



HATHITRUST

A Shared Digital Repository

Humanistic Inquiry with Large Corpora of Digitized Text and Metadata: Toward New Epistemologies?

Sayan Battacharyya

Jeremy York, Assistant Director, HathiTrust

January 9, 2015



Unless otherwise noted, these slides and their contents are licensed under a [Creative Commons Attribution Unported License](https://creativecommons.org/licenses/by/4.0/).

Outline

- Part 1
 - Brief history of digital texts
 - Overview of HathiTrust / Origins of HTRC
- Part 2
 - HathiTrust Research Center Demo
 - HathiTrust Research Center Initiatives
- Part 3
 - Discussion



A Brief History of Digital Texts



Digital text resources

- Project Gutenberg (U of Illinois) – 1971
- Thesaurus Linguae Graecae (UC Irvine)– 1972
- Oxford Text Archive (U of Oxford) – 1976
- ARTFL Project (U of Chicago)– 1982
- Perseus Digital Library (Tufts)– 1985
- Text Encoding Initiative – 1987
- Women Writers Project (Brown U) – 1988



Digital Imaging

- Yale Open Book Project (1991)
- Cornell Demonstration Project (1993)
- Library of Congress National Digital Library; American Memory (1994)
- Making of America (University of Michigan and Cornell) (1995)



HathiTrust



HathiTrust Members

Allegheny College
Arizona State University
Baylor University
Boston College
Boston University
Brandeis University
Brown University
California Digital Library
Carnegie Mellon University
Case Western Reserve University
Colby College
Columbia University
Cornell University
Dartmouth College
Duke University
Emory University
Florida State University System
Georgetown University
Getty Research Institute
Harvard University
Indiana University
Iowa State University
Johns Hopkins University
Kansas State University
Lafayette College
Universidad Complutense de Madrid
University of Alabama
University of Alberta
University of Arizona
University of British Columbia
University of Calgary
Syracuse University
Temple University
Texas A&M University
Texas Tech University

Tufts University
University of Notre Dame
University of Oklahoma
Library of Congress
Massachusetts Institute of Technology
McGill University
Michigan State University
Montana State University
Mount Holyoke College
New York Public Library
New York University
North Carolina Central University
North Carolina State University
Northeastern University
Northwestern University
Ohio State University
Oklahoma State University
Pennsylvania State University
Princeton University
Purdue University
Rutgers University
Stanford University
University of California, Berkeley
Davis
Irvine
Los Angeles
Merced
Riverside
San Diego
San Francisco
Santa Barbara
Santa Cruz
University of Chicago

University of Connecticut
University of Delaware
University of Houston
University of Illinois, Chicago
University of Illinois, Urbana Champaign
University of Iowa
University of Kansas
University of Maine
University of Maryland
University of Massachusetts, Amherst
University of Miami
University of Michigan
University of Minnesota
University of Missouri
University of Nebraska-Lincoln
University of New Mexico
University of North Carolina, Chapel Hill
University of Pennsylvania
University of Pittsburgh
University of Queensland
University of Tennessee, Knoxville
University of Texas System
University of Utah
University of Vermont
University of Virginia
University of Washington
University of Wisconsin-Madison
Utah State University
Vanderbilt University
Virginia Tech
Wake Forest University
Washington University, St. Louis
Yale University

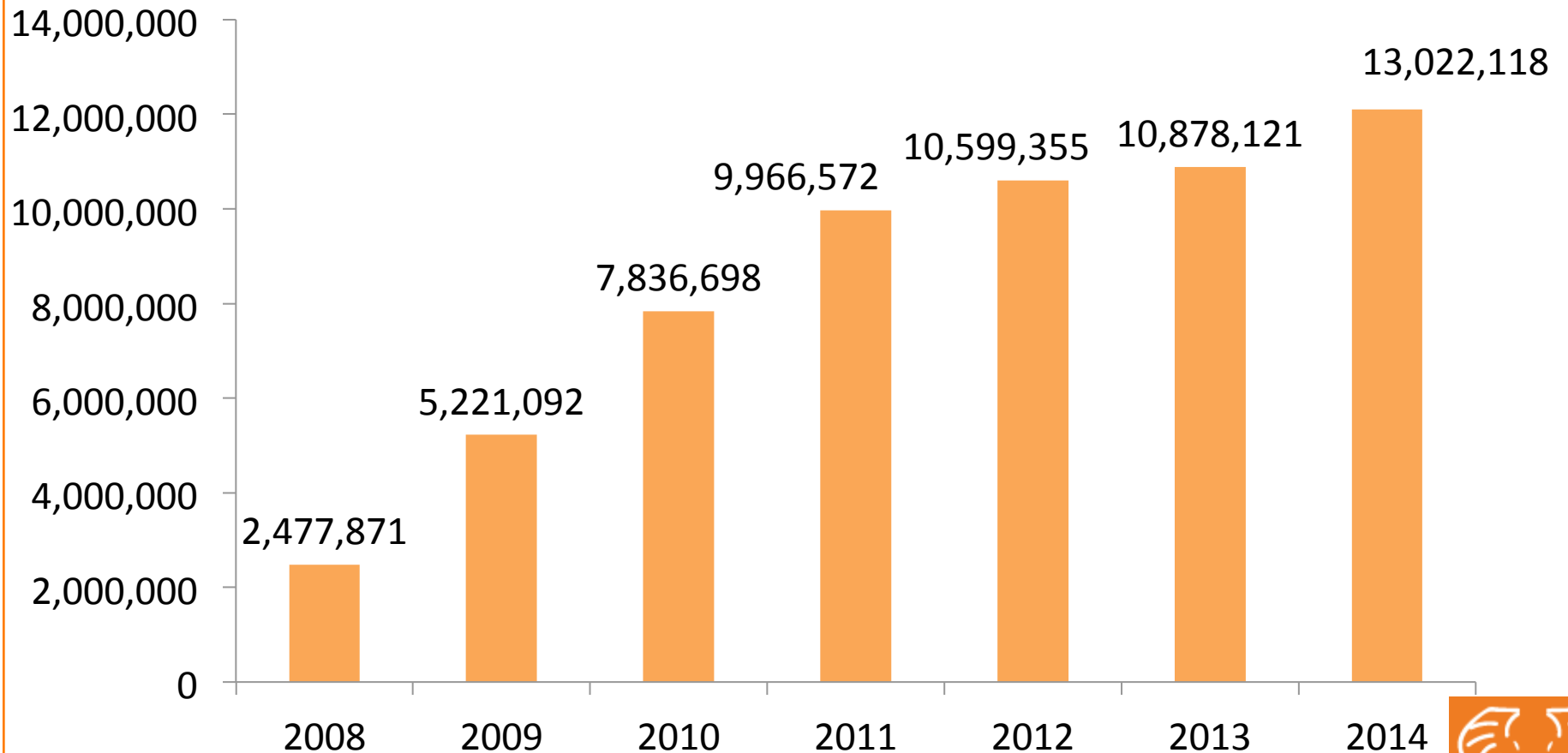


The Name

- The meaning behind the name
 - Hathi (hah-tee)--Hindi for elephant
 - Big, strong
 - Never forgets, wise
 - Secure
 - Trustworthy



