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*

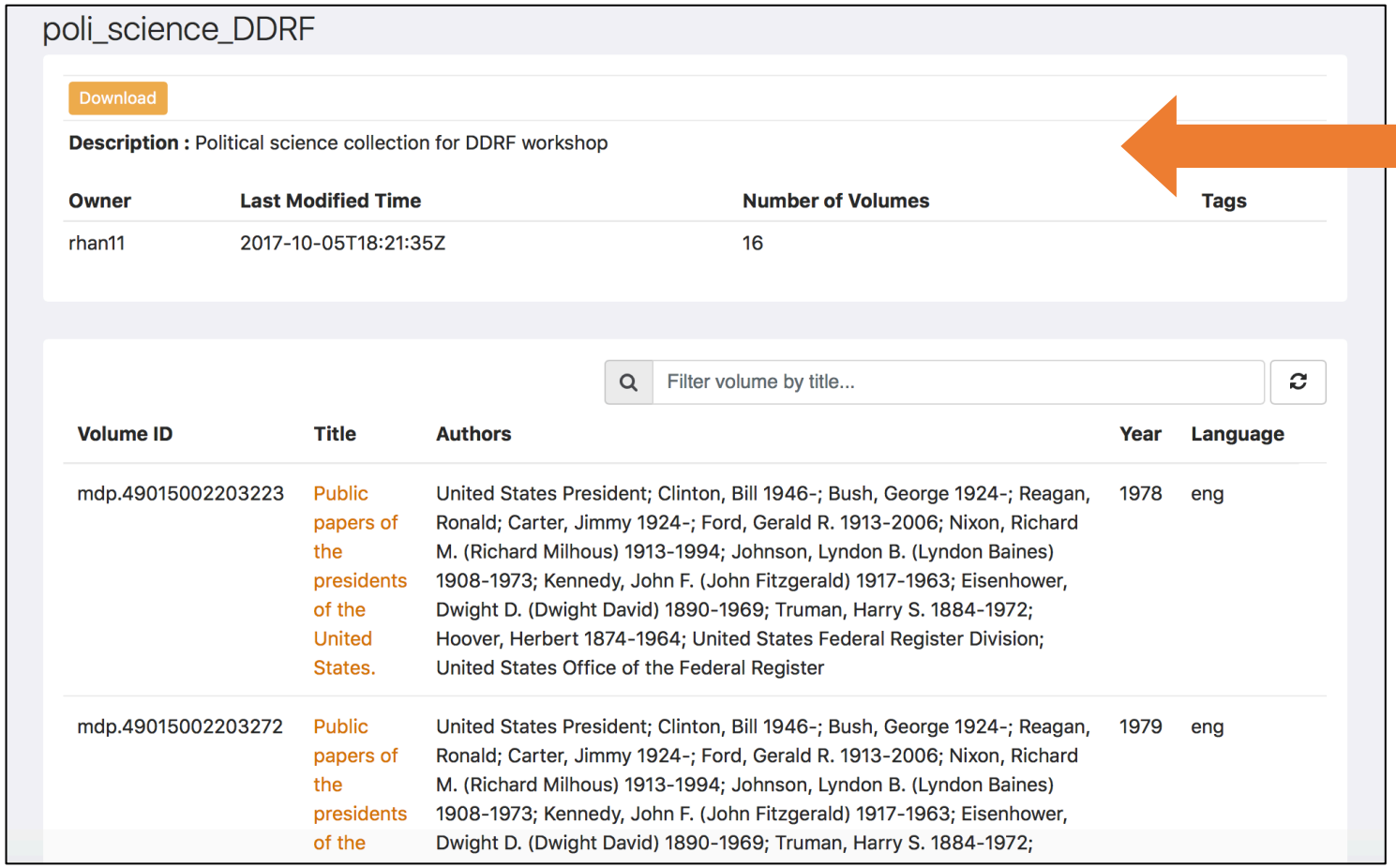
- Build a HathiTrust workset

  → *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

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!

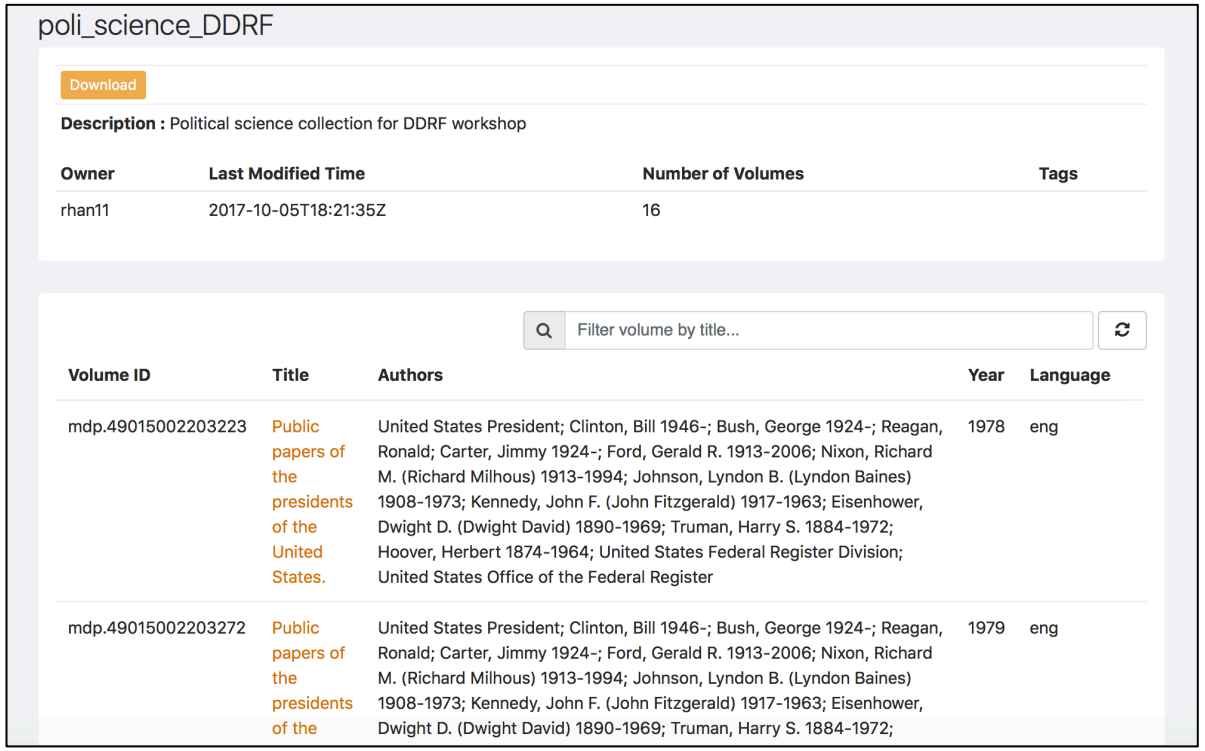
Building corpora

- 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*

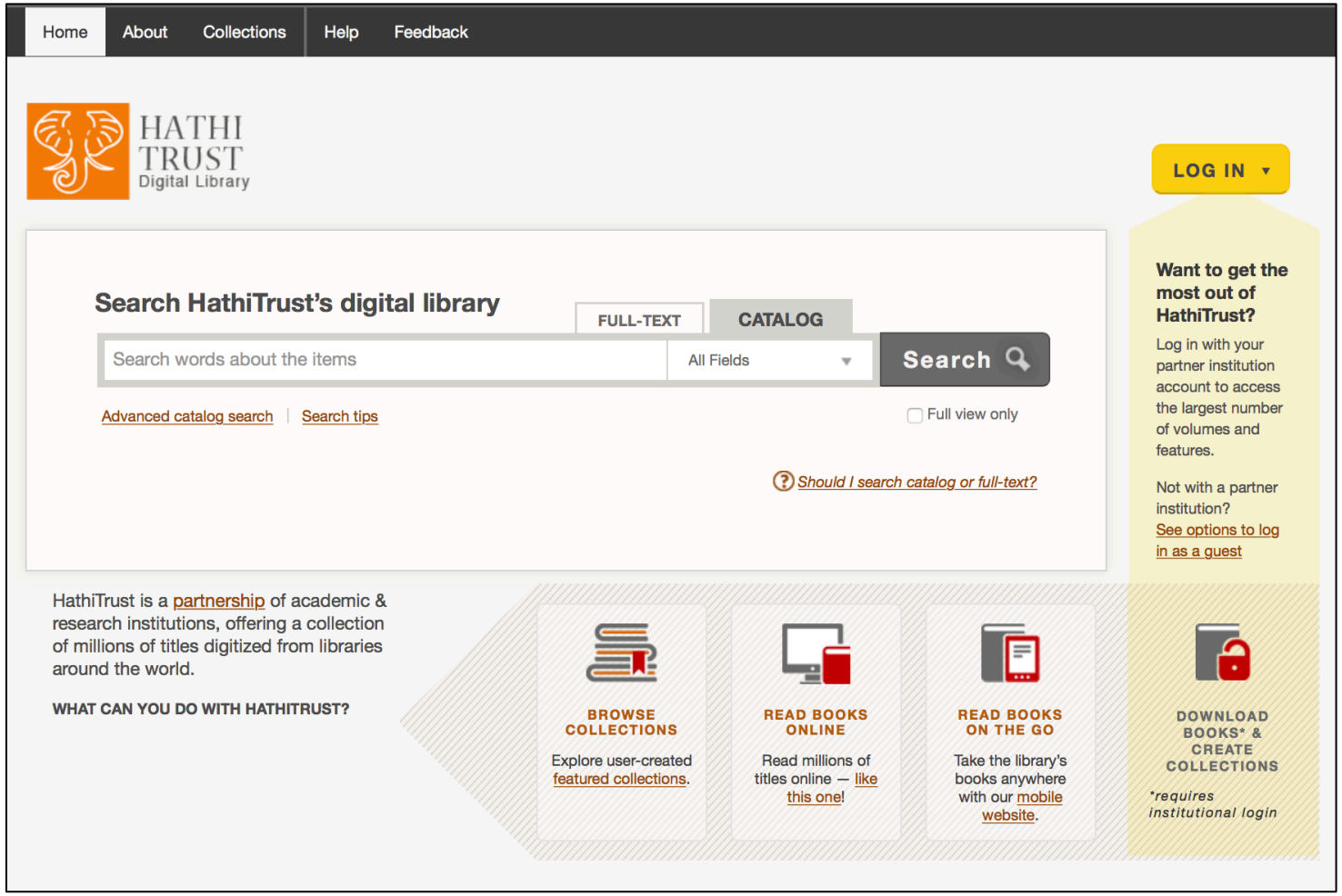
In this activity, you will log in to HTDL and create a collection containing volumes of the public papers of the presidents of the United States, and import it into HTRC Analytics as a workset.

