



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



Introduction



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



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?

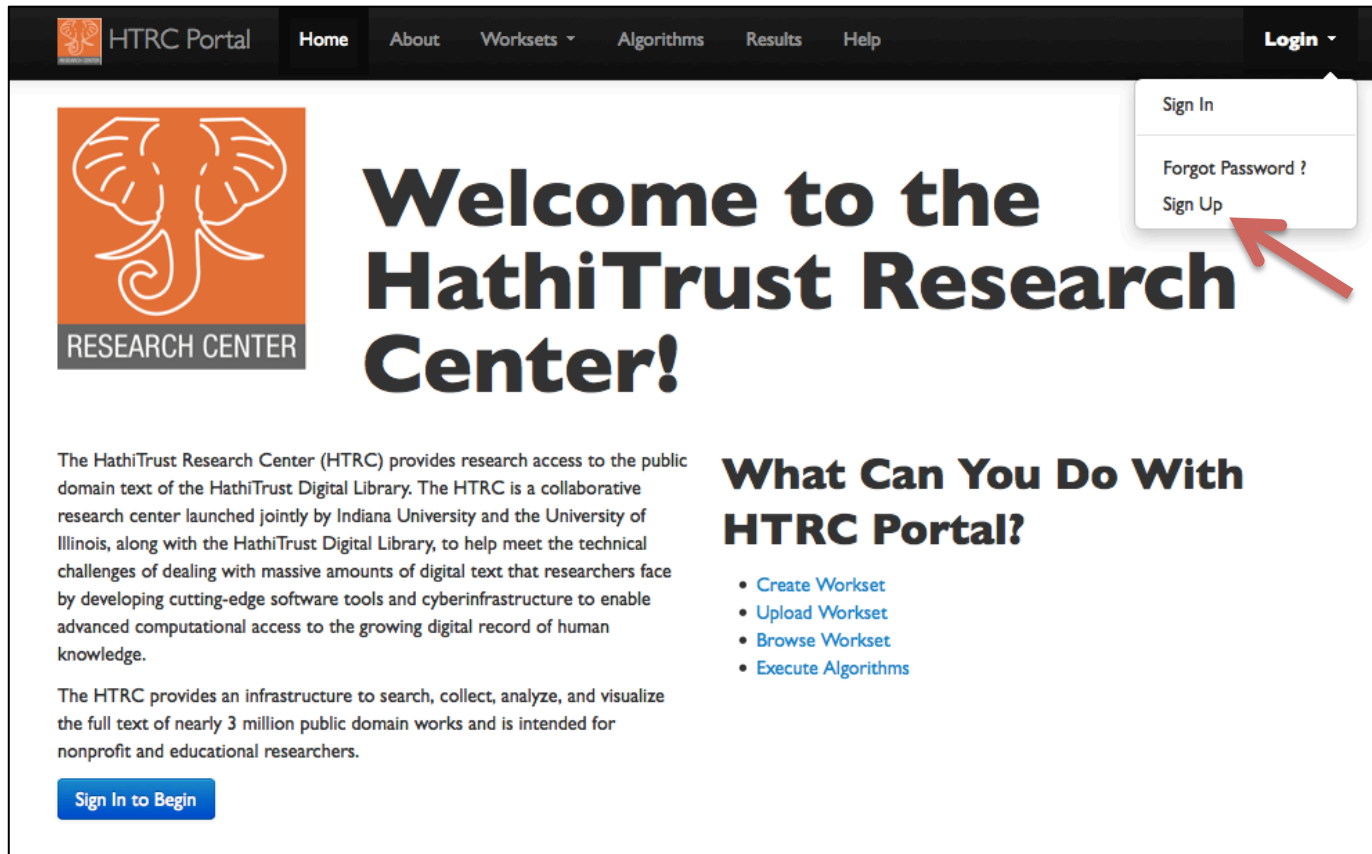


Case Study Questions (cont'd)

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



The screenshot shows the HTRC Portal website. The navigation bar at the top includes 'Home', 'About', 'Worksets', 'Algorithms', 'Results', and 'Help'. The 'Login' button is highlighted, and a dropdown menu is open, showing 'Sign In', 'Forgot Password?', and 'Sign Up'. A red arrow points to the 'Sign Up' option. The main content area features the HTRC logo (an orange elephant head) and the text 'Welcome to the HathiTrust Research Center!'. Below this, there is a paragraph describing the center's mission and a list of actions users can take: 'Create Workset', 'Upload Workset', 'Browse Workset', and 'Execute Algorithms'. A 'Sign In to Begin' button is also visible.

HTRC Portal Home About Worksets Algorithms Results Help **Login**

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.

[Sign In to Begin](#)

What Can You Do With HTRC Portal?

- [Create Workset](#)
- [Upload Workset](#)
- [Browse Workset](#)
- [Execute Algorithms](#)

Sign In
Forgot Password ?
Sign Up

<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 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

RESEARCH CENTER

HTRC Workset Builder

Sign Up | Login | Selected Items (0) | Portal

Limit your search

- Subject
- Author
- Language
- Place of Publication
- Year
- Original Location

in Full Text Search

More options

Type some keywords in the search box above to search the full text, then click "Search."
You can filter results by era, publication date, topic, language and source.
Need to craft a complex search? Choose more options below the search box.
You must be logged in to create a workset.

About Help/FAQ Contact

NB: Click “More options” for advanced search options

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



Workset Builder: Search Results

HTRC Workset Builder

Log Out [Megan Senseney] | Selected Items (0) | Manage Worksets | Portal

Limit your search

Subject

- Painting (113)[remove]
- Aesthetics (62)
- Architecture (13)
- Pre-Raphaelitism (9)
- Turner, J. M. W (9)
- Turner, J. M. W. (Joseph Mallord William), 1775-1851 (9)
- Art (4)
- Drawing (4)
- Royal Academy of Arts (Great Britain) (4)
- Drawing Study and teaching (3)
- Painting England London Catalogs (3)
- Painting Great Britain (3)
- Painting Study and teaching (2)
- Preraphaelitism (2)
- Art criticism (1)
- Art criticism Great Britain History 19th century (1)
- National Gallery (Great Britain) (1)
- Painters (1)
- Painting England London Exhibitions (1)
- Painting England London Handbooks, manuals, etc (1)

more »

Ruskin, John in Author Search

More options

Author > Ruskin, John x Subject > Painting x

Author > Ruskin, John, 1819-1900 x

Displaying items 1 - 10 of 113 [start over](#)

Sort by **relevance** Show 10 per page

« Previous 1 2 3 4 5 ... 11 12 Next »

Select items on page Deselect items on page Select all search items Deselect all search items

1. Modern Painters ... By John Ruskin ... Select

Title: Modern Painters ... By John Ruskin ...

Author: Ruskin, John, 1819-1900.