Growth of the Collection

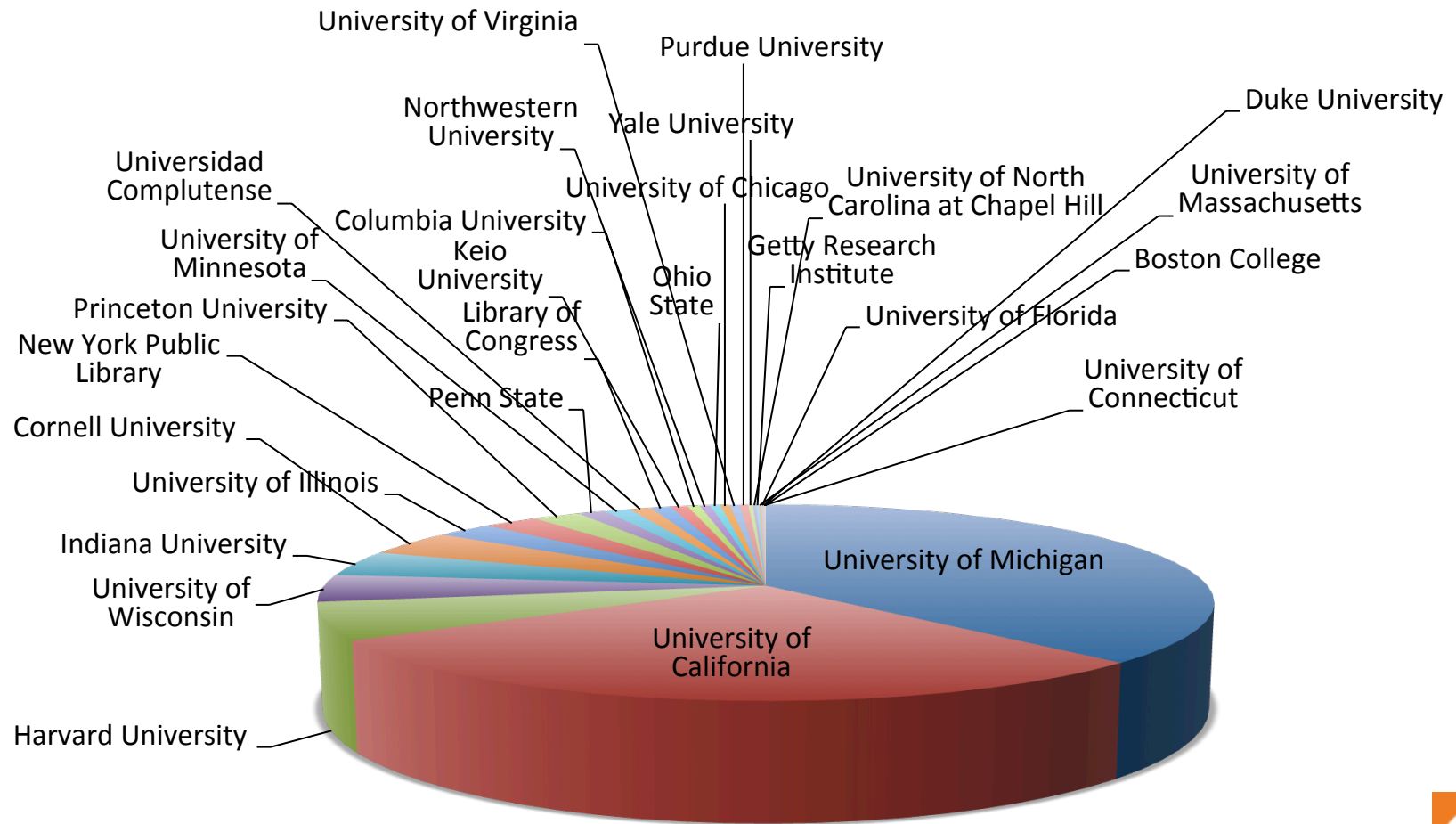


Total Numbers

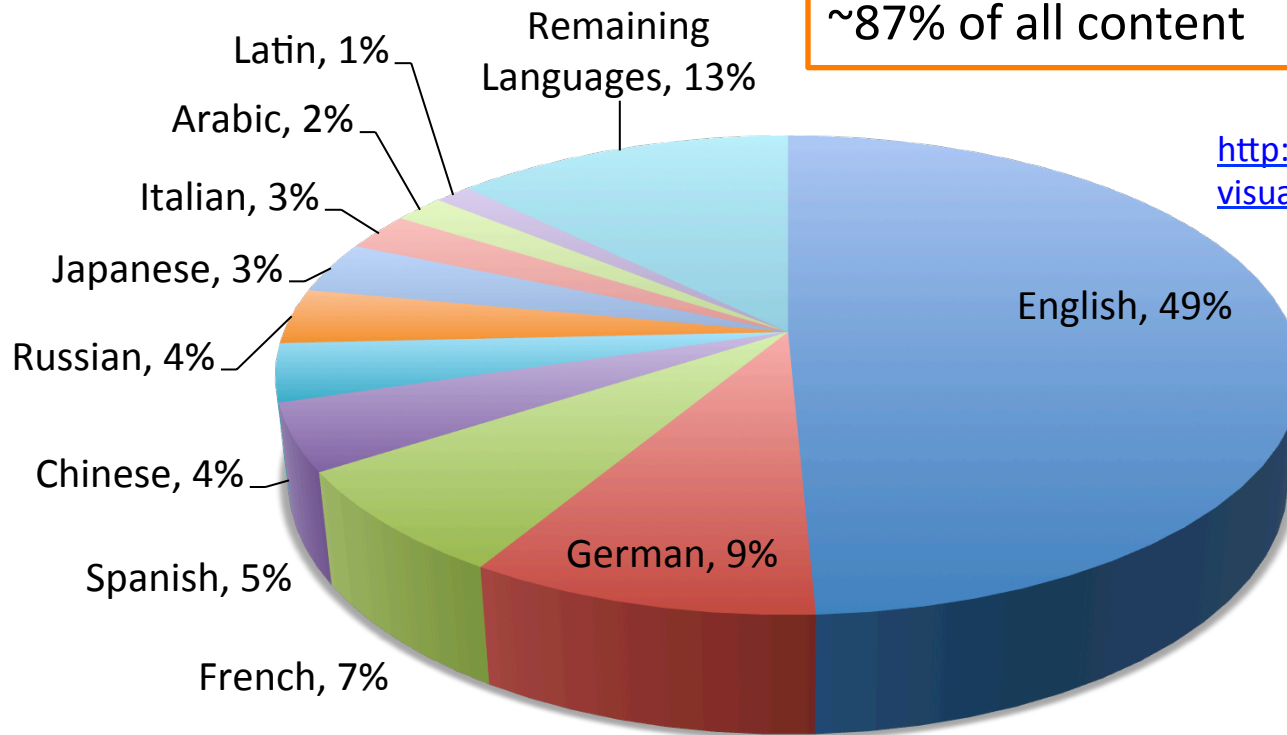
- 13 million total volumes
- 6.6 million book titles
- 340,000 serial titles
- 4.8 million volumes in the public domain (~37%)
- 576,000+ US government publications



Content Providers



Language Distribution (1)



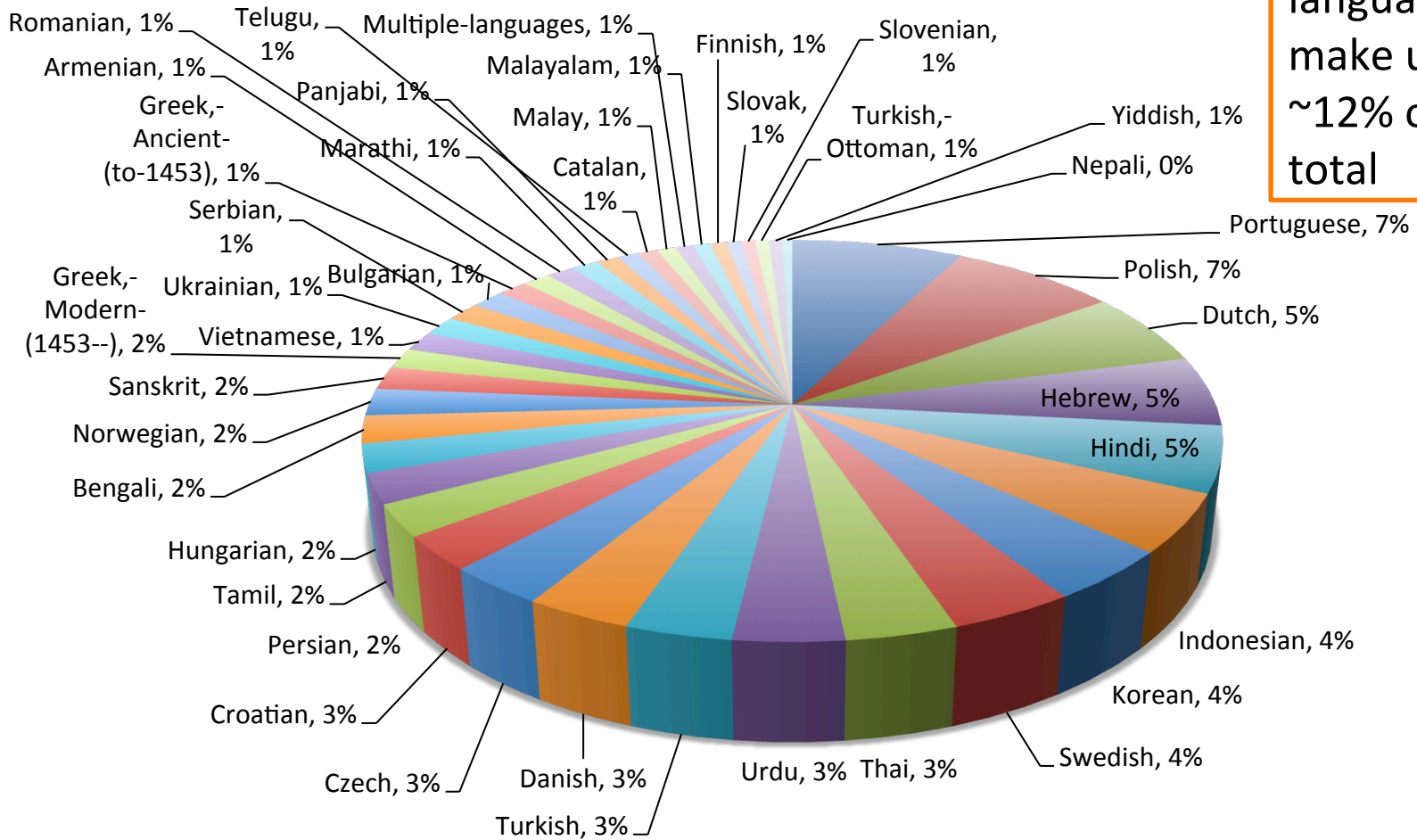
The top 10 languages make up ~87% of all content