Follow the instructions on the handout to build your workset.

Websites:

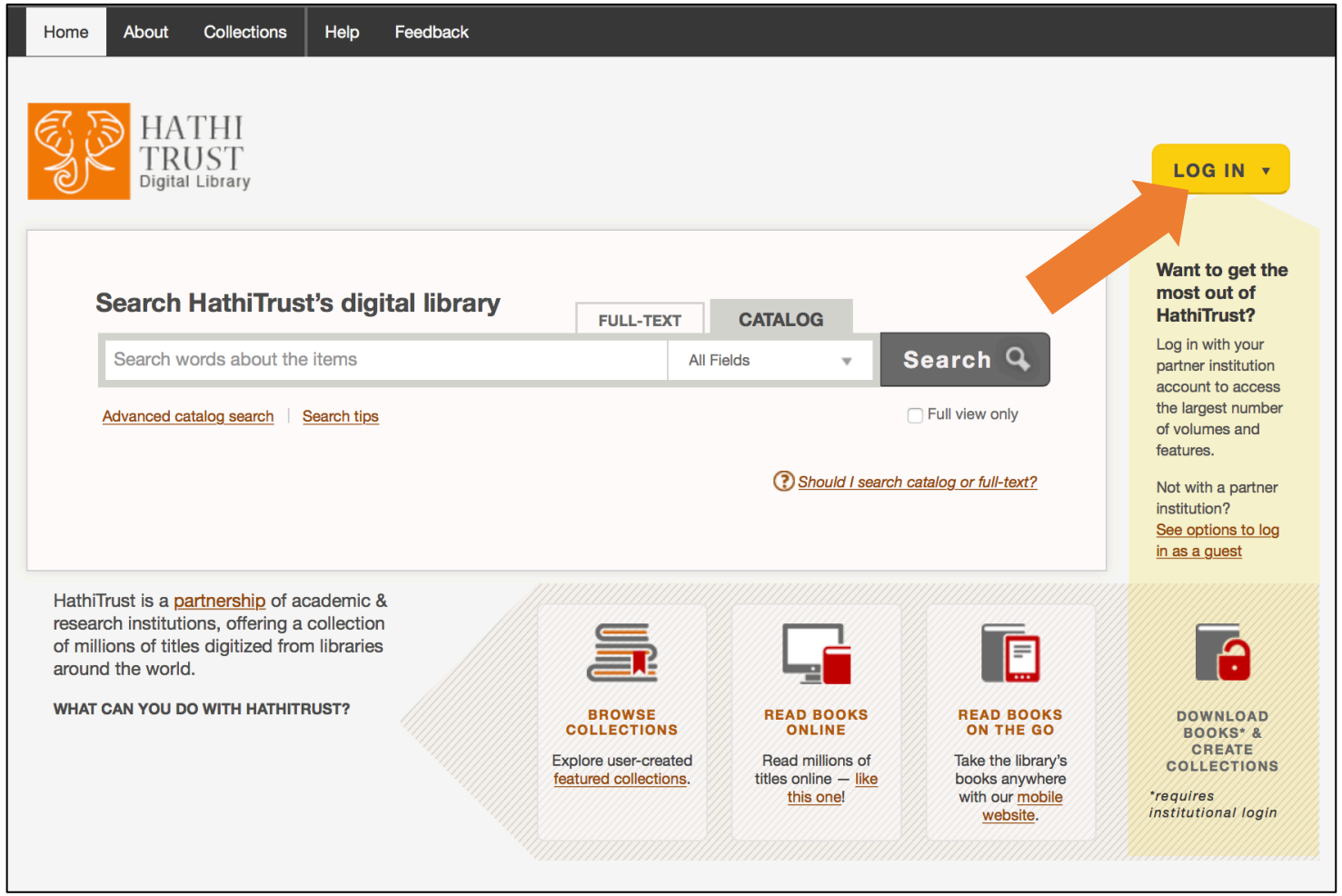
- HTDL: https://www.hathitrust.org
- HTRC Analytics: https://analytics.hathitrust.org

