HATHI TRUST
RESEARCH CENTER

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

RESEARCH CENTER

Introduction

RESEARCH CENTER

Project Case Study

- 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

RESEARCH CENTER

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?

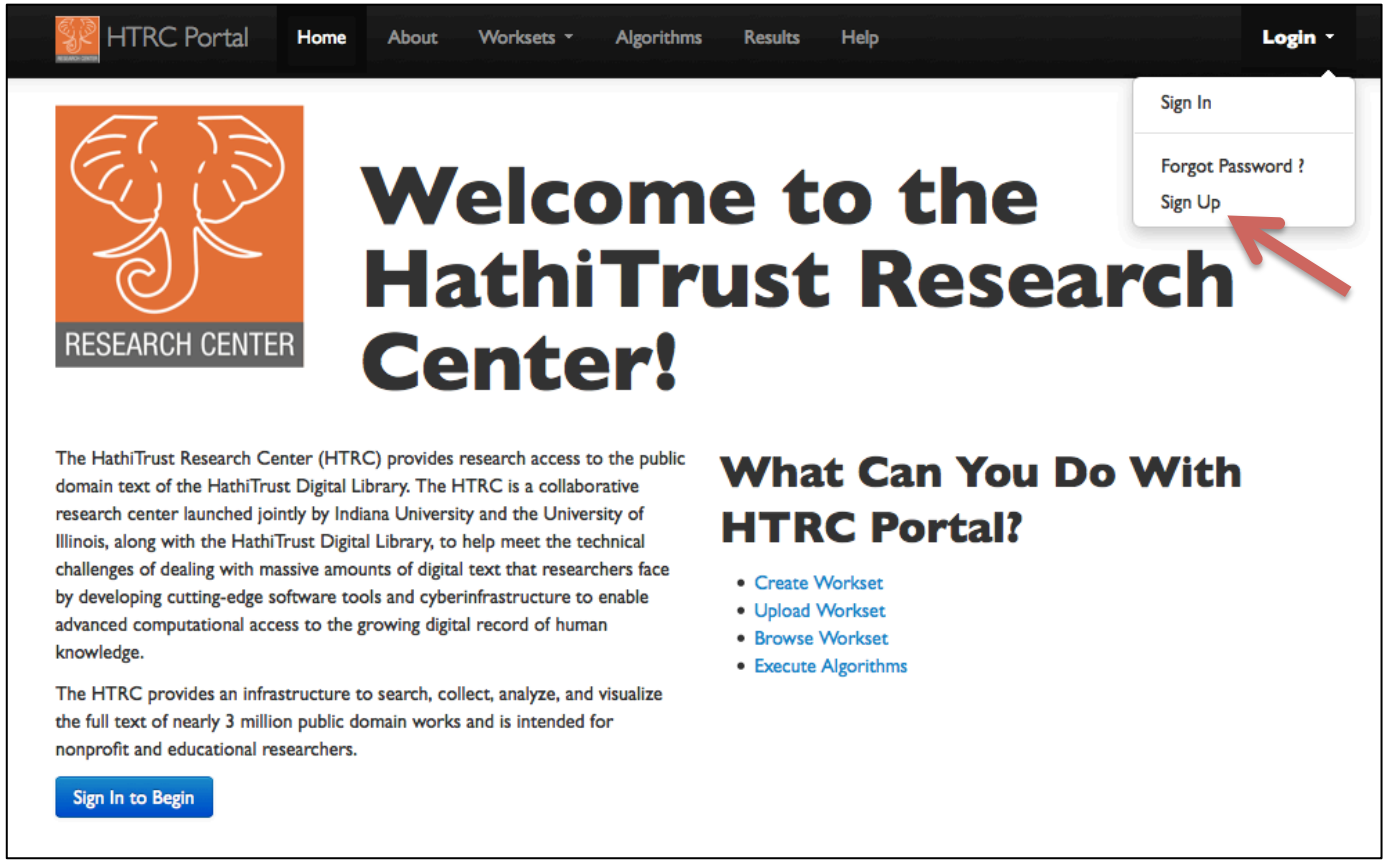
RESEARCH CENTER

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?

RESEARCH CENTER

Personal Account Creation



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

RESEARCH CENTER

Registration

User Registration (ALL fields are required)

User ID

Password

Retype-Password

First Name

Last Name

E-mail

Submit

RESEARCH CENTER

Summary of Steps for Workset Creation and Analysis

- Create a workset
- Select algorithm
  - Provide job name
  - Select workset
  - Adjust parameters
- Run
- View results

