

**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

poli_science_DDRF

[Download](#)

Description : Political science collection for DDRF workshop

Owner	Last Modified Time	Number of Volumes	Tags
rhan11	2017-10-05T18:21:35Z	16	

Filter volume by title...

Volume ID	Title	Authors	Year	Language
mdp.49015002203223	Public papers of the presidents of the United States.	United States President; Clinton, Bill 1946-; Bush, George 1924-; Reagan, Ronald; Carter, Jimmy 1924-; Ford, Gerald R. 1913-2006; Nixon, Richard M. (Richard Milhous) 1913-1994; Johnson, Lyndon B. (Lyndon Baines) 1908-1973; Kennedy, John F. (John Fitzgerald) 1917-1963; Eisenhower, Dwight D. (Dwight David) 1890-1969; Truman, Harry S. 1884-1972; Hoover, Herbert 1874-1964; United States Federal Register Division; United States Office of the Federal Register	1978	eng
mdp.49015002203272	Public papers of the presidents of the	United States President; Clinton, Bill 1946-; Bush, George 1924-; Reagan, Ronald; Carter, Jimmy 1924-; Ford, Gerald R. 1913-2006; Nixon, Richard M. (Richard Milhous) 1913-1994; Johnson, Lyndon B. (Lyndon Baines) 1908-1973; Kennedy, John F. (John Fitzgerald) 1917-1963; Eisenhower, Dwight D. (Dwight David) 1890-1969; Truman, Harry S. 1884-1972;	1979	eng

Create a collection of volumes from the HathiTrust Digital Library and prepare it for analysis in HTRC Analytics as a workset



Kludging access

“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

 See Handout p. 1

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

	Strengths	Weaknesses
Vendor database		
Library/archives digital collections		
Social media		



Evaluating sources of text data

Does the researcher already have a data source in mind?

Is the text they want to use already digitized?

Are there copyright and licensing concerns?

How technically experienced is the researcher?

What is the period, place, person of interest?

How much flexibility is needed for working with the data?

Does the researcher have funding?

What format does the researcher expect the data in?



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

poli_science_DDRF

[Download](#)

Description : Political science collection for DDRF workshop

Owner	Last Modified Time	Number of Volumes	Tags
rhan11	2017-10-05T18:21:35Z	16	

Filter volume by title...

Volume ID	Title	Authors	Year	Language
mdp.49015002203223	Public papers of the presidents of the United States.	United States President; Clinton, Bill 1946-; Bush, George 1924-; Reagan, Ronald; Carter, Jimmy 1924-; Ford, Gerald R. 1913-2006; Nixon, Richard M. (Richard Milhous) 1913-1994; Johnson, Lyndon B. (Lyndon Baines) 1908-1973; Kennedy, John F. (John Fitzgerald) 1917-1963; Eisenhower, Dwight D. (Dwight David) 1890-1969; Truman, Harry S. 1884-1972; Hoover, Herbert 1874-1964; United States Federal Register Division; United States Office of the Federal Register	1978	eng
mdp.49015002203272	Public papers of the presidents of the	United States President; Clinton, Bill 1946-; Bush, George 1924-; Reagan, Ronald; Carter, Jimmy 1924-; Ford, Gerald R. 1913-2006; Nixon, Richard M. (Richard Milhous) 1913-1994; Johnson, Lyndon B. (Lyndon Baines) 1908-1973; Kennedy, John F. (John Fitzgerald) 1917-1963; Eisenhower, Dwight D. (Dwight David) 1890-1969; Truman, Harry S. 1884-1972;	1979	eng

Workset viewed on the web

mdp.49015002221845
mdp.49015002221837
mdp.49015002221829
mdp.49015002221787
mdp.49015002221811
mdp.49015002221761
mdp.49015002221779
mdp.49015002203140
mdp.49015002203157
mdp.49015002203033
mdp.49015002203231
mdp.49015002203249
mdp.49015002203223
mdp.49015002203405
mdp.49015002203272
mdp.49015002203215

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

 *See Handout pp. 2-4*

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

The screenshot displays the HathiTrust Digital Library homepage. At the top, a dark navigation bar contains links for Home, About, Collections, Help, and Feedback. Below this is the HathiTrust logo, which features an orange elephant head icon and the text "HATHI TRUST Digital Library". To the right of the logo is a yellow "LOG IN" button with a dropdown arrow.

