

Is the Elephant Singing a Song?

Digital Humanities Research Possibilities
Using HathiTrust and the
HathiTrust Research Center

Joseph Hafner,
Associate Dean, Collection Services
Member, HathiTrust Program Steering
Committee



HathiTrust

- Membership organization
- “Collecting, organizing, preserving, communicating, and sharing the record of human knowledge”
- Large-scale digitization initiative at research libraries



HathiTrust data

- Over 16,000,000 total volumes
- Books and serials
- 751 terabytes
- Over 6,000,000 volumes in the public domain
- [Call Numbers](#) from A-Z
- [A long tail of languages](#): over 450 different languages, ancient and modern, from Aleut to Zulu
 - 50% in English

Stats updated daily at: <https://www.hathitrust.org/about>



HathiTrust Research Center

- Provides computational access and digital research assistance for the HathiTrust corpus.
- Collaborative effort at Indiana University and the University of Illinois, with support from the University of Michigan through the HathiTrust.



HTRC Mission

To provide infrastructure and tools enabling and supporting computational research on the more than 16 million volumes of the HathiTrust collection. HTRC enable scholars to fully utilize content of HathiTrust under existing copyright laws in the U.S.A. and around the world.



Non-Consumptive Research Paradigm

Not research in which a researcher reads or displays substantial portions of an in-copyright or rights-restricted volume to understand the expressive content presented within that volume.

More here: https://www.hathitrust.org/htrc_ncup



Non-Consumptive Research Paradigm

Includes such computational tasks as:

- text extraction
- textual analysis and information extraction
- linguistic analysis
- automated translation
- image analysis
- file manipulation
- OCR correction
- indexing and search

More here: https://www.hathitrust.org/htrc_ncup



Three Approaches

1. [HTRC Analytics](#) (for pre-determined web-based tools, including [Bookworm](#))
2. [Feature Extraction Services](#) (including downloadable data sets)
3. [Secure Data Capsule](#) access



Recent Updates

- HTRC Analytics **NOW** includes functionality to analyze in-copyright material:
 - Using algorithms FOR ALL
 - In a Data Capsule FOR HT MEMBERS ONLY
- In addition to access via HTRC Extracted Features FOR ALL



HTRC Analytics for All

Tool	Function	Data access
HTRC Algorithms	Web-based, click-and-run tools to perform computational text analysis on shared public worksets or those you have created	Including copyrighted items for ALL USERS
Extracted Features Dataset	Allows non-consumptive analysis on specific features extracted from the full text of the HathiTrust corpus	Including copyrighted items for ALL USERS
HathiTrust+Bookworm	A tool for visualizing and analyzing word usage trends in the HathiTrust corpus	Including copyrighted items for ALL USERS

HTRC Algorithms

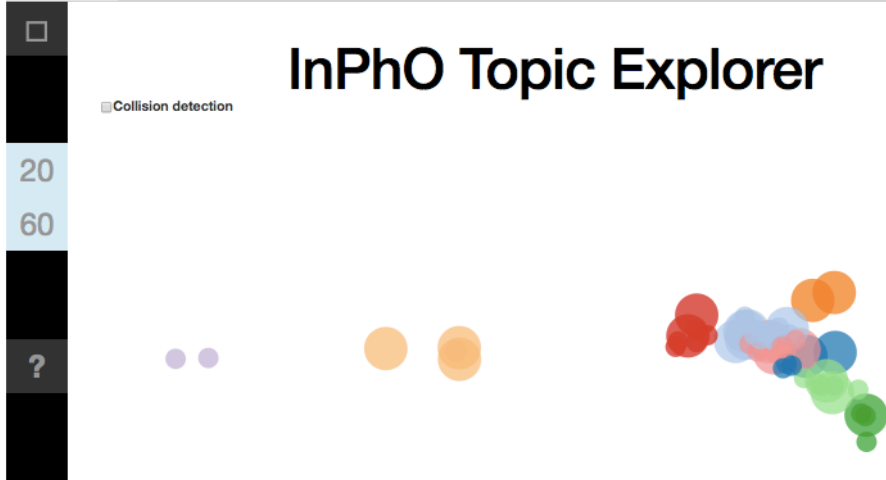
Topic Modeling

Input Parameters

Name	Value
iter	200
input_collection	poli_science_DDRF@eleanordicksonkoehl
k	20 60