RESEARCH CENTER

Workset Builder

https://htrc2.pti.indiana.edu/blacklight/

RESEARCH CENTER

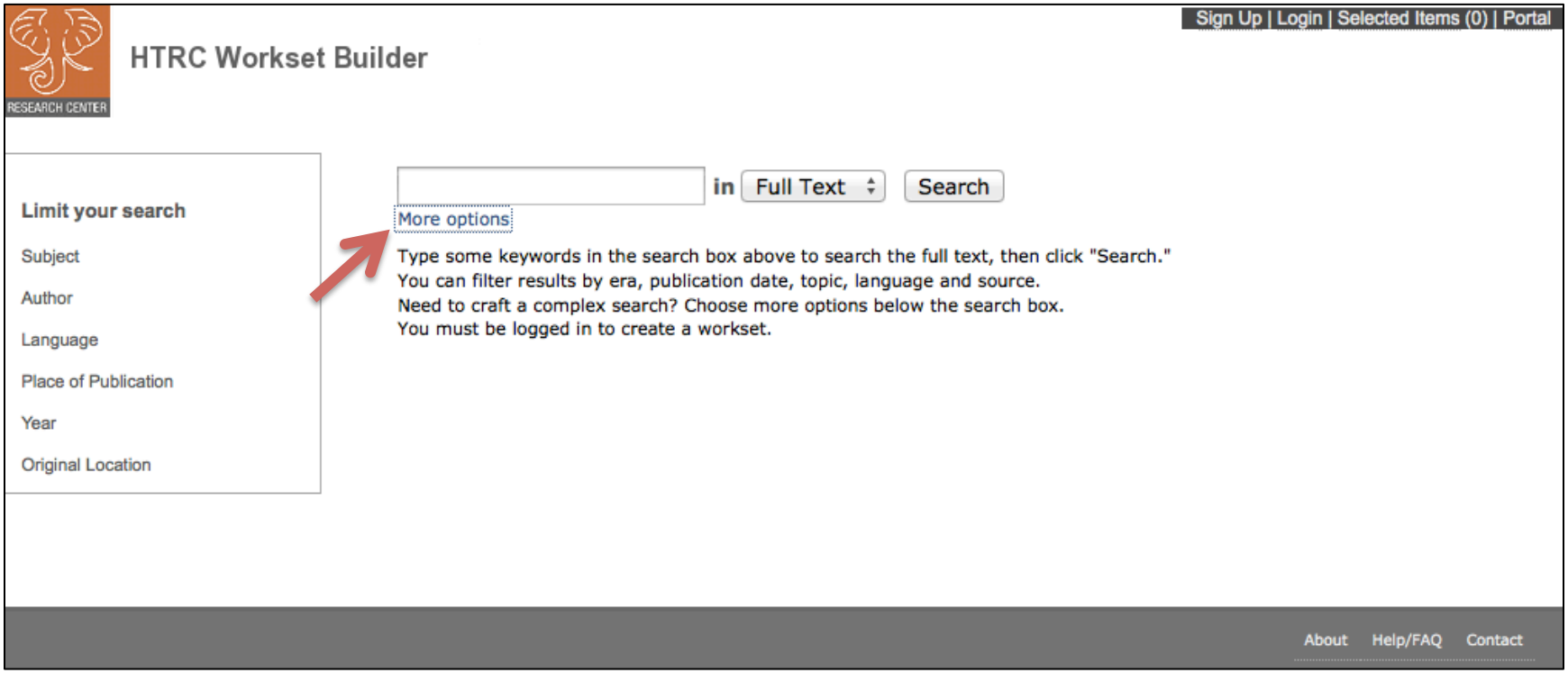
Workset Builder

- 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

RESEARCH CENTER

Workset Builder: Search

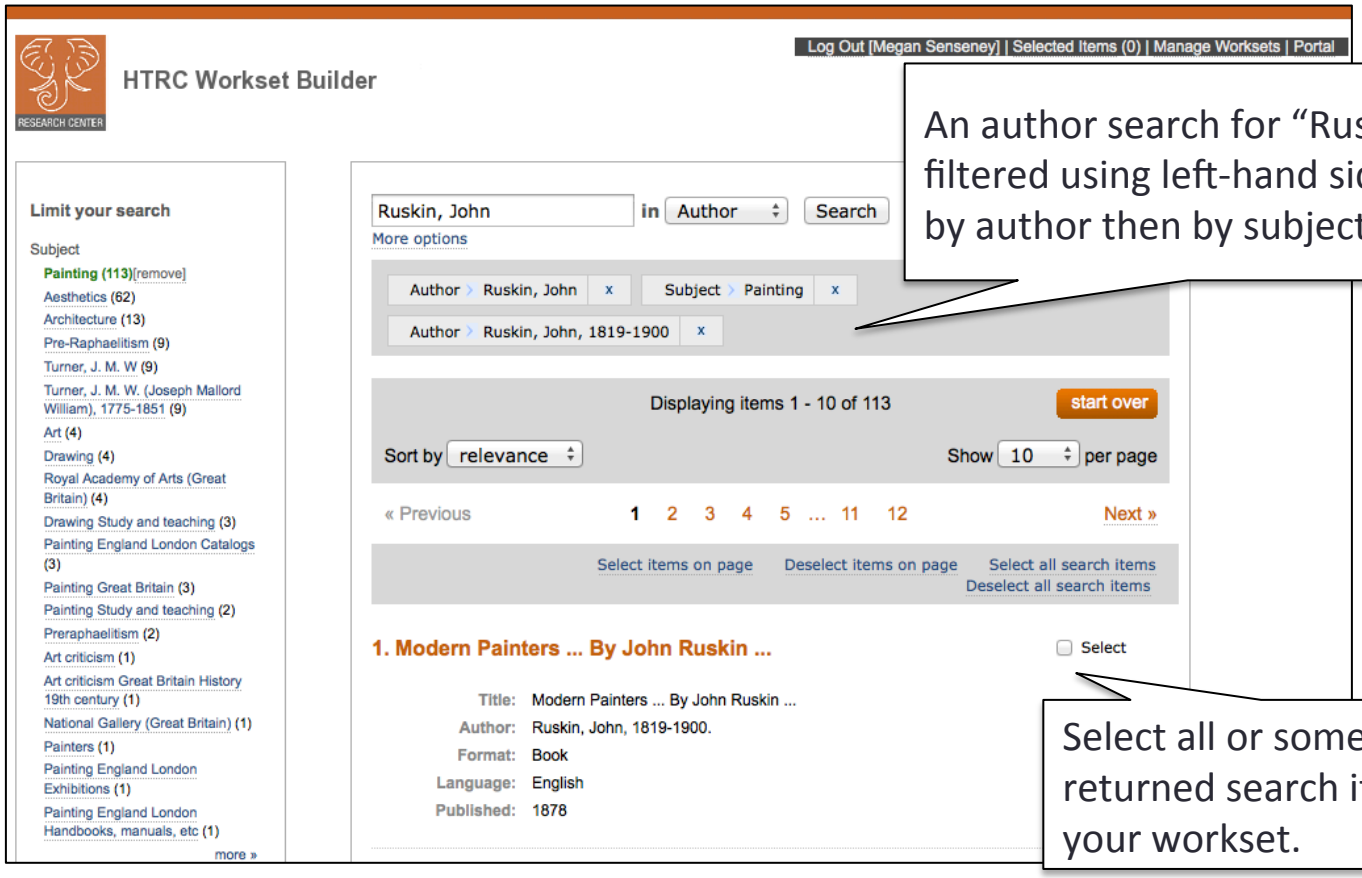


NB: Click "More options" for advanced search options

https://htrc2.pti.indiana.edu/blacklight/

RESEARCH CENTER

Workset Builder: Search Results



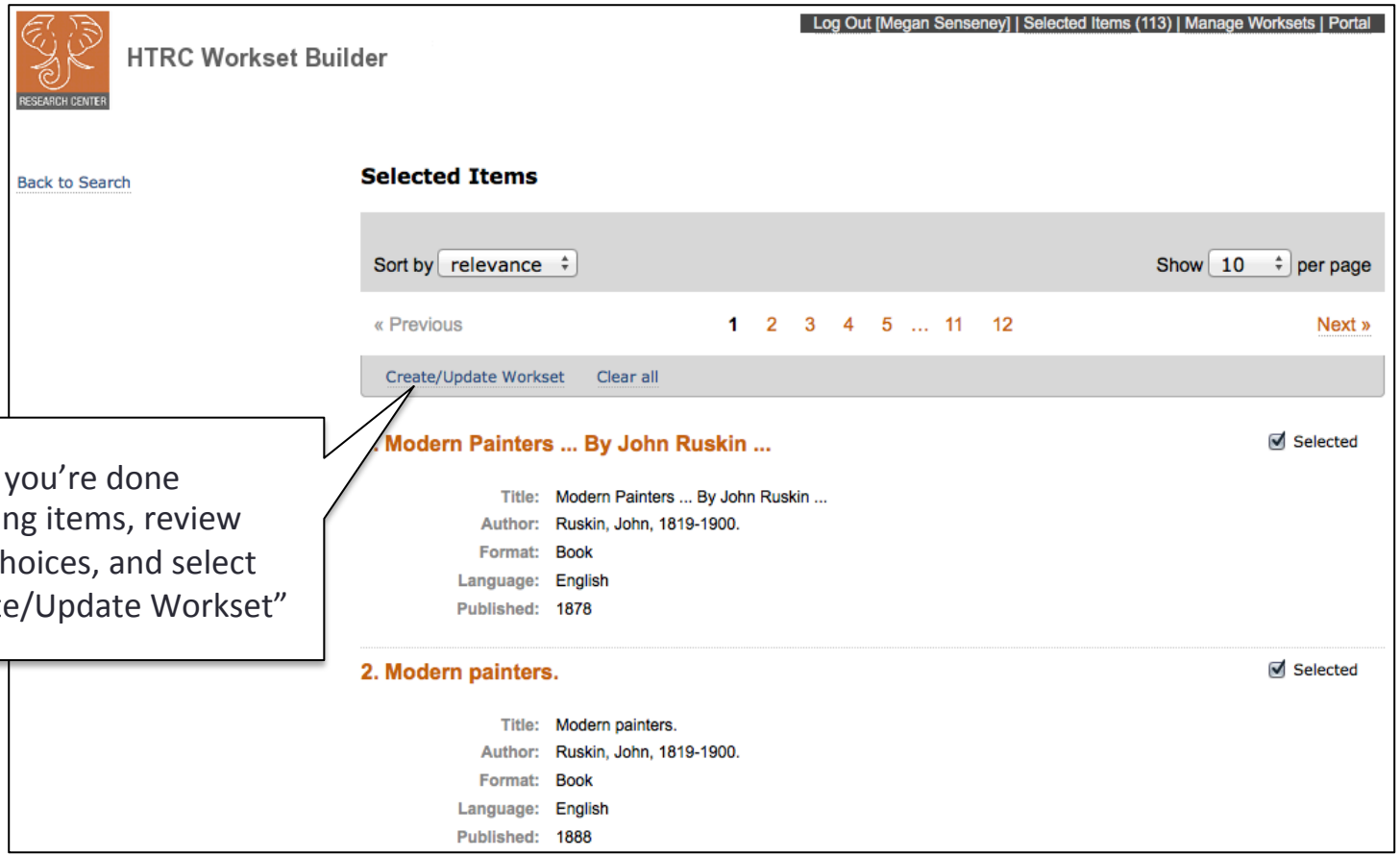
An author search for “Ruskin, John” filtered using left-hand sidebar, first by author then by subject.