Go to HTDL interface

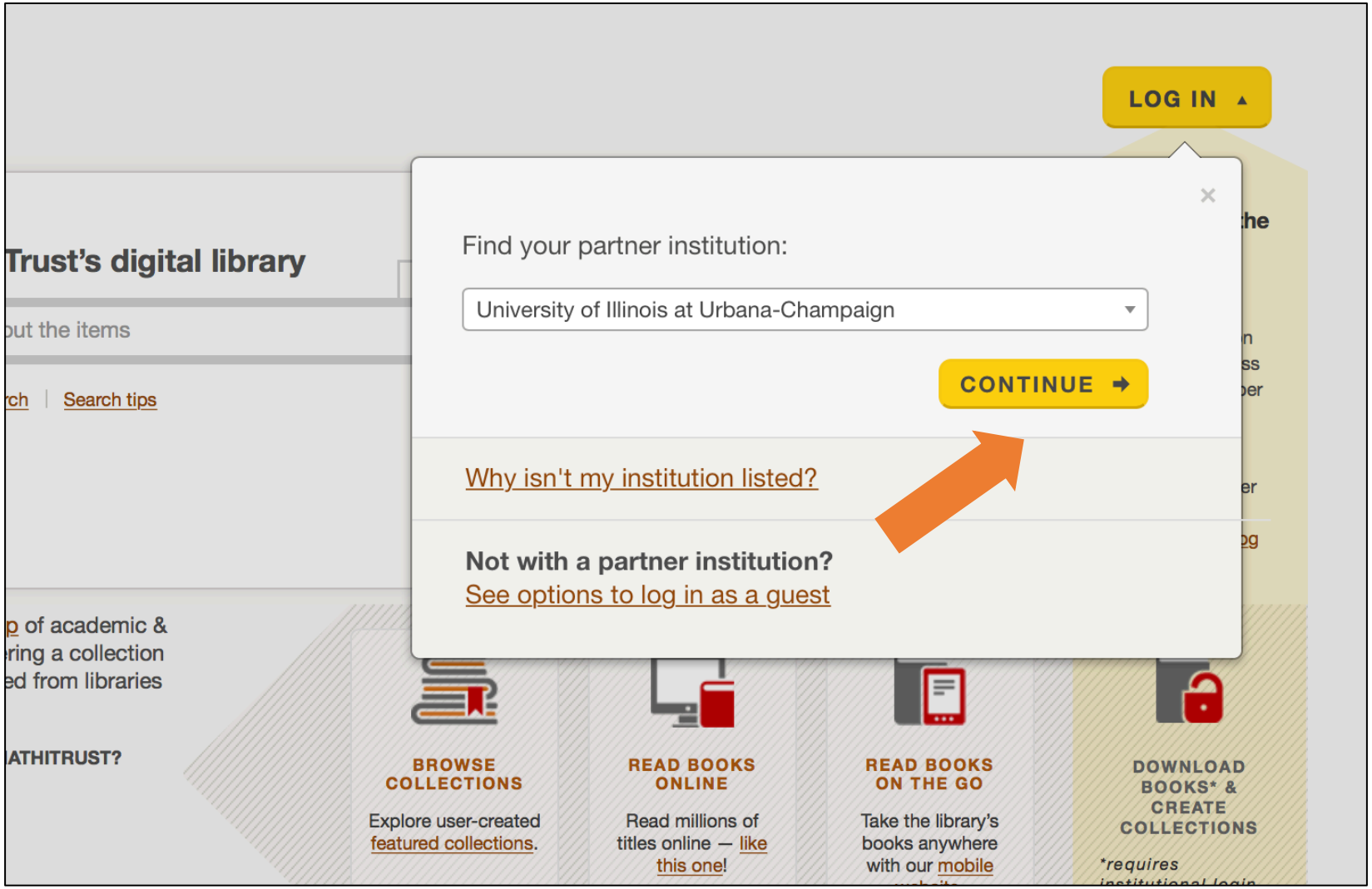


https://www.hathitrust.org

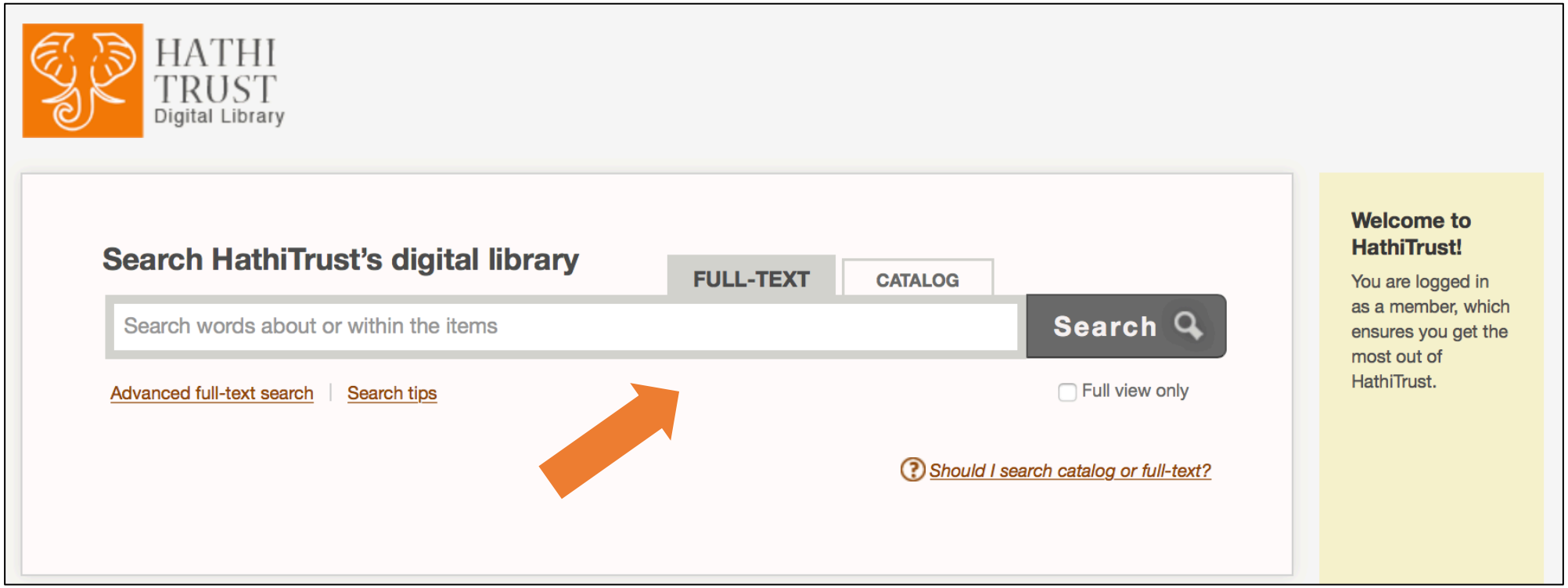
Log in


Home
About
Collections
Help
Feedback
HATHI TRUST Digital Library
LOG IN
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

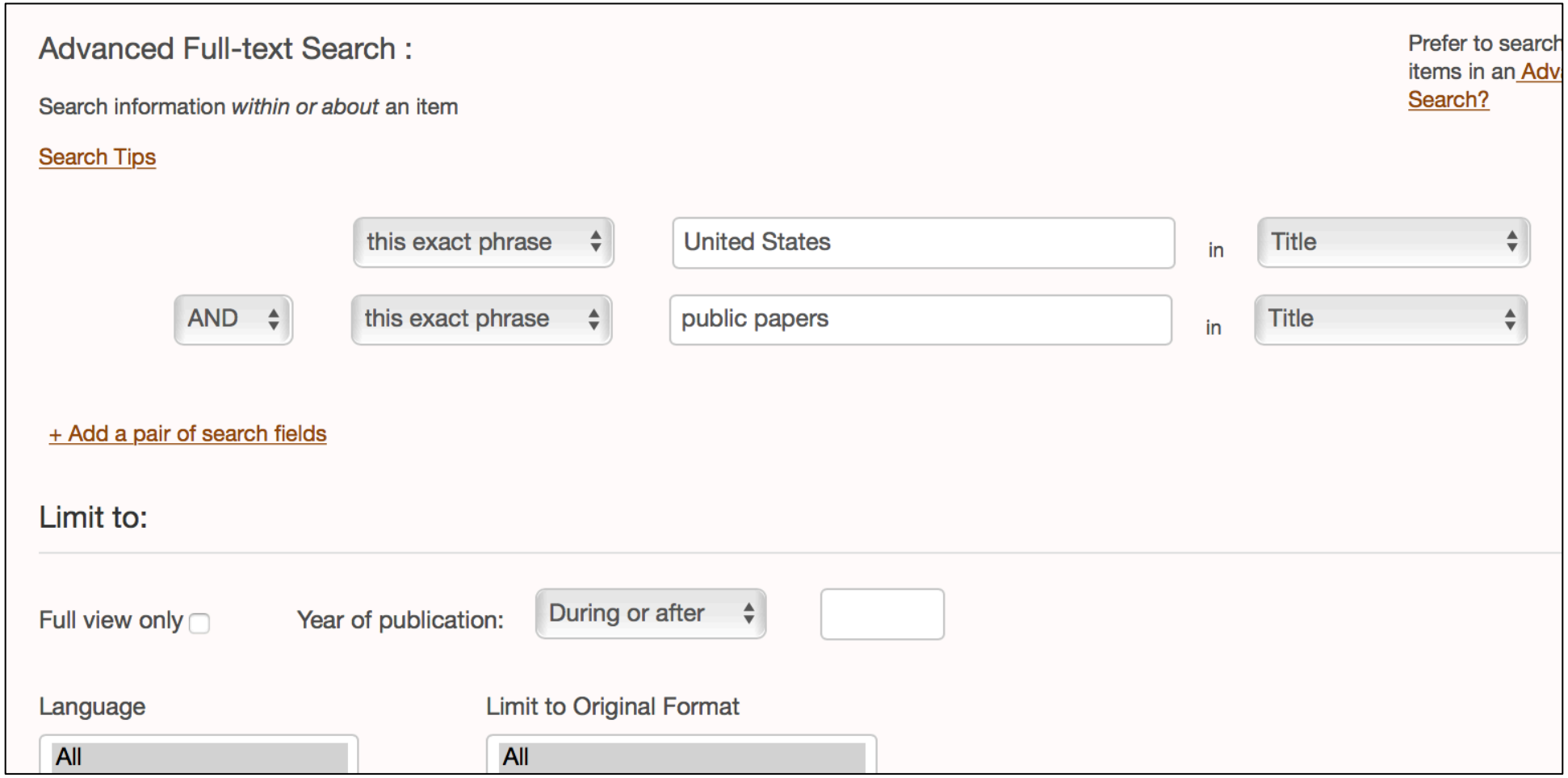

LOG IN
Find your partner institution:
University of Illinois at Urbana-Champaign
CONTINUE
Why isn't my institution listed?
Not with a partner institution?
See options to log in as a guest
Trust's digital library
Search tips
BROWSE COLLECTIONS
Explore user-created featured collections.
READ BOOKS ONLINE
Read millions of titles online — like this one!
READ BOOKS ON THE GO
Take the library's books anywhere with our mobile
DOWNLOAD BOOKS* & CREATE COLLECTIONS
*requires

Search for volumes



- Click on "Advance full-text search"