Output

topics.html cluster.csv workset.tez topics.json stdout.txt stderr.txt



Output

entities.csv stdout.txt stderr.txt

[Click here to download entities.csv](#)

vol_id	page_seq	entity	type
mdp.39015037380378	00000007	December 1, 1966	DATE
mdp.39015037380378	00000007	1967	DATE
mdp.39015037380378	00000007	Regd	MISC
mdp.39015037380378	00000007	AFRICA	ORGANIZATION
mdp.39015037380378	00000008	Algeria	LOCATION
mdp.39015037380378	00000008	MPLA	ORGANIZATION
mdp.39015037380378	00000008	Sudanese	MISC
mdp.39015037380378	00000008	1966	DATE
mdp.39015037380378	00000008	Independence four years ago	DATE
mdp.39015037380378	00000008	Ghana	LOCATION
mdp.39015037380378	00000008	Central Committee	ORGANIZATION
mdp.39015037380378	00000008	Amin Shaker	PERSON
mdp.39015037380378	00000008	November 29, 1966	DATE
mdp.39015037380378	00000008	Egyptian Gazette	MISC
mdp.39015037380378	00000008	1966	DATE
mdp.39015037380378	00000008	November	DATE

Named Entity Recognition

www.analytics.hathitrust.org

HTRC Extracted Features Dataset

```
1 {
2   "id": "aeu.ark:/13960/t5x649n2b",
3   "metadata": {
4     "schemaVersion": "1.2",
5     "dateCreated": "2015-02-12T20:51",
6     "title": "Sermons delivered on various occasions by Matthew Richey.",
7     "pubDate": "1840",
8     "language": "eng",
9     "htBibUrl": "http://catalog.hathitrust.org/api/volumes/full/htid/aeu.ark:/13960/t5x649n2b",
10    "handleUrl": "http://hdl.handle.net/2027/aeu.ark:/13960/t5x649n2b",
11    "oclc": "720289813",
12    "imprint": "J. Ryerson, 1840."
13  },
14  "features": {
15    "schemaVersion": "2.0",
16    "dateCreated": "2015-02-19T17:14",
17    "pageCount": 304,
18    "pages": [
19
```

Volume and page metadata

Feature data,
including word &
word counts

www.analytics.hathitrust.org

```
1252   "body": {
1253     "tokenCount": 433,
1254     "lineCount": 89,
1255     "emptyLineCount": 22,
1256     "sentenceCount": 13,
1257     "tokenPosCount": {
1258       "1": {
1259         "CD": 2
1260       },
1261       "2": {
1262         "CD": 2
1263       },
1264       "3": {
1265         "CD": 2
1266       },
1267       "4": {
1268         "CD": 1
1269       },
1270       "5": {
1271         "CD": 1
1272       },
1273       "6": {
1274         "CD": 1
1275       },
1276       "est": {
1277         "NN": 3
1278       },
1279       ">»": {
1280         "NN": 1
1281       },
1282       "entirely": {
1283         "RB": 1
1284       },
1285       "quality": {
1286         "NN": 1
1287       },
1288       "click6": {
1289         "FW": 1
1290     },