http://www.hathitrust.org/visualizations_languages

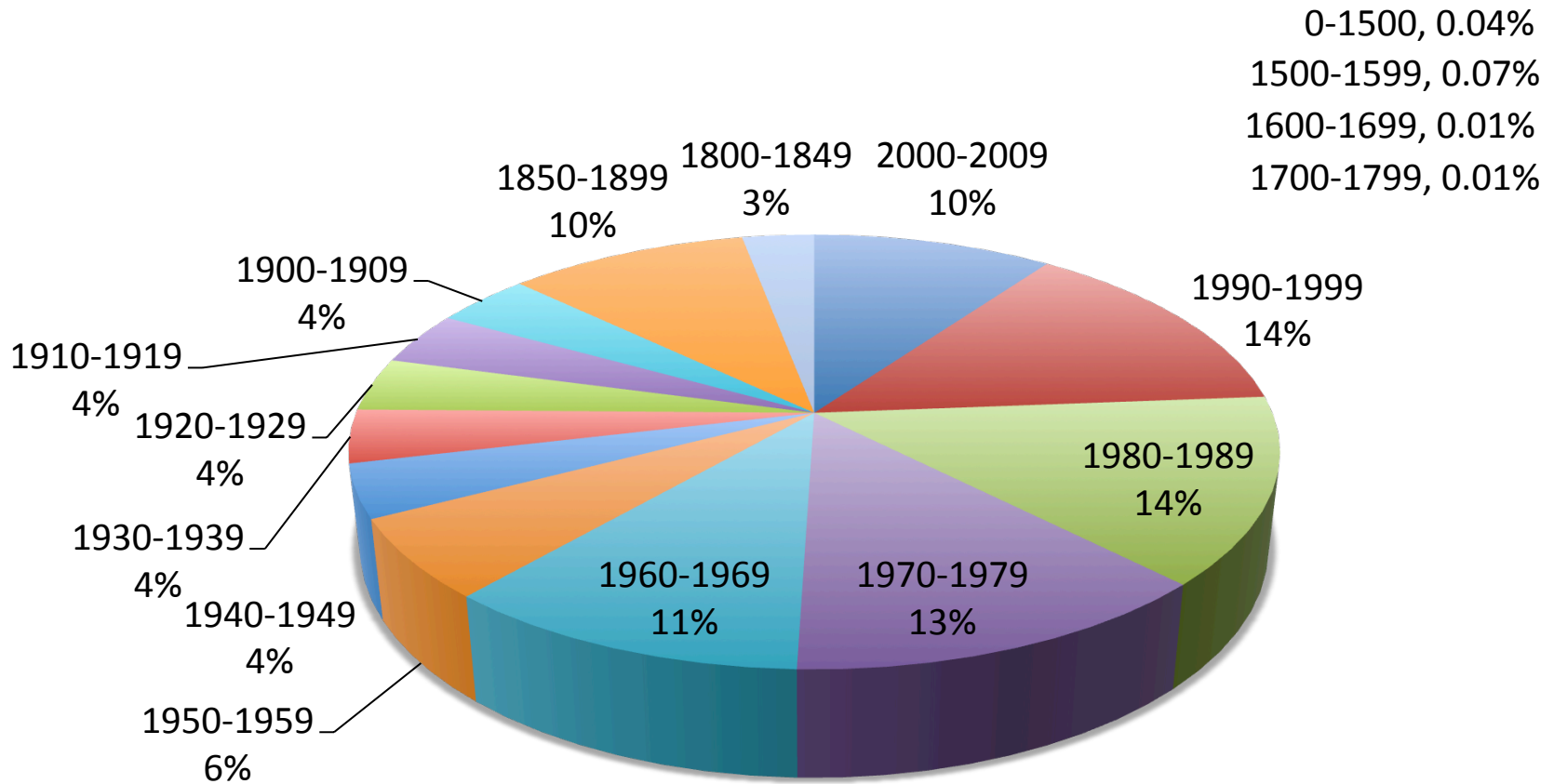


Language Distribution (2)

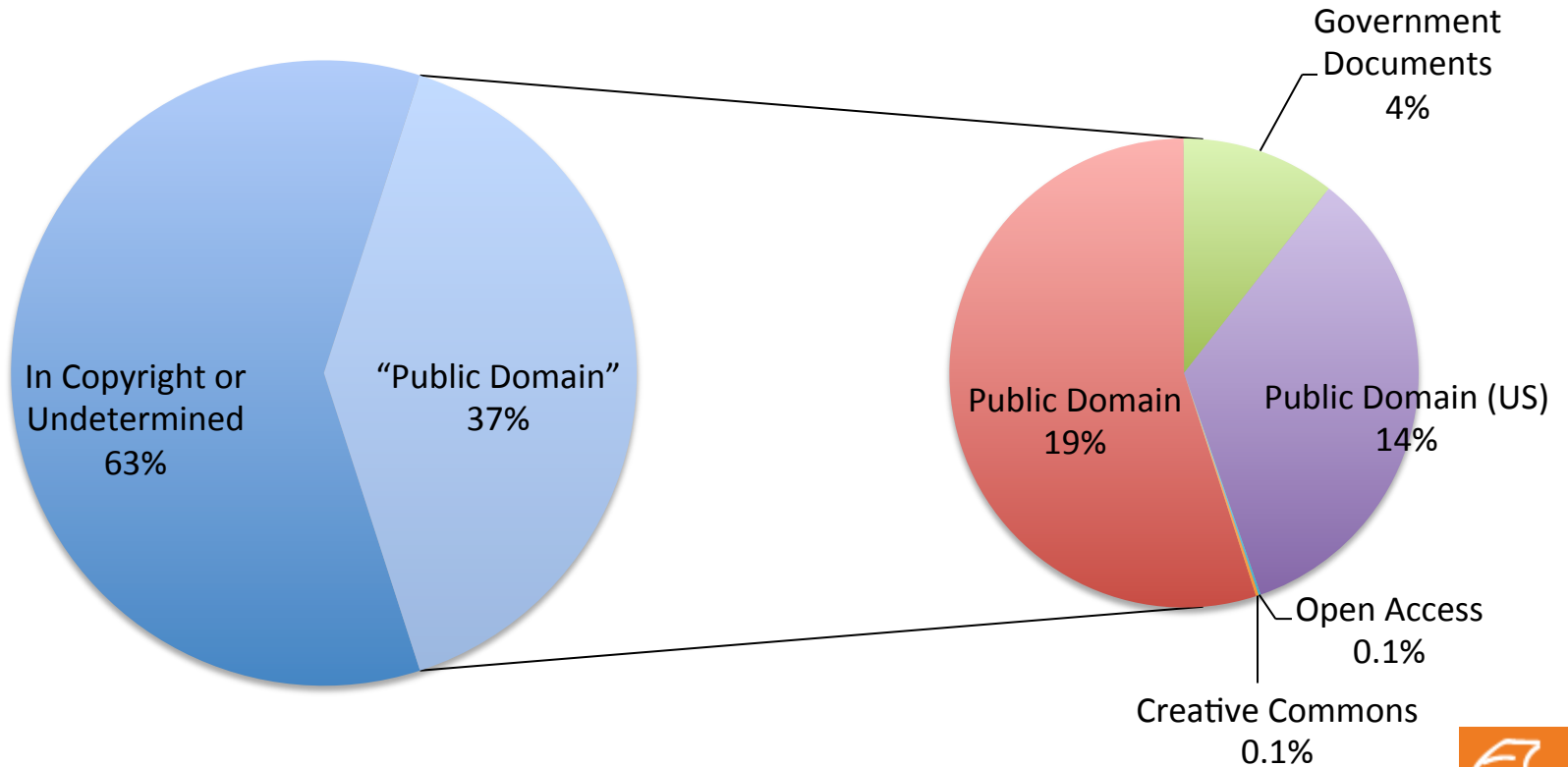
The next 40 languages make up ~12% of total



Dates



Content Distribution



Core Functionality

- Preservation functions – validation, error-checking
- Discovery
- Reading/Access
- Collections
- APIs
- Data Mining/Analysis





HATHI
TRUST
Digital Library

HATHITRUST.ORG

[LOG IN](#) ▾

Search HathiTrust's digital library

[FULL-TEXT](#)[CATALOG](#)[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? »](#)

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*



Refine Results

Subject

- [Printing](#) (14)
- [Printing, Public](#) (11)
- [Authorship](#) (8)
- [Printing Style manuals](#) (8)
- [Printing, Public United States](#) (7)
- [Government publications](#) (5)
- [United States](#) (5)
- [Authorship Handbooks, manuals, etc](#) (4)
- [Authorship Style manuals](#) (4)
- [Printing Specimens](#) (3)
- [Proofreading](#) (3)
- [Type and type-founding](#) (3)
- [Typesetting](#) (3)
- [Government publications](#)
- [United States Statistics Periodicals](#) (2)
- [Printing, Public United States Accounting Statistics Periodicals](#) (2)
- [Type and type-founding Specimens](#) (2)
- [United States, Government Printing Office Accounting Statistics Periodicals](#) (2)

Showing 1 - 20 of 29 Results for all fields:public printer AND author:government printing office

All items (29) Only full view (29)

Sort Relevance

1 2 Next »



Public printer's annual report

by United States. Government Printing Office.
Published 2003

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



Annual report of the Public Printer.

by United States. Government Printing Office.
Published 1900

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



Style manual. Issued by the public printer under authority of section 51 of an act of Congress approved January 12, 1895. November 1935.

by United States. Government Printing Office.
Published 1935

[Catalog Record](#) (view record to see multiple volumes)



Annual report.

by United States. Government Printing Office. United States. Superintendent of Documents.
Published 1853

[Catalog Record](#) (view record to see multiple volumes)

Full text and catalog search

About this Book

Annual report of the Public Printer. 1910/11.

[View full catalog record](#)

Copyright: Public Domain, Google-digitized.

Get this Book

- [Find in a library](#)
- [Download this page \(PDF\)](#)
- [Download whole book \(PDF\)](#)

Add to Collection

This item is not in any of your collections

Select Collection

Add

Share

Permanent link to this book

<http://hdl.handle.net/2027/coo>

Link to this page

<http://hdl.handle.net/2027/coo>

[Embed this book](#)

Version: 2013-02-21 10:31 UTC



THE WORK OF THE FISCAL YEAR.

The following statement shows the larger items of production for the fiscal years ended June 30, 1910 and 1911:

Class of work.	1910	1911
Ems of type set.....	1,801,010,900	1,881,721,600
Hours of time-work in composing rooms.....	¹ 274,913	292,035
Foundry square-inch production.....	¹ 12,685,357	12,419,416
Chargeable impressions of presswork, not including postal-card or money-order presses.....	836,181,344	890,592,096
Number of forms sent to press.....	172,039	171,410
Number of money-order books registered and mailed.....	399,556	448,139
Signatures gathered on machine.....	88,157,436	91,533,833
Signatures folded on machines.....	63,812,025	76,496,216
Number of copies wire-stitched.....	21,631,448	23,409,410
Number of sheets passing through ruling machines.....	22,848,486	25,206,667
Number of signatures sewed on machines.....	72,711,576	83,075,988
Number of tablets made.....	1,417,521	1,900,421
Number of cards and sheets punched and drilled.....	17,128,306	19,265,490
Number of cases made on machines.....	¹ 1,936,691	2,014,514
Number of postal cards produced.....		1,280,895,840

¹ Represent 10 months' actual production and 2 months, approximate or average production.

The expenditure for labor, exclusive of the salaries in the offices of the Public Printer and the superintendent of documents, was \$60,790.46 less in 1911 than in the preceding fiscal year.

The saving for the fiscal year was effected as the result of greater efficiency on the part of the employees and improvement in method and equipment. For example, there was an increase in the chargeable impressions of presswork of 54,410,752 and a decrease of 629 in the number of forms sent to press.