The main content area is dominated by a search interface. It includes a search bar with the placeholder text "Search words about the items" and a "Search" button with a magnifying glass icon. Above the search bar are two tabs: "FULL-TEXT" and "CATALOG". To the right of the search bar is a dropdown menu labeled "All Fields". Below the search bar, there are links for "Advanced catalog search" and "Search tips", and a checkbox for "Full view only". A help icon and the text "? Should I search catalog or full-text?" are also present.

Below the search interface, there is a section titled "WHAT CAN YOU DO WITH HATHITRUST?". It contains a paragraph: "HathiTrust is a partnership of academic & research institutions, offering a collection of millions of titles digitized from libraries around the world." Below this paragraph are four featured collection cards:

- BROWSE COLLECTIONS**: Explore user-created featured collections.
- READ BOOKS ONLINE**: Read millions of titles online — like this one!
- READ BOOKS ON THE GO**: Take the library's books anywhere with our mobile website.
- DOWNLOAD BOOKS* & CREATE COLLECTIONS**: *requires institutional login

On the right side of the page, there is a yellow sidebar with the heading "Want to get the most out of HathiTrust?". It contains the text: "Log in with your partner institution account to access the largest number of volumes and features. Not with a partner institution? See options to log in as a guest".



Log in

The screenshot shows the HathiTrust Digital Library homepage. At the top, there is a navigation menu with links for Home, About, Collections, Help, and Feedback. Below the menu is the HathiTrust logo, which consists of an orange elephant head icon and the text 'HATHI TRUST Digital Library'. The main content area features a search bar with the text 'Search HathiTrust's digital library'. The search bar has a dropdown menu for 'All Fields' and a 'Search' button. Below the search bar, there are links for 'Advanced catalog search' and 'Search tips', and a checkbox for 'Full view only'. A yellow callout box on the right side of the page contains the text 'Want to get the most out of HathiTrust?' and 'Log in with your partner institution account to access the largest number of volumes and features.' Below this, it says 'Not with a partner institution? See options to log in as a guest'. An orange arrow points from the callout box to the 'LOG IN' button. At the bottom of the page, there is a section titled 'WHAT CAN YOU DO WITH HATHITRUST?' with four columns of content: 'BROWSE COLLECTIONS', 'READ BOOKS ONLINE', 'READ BOOKS ON THE GO', and 'DOWNLOAD BOOKS* & CREATE COLLECTIONS'. The 'DOWNLOAD BOOKS*' section includes a note: '*requires institutional login'.

Home About Collections Help Feedback

 **HATHI TRUST**
Digital Library

Search HathiTrust's digital library

FULL-TEXT CATALOG

Search words about the items All Fields Search

[Advanced catalog search](#) | [Search tips](#) Full view only

[? Should I search catalog or full-text?](#)

Want to get the most out of HathiTrust?

Log in with your partner institution account to access the largest number of volumes and features.

Not with a partner institution?
[See options to log in as a guest](#)

HathiTrust is a [partnership](#) of academic & research institutions, offering a collection of millions of titles digitized from libraries around the world.

WHAT CAN YOU DO WITH HATHITRUST?

- BROWSE COLLECTIONS**
Explore user-created [featured collections](#).
- READ BOOKS ONLINE**
Read millions of titles online — [like this one!](#)
- READ BOOKS ON THE GO**
Take the library's books anywhere with our [mobile website](#).
- DOWNLOAD BOOKS* & CREATE COLLECTIONS**
**requires institutional login*



Log in

Trust's digital library

LOG IN

Find your partner institution:

University of Illinois at Urbana-Champaign

CONTINUE →

[Why isn't my institution listed?](#)

Not with a partner institution?
[See options to log in as a guest](#)

BROWSE COLLECTIONS
Explore user-created featured collections.

READ BOOKS ONLINE
Read millions of titles online — like this one!

READ BOOKS ON THE GO
Take the library's books anywhere with our mobile website.