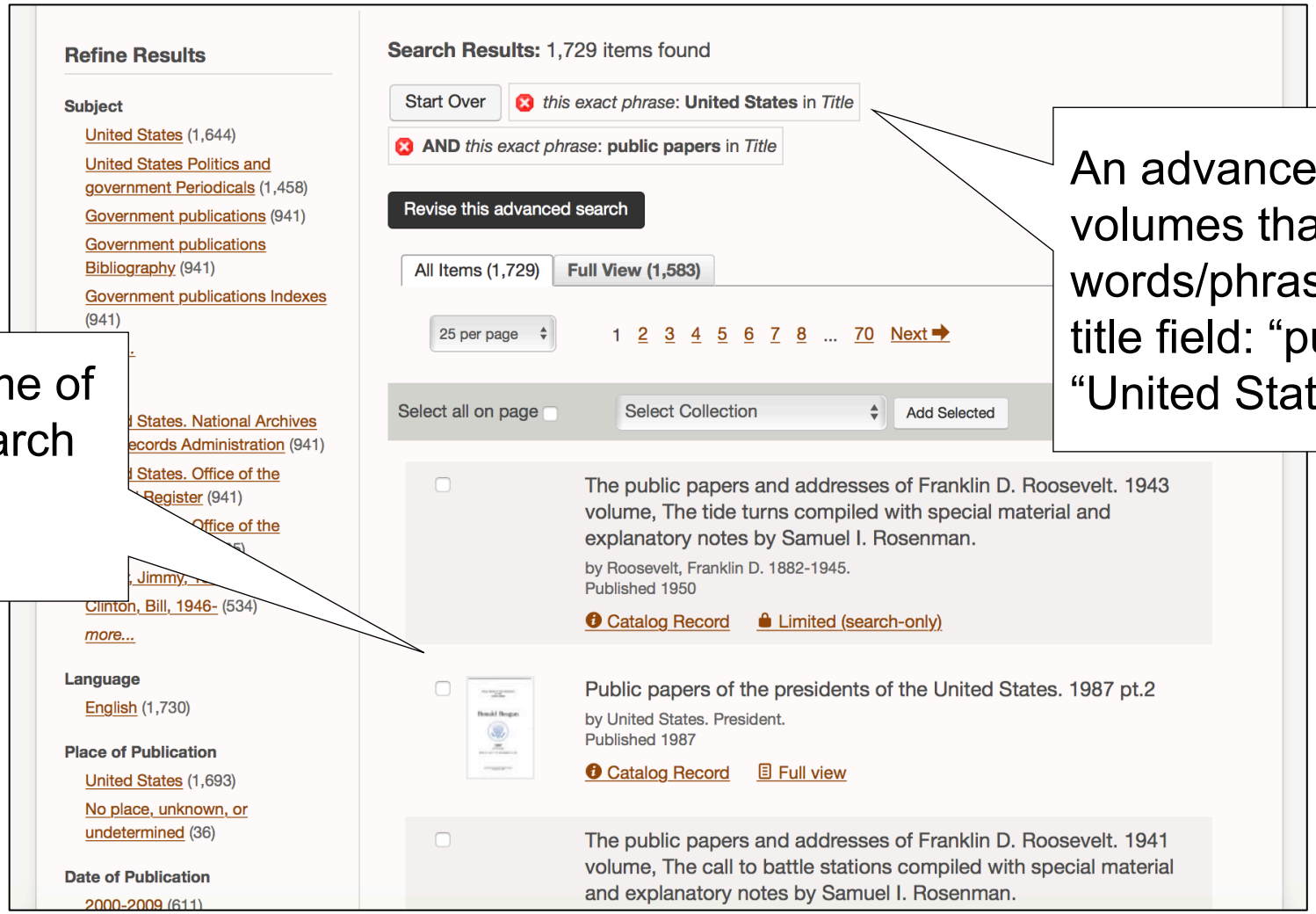
Search for volumes



Filter results and select volumes

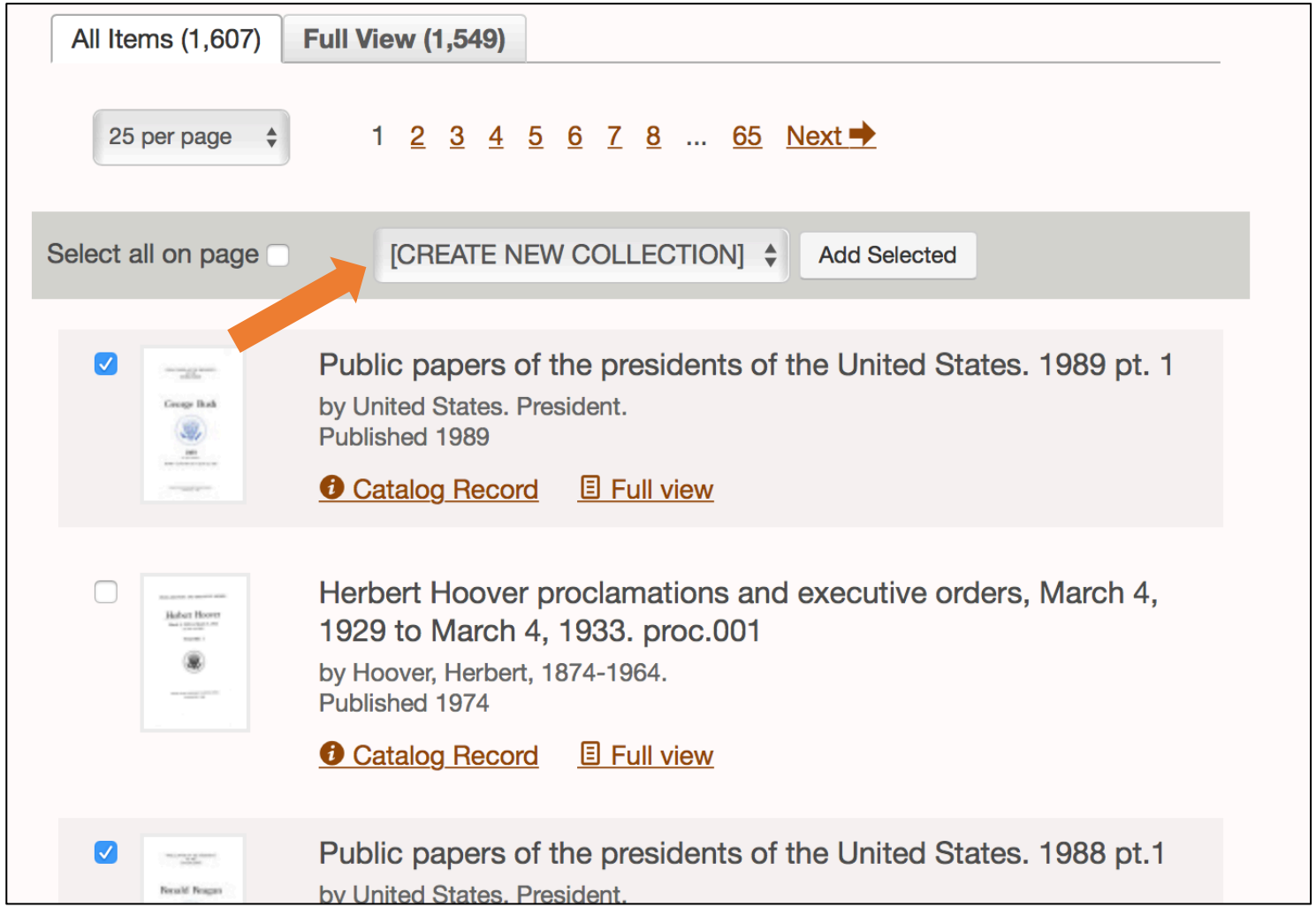
Filter results on the left sidebar

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



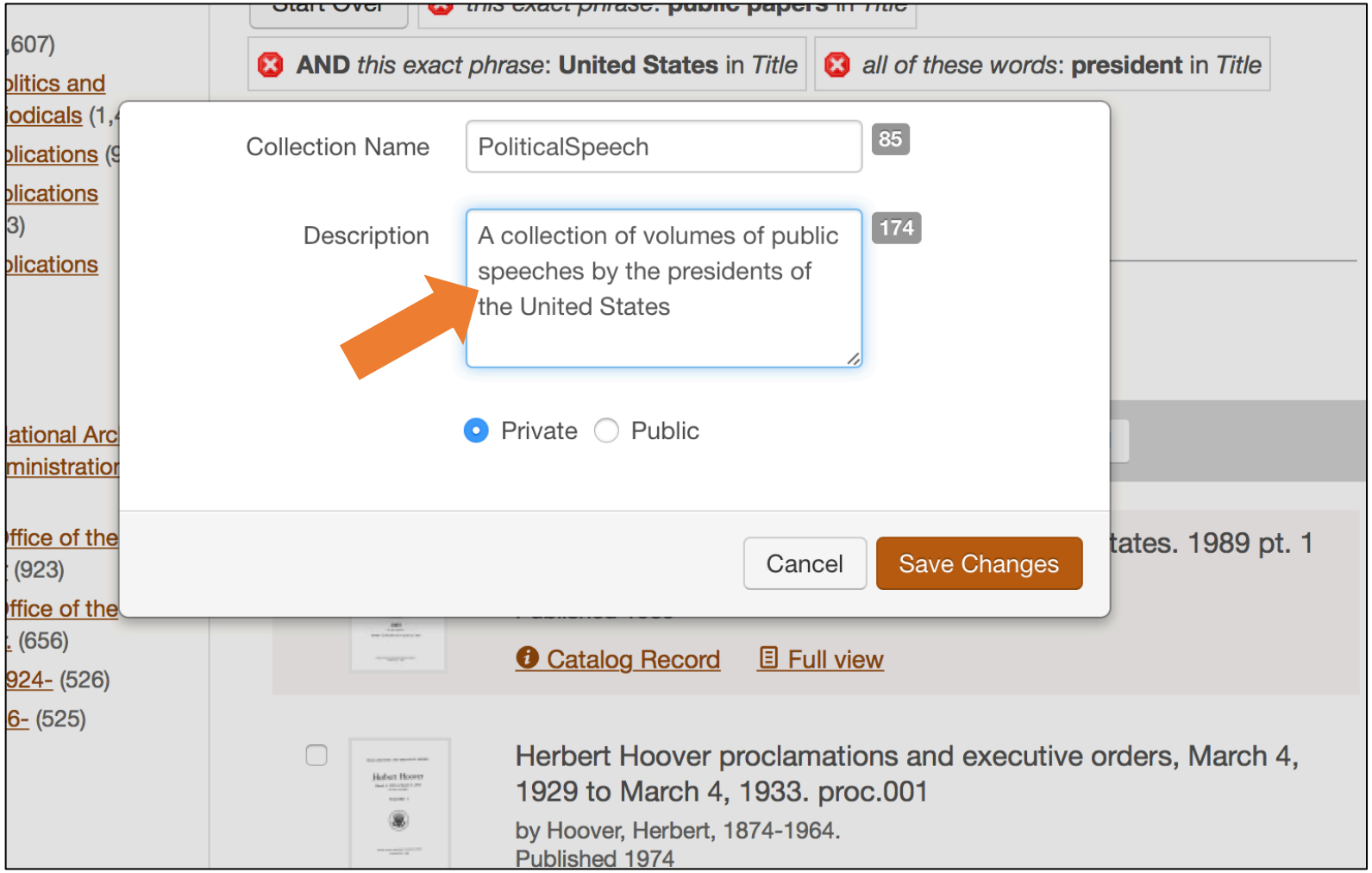
An advanced search for volumes that contain all the words/phrases below in the title field: “public papers” and “United States”