```

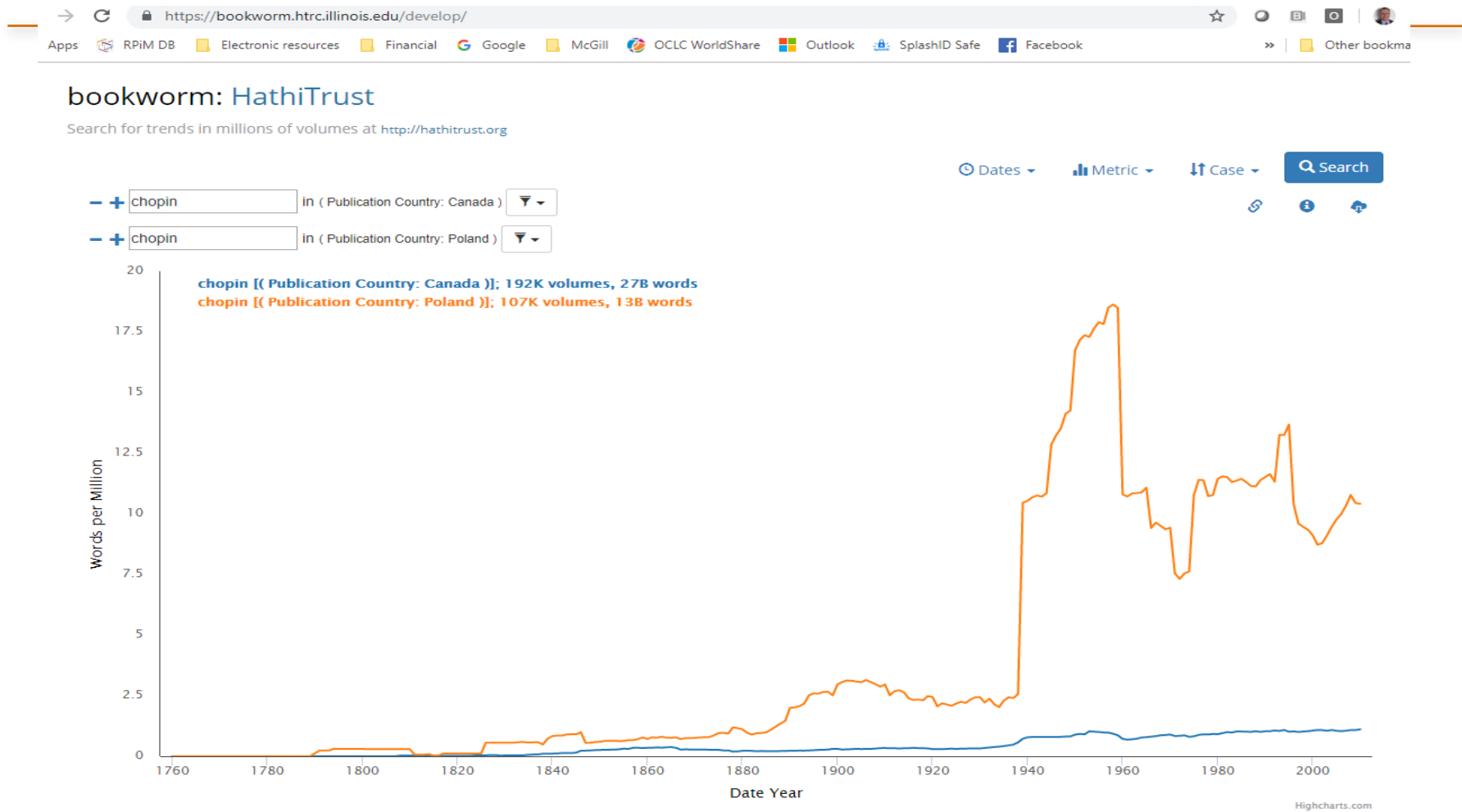
HathiTrust+Bookworm

- Good place to start with digital humanities research questions for the HathiTrust
- The NEH-funded [HT+BW \(HathiTrust+Bookworm\) project](#), a collaboration between the [HathiTrust Research Center \(HTRC\)](#), the Baylor College of Medicine and Northeastern University.
- [HathiTrust+Bookworm](#) uses textual data from the [HathiTrust Digital Library](#) and allows you to track changes in word use based on publication country, genre of works, language, dates and more.
- [HathiTrust+Bookworm](#) is an online tool for visualizing trends in language over time. [HathiTrust+Bookworm](#) is useful for plotting and analyzing usage trends in collections of texts.
- Link to a LibGuide from the University of Illinois about Bookworm:
<https://guides.library.illinois.edu/c.php?g=348508&p=2347886>