Format: Book

Language: English

Published: 1878

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

The screenshot displays the HTRC Workset Builder interface. At the top left is the HTRC Research Center logo. The top right navigation bar includes links for 'Log Out [Megan Senseney]', 'Selected Items (113)', 'Manage Worksets', and 'Portal'. The main content area is titled 'Selected Items' and features a 'Sort by' dropdown set to 'relevance' and a 'Show 10 per page' option. Below this is a pagination control with links for '« Previous', page numbers '1 2 3 4 5 ... 11 12', and 'Next »'. A toolbar contains 'Create/Update Workset' and 'Clear all' buttons. Two items are listed, both marked as 'Selected' with a checked checkbox. The first item is 'Modern Painters ... By John Ruskin ...' with metadata: Title: Modern Painters ... By John Ruskin ...; Author: Ruskin, John, 1819-1900.; Format: Book; Language: English; Published: 1878. The second item is '2. Modern painters.' with metadata: Title: Modern painters.; Author: Ruskin, John, 1819-1900.; Format: Book; Language: English; Published: 1888. A callout box on the left contains the text: 'When you're done selecting items, review your choices, and select "Create/Update Workset"'.

HTRC Workset Builder

Log Out [Megan Senseney] | Selected Items (113) | Manage Worksets | Portal

RESEARCH CENTER

Back to Search

Selected Items

Sort by relevance Show 10 per page

« Previous 1 2 3 4 5 ... 11 12 Next »

Create/Update Workset Clear all

Modern Painters ... By John Ruskin ... Selected

Title: Modern Painters ... By John Ruskin ...
Author: Ruskin, John, 1819-1900.
Format: Book
Language: English
Published: 1878

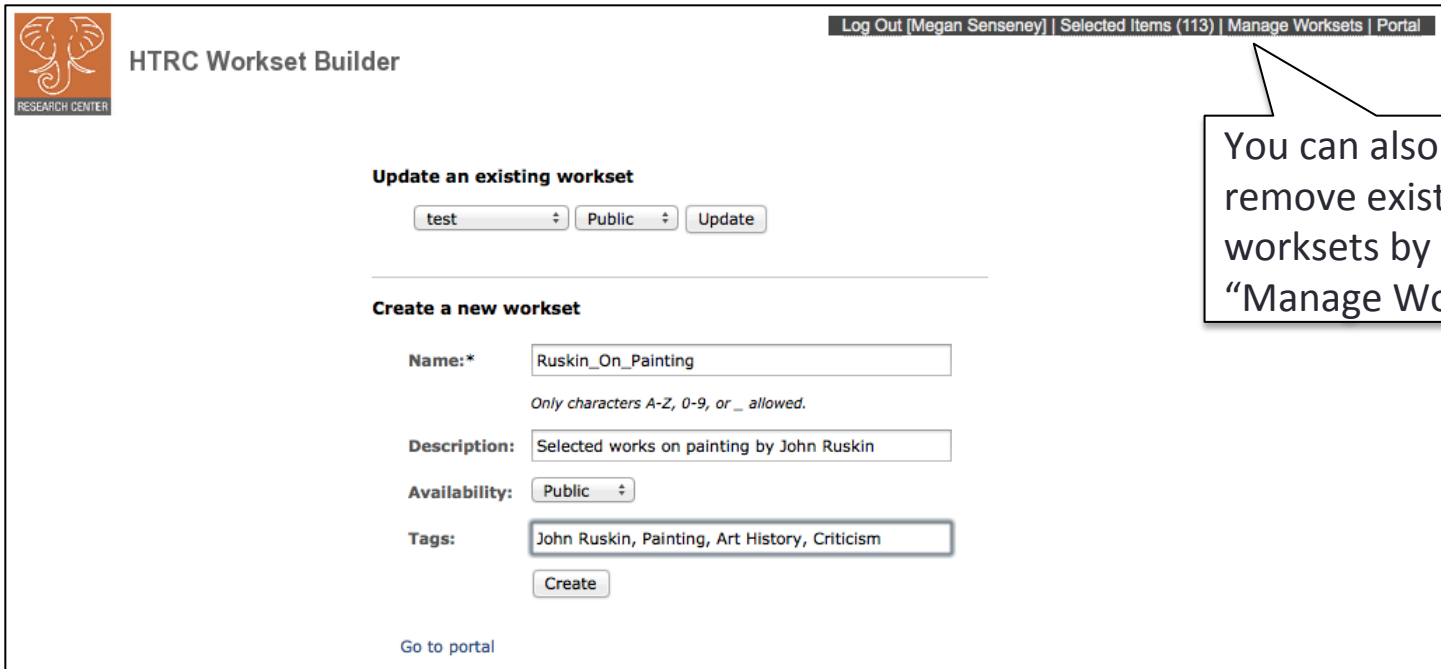
2. Modern painters. Selected

Title: Modern painters.
Author: Ruskin, John, 1819-1900.
Format: Book
Language: English
Published: 1888

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



Workset Builder: Workset Metadata



The screenshot shows the HTRC Workset Builder interface. At the top right, there is a navigation bar with links: "Log Out [Megan Senseney] | Selected Items (113) | Manage Worksets | Portal". The main content area is divided into two sections:

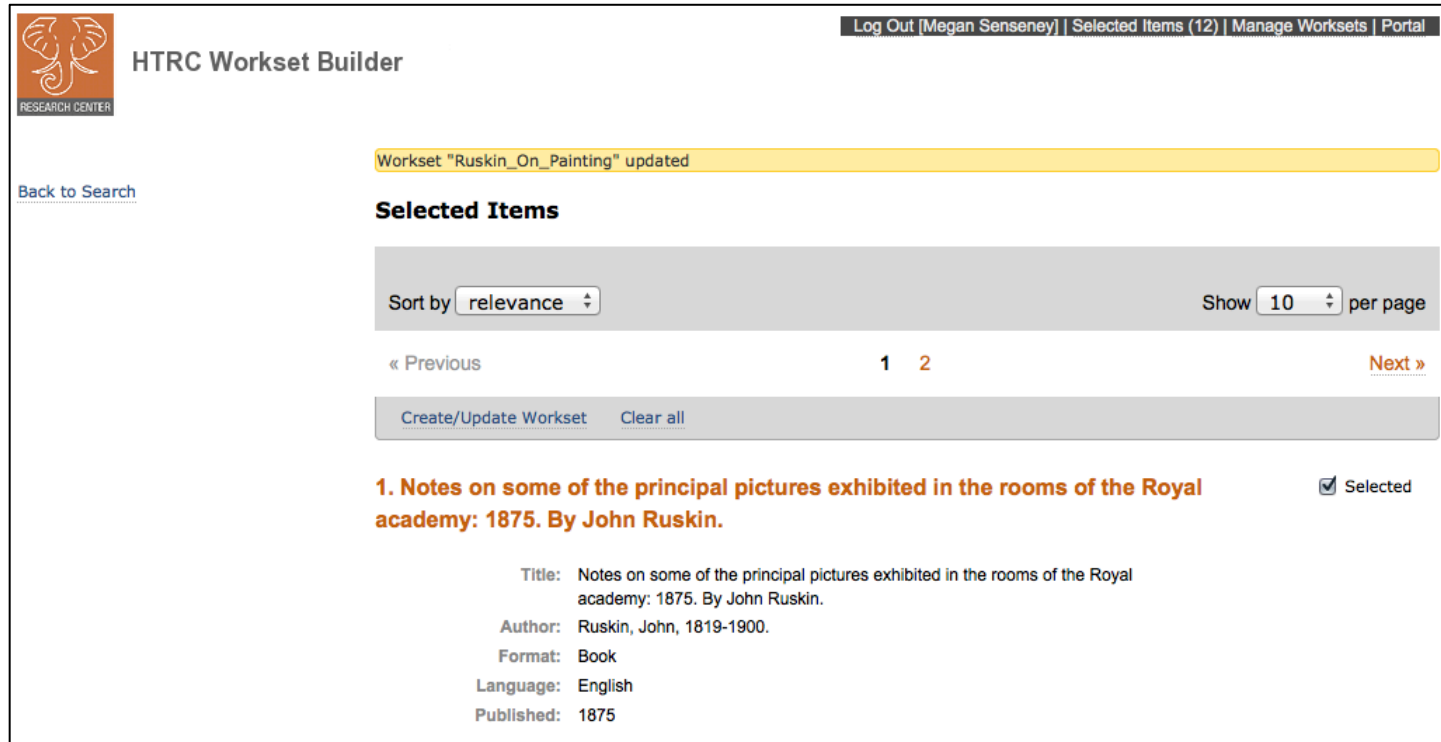
- Update an existing workset:** This section contains a dropdown menu with "test" selected, a "Public" dropdown menu, and an "Update" button.
- Create a new workset:** This section contains several input fields:
 - Name:** A text box containing "Ruskin_On_Painting". Below it, a note reads "Only characters A-Z, 0-9, or _ allowed."
 - Description:** A text box containing "Selected works on painting by John Ruskin".
 - Availability:** A dropdown menu with "Public" selected.
 - Tags:** A text box containing "John Ruskin, Painting, Art History, Criticism".Below these fields is a "Create" button.

At the bottom left of the form, there is a link "Go to portal".

You can also alter or remove existing worksets by choosing "Manage Worksets"

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

Workset Builder: Management



The screenshot displays the HTRC Workset Builder interface. At the top right, there are navigation links: "Log Out [Megan Senseney] | Selected Items (12) | Manage Worksets | Portal". The HTRC Research Center logo is in the top left. A yellow notification bar states "Workset 'Ruskin_On_Painting' updated". Below this, a "Back to Search" link is visible. The "Selected Items" section includes a "Sort by" dropdown set to "relevance" and a "Show 10 per page" dropdown. Navigation links for "« Previous", "1 2", and "Next »" are present. A "Create/Update Workset" and "Clear all" button bar is also shown. The first item in the list is "1. Notes on some of the principal pictures exhibited in the rooms of the Royal academy: 1875. By John Ruskin.", which is marked as "Selected" with a checked checkbox. The item's metadata is displayed below the title:

Title: Notes on some of the principal pictures exhibited in the rooms of the Royal academy: 1875. By John Ruskin.
Author: Ruskin, John, 1819-1900.
Format: Book
Language: English
Published: 1875