Add volumes to collection

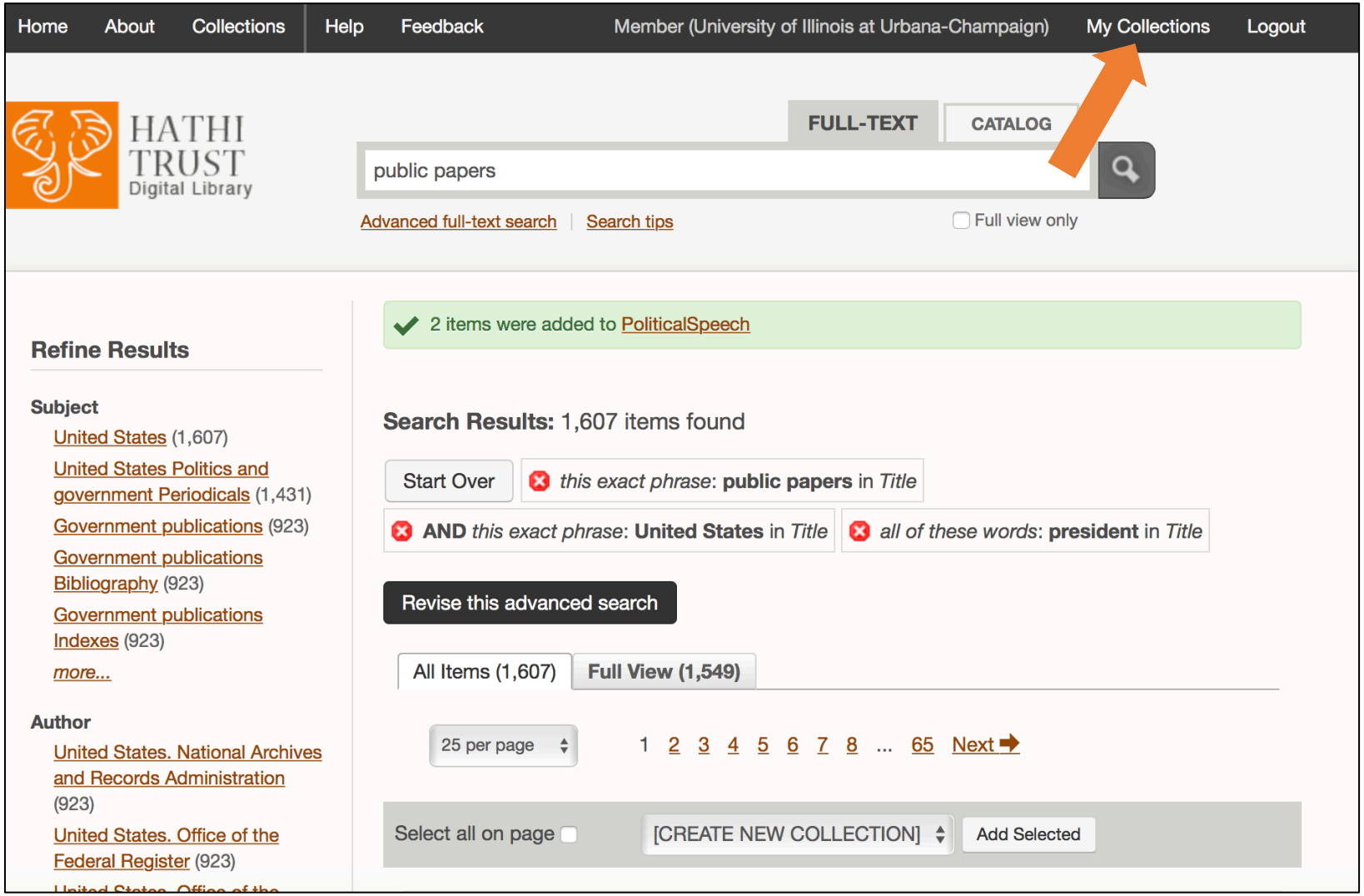


Once texts are selected, click "Select Collection" → choose "[CREATE NEW COLLECTION]" →click "Add Selected"