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

RESEARCH CENTER

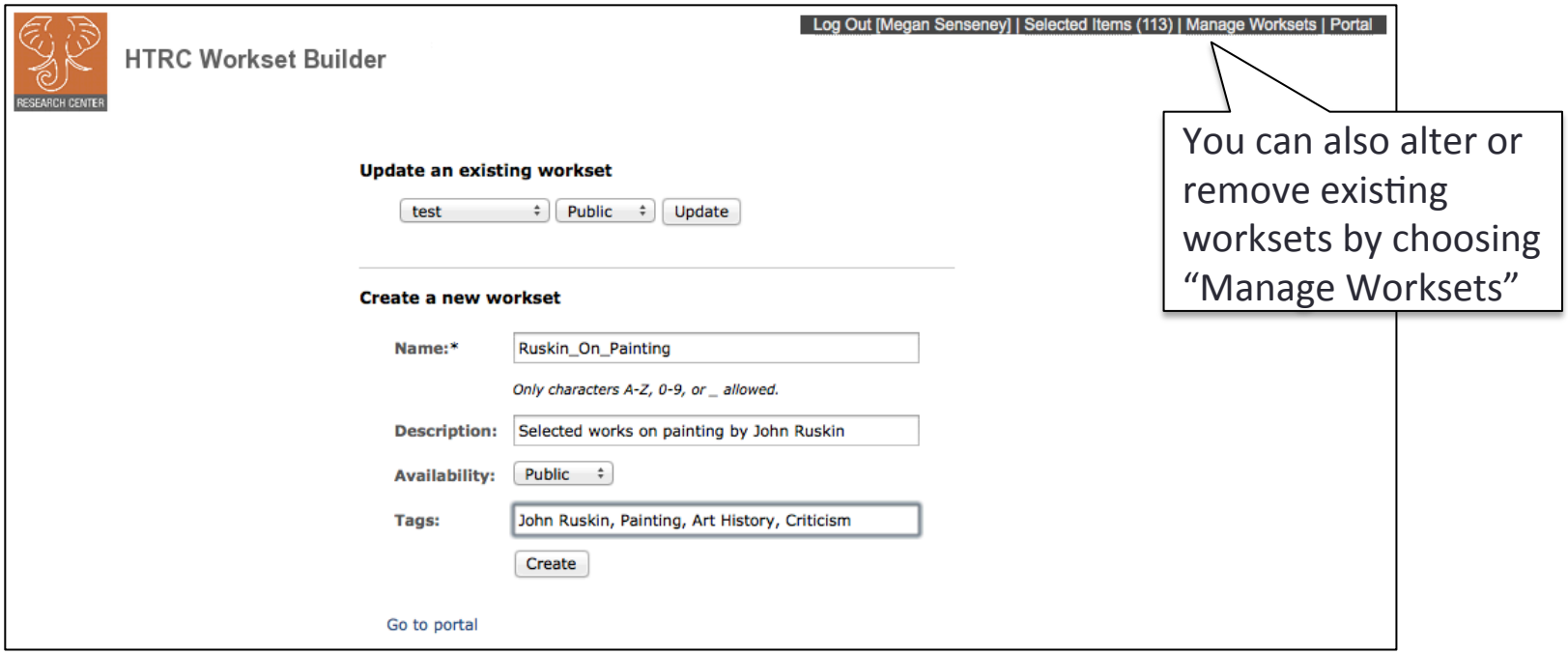
Workset Builder: Create Workset



When you're done selecting items, review your choices, and select "Create/Update Workset"

RESEARCH CENTER

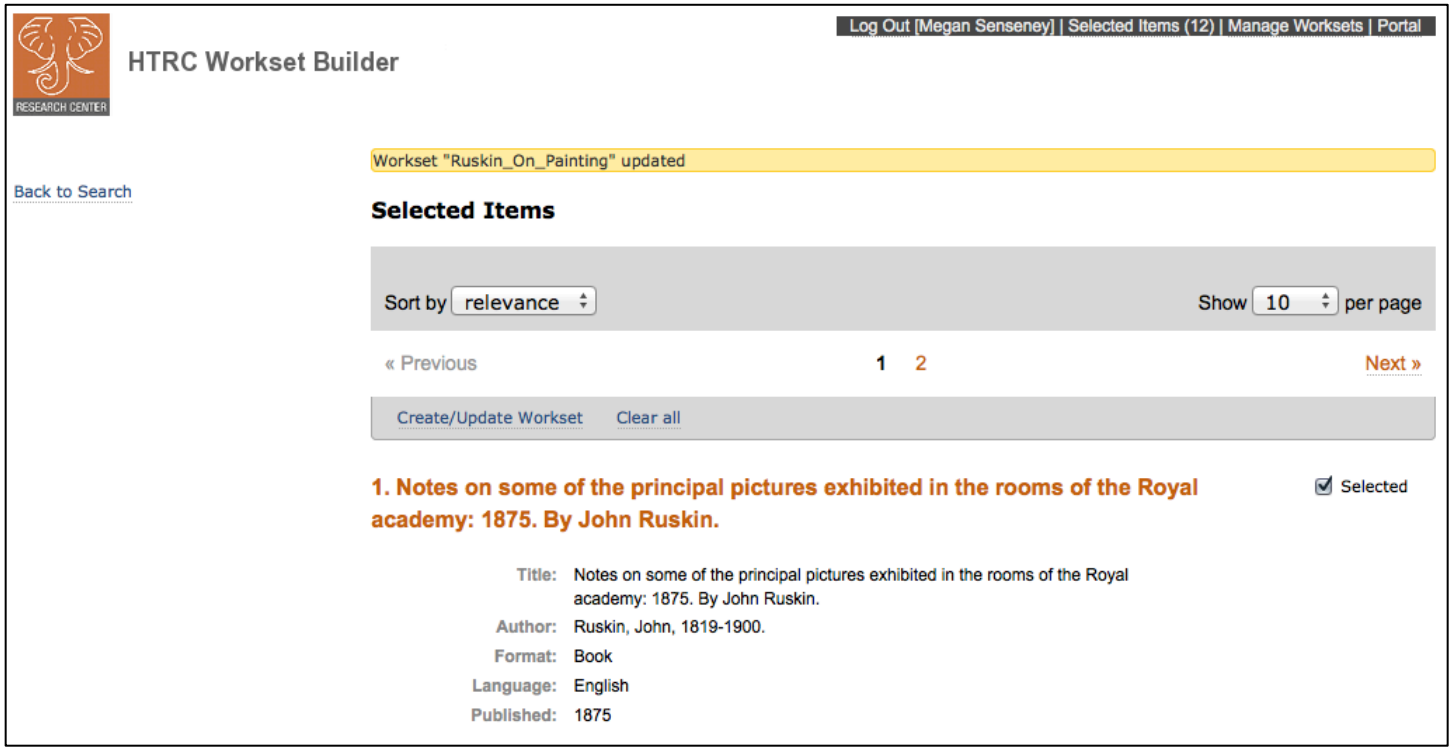
Workset Builder: Workset Metadata



NB: you have the option to add selected items to an existing workset or to create a new workset.