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

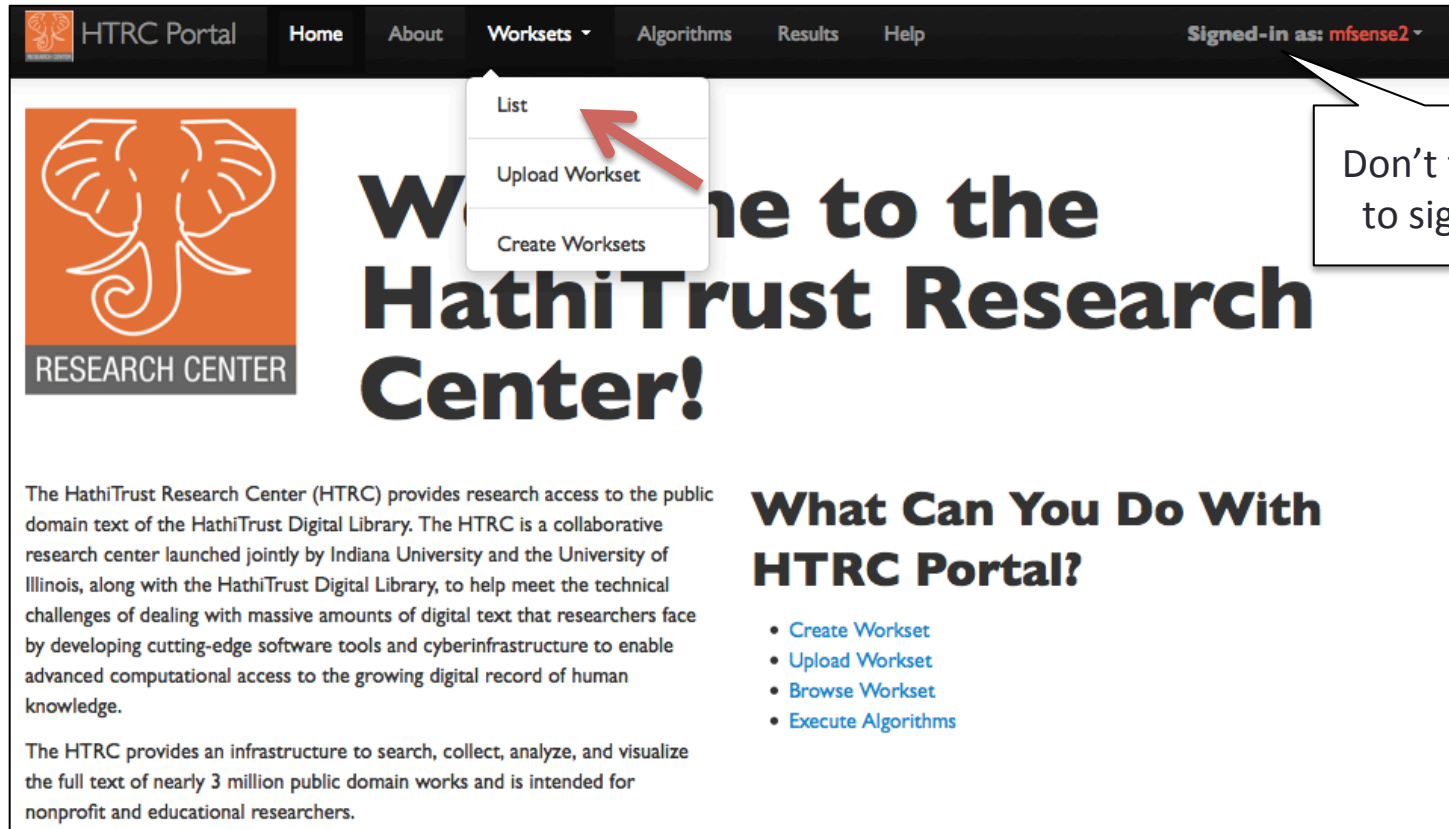


HTRC Portal

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



Portal: Workset Menu



The screenshot shows the HTRC Portal interface. At the top, there is a navigation bar with the following items: HTRC Portal (with logo), Home, About, Worksets (with a dropdown arrow), Algorithms, Results, and Help. On the right side of the navigation bar, it says "Signed-in as: mfsense2". The "Worksets" dropdown menu is open, showing three options: "List", "Upload Workset", and "Create Worksets". A red arrow points to the "List" option. Below the navigation bar, there is a large orange elephant logo with the text "RESEARCH CENTER" underneath it. To the right of the logo, the text reads "Welcome to the HathiTrust Research Center!". Below this, there is a paragraph of text describing the HTRC and its mission. To the right of this paragraph, there is a section titled "What Can You Do With HTRC Portal?" with a bulleted list of actions: "Create Workset", "Upload Workset", "Browse Workset", and "Execute Algorithms". A speech bubble on the right side of the screenshot says "Don't forget to sign in!".

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

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](#)

Don't forget to sign in!

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



Portal: Workset Details

The screenshot shows the HTRC Portal interface. The top navigation bar includes 'Home', 'About', 'Worksets', 'Algorithms', 'Results', and 'Help'. The user is signed in as 'mfsense2'. The main content area is titled 'Workset Details' and features a list of available worksets on the left and detailed information for the selected workset on the right. The selected workset is 'Ruskin_On_Painting'. The details include the name, description, author, last modified by, last modified time, and number of volumes. There are buttons for 'Edit Workset' and 'Download CSV File'. Below the details is a pagination control with 'First', 'Prev', 'Next', and 'Last' buttons. The main content area displays the first item in the workset, which is a list of modern painters by a graduate of Oxford. The item details include volume ID, author (Ruskin, John, 1819-1900), page count (414), and word count (154101). The second item is a list of notes on principal pictures exhibited in the rooms of The Royal Academy and The Society of Painters in Water Colours, no. 2, 1856 ...

Available Worksets +

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.



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 hyphenated 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)

Please select a collection for analysis

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

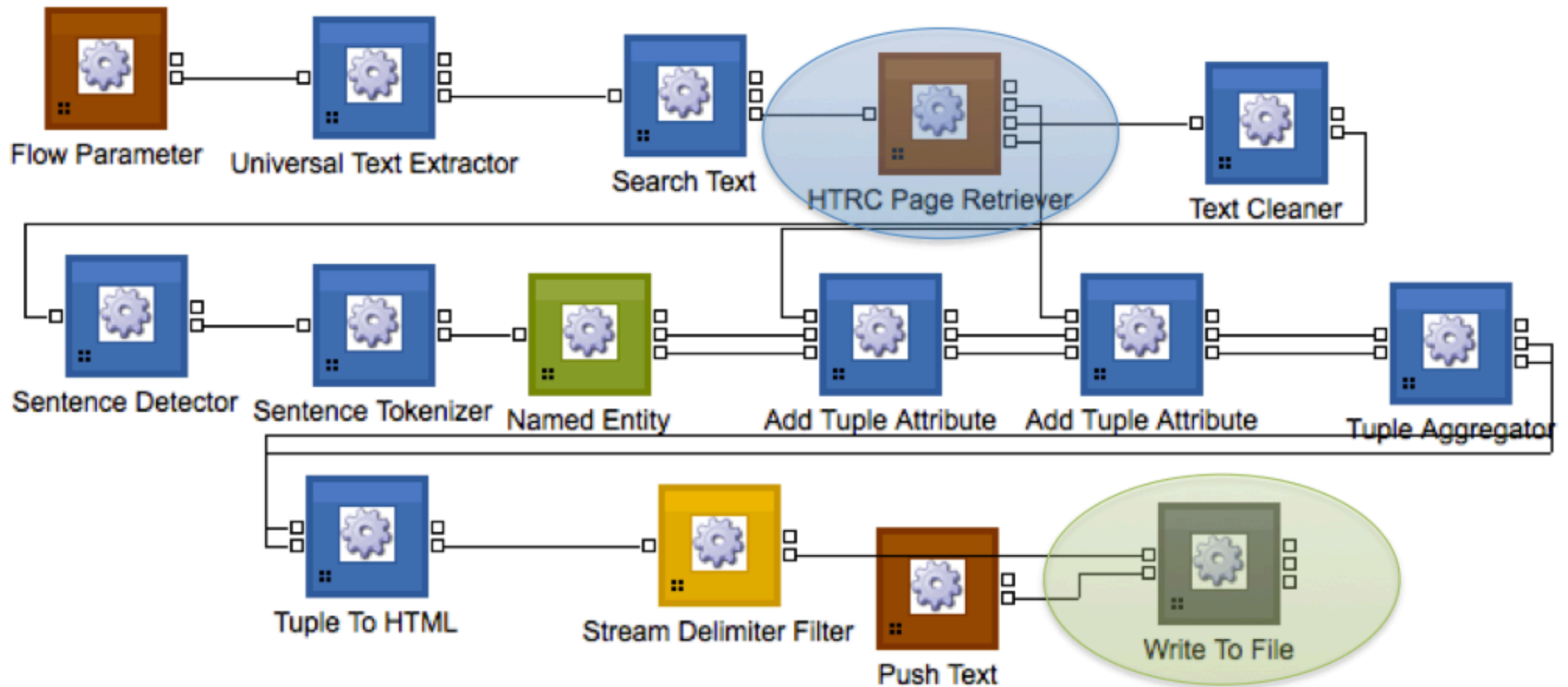
Please provide the number of topics to be created (default: 10)
 (optional)

Choose among 10 possible algorithms and set parameters.

