it go?

- What kind of search criteria did you use?

- Did you find any challenges?
Case study: *Inside the Creativity Boom*

Building a creativity corpus

- Searched across full text of HTDL for creativ*
- Made initial list of over million volumes
- Deduplicated
  - Kept different editions of same work; discard multiple copies of same edition
- Ended up with refined list (workset) of volumes
Discussion

- What expertise do librarians already have to help with building a corpus for textual analysis?
Questions?
References