RESEARCH CENTER

Workset Builder: Management



Example: after creating the Ruskin workset, I went back and removed all duplicate titles, reducing workset from 113 items to 12.

RESEARCH CENTER

HTRC Portal

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

RESEARCH CENTER

Portal: Workset Menu

HTRC Portal | Home | About | Worksets | Algorithms | Results | Help | Signed-in as: mfsense2

List
Upload Workset
Create Worksets

Don't forget to sign in!

RESEARCH CENTER

Welcome to the HathiTrust Research Center!

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 Can You Do With HTRC Portal?

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

Click "List" in the Workset menu to review your workset in the Portal

RESEARCH CENTER

Portal: Workset Details

HTRC Portal | Home | About | Worksets | Algorithms | Results | Help | Signed-in as: mfsense2

Available Worksets +

- Ruskin_On_Painting
- 1darwin-test
- 2darwin-english
- 2vesalius
- Agrippa
- Anarchism
- AncientGreek
- Austen_Dickens_Labels
- Author_Twain
- BestCoreComplex
- BigLaw
- Bleakhouse
- Cicero_Orations_Letters
- Coffee_Books
- Cornell_HIS_2293_1
- Dickens_as_Authors
- Dickens_yo
- Diderot-test
- Digital_Preservation
- DocSouthMatch
- EAPoeAsAuthor
- ECCOmatch
- EEBOmatch
- Early_English_Books

Workset Details

Name:	Ruskin_On_Painting
Description:	Selected works on painting by John Ruskin
Author:	mfsense2
Last Modified By:	mfsense2
Last Modified Time:	2014-06-06T14:59:39-04:00
Number of Volumes:	12

Edit Workset | Download CSV File

First | Prev | Next | Last

1. Modern painters. By a graduate of Oxford.

Volume Id:	hvd.hw2hv9
Author:	Ruskin, John, 1819-1900 M
Page Count:	414
Word Count:	154101

2. Notes on some of the principal pictures exhibited in the rooms of The Royal Academy, and The Society of Painters in Water Colours, no. 2, 1856 ...

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.

RESEARCH CENTER

Portal: Algorithms

HTRC Portal | Home | About | Worksets | Algorithms | Results | Help | Signed-in as: mfsense2

Available Algorithms

- Marc_Downloader
- Meandre_Classification_NaiveBayes
- Meandre_Dunning_LogLikelihood_to_Tagcloud
- Meandre_OpenNLP_Date_Entities_To_Simile
- Meandre_OpenNLP_Entities_List
- Meandre_Spellcheck_Report_Per_Volume
- Meandre_Tagcloud
- Meandre_Tagcloud_with_Cleaning
- Meandre_Topic_Modeling
- Simple_Deployable_Word_Count

Algorithm Parameters

Name: Meandre_Topic_Modeling

Description: This analysis performs topic analysis in the style of LDA and its variants using Mallet. Loads each page of each volume from HTRC. Removes the first and last line of each page. Joins hypenated words that occur at the end of the line. Removes all tokens that don't consist of alphanumeric characters. Filters stop words. Replaces "not " with "not_" to deal with negations. Creates a topic model using Mallet. Displays the top 200 tokens in a tag cloud. Reference: For more information, see http://mallet.cs.umass.edu. NOTE: The volume limit is 1000.

Version: 1.1

Author: Loretta Auvil

Please Input Job Name:

(required)

Ruskin_Painting_TM

Please select a collection for analysis

Ruskin_On_Painting@mfsense2

Please provide the number of tokens to be displayed in the tagcloud (default: 200)

200 (optional)

Please provide the number of topics to be created (default: 10)

10 (optional)

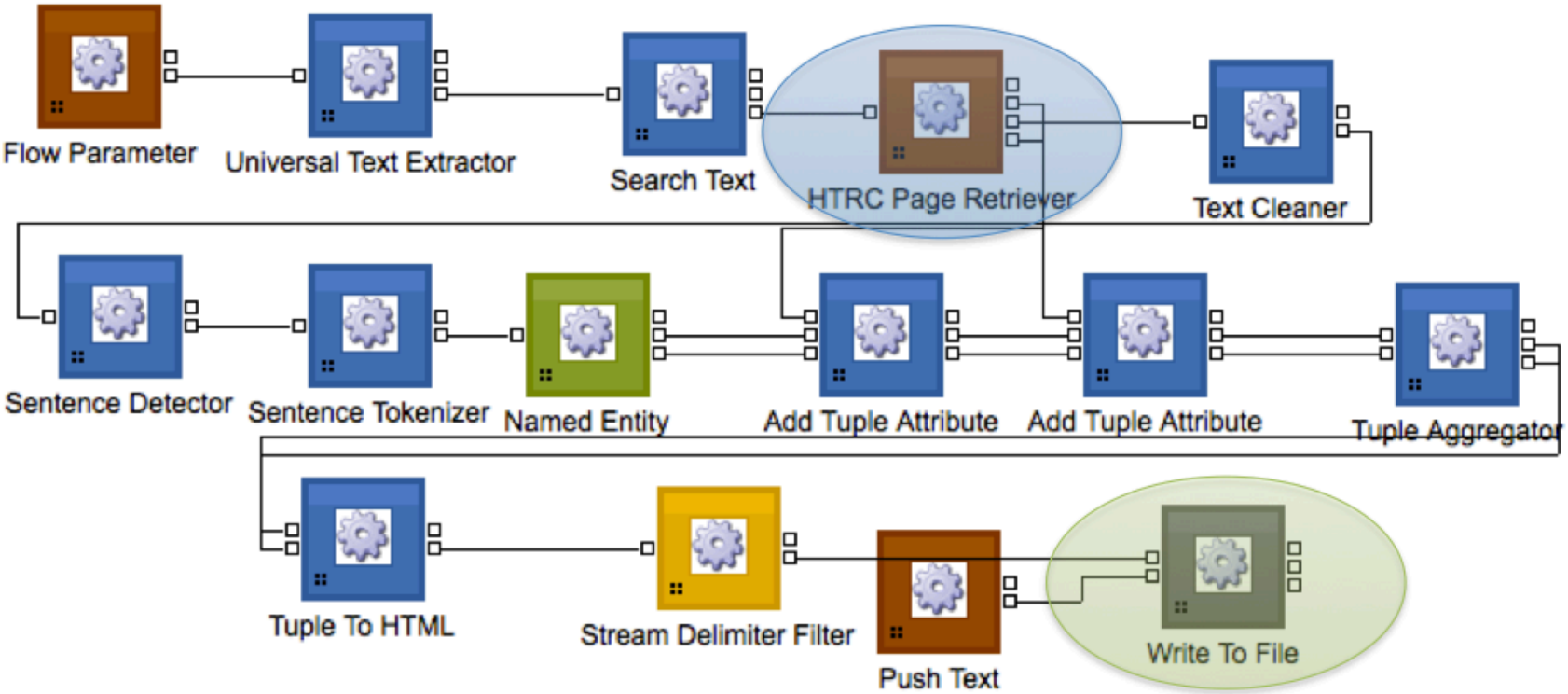
Submit

Choose among 10 possible algorithms and set parameters.