<https://bookworm.htrc.illinois.edu/>



HathiTrust+Bookworm



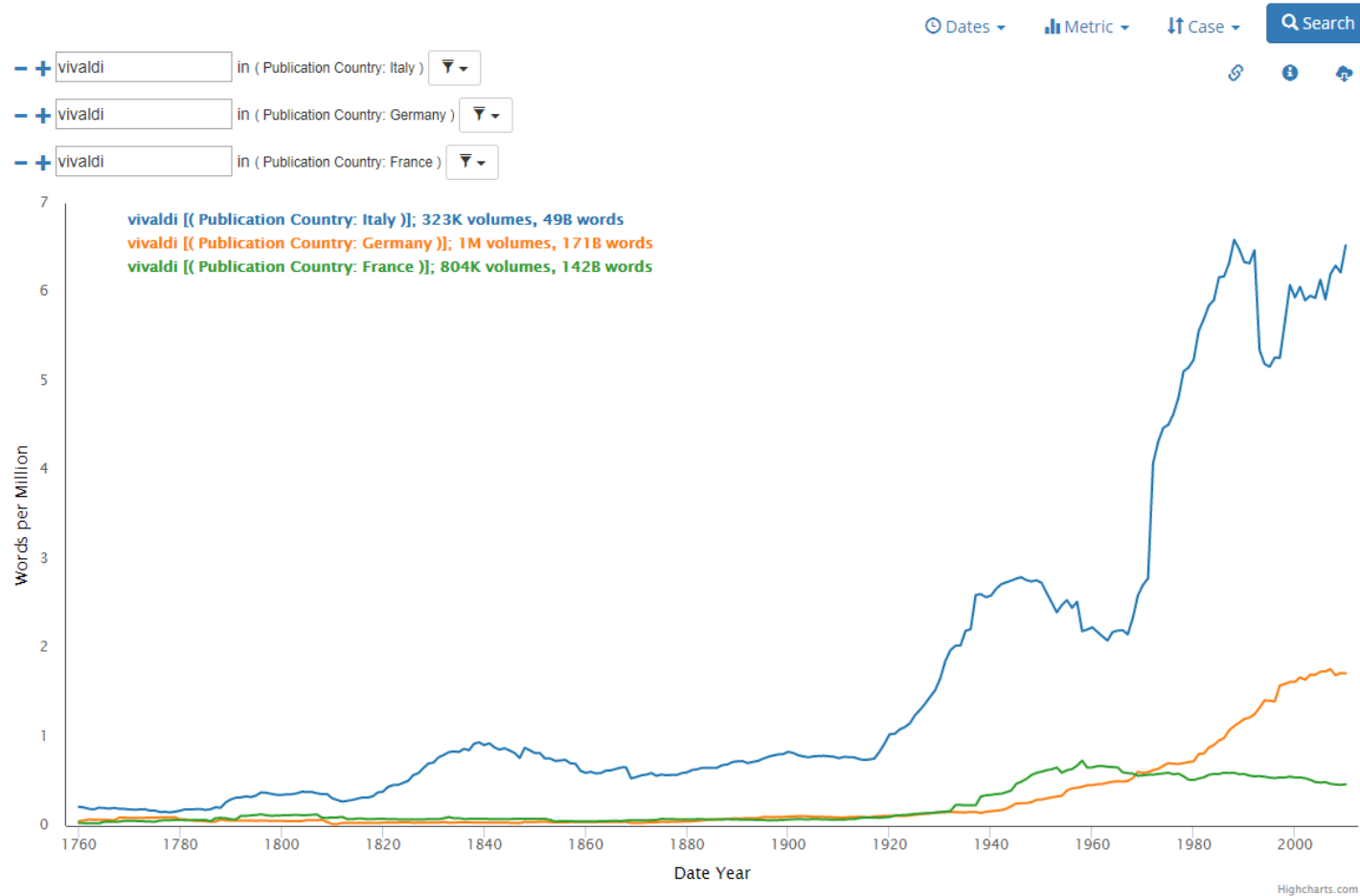
<https://bookworm.htrc.illinois.edu/>



HathiTrust+Bookworm

bookworm: [HathiTrust](#)

Search for trends in millions of volumes at <http://hathitrust.org>



<https://bookworm.htrc.illinois.edu/>



Some examples of ways to narrow your search or change the variables:

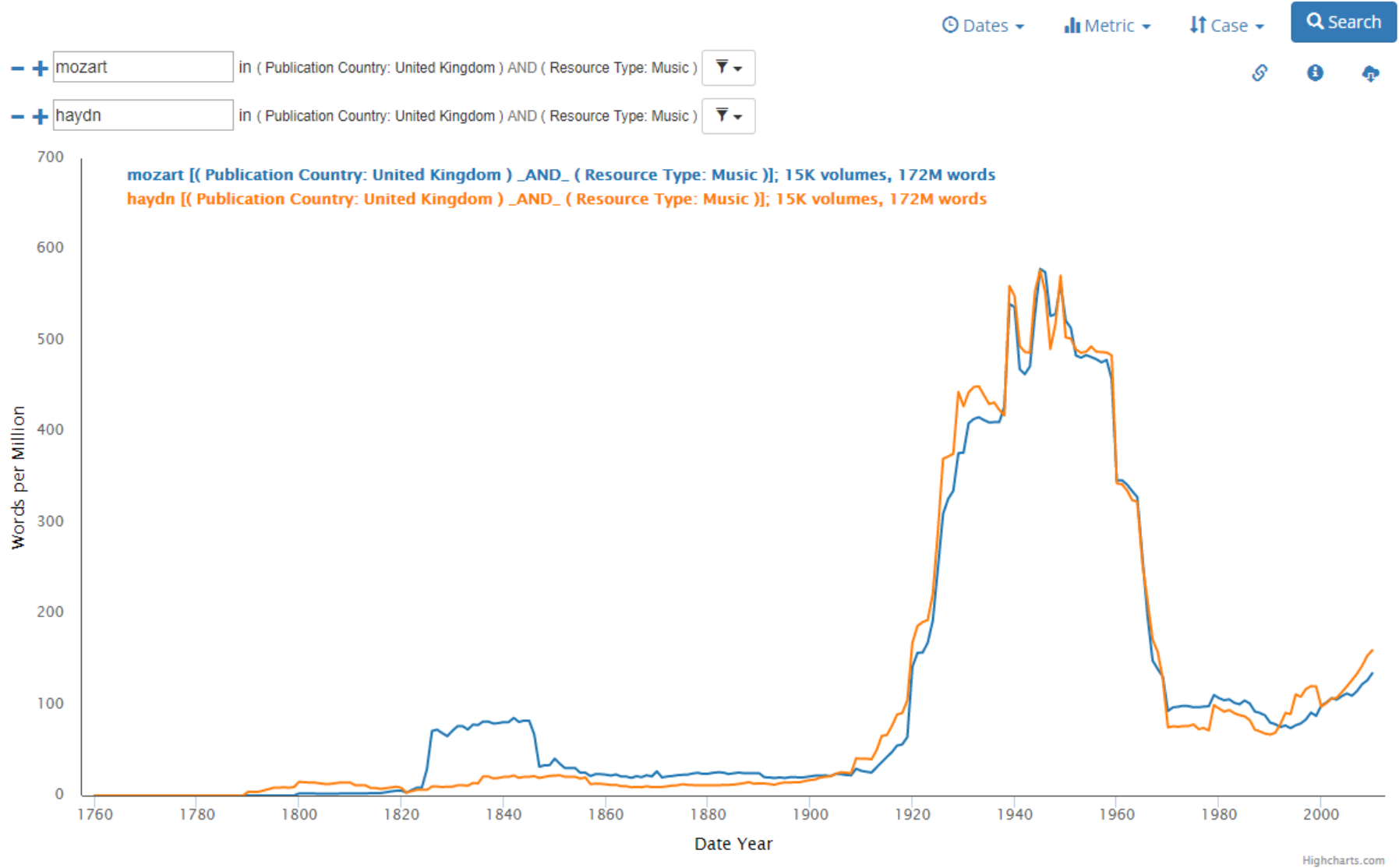
Publication Country	<input type="text" value="× Italy"/>
Publication State	<input type="text" value="All Texts"/>
Subclass	<input type="text" value="All Texts"/>
Narrow Class	<input type="text" value="All Texts"/>
Class	<input type="text" value="All Texts"/>
Resource Type	<input type="text" value="All Texts"/>
Target Audience	<input type="text" value="All Texts"/>
Scanner	<input type="text" value="All Texts"/>
First Author Birth	<input type="text" value="All Texts"/>
First Author Name	<input type="text" value="All Texts"/>
Contributing Library	<input type="text" value="All Texts"/>
Literary Form	<input type="text" value="All Texts"/>
Cataloging Source	<input type="text" value="All Texts"/>
First Author Death	<input type="text" value="All Texts"/>
First Place	<input type="text" value="All Texts"/>
First Publisher	<input type="text" value="All Texts"/>
Is Gov	<input type="text" value="All Texts"/>
Subject Places	<input type="text" value="All Texts"/>