Read on-screen or download

American Guide Series

Owner

sooty

Description

Books and pamphlets about the U.S. published under the auspices of the Federal Writers' Project (FWP)





Status

public

Search in this collection

Sort by: 25 per page Previous [1](#) [2](#) [3](#) [4](#) [5](#) [6](#) ... [9](#) [10](#) [Next](#)

Select all on page

-  **3 hikes thru the Wissahickon, compiled by the Federal Writers' Project, Works Progress Administration.**
by Federal Writers' Project (Pa.)
Published 1936
[Catalog Record](#) [Full view](#) **In my collections:** ---
-  **Alabama; a guide to the deep South, compiled by workers of the Writers' Program of the Work Projects Administration in the State of Alabama ... Sponsored by the Alabama State Planning Commission.**
by Writers' Program (Ala.)
Published 1941
[Catalog Record](#) [Full view](#) **In my collections:** ---
-  **Alabama; a guide to the Deep South.**
by Writers' Program (Ala.)
Published 1949
[Catalog Record](#) [Full view](#) **In my collections:** ---
-  **Arizona, the Grand Canyon State; a State guide, compiled by workers of the Writers' Program of the Work Projects Administration in the State of Arizona. Completely rev. by Joseph Miller; edited by Henry G. Alsberg.**
In my collections: ---

Build and share a collection

APIs

- Bibliographic API
 - Volume and rights information
 - MARC records
 - http://www.hathitrust.org/bib_api
- OAI
 - <http://www.hathitrust.org/data>
- “Hathifiles”
 - <http://www.hathitrust.org/hathifiles>
- Data API
 - Volume and rights information
 - Page images
 - OCR
 - http://www.hathitrust.org/data_api



View:    

public printer government printing office



Search Results

Save Share

Your search for **public printer government printing office** returned 72 results.

Items per page: 10

Sort by: Relevance

1 2 3 ... 7 8

Refine

By Format

text 67
image 3

Contributing Institution

University of California 14
University of Michigan 11
Purdue University 11
United States Government Printing Office (GPO) 7

TEXT

Annual report of the Public Printer

United States. Government Printing Office

Includes [1st]-26th, and [32d] annual report of the Superintendent of Documents for the years 1895-1920, 1927 (also issued separately). Publication suspended July 1939-June 1946. 1921-28.

[View Object](#) 



Link takes you to HathiTrust



TEXT

Style manual. Issued by the public printer under authority of section 51 of an act of Congress approved January 12, 1895. November 1935

United States. Government Printing Office

At head of title: United States Government printing office.



Records loaded into DPLA, local library catalogs, and commercial databases

Data Mining/Analysis

- Dataset Distribution
 - <http://www.hathitrust.org/datasets>
- Research Center



Examples of uses

- Oxford English Dictionary research

@bgzimmer Ben Zimmer 7/4/11

@armavirumque Problem is "cut the mustard" (OED 1891) predates "muster." Earliest I've seen for "muster" is 1912.<http://bit.ly/kOy3aD>

- Thesis research
- Islamic Manuscripts
- Local/Family History



Projects (1)

- Burton, Vernon. “The South as ‘Other,’ the Southerner as ‘Stranger.’”
 - Explore how attitudes expressed in print about slavery, southerners, and non-southerners have changed over both time and space.
- Ted Underwood, Associate Professor of English at the University of Illinois, Urbana-Champaign.
 - Using public domain texts received from HathiTrust to explore changing relationships in literary genres from 1700-1899.
- Andrew Piper, Associate professor of German literature at McGill University.
 - Analyzing linguistic patterns in German texts from 1700-1900



Projects (2)

- Amanda Watson, librarian at New York University.
 - Studying How poetry anthologies in selected texts reflect the rise and fall of poets' reputations over the course of the 19th century.
- Glenn Worthey, Digital Humanities Librarian at Stanford University Libraries.
 - Performing spatio-temporal investigation into the history of Brazilian Portuguese, to be accomplished by text-mining methods (n-gram analysis, etc.).
- Matthew Wilkens, Assistant professor of English, University of Notre Dame.
 - American Council of Learned Societies (ACLS) fellowship for project “Literary Geography at Scale.”



How to find out more

- About: <http://www.hathitrust.org/about>
- Resources: <http://www.hathitrust.org/resources>
- Twitter: <http://twitter.com/hathitrust>
- Facebook: <http://www.facebook.com/hathitrust>
- Monthly newsletter:
 - <http://www.hathitrust.org/updates>
 - RSS http://www.hathitrust.org/updates_rss
- Contact us: feedback@issues.hathitrust.org
- Blogs: <http://www.hathitrust.org/blogs>
 - Large-scale Search
 - Perspectives from HathiTrust



Introduction to the HathiTrust Research Center (HTRC)

Sayan Bhattacharyya



With grateful acknowledgements to: Harriett Green, Erica Parker, Loretta Auvil, Boris Capitanu, Ted Underwood, Peter Organisciak, and other members of the Illinois and Indiana HTRC teams

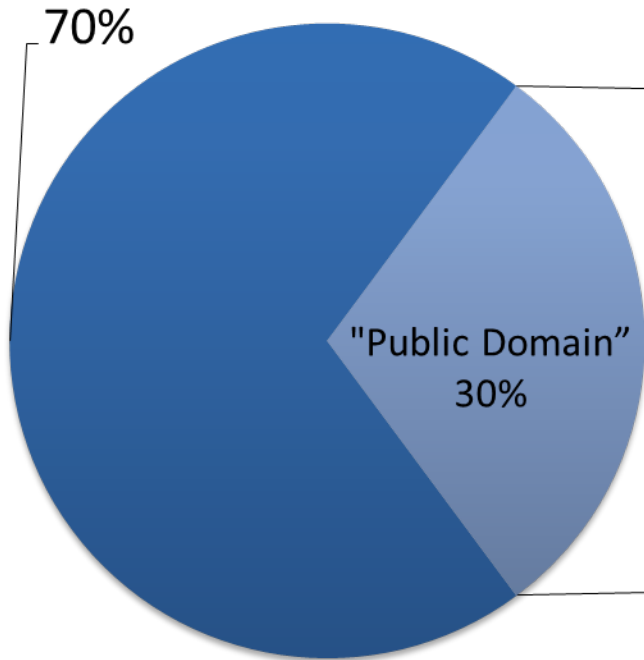
Workshop Outline

- **Overview** of the HathiTrust and the HathiTrust Research Center
- **How to use the HTRC Portal**
 - *Create/manage your own custom set of HathiTrust materials*
 - How to use HTRC Portal Workset Builder
 - *Textual analytics on a workset:*
 - How to use HTRC Portal Algorithms, HTRC Bookworm, Data Capsule
- **Opportunities** to connect you and your research with the HathiTrust Research Center

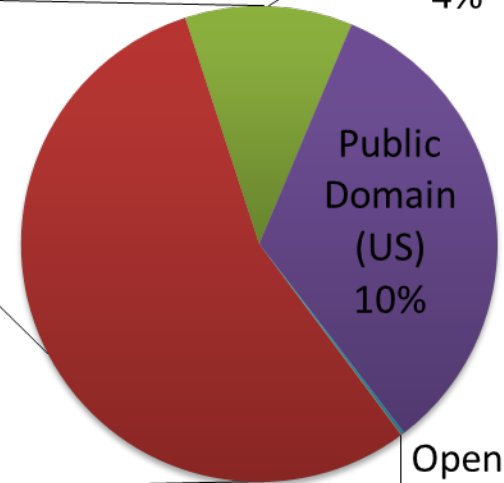


Content Distribution

In-copyright or
undetermined



Public Domain
(worldwide)
15%



U.S. Federal
Government
Documents
(worldwide)
4%

Creative Commons
.01%

Open Access
.1%



HathiTrust Collection Builder

out Collections Help Feedback Hi Megan Finn Senseney! My Collections

HATHI TRUST Digital Library

FULL-TEXT CATALOG

Search [input] [magnifying glass icon]

Full view only

Collection Name 100

Description 255

Private Public

Cancel Add

Collection can be searched

[Create new collection](#)

Find a collection [input]

Recently Updated [input] (all items) [dropdown]

Collection Title [dropdown]

Showing 1-50 of 1468 of all collections

Previous 1 2 3 4 ... 30 Next


2 items
last updated: 10/11/10

erations'

r: quoddy

Featured Collection

[Records of the American Colonies](#)



Published documents--leg court proceedings, records correspondence, etc.--from original colonies and their predecessors.



HTRC Portal



HTRC Portal

[Home](#)

[About](#)

[Worksets](#)

[Algorithms](#)

[Results](#)

[Help](#)

[Login](#)



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](#)

[About](#) | [Contact](#)



www.hathitrust.org/htrc

Overview

The HathiTrust Research Center (HTRC) enables computational access for nonprofit and educational users to published works in the public domain and, in the future, on limited terms to works in-copyright from the [HathiTrust](#).

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.

Leveraging data storage and computational infrastructure at Indiana University and the University of Illinois at Urbana-Champaign, the HTRC will provision a secure computational and data environment for scholars to perform research using the HathiTrust Digital Library. The center will break new ground in the areas of text mining and non-consumptive research, allowing scholars to fully utilize content of the HathiTrust Library while preventing intellectual property misuse within the confines of current U.S. copyright law.

Quick Links

- [HTRC Portal and Work Set Builder](#)
- [User Community](#)

Get Involved

- htrc-announce-l@list.indiana.edu General announcements about HTRC workshops, updates, new tools, and larger community issues.
- htrc-usergroup-l@list.indiana.edu Submit recommendations, development issues, technical discussion about HTRC.
- htrc-uncamp-l@list.indiana.edu Logistics and Announcements specific to HTRC UnCamp.