Each algorithm include a basic description of it's use, purpose, and any computational limits

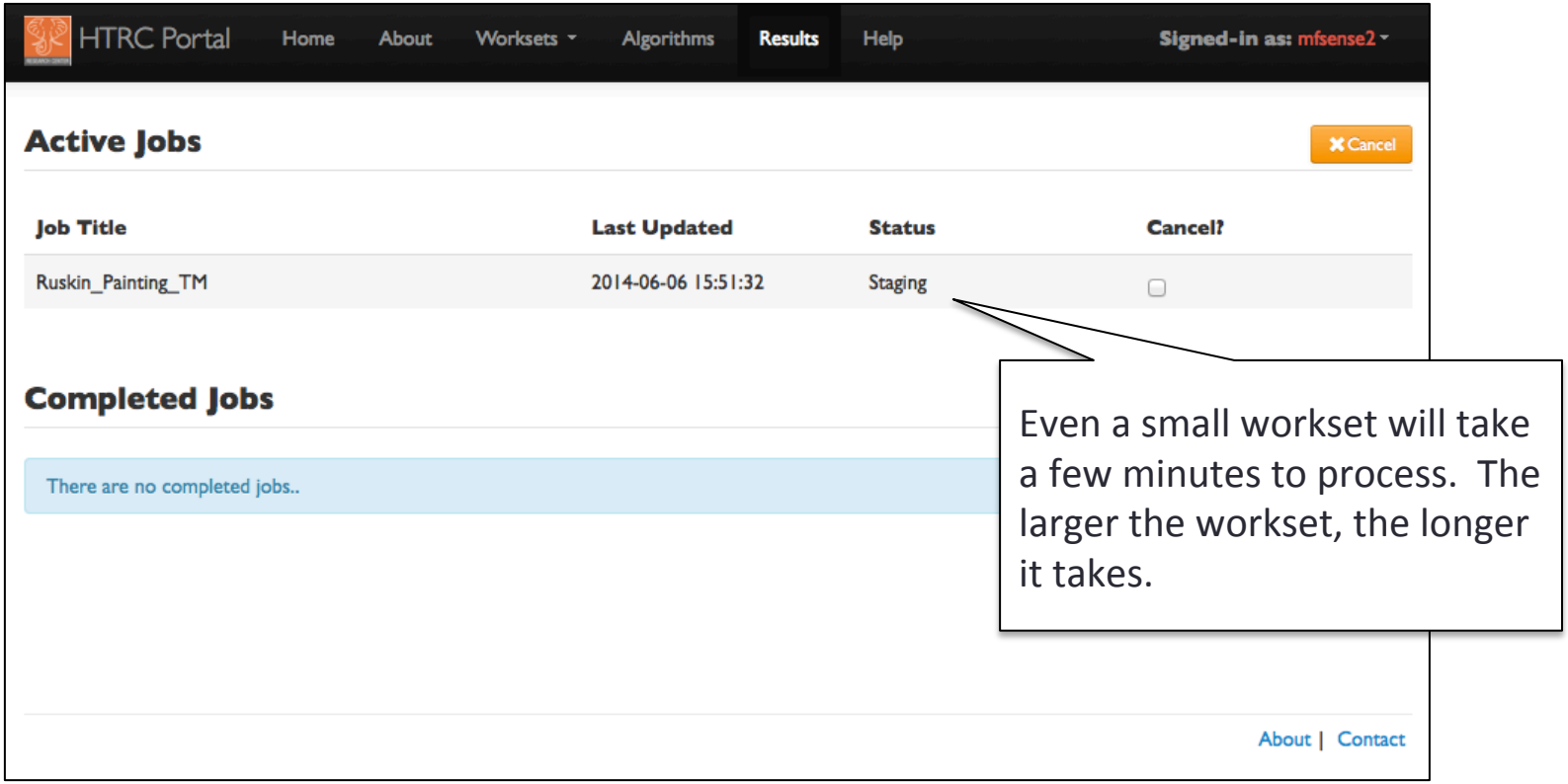
RESEARCH CENTER

Meandre Data Flow



RESEARCH CENTER

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.

RESEARCH CENTER

Job Details: Topic Modeling

HTRC Portal | Home | About | Worksets | Algorithms | Results | Help | Signed-in as: mfsense2

Job Details

Job Title: Ruskin_Painting_TM
Algorithm Name: Meandre_Topic_Modeling
Last Updated: 2014-06-07 13:14:12
Results:

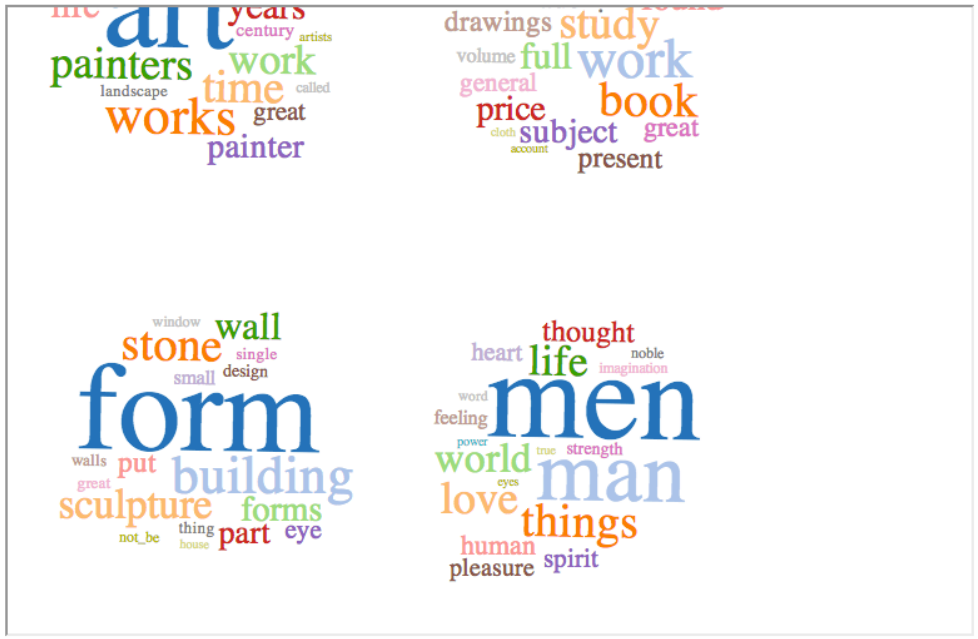
- topic_top_words.xml
- stderr.txt
- stdout.txt
- topic_tagclouds.html

Job Parameters:

Name	Value
n_top_tokens	200
num_topics	10
input_collection	Ruskin_On_Painting@mfsense2

Job Id:
d7269ad6-d576-448c-89d1-0353a244370a
Status:
Finished

View Results



Topic modeling data is now available as a word cloud or as a structured XML document

RESEARCH CENTER

Job Details: Named Entities

HTRC Portal | Home | About | Worksets | Algorithms | Results | Help | Signed-in as: mfsense2

Job Details

Job Title: Ruskin_NamedEntitles
Algorithm Name: Meandre_OpenNLP_Entities_List
Last Updated: 2014-06-07 16:17:57
Results:

stderr.txt

stdout.txt

named_entities_list.html

Job Parameters:

Name	Value
entity_types	person, location, organization
input_collection	Ruskin_On_Painting@mfsense2

Job Id:
45fffa84-c1a3-4b08-858e-f50a86b500b4
Status:
Finished

View Results

sentenceId	text	type	textStart	volume_id	page_id
4	ELDER	organization	15	uc1.b3820154	7
4	Oil Painting	organization	176	uc1.b3820154	10
4	Lord Lindsay	person	107	uc1.b3820154	10
4	Charles Eastlake	person	148	uc1.b3820154	10
2	Royal Academy of England	organization	247	uc1.b3820154	15
2	England	location	264	uc1.b3820154	15
4	Rome	location	33	uc1.b3820154	17
10	Spring	location	13	uc1.b3820154	17
13	Hunt	person	3	uc1.b3820154	19
5	E. Turck	person	6	uc1.b3820154	20
6	Cinderella	person	35	uc1.b3820154	20
10	White Owl	organization	4	uc1.b3820154	20
17	Mr. Redgrave	person	20	uc1.b3820154	20
2	Marie Antoinette	person	20	uc1.b3820154	21
4	Ward	person	3	uc1.b3820154	21
13	Lewis	person	3	uc1.b3820154	21
15	Mr. Cooper	person	0	uc1.b3820154	21
15	Mr. Lewis	person	126	uc1.b3820154	21
16	Academy	organization	49	uc1.b3820154	21
1	Colour Society	organization	177	uc1.b3820154	22
1	Mr. Lewis	person	16	uc1.b3820154	22
4	Academy	organization	64	uc1.b3820154	22
4	Mr. Lewis	person	31	uc1.b3820154	22
5	Colour Society	organization	82	uc1.b3820154	22

Named entities for person, organization, and location in a tabular format.