Based on information from the metadata / MARC records

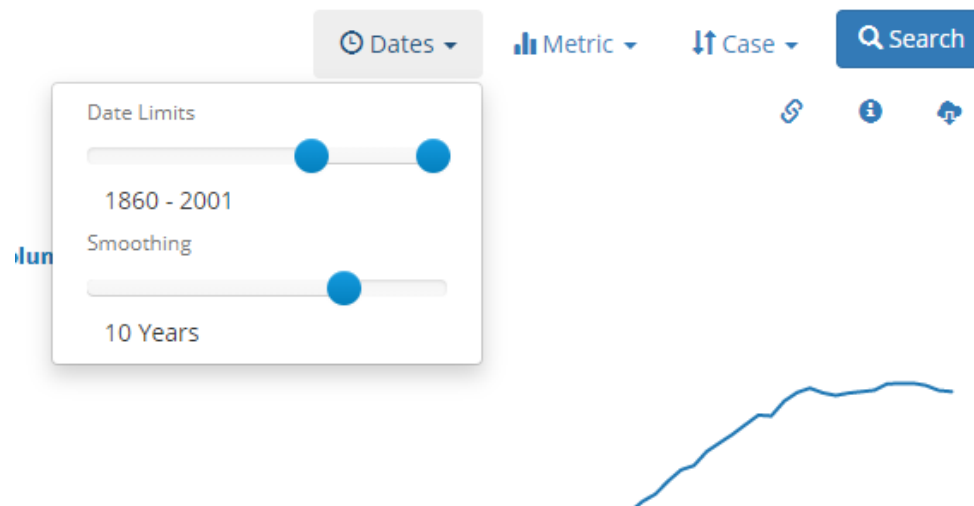


bookworm: HathiTrust

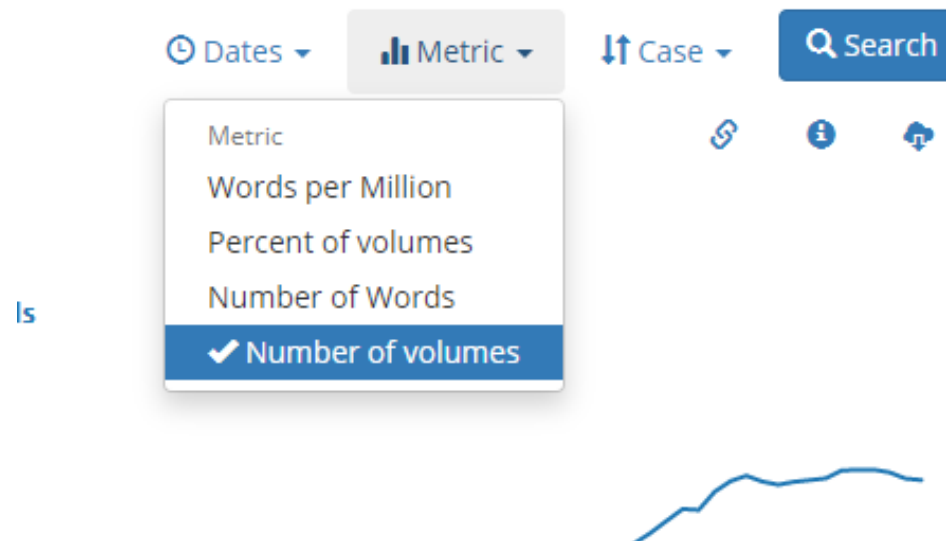
Search for trends in millions of volumes at <http://hathitrust.org>