DOWNLOAD BOOKS* & CREATE COLLECTIONS
*requires institutional login



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](#)

Prefer to search items in an [Advanced Search?](#)

in

in

[+ Add a pair of search fields](#)

Limit to:

Full view only Year of publication:

Language Limit to Original Format



Filter results and select volumes

Filter results on the left sidebar

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

The screenshot shows a search results interface. On the left is a 'Refine Results' sidebar with categories: Subject, Language, Place of Publication, and Date of Publication. The main area shows 'Search Results: 1,729 items found' with an advanced search query: 'this exact phrase: United States in Title' AND 'this exact phrase: public papers in Title'. Below the search bar are options to 'Revise this advanced search', view counts for 'All Items (1,729)' and 'Full View (1,583)', a '25 per page' dropdown, and a pagination list from 1 to 70. At the bottom of the search bar area are 'Select all on page' and 'Add Selected' buttons. The results list includes three items, each with a checkbox, a title, author information, and publication year. The first item is 'The public papers and addresses of Franklin D. Roosevelt. 1943 volume...' by Samuel I. Rosenman, published 1950, with 'Catalog Record' and 'Limited (search-only)' links. The second item is 'Public papers of the presidents of the United States. 1987 pt.2' by United States. President, published 1987, with 'Catalog Record' and 'Full view' links. The third item is 'The public papers and addresses of Franklin D. Roosevelt. 1941 volume...' by Samuel I. Rosenman.

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



Add volumes to collection

The screenshot shows a digital library interface. At the top, there are two tabs: "All Items (1,607)" and "Full View (1,549)". Below the tabs is a pagination control showing "25 per page" and a list of page numbers from 1 to 65, with a "Next" button. A grey bar contains a "Select all on page" checkbox, a "[CREATE NEW COLLECTION]" dropdown menu, and an "Add Selected" button. An orange arrow points from the "Select all on page" checkbox to the "[CREATE NEW COLLECTION]" dropdown. Below this bar is a list of items. The first item is "Public papers of the presidents of the United States. 1989 pt. 1" by United States. President, published 1989, with a checked checkbox and a thumbnail. The second item is "Herbert Hoover proclamations and executive orders, March 4, 1929 to March 4, 1933. proc.001" by Hoover, Herbert, 1874-1964, published 1974, with an unchecked checkbox and a thumbnail. The third item is "Public papers of the presidents of the United States. 1988 pt.1" by United States. President, with a checked checkbox and a thumbnail. Each item has links for "Catalog Record" and "Full view".

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



Add collection metadata

Start Over ✖ this exact phrase: public papers in Title

✖ AND this exact phrase: United States in Title ✖ all of these words: president in Title

Collection Name 85

Description 174

Private Public

Cancel Save Changes

[Catalog Record](#) [Full view](#)

Herbert Hoover proclamations and executive orders, March 4, 1929 to March 4, 1933. proc.001
by Hoover, Herbert, 1874-1964.
Published 1974



View your collection

The screenshot shows the HathiTrust Digital Library interface. At the top, there is a navigation bar with links for Home, About, Collections, Help, Feedback, Member (University of Illinois at Urbana-Champaign), My Collections, and Logout. An orange arrow points to the 'My Collections' link. Below the navigation bar is the HathiTrust logo and a search bar containing the text 'public papers'. To the right of the search bar are buttons for 'FULL-TEXT' and 'CATALOG', and a search icon. Below the search bar, there are links for 'Advanced full-text search' and 'Search tips', and a checkbox for 'Full view only'. A green notification bar at the top of the results area states '2 items were added to PoliticalSpeech'. The main content area displays 'Search Results: 1,607 items found'. Below this, there are buttons for 'Start Over' and a search filter 'this exact phrase: public papers in Title'. There are also filters for 'AND this exact phrase: United States in Title' and 'all of these words: president in Title'. A 'Revise this advanced search' button is present. At the bottom of the results area, there are buttons for 'All Items (1,607)' and 'Full View (1,549)'. A pagination control shows '25 per page' and a sequence of page numbers from 1 to 65, with a 'Next' button. At the very bottom, there is a 'Select all on page' checkbox, a '[CREATE NEW COLLECTION]' dropdown, and an 'Add Selected' button.