RESEARCH CENTER

Job Details: Dunning Log Likelihood

HTRC Portal | Home | About | Worksets | Algorithms | Results | Help | Signed-in as: mfsense2

Job Details

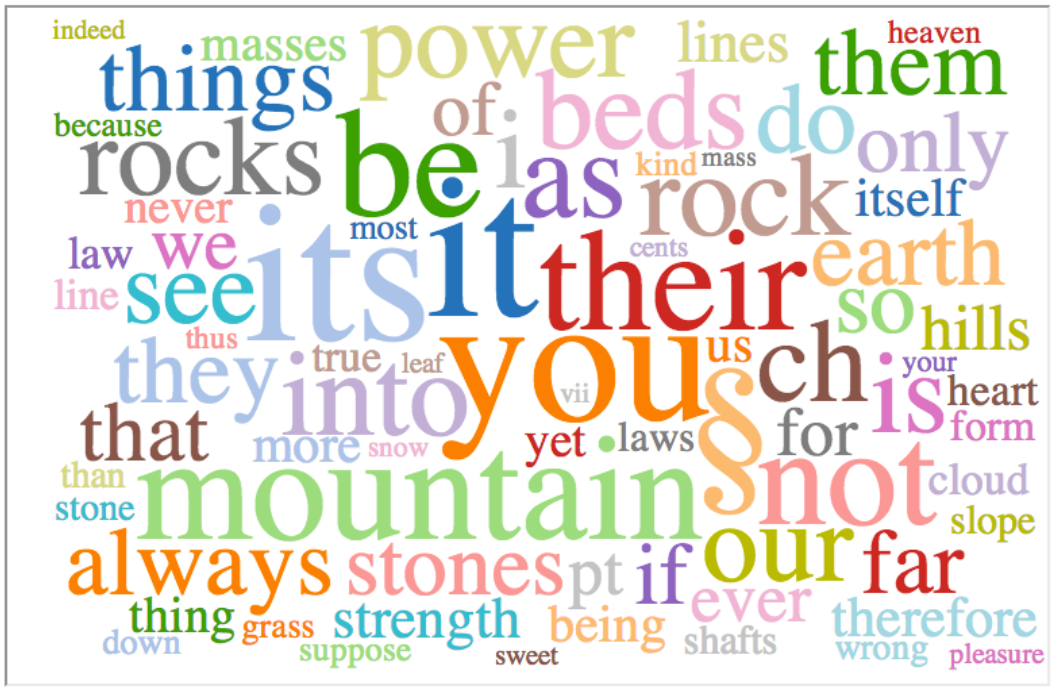
Job Title: RuskinCompared_DLL
Algorithm Name:
Meandre_Dunning_LogLikelihood_to_Tagcloud
Last Updated: 2014-07-06 11:37:53
Results:

- stderr.txt
- dunning_tagcloud_under.html
- dunning_over_represented.csv.txt
- stdout.txt
- dunning_tagcloud_over.html
- dunning_under_represented.csv.txt

Job Parameters:

Name	Value
n_top_tokens	200
input_collection_analysis	Ruskin_On_Painting@mfsense2
input_collection_reference	BooksOnPainting_EnglishLanguage@mfsense2

View Results



How Ruskin compares to 200 other works on painting.

RESEARCH CENTER

A Note on Performance

Algorithm	Number of Volumes	Execution Time (Minutes)
Meandre_Classification_NaiveBayes	1000	85.28
Meandre_Dunning_LogLikelihood_to_Tagcloud	1000	34.35
Meandre_OpenNLP_Date_Entities_to_Simile	100	21.93
Meandre_OpenNLP_Entities_List	100	26.22
Meandre_Spellcheck_Report_Per_Volume	100	16.02
Meandre_Tagcloud	1000	2.00
Meandre_Tagcloud_with_Cleaning	1000	6.79
Meandre_Topic_Modeling	1000	84.90