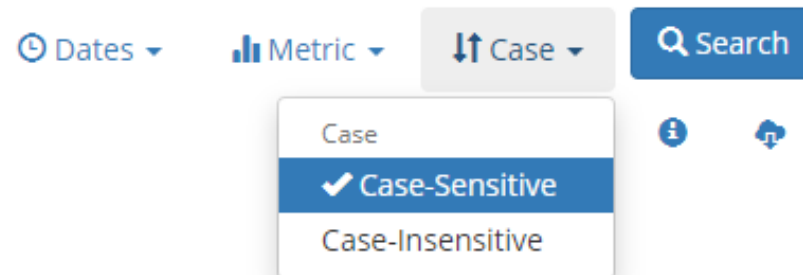
You can change the date limits and how they “smooth”



You can change the metric



You can change the
case to be sensitive or
insensitive



bookworm: HathiTrust

Search for trends in millions of volumes at <http://hathitrust.org>

⌚ Dates ▾

📊 Metric ▾

↕ Case ▾

Search

🔗 ⓘ 📄

- +

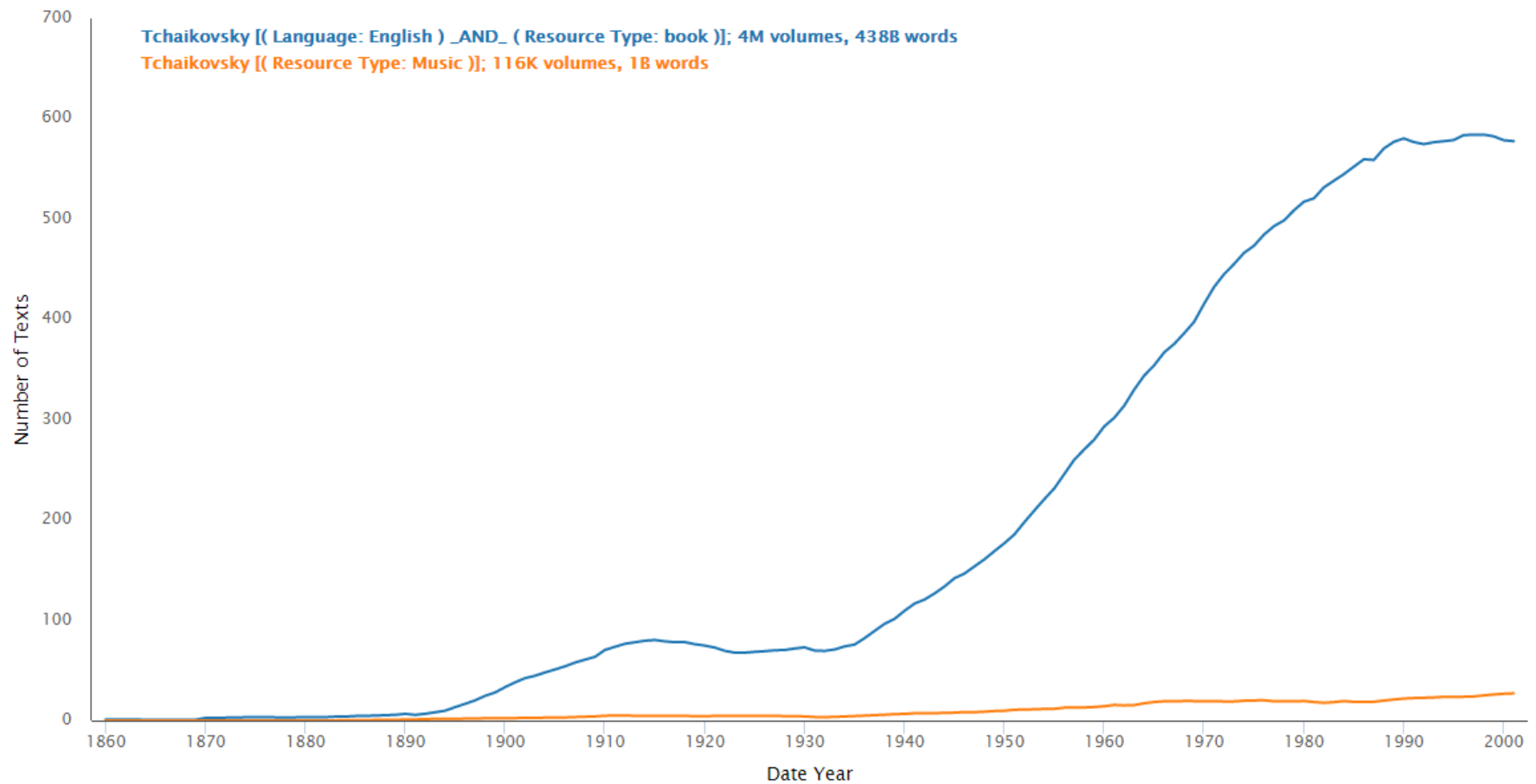
in (Language: English) AND (Resource Type: book)