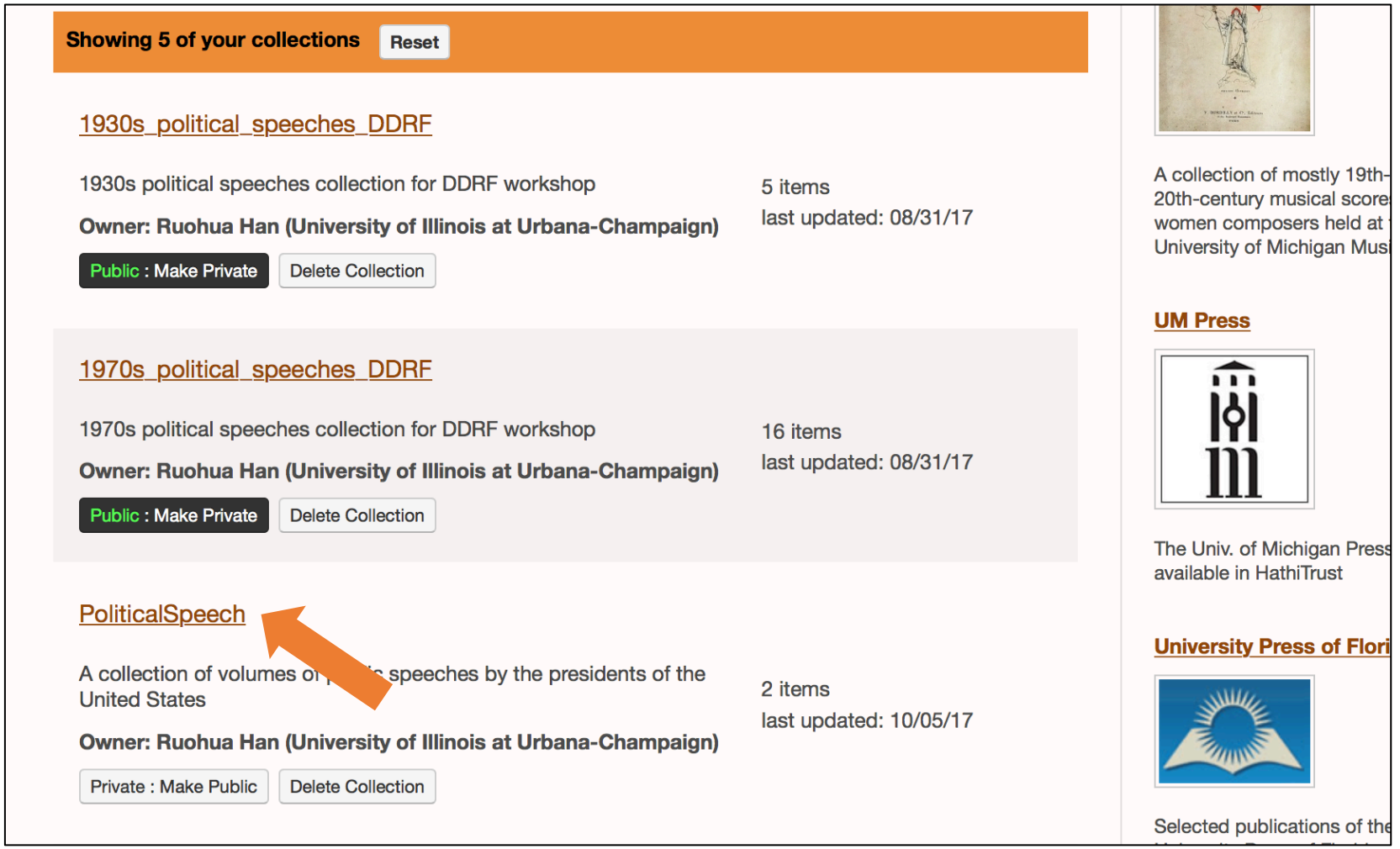
Add collection metadata



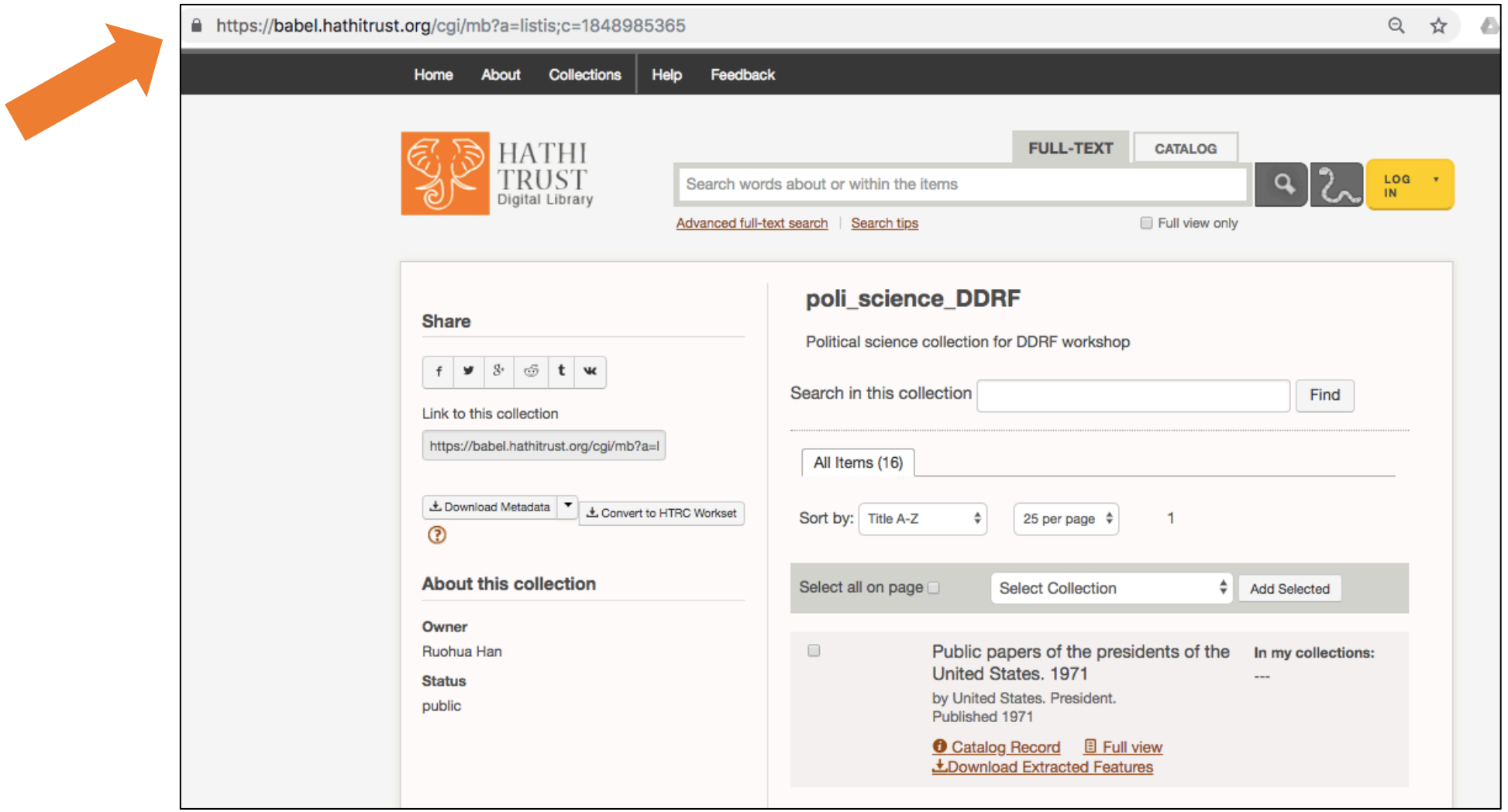
View your collection



View your collection



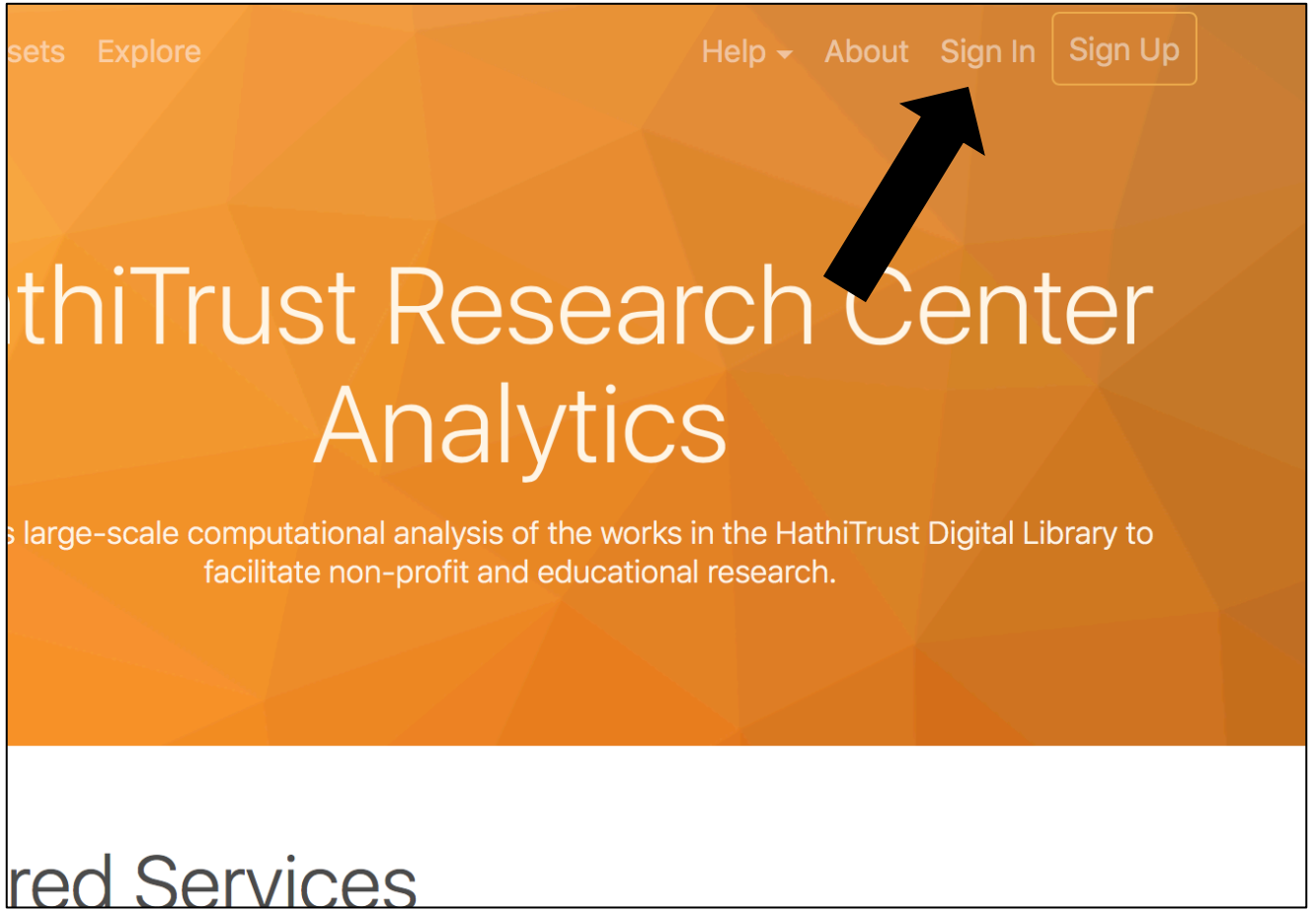
Grab the collection URL



Go to HTRC Analytics

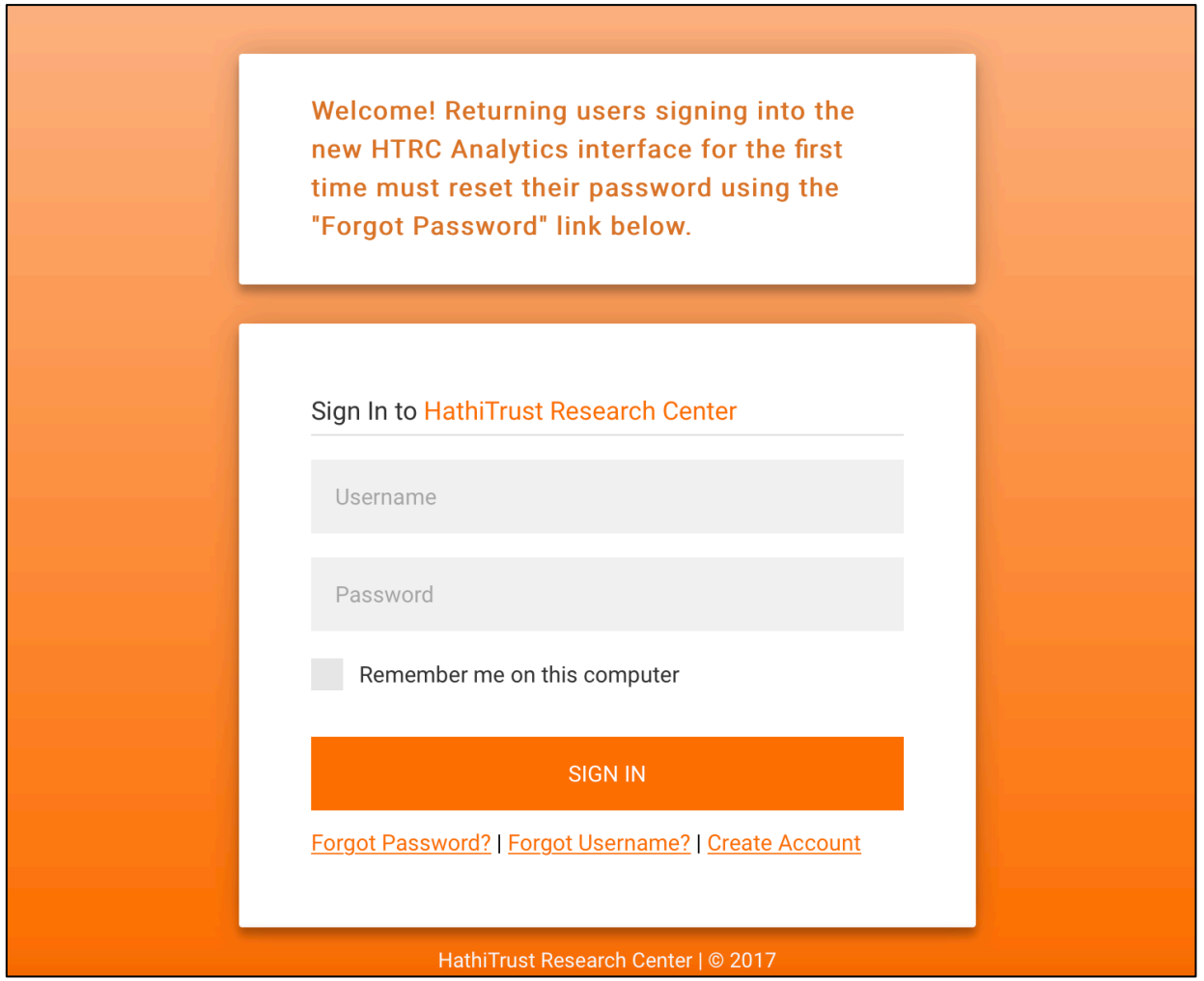
https://analytics.hathitrust.org