Home About Collections Help Feedback Member (University of Illinois at Urbana-Champaign) **My Collections** Logout

HATHI TRUST Digital Library

FULL-TEXT CATALOG

public papers

Advanced full-text search | Search tips Full view only

✓ 2 items were added to [PoliticalSpeech](#)

Refine Results

Subject

- [United States](#) (1,607)
- [United States Politics and government Periodicals](#) (1,431)
- [Government publications](#) (923)
- [Government publications Bibliography](#) (923)
- [Government publications Indexes](#) (923)
- [more...](#)

Author

- [United States. National Archives and Records Administration](#) (923)
- [United States. Office of the Federal Register](#) (923)
- [United States. Office of the](#)

Search Results: 1,607 items found

Start Over this exact phrase: **public papers** in Title

AND this exact phrase: **United States** in Title all of these words: **president** in Title

Revise this advanced search

All Items (1,607) **Full View (1,549)**

25 per page 1 2 3 4 5 6 7 8 ... 65 Next ➔

Select all on page [CREATE NEW COLLECTION] Add Selected



View your collection

Showing 5 of your collections [Reset](#)

[1930s political speeches DDRF](#)

1930s political speeches collection for DDRF workshop 5 items
last updated: 08/31/17

Owner: Ruohua Han (University of Illinois at Urbana-Champaign)

[Public : Make Private](#) [Delete Collection](#)

[1970s political speeches DDRF](#)

1970s political speeches collection for DDRF workshop 16 items
last updated: 08/31/17

Owner: Ruohua Han (University of Illinois at Urbana-Champaign)

[Public : Make Private](#) [Delete Collection](#)

[PoliticalSpeech](#)

A collection of volumes of public speeches by the presidents of the United States 2 items
last updated: 10/05/17

Owner: Ruohua Han (University of Illinois at Urbana-Champaign)

[Private : Make Public](#) [Delete Collection](#)



A collection of mostly 19th-20th-century musical score women composers held at University of Michigan Mus

[UM Press](#)



The Univ. of Michigan Press available in HathiTrust

[University Press of Florida](#)



Selected publications of the



Grab the collection URL



https://babel.hathitrust.org/cgi/mb?a=listis;c=1848985365

Home About Collections Help Feedback

HATHI TRUST Digital Library

FULL-TEXT CATALOG

Search words about or within the items

Advanced full-text search | Search tips

Full view only

Share

Link to this collection

<https://babel.hathitrust.org/cgi/mb?a=l>

Download Metadata Convert to HTRC Workset

About this collection

Owner
Ruohua Han

Status
public

poli_science_DDRF

Political science collection for DDRF workshop

Search in this collection Find

All Items (16)

Sort by: Title A-Z 25 per page 1

Select all on page Select Collection Add Selected

Public papers of the presidents of the United States. 1971 In my collections: ---
by United States. President.
Published 1971

[Catalog Record](#) [Full view](#)
[Download Extracted Features](#)



Go to HTRC Analytics

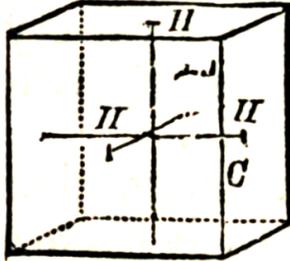
HTRC Analytics Algorithms Data Capsules Worksets Datasets Explore Help About Sign In Sign Up



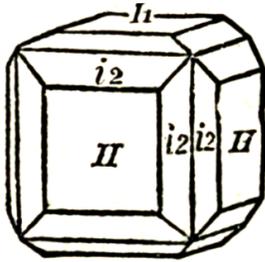
HathiTrust Research Center Analytics

Supports large-scale computational analysis of the works in the HathiTrust Digital Library to facilitate non-profit and educational research.