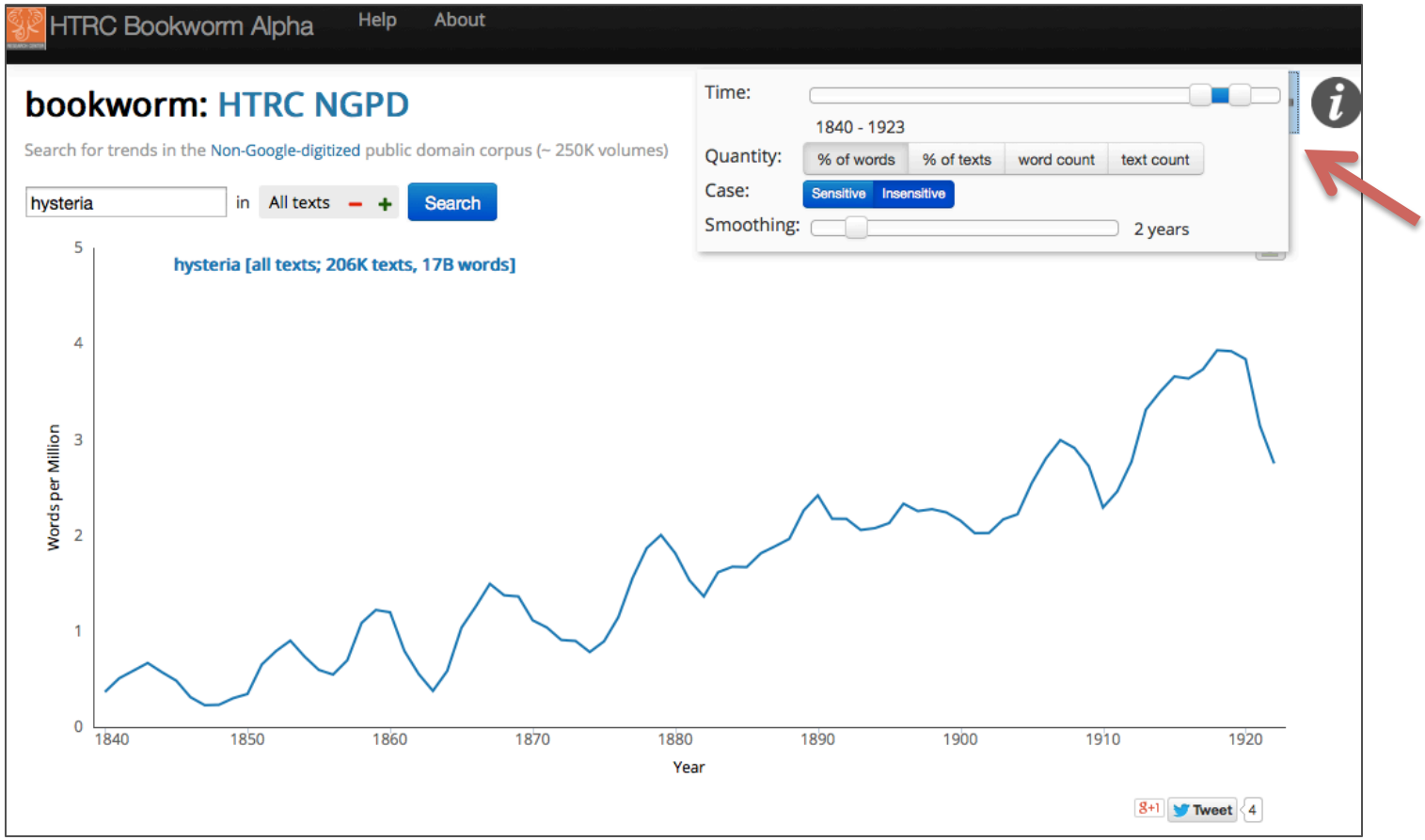
RESEARCH CENTER

ALPHA

Other HTRC Tools

RESEARCH CENTER

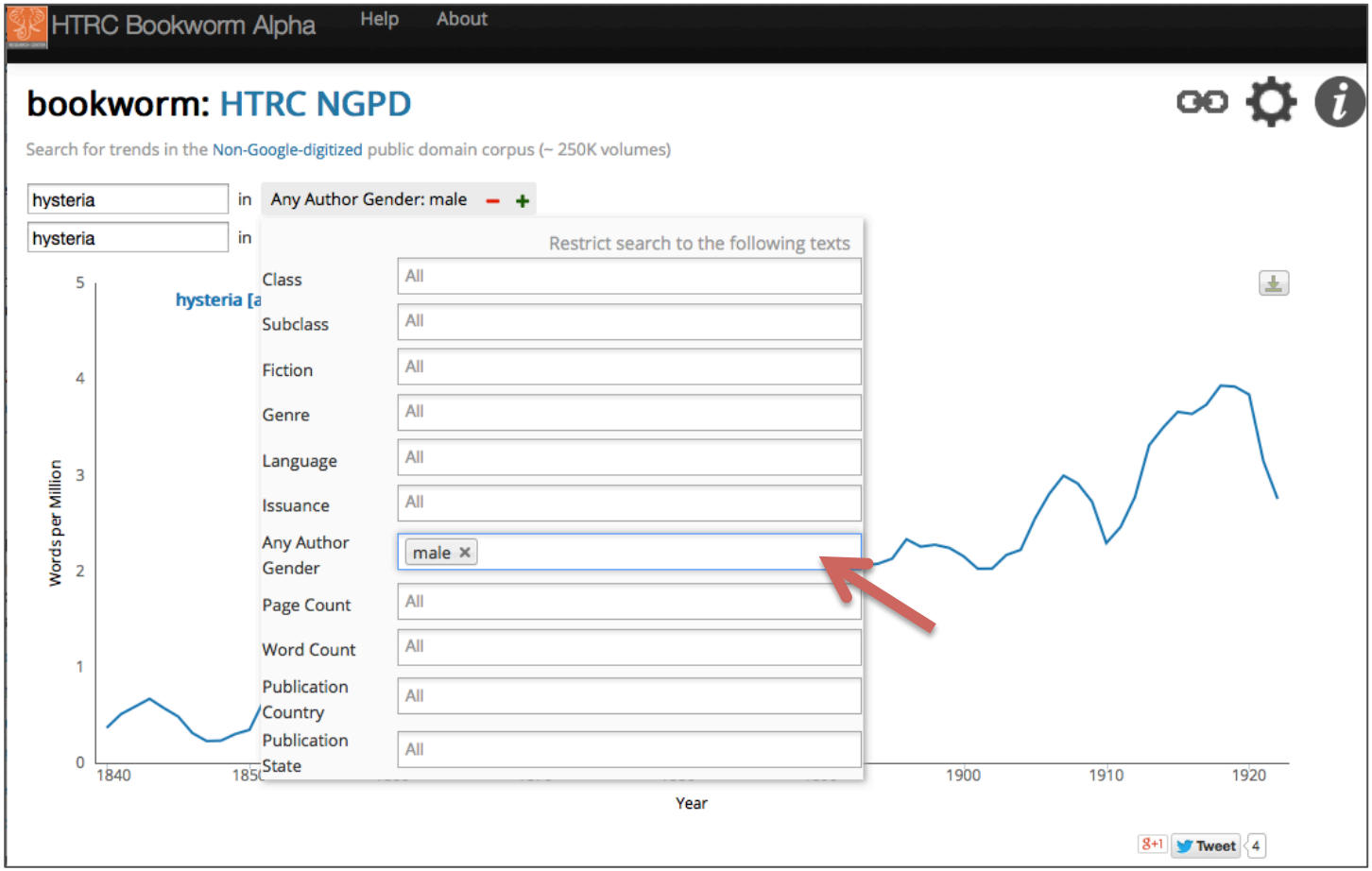
HTRC Bookwork Instance: Settings



http://sandbox.htrc.illinois.edu/bookworm/

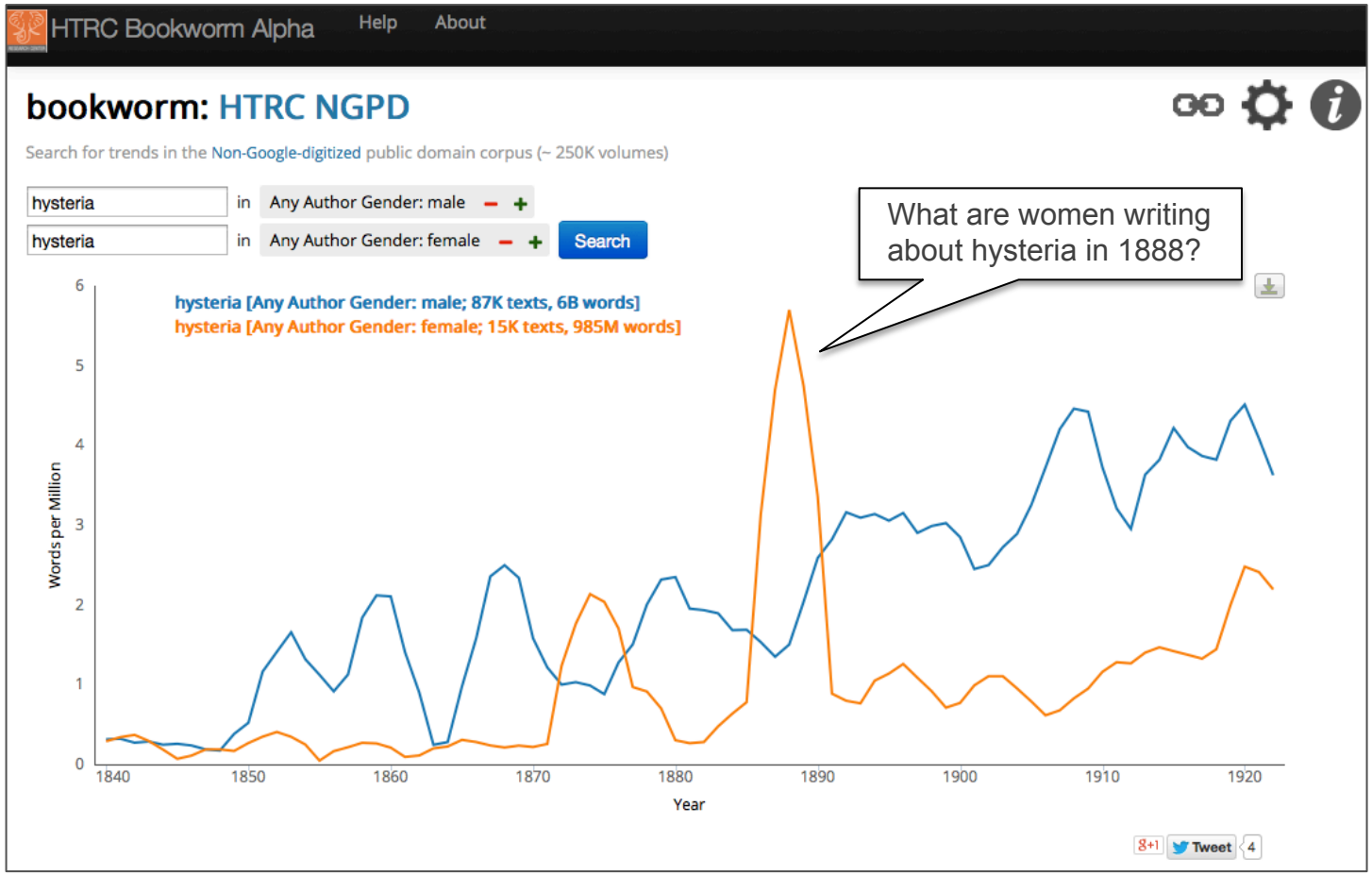
RESEARCH CENTER

HTRC Bookworm Instance: Search



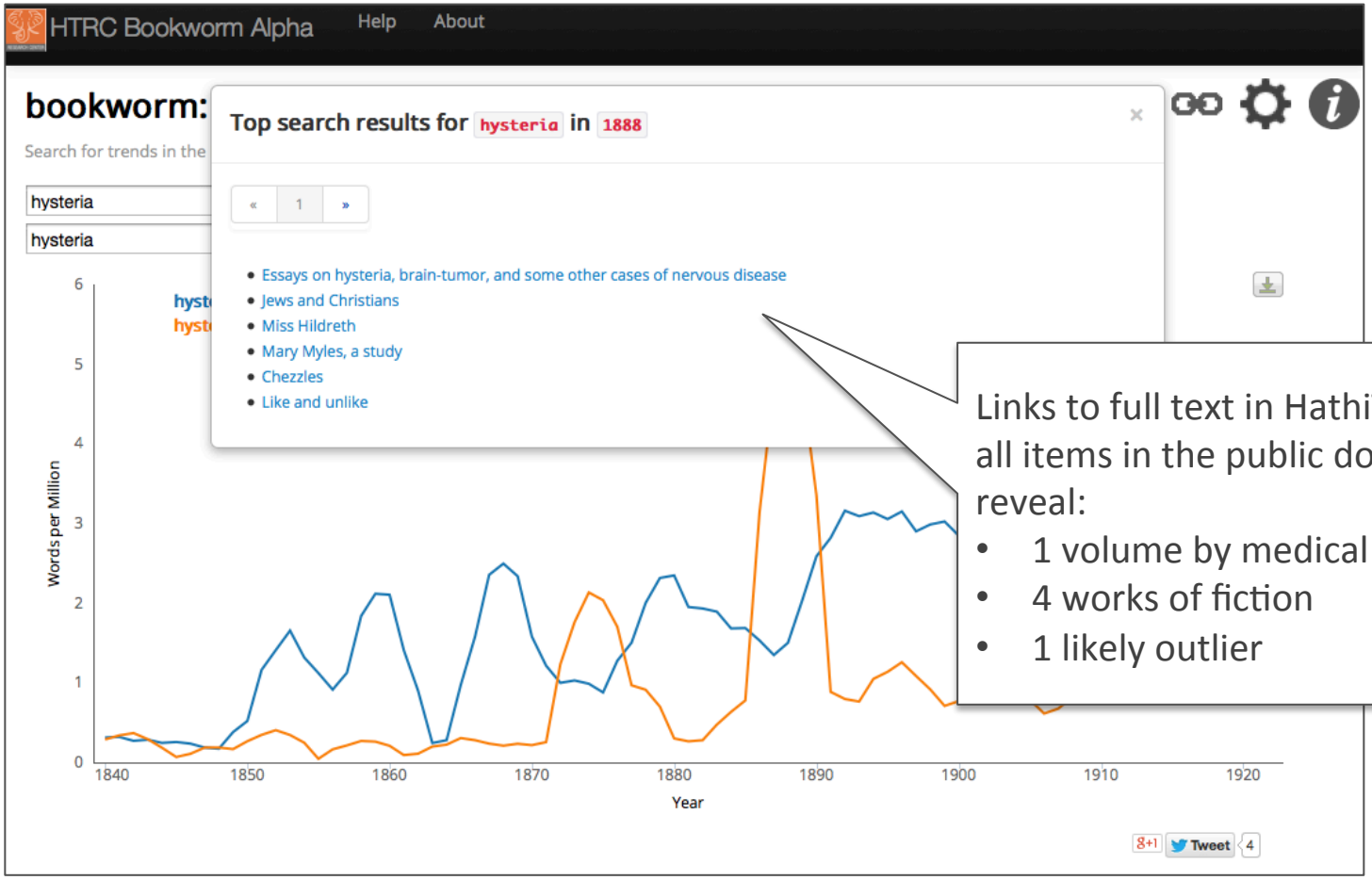
RESEARCH CENTER

HTRC Bookworm Instance: Results



RESEARCH CENTER

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

RESEARCH CENTER

HTRC Feature Extraction: Introduction

https://sandbox.htrc.illinois.edu/HTRC-UI-Portal2/

RESEARCH CENTER

HTRC Feature Extraction: About

HTRC Sandbox Portal | Home | About | Worksets | Algorithms | Results | Help | Data | Login

You are on the HTRC Sandbox. This is where new functionality is introduced on a smaller **public domain** subset of the HathiTrust.
Read more about the Sandbox or visit the **main HTRC Portal**.

Features extracted from the HTRC

Note that this is an **alpha** data release. Please send feedback to htrc-support-l@list.indiana.edu.

A great deal of fruitful research can be performed using non-consumptive pre-extracted features. For this reason, HTRC has put together a select set of page-level features extracted from the HathiTrust's non-Google-digitized public domain volumes. The source texts for this set of feature files are primarily in English.

Features are notable or informative characteristics of the text. We have processed a number of useful features, including part-of-speech tagged token counts, header and footer identification, and various line-level information. This is all provided *per-page*. Providing token information at