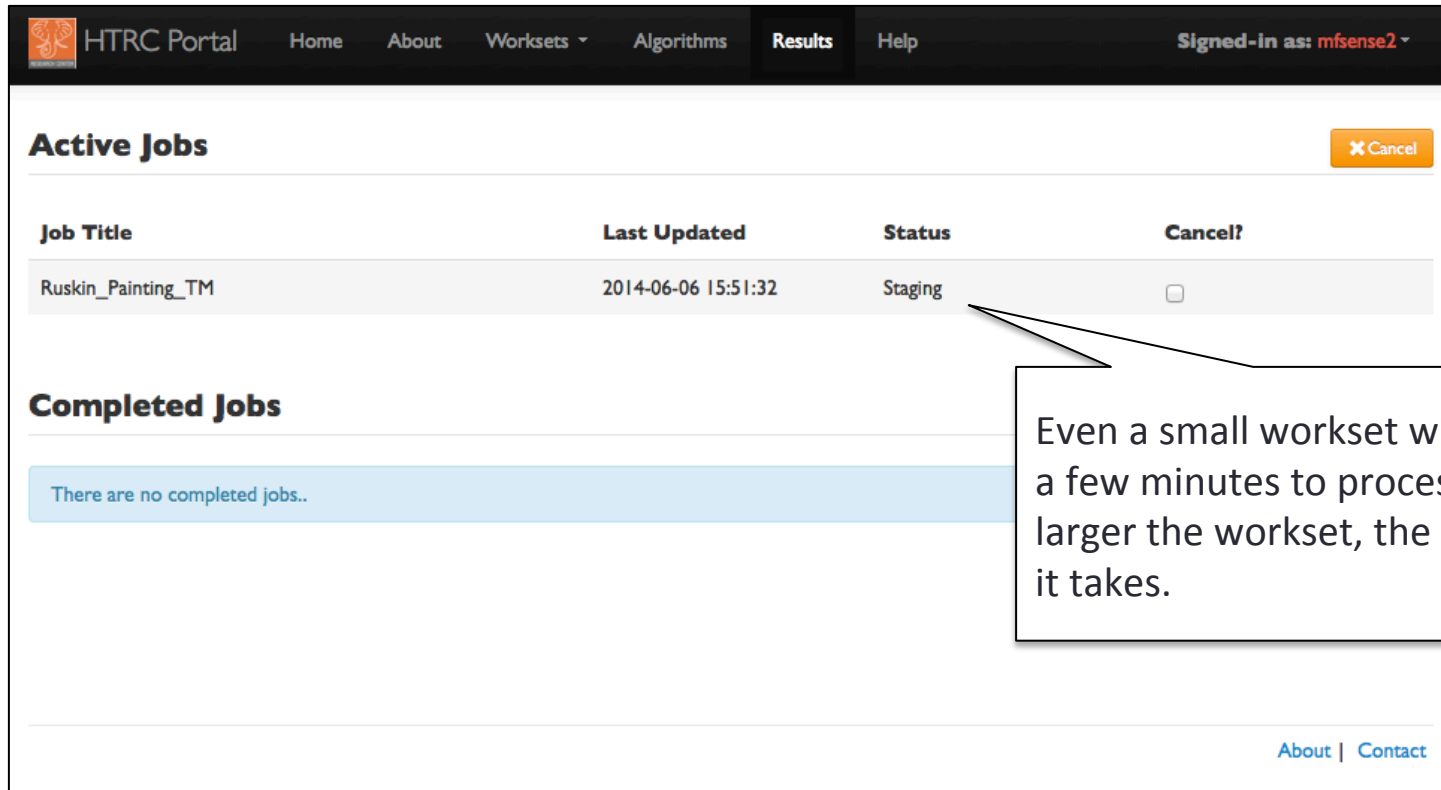
Each algorithm include a basic description of it's use, purpose, and any computational limits



Meandre Data Flow



Portal: Results



The screenshot shows the HTRC Portal interface. The top navigation bar includes links for Home, About, Worksets, Algorithms, Results (selected), and Help. The user is signed in as 'mfsense2'. The main content area is divided into two sections: 'Active Jobs' and 'Completed Jobs'. The 'Active Jobs' section contains a table with one job entry: 'Ruskin_Painting_TM', last updated on '2014-06-06 15:51:32', with a status of 'Staging' and a 'Cancel?' checkbox. A callout box points to the 'Staging' status with the text: 'Even a small workset will take a few minutes to process. The larger the workset, the longer it takes.' The 'Completed Jobs' section shows a message: 'There are no completed jobs..'. At the bottom right, there are links for 'About' and 'Contact'.

Job Title	Last Updated	Status	Cancel?
Ruskin_Painting_TM	2014-06-06 15:51:32	Staging	<input type="checkbox"/>

Refresh your browser to track status updates.



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

The image displays four word clouds representing different topics. The top-left cloud includes terms like 'painters', 'works', 'time', 'great', 'painter', 'landscape', 'called', 'century', and 'artists'. The top-right cloud features 'drawings', 'study', 'work', 'book', 'price', 'subject', 'great', 'present', 'volume', 'full', 'general', 'cloth', and 'account'. The bottom-left cloud is dominated by 'form', 'stone', 'wall', 'building', 'sculpture', 'forms', 'part', 'eye', 'wells', 'put', 'great', 'not_be', 'thing', 'house', 'single', 'design', and 'small'. The bottom-right cloud contains 'men', 'man', 'things', 'love', 'world', 'strength', 'true', 'eyes', 'power', 'feeling', 'word', 'heart', 'life', 'noble', 'imagination', 'thought', 'human', 'pleasure', and 'spirit'.

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



Job Details: Named Entities

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

Job Details

Job Title: Ruskin_NamedEntities
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.



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

Word cloud visualization of the results, showing words like 'mountain', 'stones', 'strength', 'being', 'therefore', 'power', 'lines', 'them', 'rocks', 'be', 'as', 'rock', 'earth', 'see', 'it', 'their', 'into', 'you', 'such', 'is', 'form', 'cloud', 'slope', 'always', 'stones', 'pt', 'if', 'our', 'far', 'thing', 'grass', 'suppose', 'sweet', 'shafts', 'wrong', 'pleasure', 'indeed', 'masses', 'of', 'beds', 'do', 'only', 'itself', 'never', 'we', 'most', 'cent', 'law', 'line', 'thus', 'they', 'true', 'leaf', 'vii', 'yet', 'laws', 'for', 'heart', 'than', 'stone', 'more', 'snow', 'that', 'mountain', 'not', 'form', 'cloud', 'slope', 'down', 'grass', 'suppose', 'sweet', 'shafts', 'wrong', 'pleasure'.

How Ruskin compares to 200 other works on painting.