Sign in / Sign up

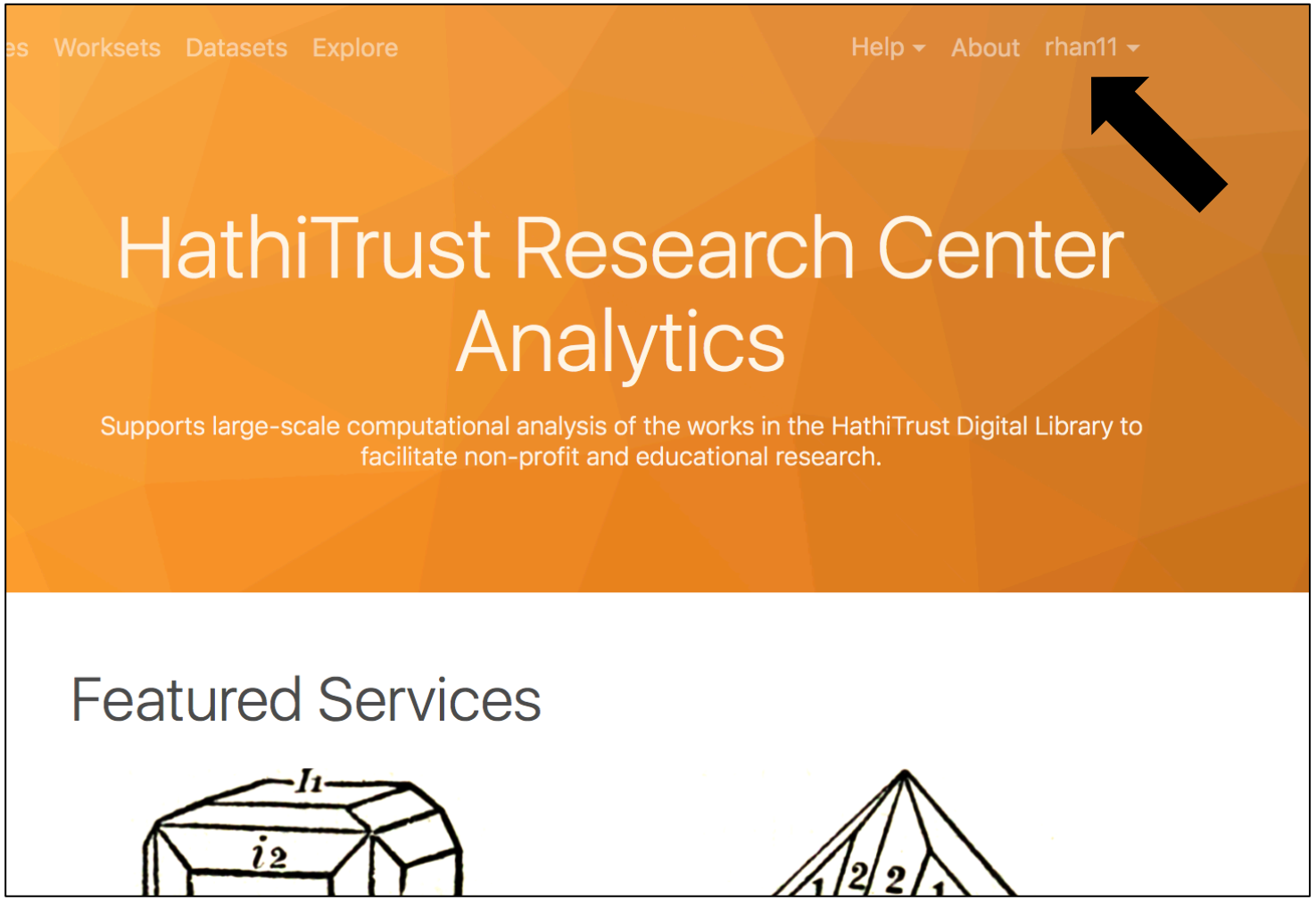


https://analytics.hathitrust.org

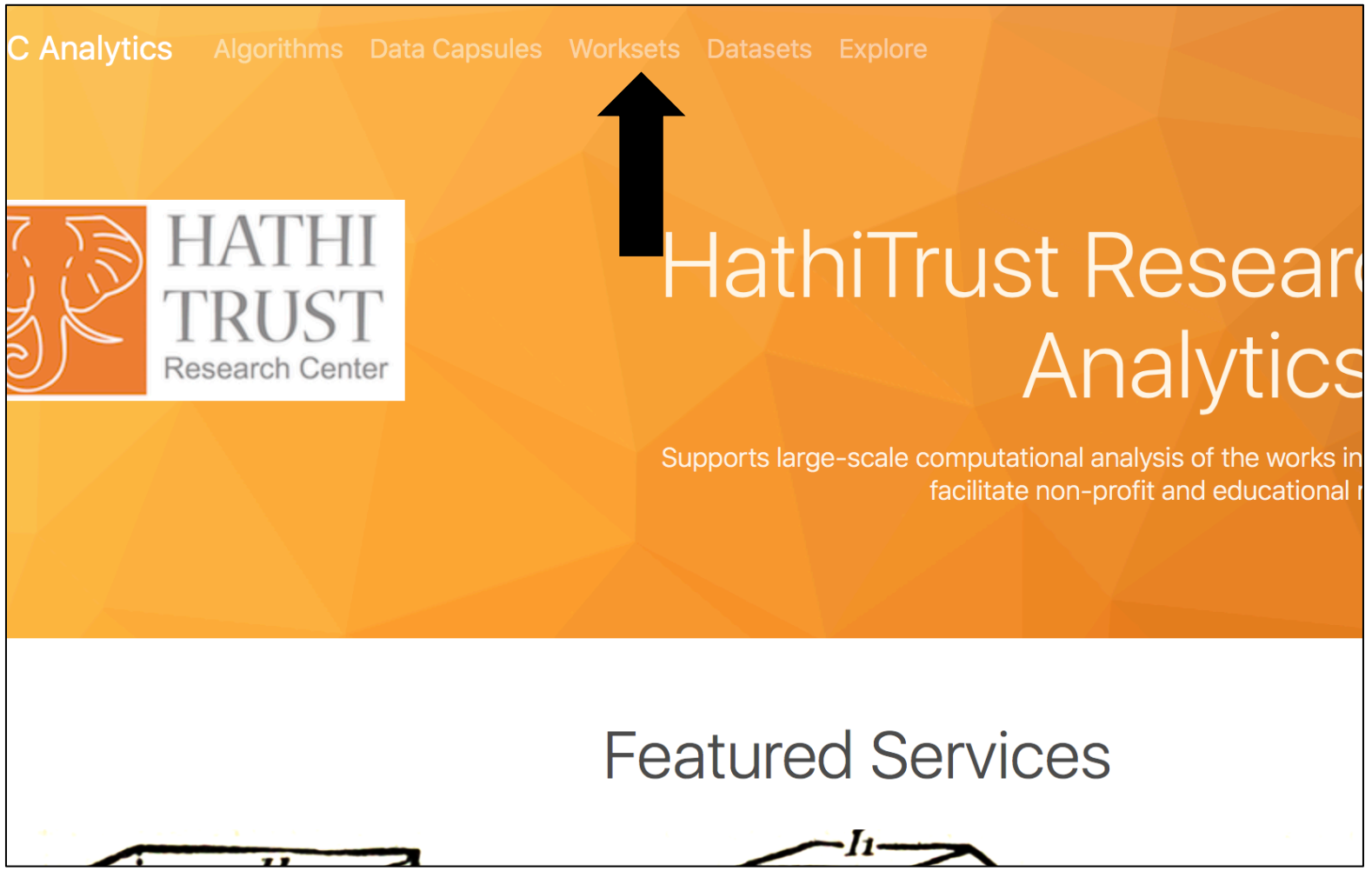
Sign in


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

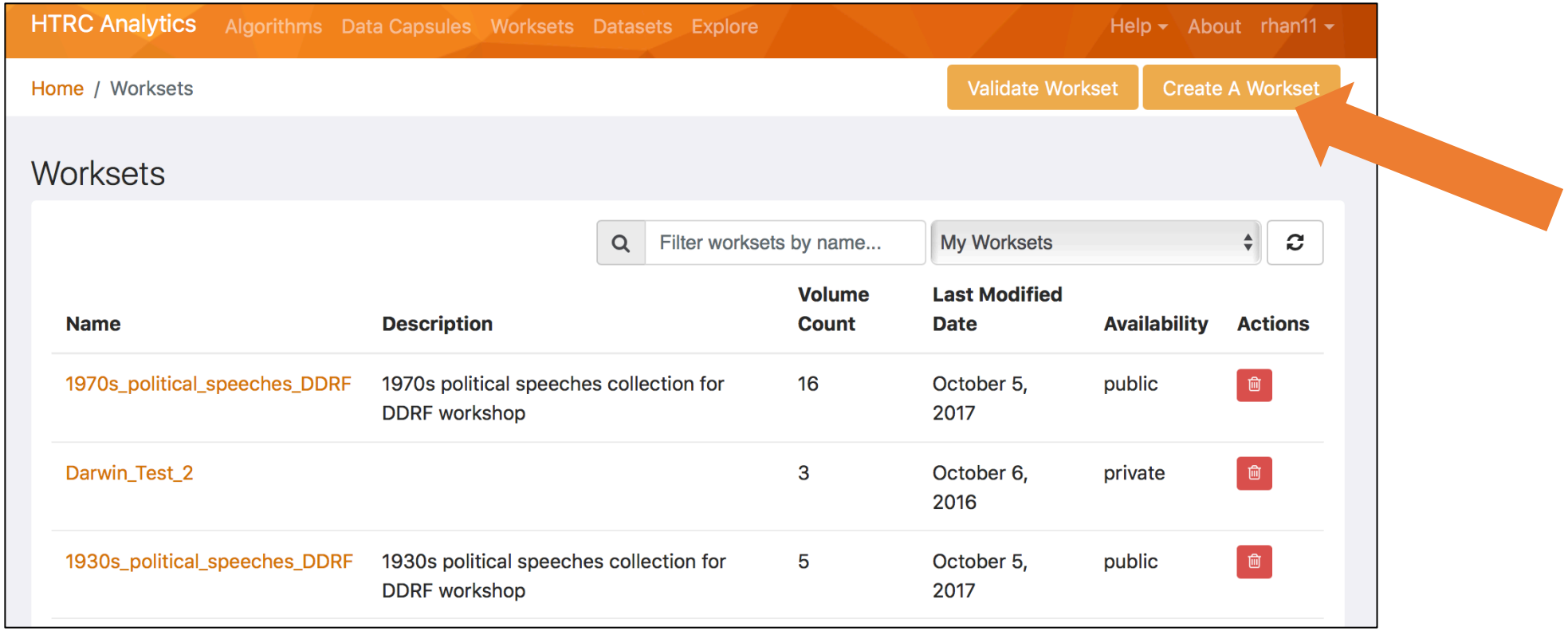
Sign in


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