Featured Services



Extracted Features



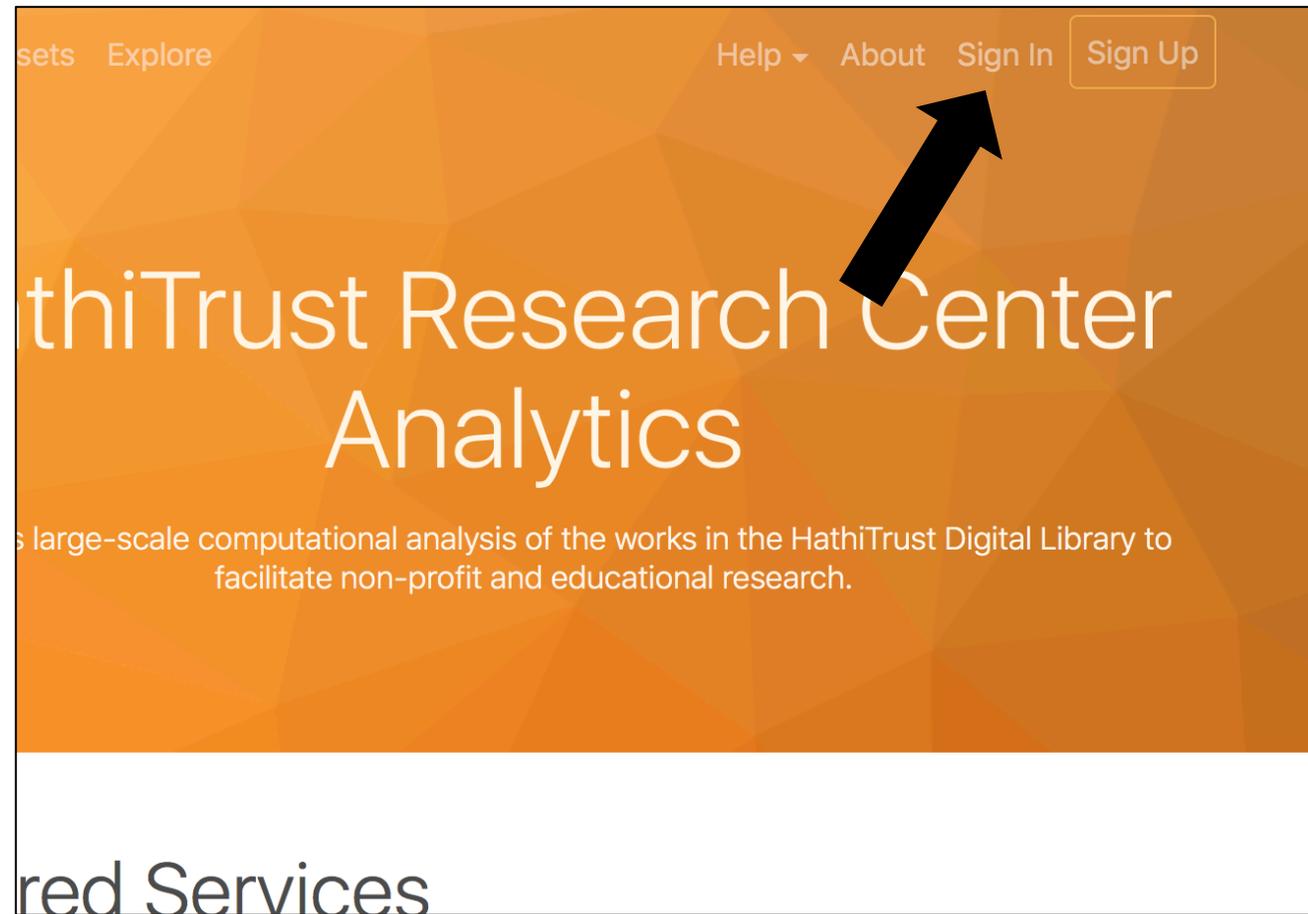
Text Analysis Algorithms



Data Capsules



Sign in / Sign up



Sign in

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](#)