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

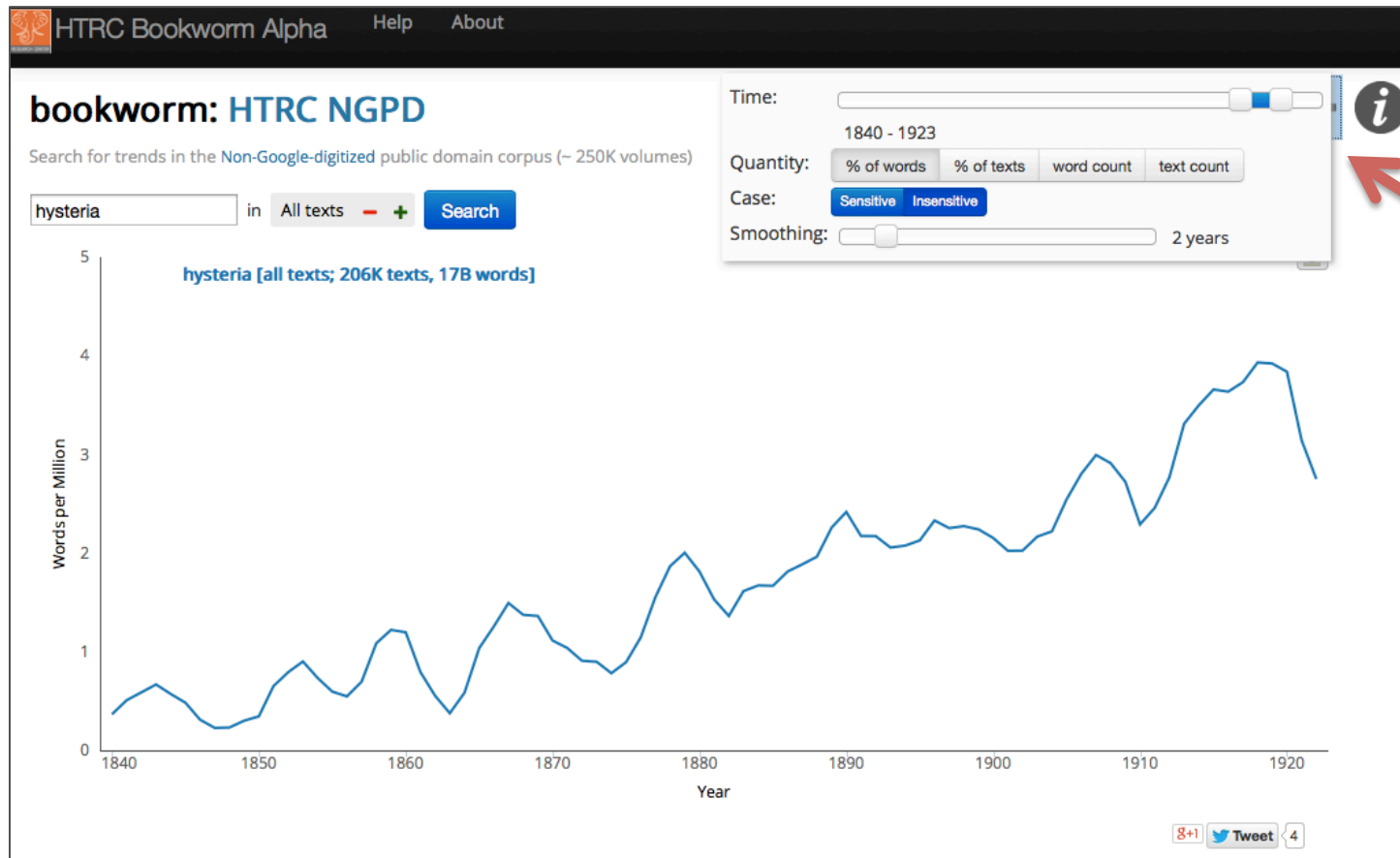


ALPHA

Other HTRC Tools