Learn More About

- [Governance](#)
- [Architecture and Organization](#)
- [Access and Use](#)
- [Timeline and Deliverables](#)
- [Partnering in Research](#)
- [Collections and Tools](#)



Log in to the HTRC Portal, <https://htrc2.pti.indiana.edu>



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](#)



Create a login id (i.e. username)



production-portal requests access to default;

Authorize Deny

Username:

Password:



How to create a workset



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



Log In Again to Workset Builder



WSO₂ Identity Server

production-portal requests access to default;

Authorize

Deny

Username:

Password:

Login

Cancel



Workset Builder

(currently works only on non-copyrighted material not digitized by Google)



HTRC Workset Builder

[Log Out](#) | [Home](#) | [Green](#) | [Selected Items](#) | [Worksets](#) | [Portals](#)

Limit your search

[Subject](#)

[Author](#)

[Language](#)

[Place of Publication](#)

[Year](#)

[Original Location](#)

Successfully authenticated from HTRC account.

in [Full Text](#)

[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](#)



Why worksets?

- Two reasons:
 - Organizational:
 - The “virtual study carrel” idea:
 - Gather together material of interest to *you in one place*
 - Accomplished by “slicing and dicing” using search and metadata criteria
 - Algorithmic:
 - Delimitation of the “scope” of the analysis
 - You don’t want to run your analysis on the whole of the HathiTrust collection — you want to run it only on material that is interesting/relevant to you



Search (on metadata and on full text) as means for building a workset



HTRC Workset Builder

[Help](#) | [Home](#) | [Advanced Search](#) | [Storage Accounts](#) | [Print](#)

Search tips

- Select "match all" to require all fields.
- Select "match any" to find at least one field.
- Combine keywords and attributes to find specific items.
- Use quotation marks to search as a phrase.
- Use "*" before a term to make it wildcard. (Otherwise, results matching only some of your terms may be included).
- Use "-" before a word or phrase to exclude.
- Use "OR", "AND", and "NOT" to create complex boolean logic. You can use parentheses in your complex expressions.
- Truncation and wildcards are not supported - word-stemming is done automatically.

More Search Options

Find items that match: all of the fields below:

Full Text:
france AND war

Title:

Author:

Subject:

Publish Date:

AND have these attributes:

Sort results by: relevance

Clear all Search



Select desired items

Limit your search

Subject
Author
Language
Place of Publication
Year
Original Location

All items on this page were successfully selected.

In Full Text Search

More options

Full Text: france AND war

Displaying items 1 - 10 of 1,174,041 [Start over](#)

Sort by: relevance Show 10 per page

Previous 1 2 3 4 5 117,484 117,485 Next

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

- List of Etonians who fought in the great war, 1914-1919.** [Select](#)
Title: List of Etonians who fought in the great war, 1914-1919
Author: Eton College., Vaughan, Edward L. (Eton)
Language: English
Published: 1921
- List of Etonians who fought in the great war, 1914-1919.** [Select](#)
Title: List of Etonians who fought in the great war, 1914-1919
Author: Eton College., Vaughan, Edward L. (Eton)
Language: English
Published: 1921
- What France did for us, by Edmund Hacket, Major, U.S.A. Pub. by La France, and American magazine.** [Select](#)
Title: What France did for us, by Edmund Hacket, Major, U.S.A. Pub. by La France, and American magazine
Author: Hacket, Edmund Francis, 1880-
Language: English
Published: 1920
- Finances and costs of the present European war. Prepared by the War college division, General staff corps, as a supplement to the Statement of a proper military policy for the United States. Army war college: Washington, November, 1915.** [Select](#)
Title: Finances and costs of the present European war. Prepared by the War college division, General staff corps, as a supplement to the Statement of a proper military policy for the United States. Army war college. Washington, November, 1915.



Put them in a workset

Selected Items

Sort by **relevance** ▾

Show **10** ▾ per page

[Create/Update Workset](#) [Clear all](#)

1. Finances and costs of the present European war. Prepared by the War college division, General staff corps, as a supplement to the Statement of a proper military policy for the United States. Army war college: Washington, November, 1915.

Selected

Title: Finances and costs of the present European war. Prepared by the War college division, General staff corps, as a supplement to the Statement of a proper military policy for the United States. Army war college: Washington, November, 1915.

Author: United States. War Dept. General Staff., United States. Army war college, Washington, D.C.

Language: English

Published: 1916

2. Finances and costs of the present European war. Prepared by the War college division, General staff corps, as a supplement to the Statement of a proper military policy for the United States. Army war college: Washington, November, 1915.

Selected

Title: Finances and costs of the present European war. Prepared by the War college division, General staff corps, as a supplement to the Statement of



Analysis in the HTRC Portal

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

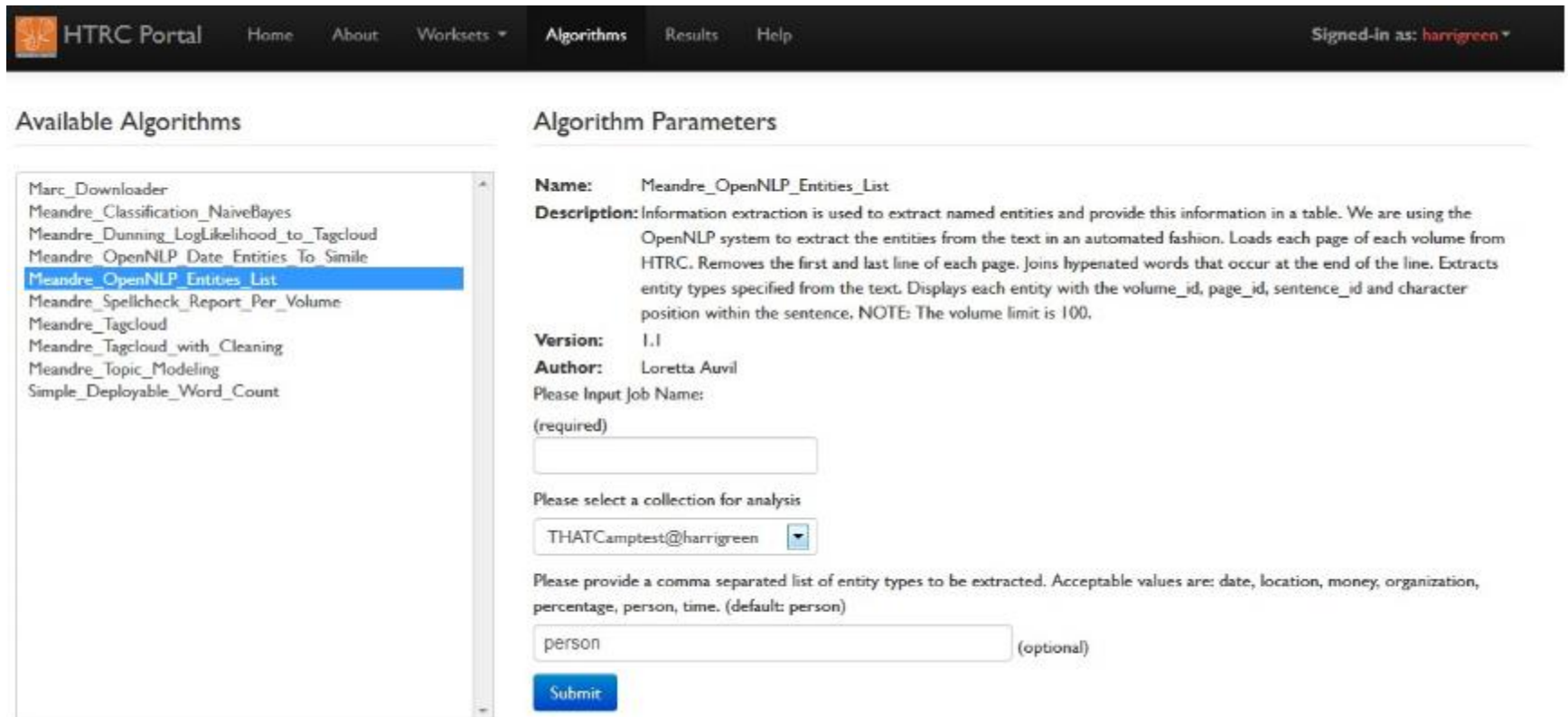
Algorithm Parameters

Please select an algorithm in the list to display information.



Choose Algorithm

Note: Enter a name of your choosing in the blank field for “Job Name.” This is the same name that will show up later as “Job Title” when looking at the results.



The screenshot shows the HTRC Portal interface. At the top, there is a navigation bar with the following items: HTRC Portal (with a logo), Home, About, Worksets, Algorithms (highlighted), Results, and Help. On the right side of the navigation bar, it says "Signed-in as: harrigreen".

The main content area is divided into two columns:

- Available Algorithms:** A list of algorithms is shown in a scrollable box. The selected algorithm is "Meandre_OpenNLP_Entities_List". Other algorithms include Marc_Downloader, Meandre_Classification_NaiveBayes, Meandre_Dunning_LogLikelihood_to_Tagcloud, Meandre_OpenNLP_Date_Entities_To_Simile, Meandre_Spellcheck_Report_Per_Volume, Meandre_Tagcloud, Meandre_Tagcloud_with_Cleaning, Meandre_Topic_Modeling, and Simple_Deployable_Word_Count.
- Algorithm Parameters:** This section provides details for the selected algorithm:
 - Name:** Meandre_OpenNLP_Entities_List
 - Description:** Information extraction is used to extract named entities and provide this information in a table. We are using the OpenNLP system to extract the entities from the text in an automated fashion. 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. Extracts entity types specified from the text. Displays each entity with the volume_id, page_id, sentence_id and character position within the sentence. NOTE: The volume limit is 100.
 - Version:** 1.1
 - Author:** Loretta Auvil
 - Please Input Job Name:** (required) [input field]
 - Please select a collection for analysis:** THATCampTest@harrigreen [dropdown menu]
 - Please provide a comma separated list of entity types to be extracted. Acceptable values are: date, location, money, organization, percentage, person, time. (default: person)** [input field containing "person"] (optional)
 - Submit** [button]



Choose Collection(s) for Analysis

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

THATCampTest@harrigreen
1darwin-test@sheilahoover
2darwin-english@sheilahoover
2vesalius@sheilahoover
Agrippa@rkfritz
Anarchism@rsvarne
AncientGreek@miao
Austen_Dickens_Labels@lauvil
Author_Twain@jthron
BestCoreComplex@luraj
BigLaw@plale
Bleakhouse@sheilahoover
Cicero_Orations_Letters@ajs05r
Coffee_Books@rsvarne
Cornell_HIS_2293_1@eeb36
Dickens_as_Authors@skowalczyk
Dickens_yo@sbmarks
Diderot-test@sheilahoover
Digital_Preservation@skowalczyk
DocSouthMatch@mfall3

THATCampTest@harrigreen

Please provide a comma separated list of entity types to be extracted. Acceptable values are: date, location, money, organization, percentage, person, time. (default: person)

person (optional)

Submit

named entities and provide this information in a table. We are using the from the text in an automated fashion. Loads each page of each volume from each page. Joins hyphenated words that occur at the end of the line. Extracts lays each entity with the volume_id, page_id, sentence_id and character volume limit is 100.