the page level makes it possible to separate text from paratext. (An example of the latter may be: thirty pages of publishers' ads at the back of a book). We have also decided to break each page into a collection of three parts: header, body, and footer. The specific features that we extract from the text are described in more detail below.

The primary pre-calculated feature that we are providing is the token (unigram) count, on a per-page basis. Token counts are specific to the part-of-speech usage for that token's term, so that a term used as both a noun and a verb, for example, will have a separate count provided for each of these two modalities of its use. We also include line information, such as the number of lines with text on each page, and counts of characters that start and end lines on each page. This information can illuminate genre and volume structure: for instance, it helps distinguish poetry from prose, or body text from an index.

The present release is an alpha release, and we would love to hear about how you use it, or what else you would like to see!

Review sample data or jump straight to the downloads.

Sample JSON **Documentation**

Researchers performing text analysis usually process documents into a *machine-readable* set of features.

RESEARCH CENTER

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

RESEARCH CENTER

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

RESEARCH CENTER

Hands-On Exercise

RESEARCH CENTER

Discussion

RESEARCH CENTER

GRADUATE SCHOOL OF LIBRARY AND INFORMATION SCIENCE

The iSchool at Illinois

RESEARCH CENTER

Special thanks to Loretta Auvil, Stacy Kowalczyk, and Peter Organisciak.