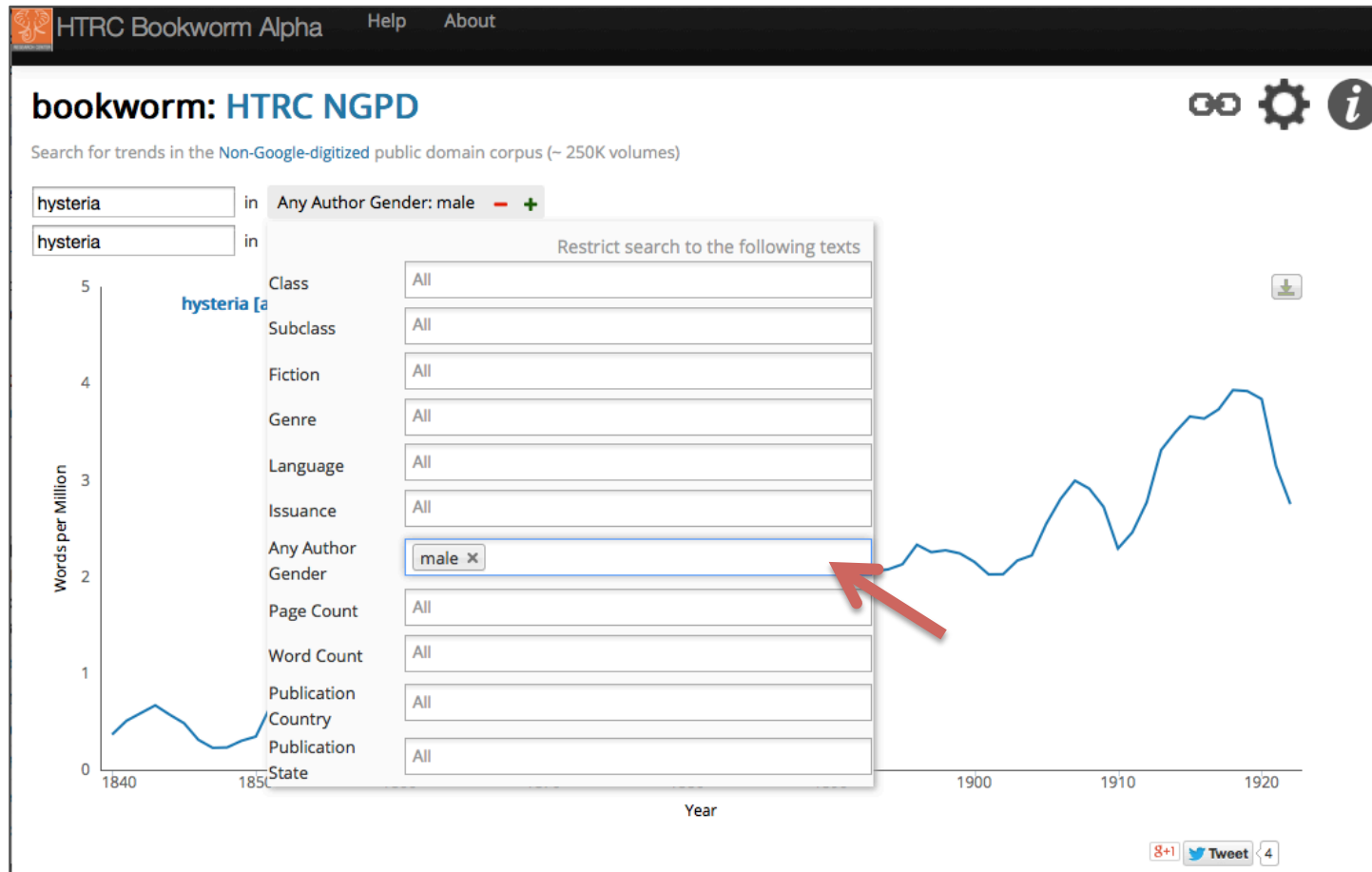
HTRC Bookwork Instance: Settings



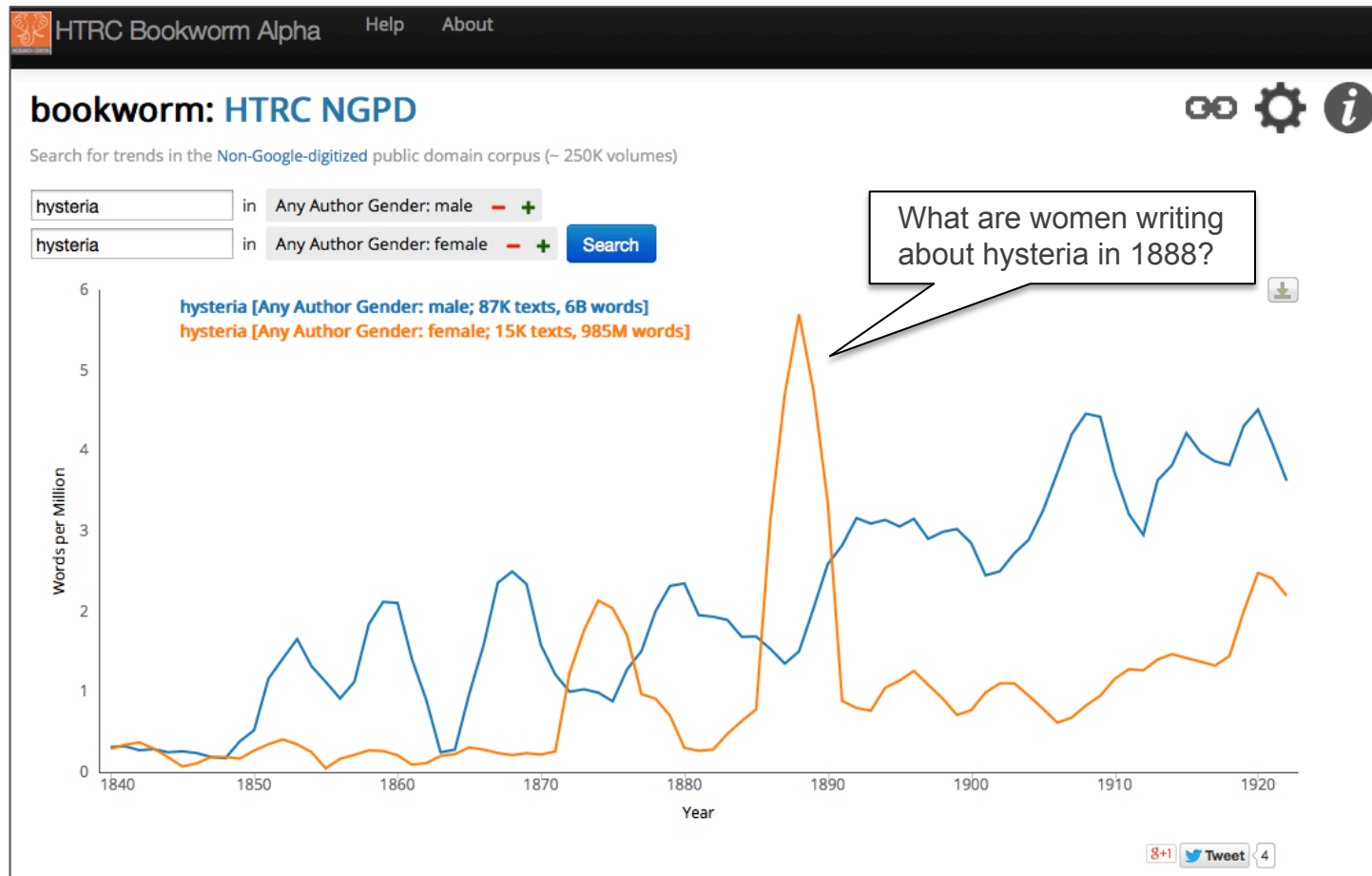
<http://sandbox.htrc.illinois.edu/bookworm/>