Run the Analysis...

Active Jobs

Cancel

Job Title	Last Updated	Status	Cancel?
Harriettgtest	2014-04-22 13:05:58	Staging	<input type="checkbox"/>

Completed Jobs

There are no completed jobs..



Results!

Active Jobs

✖ Cancel

Job Title	Last Updated	Status	Cancel?
harriettgtest2	2014-04-22 14:01:02	Staging	<input type="checkbox"/>

Completed Jobs

🗑 Delete Selected

💾 Save Selected

Job Title	Last Updated	Status	Delete/Save?	Saved?
Harriettgtest	2014-04-22 13:11:22	Finished	<input type="checkbox"/>	unsaved



View Results

Job Details

Job Title: Harriettgtest

Algorithm Name: Meandre_Tagcloud

Last Updated: 2014-04-22 13:11:22

Results:

[stderr.txt](#)

[stdout.txt](#)

[tagcloudtokencounts.html](#)

[tagcloudtokencounts.csv.txt](#)

Job Parameters:

Name	Value
input_collection	Monster@claireystew

Job Id:

c815b289-aed2-494f-beeb-2952be44579d

Status:

Finished

View Results



Topic modeling vs. Dunning log-likelihood

- Topic modeling is useful when you want to get a sense of the contents of your workset.
- Dunning log-likelihood algorithm is useful when you want to do a focused comparison/contrast between *two* worksets.
 - [If interested in the gory details of how the Dunning log-likelihood algorithm works, see: [this blog post by the researcher Ben Schmidt of Northeastern University](#).
(We won't cover the details today.)]



Making sense of the results

Based on what we know of the plots of these two novels, do the generated results make sense?

Plot summaries (abbreviated, from the Dickens Fellowship website):

Bleak House: A prolonged **law case** concerning the distribution of an estate, which brings misery and ruin to the suitors but great profit to the **lawyers**, is the foundation for this story. Bleak House is the home of John Jarndyce, principal member of the family involved in the **law case**.

Little Dorrit: Here Dickens plays on the theme of **imprisonment**, drawing on his own experience as a boy of visiting his father in a **debtors' prison**. William Dorrit is **locked up for years in that prison**, attended daily by his daughter, Little Dorrit. Her unappreciated self-sacrifice comes to the attention of Arthur Clennam, recently returned from China, who helps bring about her father's release but is **himself incarcerated for a time**.



Topic modeling *Bleak House*

- For reference, here is a partial snapshot of the results of topic modeling *Bleak House* (number of tokens/topic = 20):



Using the Portal for Research/ Teaching

- Two common use cases:
 - Topic modeling
 - If you have a set of books and want to see what common themes run through them, run topic modeling on them
 - Dunning log-likelihood
 - If you have *two* sets of books and want to compare and contrast them
- Both these algorithms are available on the portal for use with worksets.



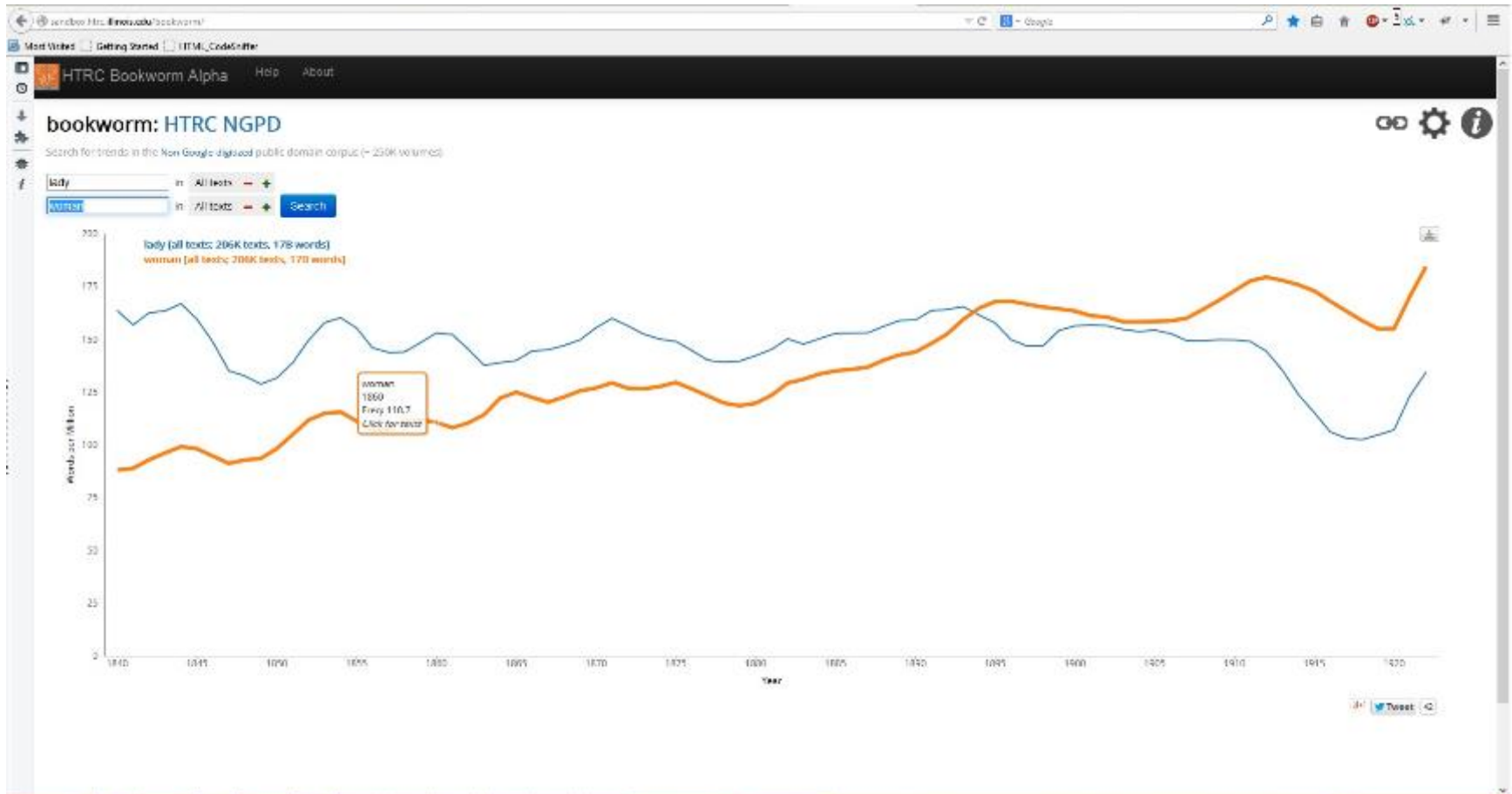
Are these algorithms positivist?

- No.
 - Depending on what parameters are selected when running the algorithms, you will get very different results.
 - There is no assumption that there is knowledge “out there” that is being “discovered”
 - Knowledge is *constructed* out of text by the algorithms, which are a form of subjectivity.
- Currently you cannot change/tune parameters on the portal, but in the future, you may be able to.



The Future: HTRC Bookworm

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



Use for trend analysis: like Google N-gram Viewer, but using metadata to allow analysis on focused worksets.

Looking into the future

- Tools for non-consumptive text analysis on copyrighted texts:
 - HTRC Data Capsule:
<https://wiki.htrc.illinois.edu/display/COM/HTRC+Data+Capsule>
 - Extracted Features:
<https://sandbox.htrc.illinois.edu/HTRC-UI-Portal2/FeatureAction>
- HathiTrust + Bookworm Project:
<https://htrcbookworm.wordpress.com/>
- Workset Creation for Scholarly Analysis (WCSA) study:
<http://worksets.htrc.illinois.edu/worksets/>
- User guides developed at <http://uiuc.libguides.com/htrcguide>



Non-consumptive reading via Feature Extraction

*What can you do with (1)1 million **books**...
...if they are under copyright?*

*You can read book **fragments** at the HathiTrust
Research Center via Feature Extraction...*

The Hathi Trust's (1)1 million books

- ❖ abstracted, bird's-eye view of many texts in aggregated form
- ❖ offers possibility for making sense of mankind's cultural legacy
 - in the form of digitized text
 - available through the world's great research libraries