HathiTrust Research Center | © 2017



Sign in

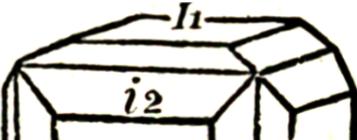
es Worksets Datasets Explore

Help About rhan11

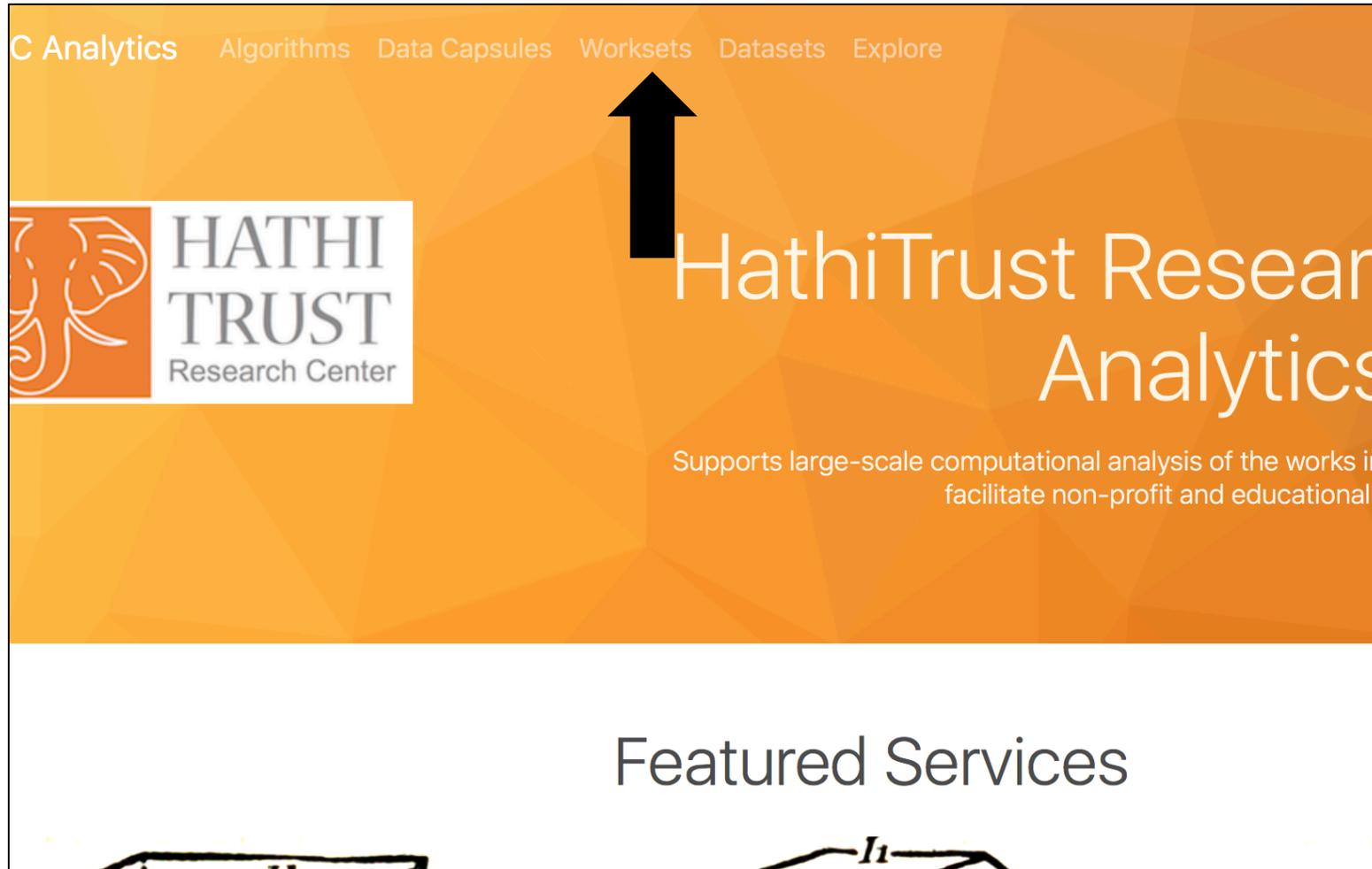
HathiTrust Research Center Analytics

Supports large-scale computational analysis of the works in the HathiTrust Digital Library to facilitate non-profit and educational research.