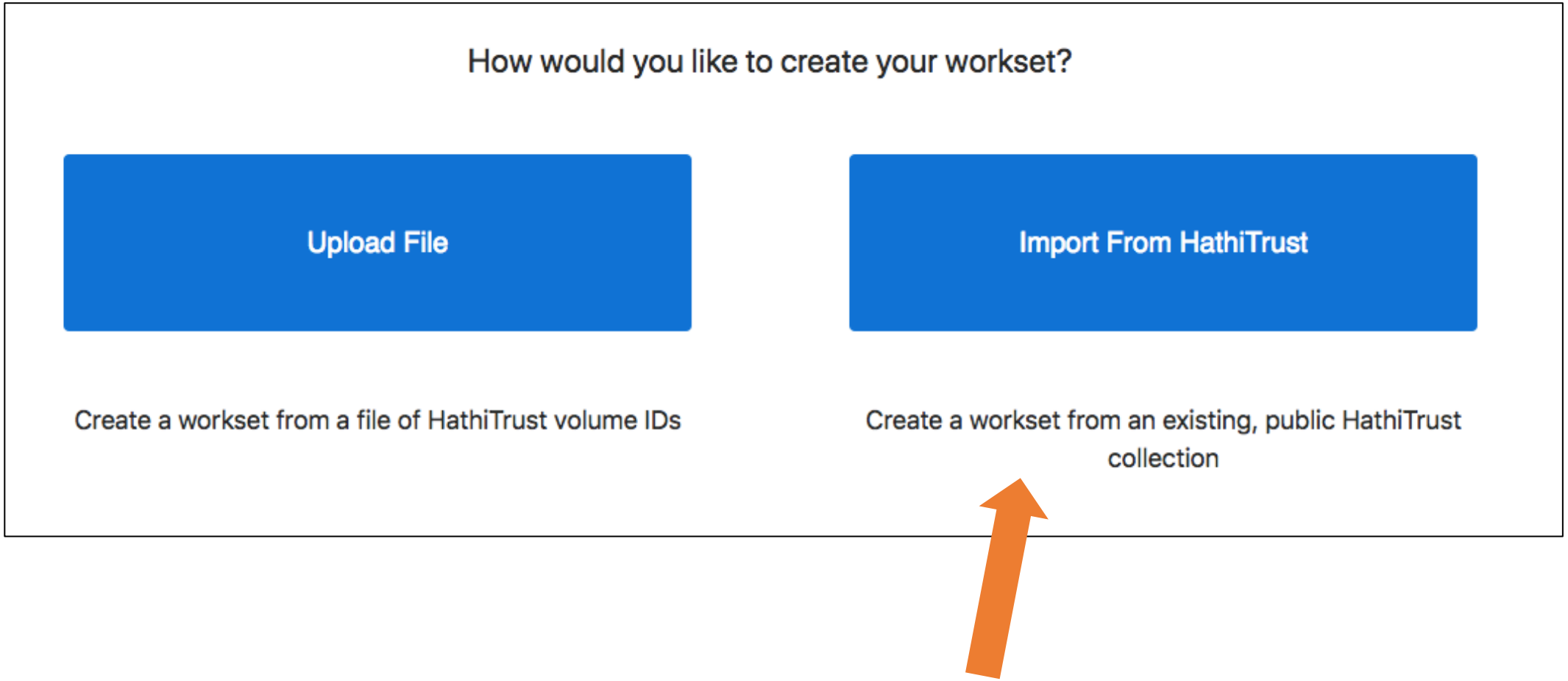
Go to Worksets page



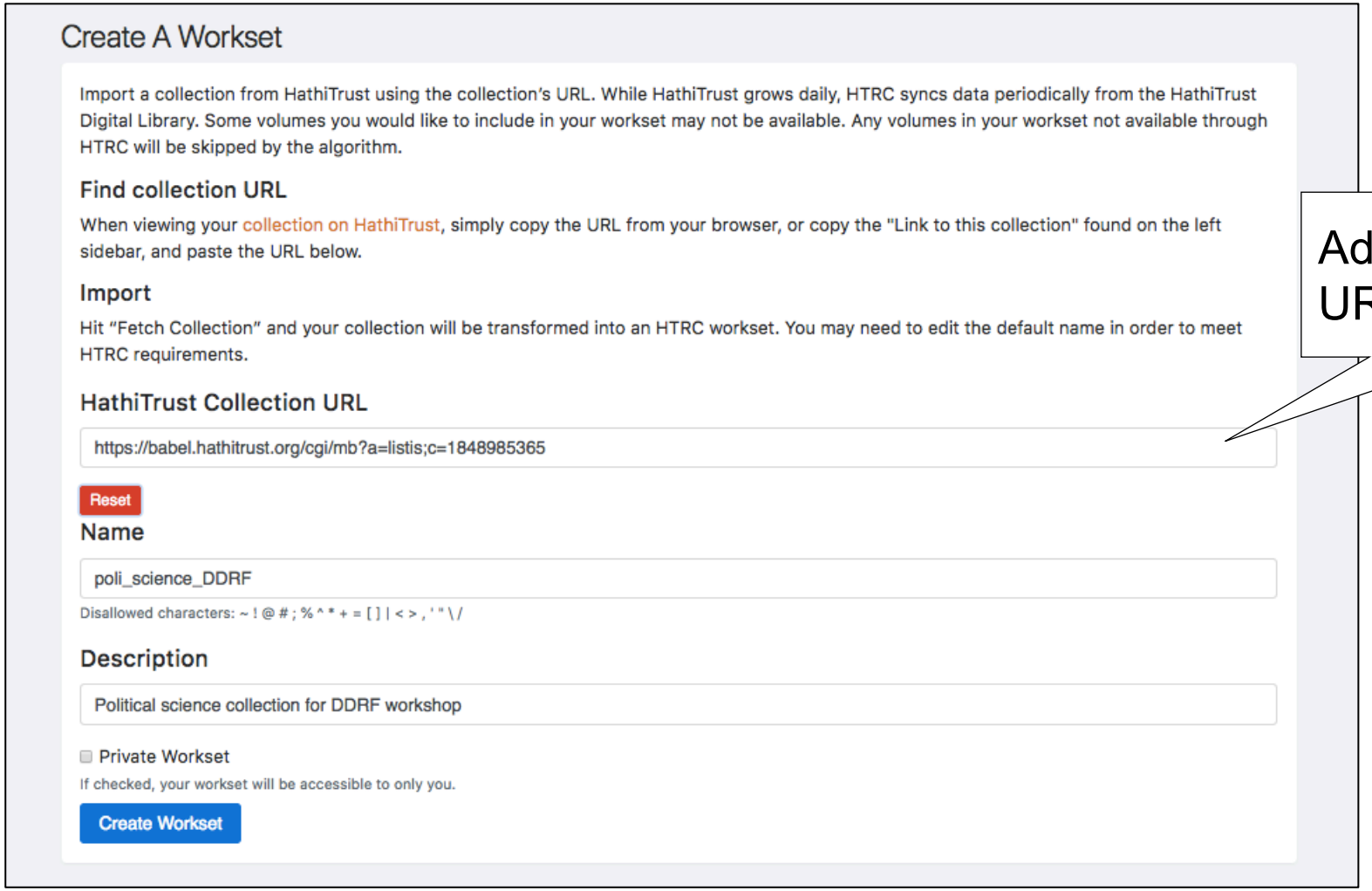
Choose to create a workset



Choose creation method

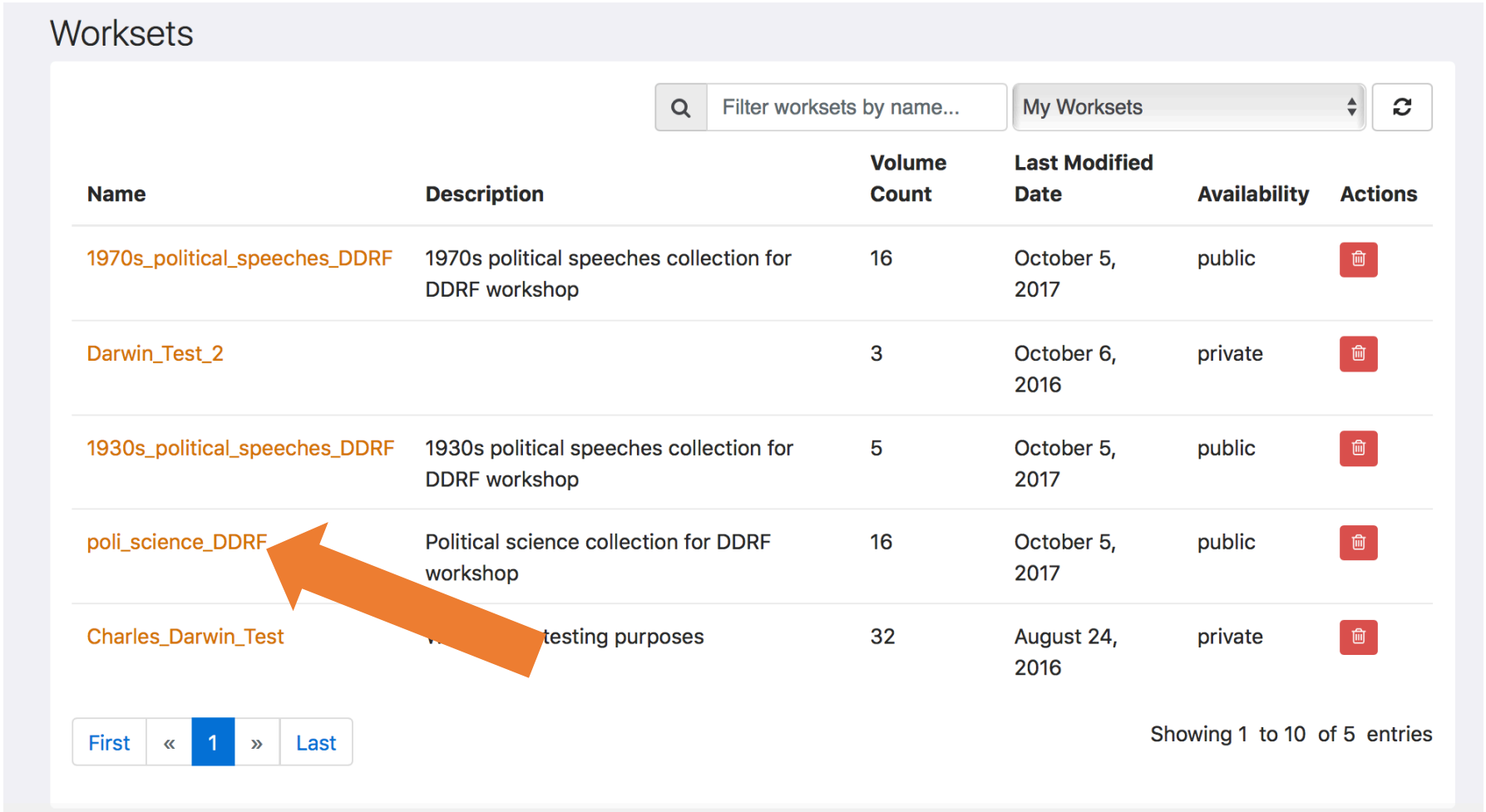


Input workset information



Add collection URL here

View created workset



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/ .