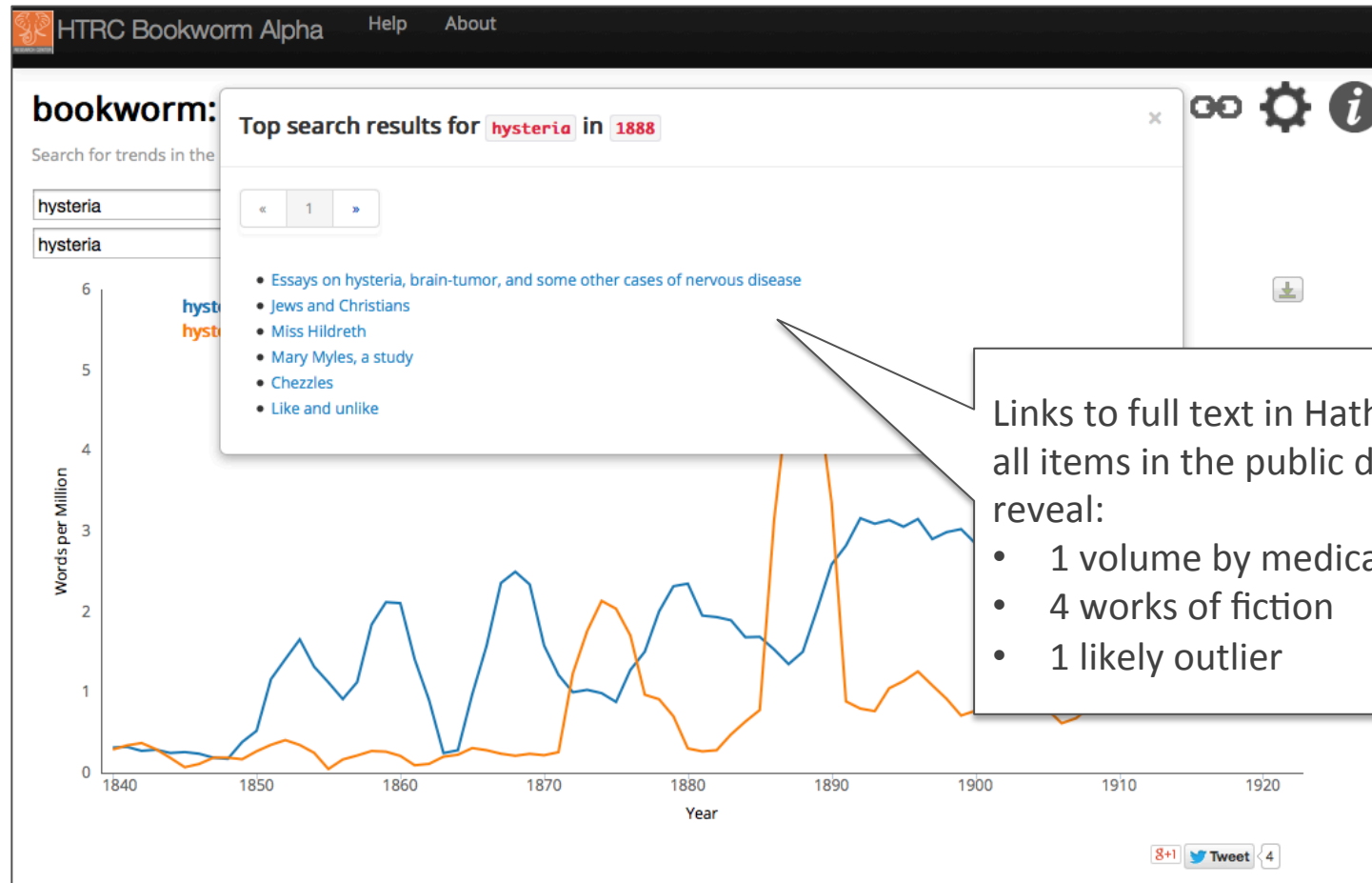
HTRC Bookworm Instance: Search



HTRC Bookworm Instance: Results



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



HTRC Feature Extraction: Introduction

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**. [Read more about the Sandbox](#) or visit the [main HTRC Portal](#).

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?

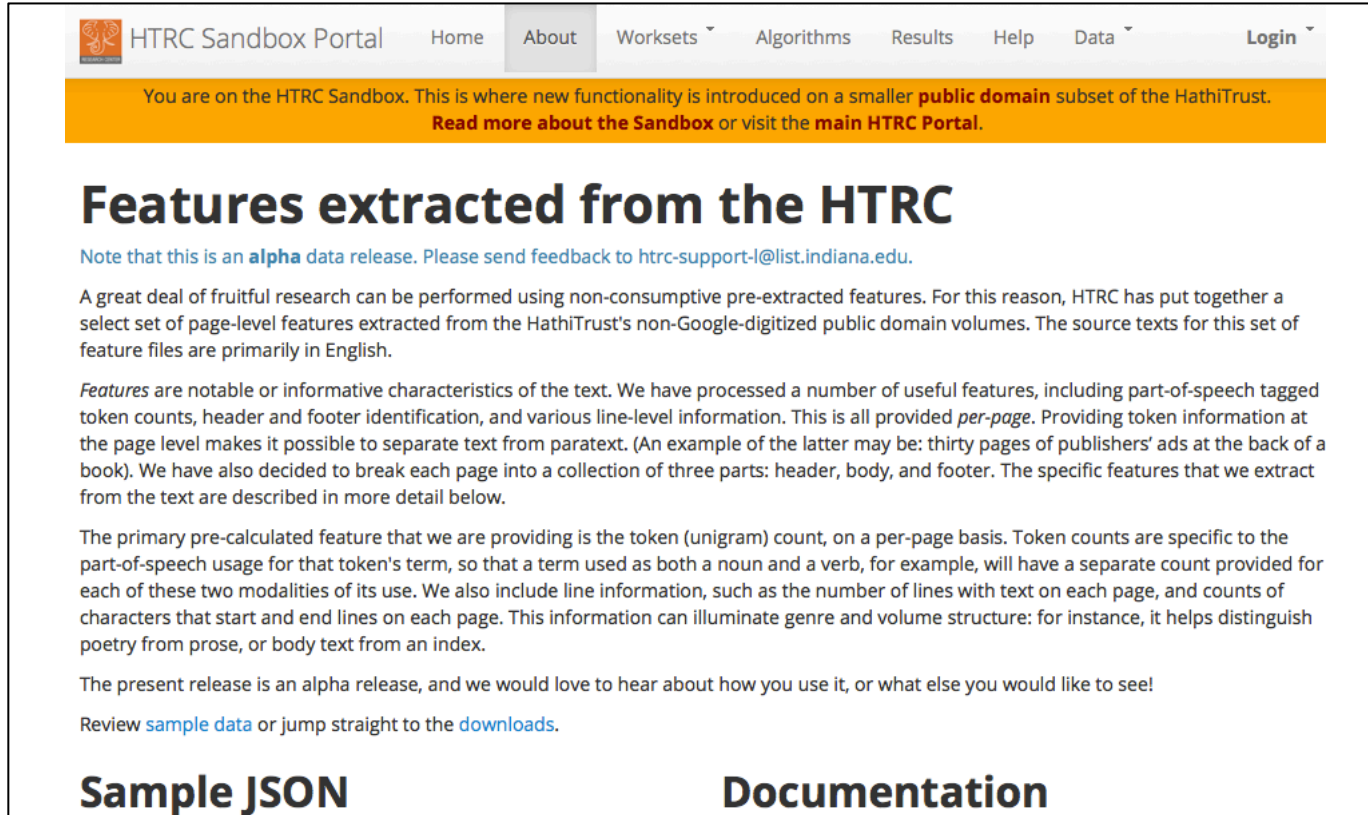
The HTRC Sandbox is distinct from the [main production portal](#) of the HTRC. The HTRC Sandbox is meant to be an arena for users to try out experiments and do exploratory work. The dataset available on the sandbox is a much smaller subset of that associated with the HTRC's main production portal. The 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 programmatically as

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



HTRC Feature Extraction: About



The screenshot shows the HTRC Sandbox Portal website. The navigation bar includes links for Home, About, Worksets, Algorithms, Results, Help, Data, and Login. A yellow banner below the navigation bar states: "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](#)." The main heading is "Features extracted from the HTRC". Below this, a note says: "Note that this is an **alpha** data release. Please send feedback to htrc-support-l@list.indiana.edu." The text explains that a great deal of fruitful research can be performed using non-consumptive pre-extracted features. It describes the features as notable or informative characteristics of the text, including part-of-speech tagged token counts, header and footer identification, and various line-level information. It also mentions that the primary pre-calculated feature is the token (unigram) count, on a per-page basis. The present release is an alpha release, and the page encourages users to review [sample data](#) or jump straight to the [downloads](#). At the bottom of the page, there are two buttons: "Sample JSON" and "Documentation".

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



GRADUATE SCHOOL OF **LIBRARY AND
INFORMATION SCIENCE**
The iSchool at Illinois



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