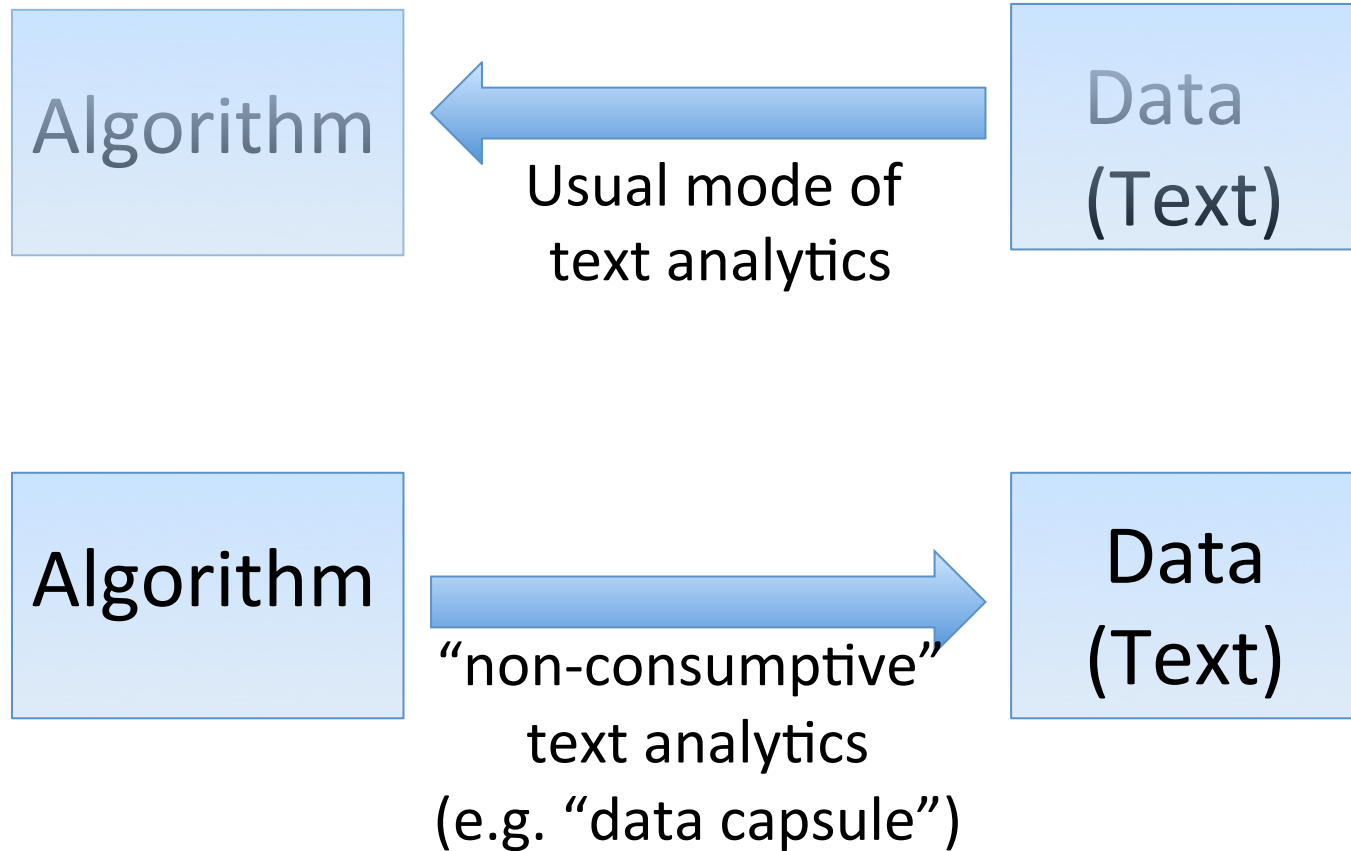
But...

- ❖ constrained by legal issues related to intellectual property:
 - not all material is in the public domain
 - ❖ direct consumption of non-public-domain material is prohibited

Solution: Textual processing
without downloading of text

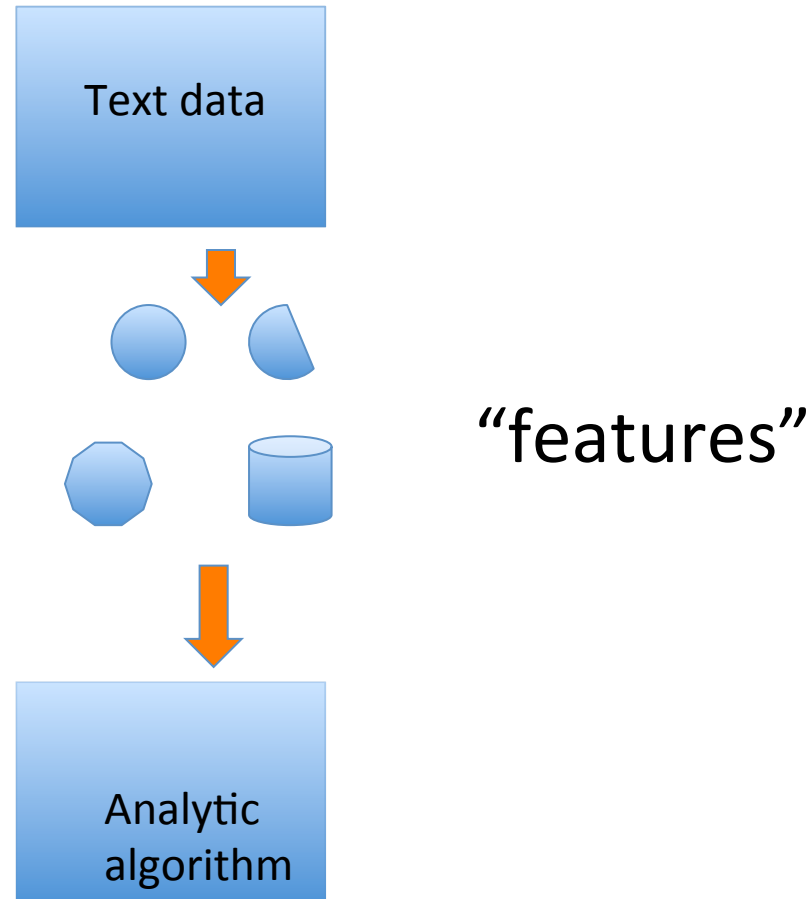
- How can this be accomplished?
 - “Non-consumptive” reading
 - Instead of bringing the data to the algorithm:
 - “Data capsule” approach
 - » bring the algorithm to the text
 - and/or**
 - “Features” approach
 - » bring certain relevant features of the text to the algorithm

Different modes of distant reading

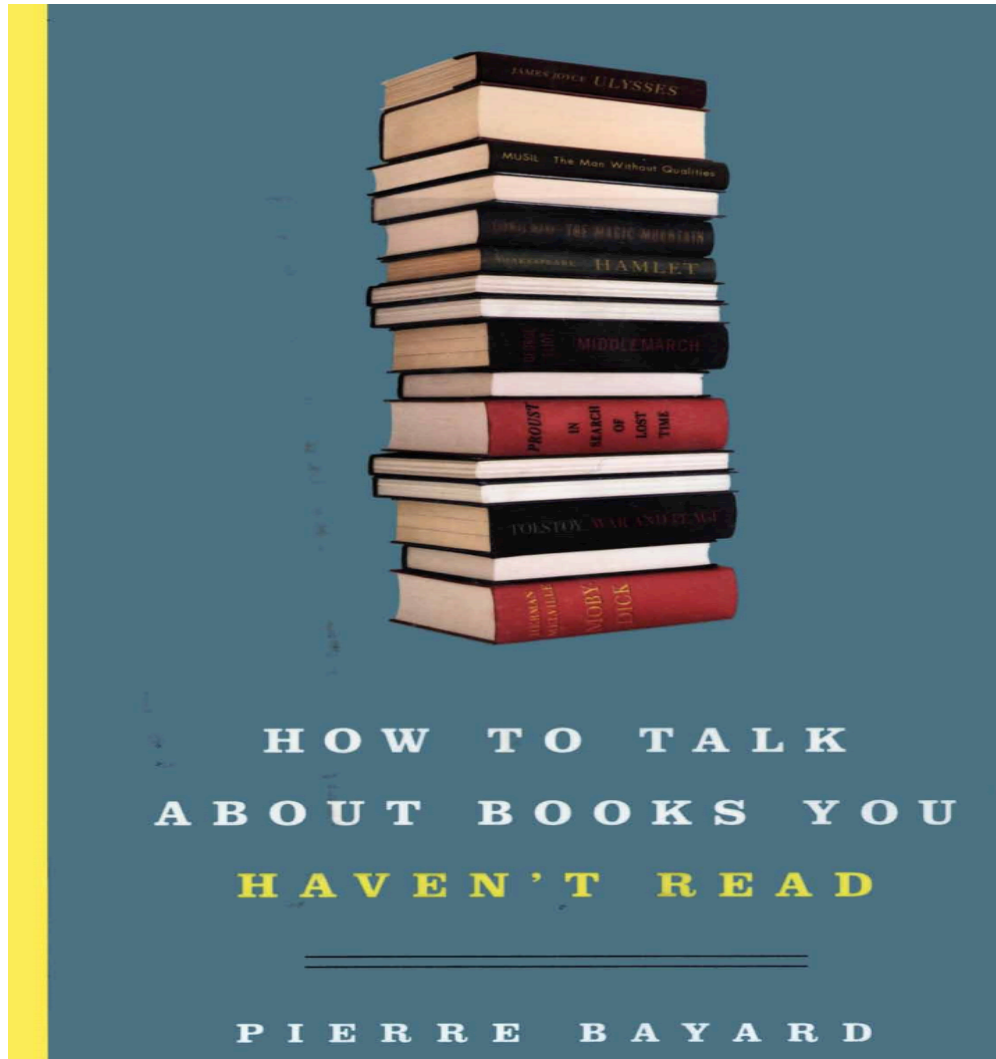


Different modes of distant reading (contd.)

Textual processing using “features”:



How to talk about books you haven't read!