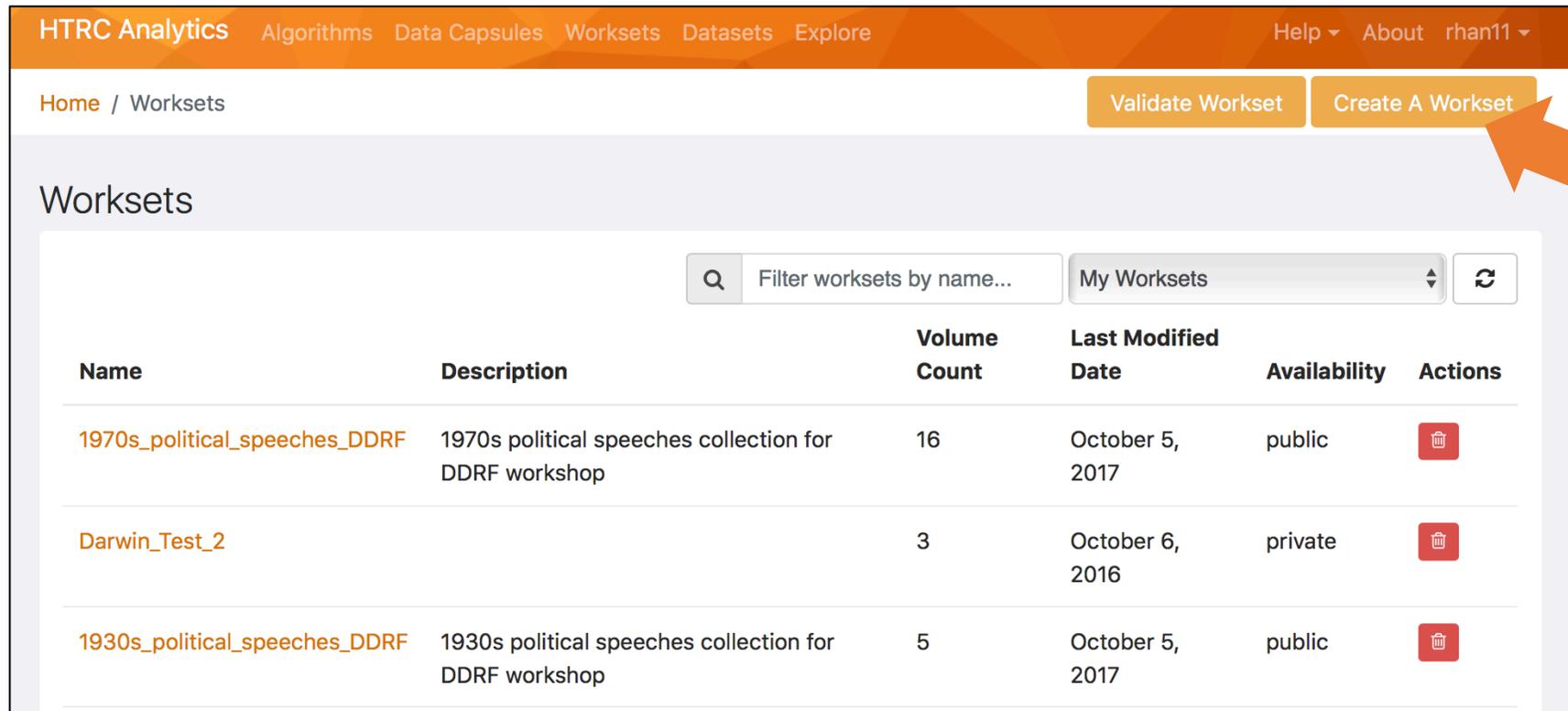
Featured Services



Go to Worksets page



Choose to create a workset



The screenshot shows the HTRC Analytics interface. The top navigation bar includes 'HTRC Analytics', 'Algorithms', 'Data Capsules', 'Worksets', 'Datasets', and 'Explore'. On the right, there are links for 'Help', 'About', and the user 'rhan11'. Below the navigation bar, the breadcrumb 'Home / Worksets' is visible. Two buttons are present: 'Validate Workset' and 'Create A Workset'. An orange arrow points to the 'Create A Workset' button. The main content area is titled 'Worksets' and contains a search bar 'Filter worksets by name...', a dropdown menu set to 'My Worksets', and a refresh icon. Below this is a table with the following data:

Name	Description	Volume Count	Last Modified Date	Availability	Actions
1970s_political_speeches_DDRF	1970s political speeches collection for DDRF workshop	16	October 5, 2017	public	
Darwin_Test_2		3	October 6, 2016	private	
1930s_political_speeches_DDRF	1930s political speeches collection for DDRF workshop	5	October 5, 2017	public	



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

Name

Disallowed characters: ~ ! @ # ; % ^ * + = [] | < > , ' " \ /

Description

Private Workset
If checked, your workset will be accessible to only you.

Add collection URL here



View created workset

Worksets

Filter worksets by name... My Worksets

Name	Description	Volume Count	Last Modified Date	Availability	Actions
1970s_political_speeches_DDRF	1970s political speeches collection for DDRF workshop	16	October 5, 2017	public	
Darwin_Test_2		3	October 6, 2016	private	
1930s_political_speeches_DDRF	1930s political speeches collection for DDRF workshop	5	October 5, 2017	public	
poli_science_DDRF	Political science collection for DDRF workshop	16	October 5, 2017	public	
Charles_Darwin_Test	testing purposes	32	August 24, 2016	private	

First « 1 » Last

Showing 1 to 10 of 5 entries



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

- Padilla, T. (2015). Kludging: Web to TXT. Retrieved August 16, 2017, from <http://www.thomaspadilla.org/2015/08/03/kludge/> .