Feature-extraction Motivation

Motivation:

- Almost of all post-1923 publications under copyright
- No bulk downloads of non-copyrighted material permitted



Need for allowing textual analytics without necessitating downloading of full text

Features: What they are

- New HTRC service in alpha release
 - (pilot: 250,000 volumes)
- A dataset of "features"
 - Features are:
 - *notable or informative characteristics of the text*
 - Features are (mostly) fragments of text

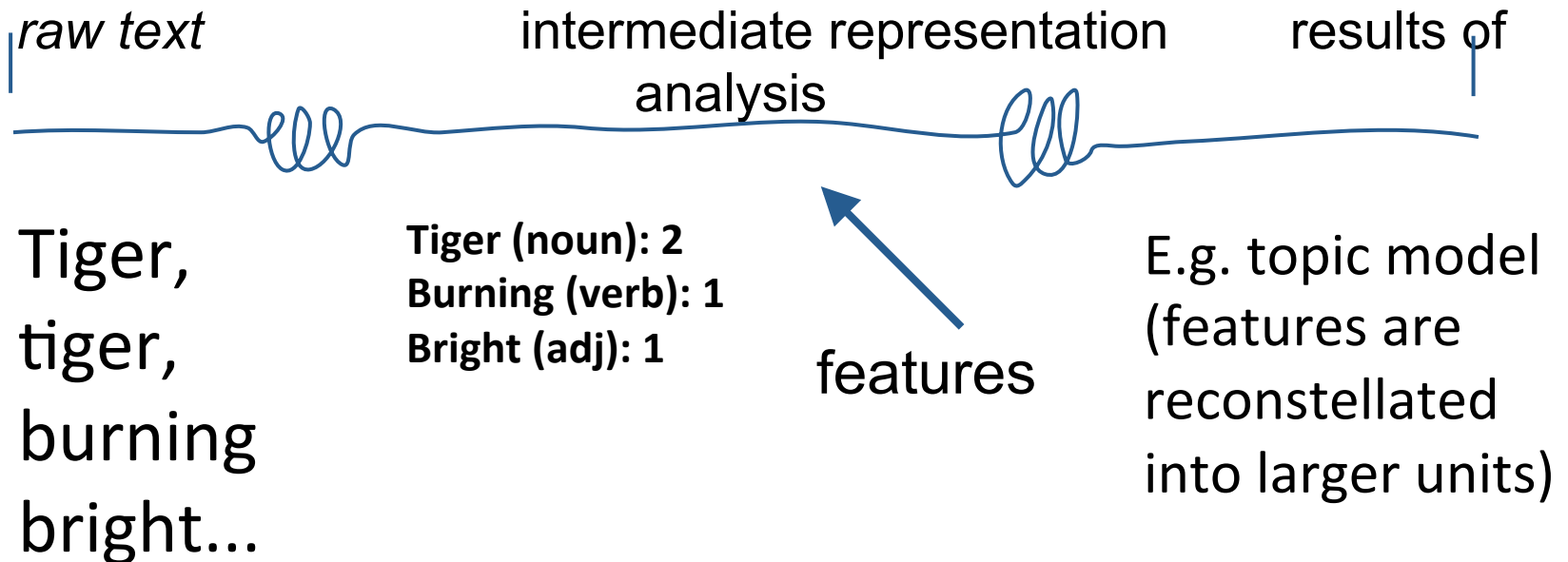
Features currently provided:

- ***Format: Per-page features, packaged (as JSONs) in one file per volume, with page section (header, footer and body) identification***
 - counts of part-of-speech-tagged words
(**bag of words, per-page**, with frequencies)
 - various line-level information:
 - number of lines containing characters of any kind in a page section
 - counts of the initial character and final character of each line in a page section
 - etc.

Feature-extraction

- Features are a “translation” of text
 - from language that humans understand
 - to machine-readable fragments

Text as data as text:



Features

Available at <http://bit.ly/HTRC-Features>

A library for help with reading with features:

github.com/organisciak/
htrc-feature-reader

Ted Underwood in 'Theorizing Research Practices We Forgot to Theorize Twenty Years Ago', *Representations*,
Summer 2014:

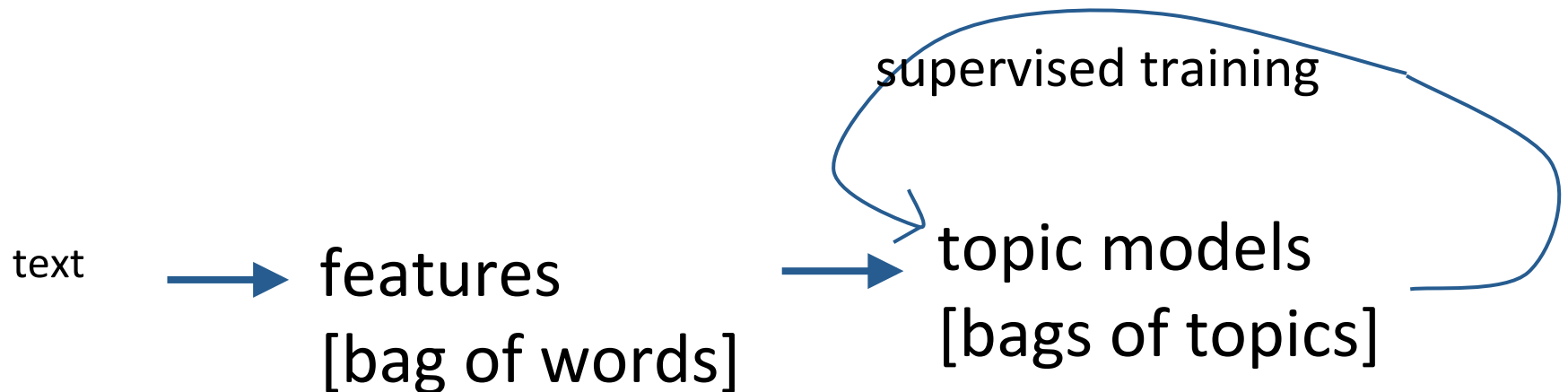
- “Humanists are gearing up to have a conversation about digital research methods; a new kind of interdisciplinary conversation [between humanists and computer scientists is about to begin, one in which] a rare opportunity is emerging for a genuinely productive exchange between scientific methodology and humanistic theory.”

(At least) two opportunities here for “interdisciplinary conversation” about digital research methods

- 1) A conversation about how the hermeneutical cycle can work at the level of a collection (rather than a single text) too large to be surveyed by a single reader
- 2) **“Fragments” as data structures and as anxiety-producing artifacts in the history and future of the humanities**

Hermeneutic cycle for text collections

In *supervised* topic models (e.g. Blei and McAuliffe, 2007), which uses a supervised learning technique, topics generated are shaped by prior assumptions of the modeler (as communicated in the supervised training phase).



Humanism and (reading) fragments

Fragments of text and reading them have fascinated humanists (and caused them anxiety) at least since the early modern period (when, after the Renaissance, “humanism” in its current form first appeared).

Humanist anxiety produced by text fragments (early modern Europe)

(John Donne, from the poem 'An Anatomy of the World',
written in 1611)

“But this were light, did our less volume hold
All the old text; or had we changed to gold
Their silver; or disposed into less glass
Spirits of virtue, which then scattered was...”

(Text and other) fragments (contd.)

“Shall I at least set my lands in order?

London Bridge is falling down falling down falling down...

These fragments I have shored against my ruins

Why then Ile fit you. Hieronymo's mad againe

Datta. Dayadhvam. Damyata...”

(from T.S. Eliot, *The Waste Land*. 1922)

Anxiety for a lost whole, the classical past in text fragments

“Grenfell gets so anxious to recover even scraps
It’s brought the poor chap almost close to a collapse...
He heard Apollo yammering for scraps and tatters
Of some lost Sophoclean play called *The Tracking Satyrs*.”

(Tony Harrison, *The Trackers of Oxyrhynchus: The Delphi Text*,
written in 1990;
about the lost papyri fragments recovered in Oxyrhynchus, Egypt,
in 1912)

Humanist anxiety
produced by
(text and other) fragments
(Augustan England)

“The ruins of Rome provided the humanists with a powerful image of the kind of desolation inevitably wrought by innovation, novelty and wilful change.”

Paul Fussell,
The Rhetorical World of Augustan Humanism,
1965

Posthumanism and reading via fragments

- Today, technological change and innovation appear to many humanists as a similar threat: “*desolation... wrought by innovation.*”
 - ❖ Anxiety about being rendered marginalized
 - ❖ Anxiety about creeping “scientism” and utilitarianism
 - ❖ Anxiety/fear of the digital
 - ❖ Anxiety about a post-humanist/ anti-humanist) future

Text fragments (as a metaphor)
are paradigmatic of much in the
digital humanities

Can computationalism be (neo)humanist?

“Ever since Leibniz, the deployment of one strand of computationalism has been to excise the element of ambiguity which is part of the human experience.”

— David Golumbia, *The Cultural Logic of Computation*, 2009

This need not always be so.

(1) Fragments can be recombined
into ludic, ambiguous,
open-ended wholes

Computationalism as neo-humanism

By unlocking knowledge resources and
accommodating ambiguity, play, and open-
ended interpretation, computationalism can
also be a neo-humanism.

Rosi Braidotti, *The Posthuman* (2013)

(2) Non-consumptive reading is post-humanist:

Interfaces can have agency

Re-imagining discursive practices

“Discursive practices are not human-based activities but rather specific material (re)configurings of the world through which local determinations of boundaries, properties and meanings are differentially enacted.”

— Karen Barad, in ‘Posthumanist Performativity,’
Signs, Vol. 28, no. 3 (2003)

How you can get involved

HathiTrust Research Center Announcements:

htrc-announce-l@list.indiana.edu

HathiTrust Research Center User Group:

htrc-usergroup-l@list.indiana.edu



Resources

Guide to the HTRC Portal:

<http://uiuc.libguides.com/htrcguide>

HathiTrust + Bookworm:

<https://htrcbookworm.wordpress.com/>

Getting started with the HTRC Data Capsule:

<http://bit.ly/1rWHPfH>

Detailed Data Capsule guide: <http://bit.ly/1BzP9O1>

Feature Extraction:

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



More resources:

The HathiTrust Research Center Publications and Presentations page:
<https://wiki.htrc.illinois.edu/display/OUT/HTRC+Publications,+Presentations>

Questions?

Contact the HathiTrust Research Center

Support Team at

htrc-support-1@list.indiana.edu

We are happy to answer your questions
and to help you use our resources!