⌵ ▾

- +

in (Resource Type: Music)

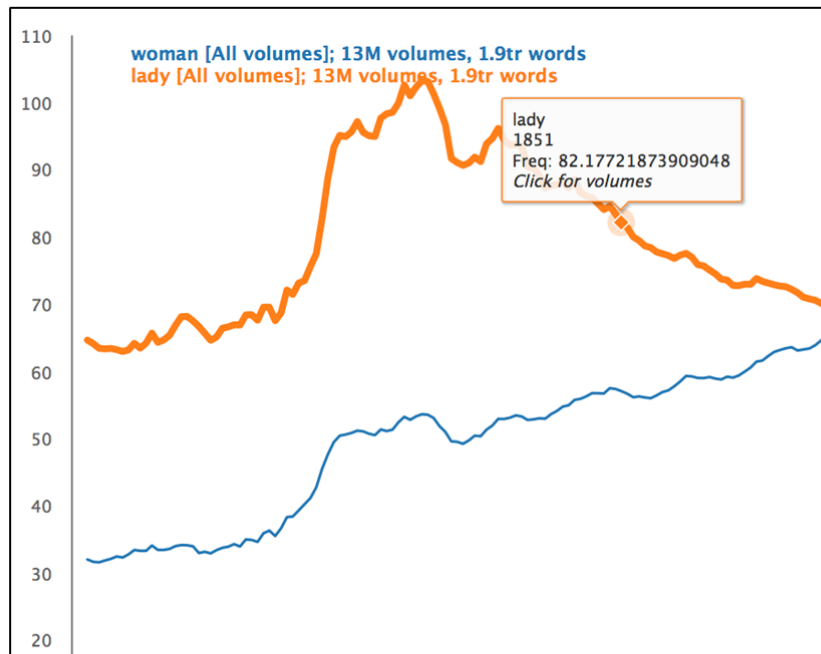
⌵ ▾



Highcharts.com



Bookworm interface



About this Book

Lady Di's minuet : a comedieta : in one act : adapted from the ... Carlingford, Chichester Samuel Parkinson Fortescue, baron, 1823-1893.

View full catalog record

Rights: Public Domain. Google-digitized.

Get this Book

Find in a library
Download this page (PDF)
Download whole book (PDF)
Partner login required

Text Only Views

Go to the [text-only view of this item.](#)

See the [HathiTrust Accessibility](#) page for more information.

Add to Collection

Login to make your personal collections permanent

Select Collection

Add

Share

Permanent link to this book

Jump to 3 Go

Search words about the items

Search in this text Find

LADY DI'S MINUET.

Enter Sir JOHN WILDUCK, speaking as he enters.

Sir J. Tell Lord Mulligatawny that Sir John Wilduck is in the drawing-room. Come, that's settled! I must make an end of it to-day. I don't know what to make of this Mulligatawny—a man that takes such a desperate fancy to one of a sudden, all about a shooting adventure—and will have one marry his daughter, whether one likes it or not. Every morning I come here, with my mind made up to break the thing off; but the moment Mulligatawny sees me, he rushes at me, seizes me by the hand, and calls me his "Dear Sir John—his good Sir John!" I should like to know how I'm to tell such a father as that—"Your daughter's not the thing for me; look out for another son-in-law." Accordingly, I hesitate—I put it off to the next time. The day's gone by, and if this goes on, I shall find myself a married man before I know where I am; not that there is anything to be said against Lady Diana—she's pretty, witty, rich! Yes, by-the-bye, she has one fault—she's too short—not like my cousin Louisa, with her five feet eight. I forgot my rule—I never fall in love under my own height. How can two people step well together in harness, if one's a foot taller than the other? And then, they call it a good match. But Louisa's up to my shoulder already, and growing visibly—and the taller she grows, the better I like her; besides, our marriage is settled between the families. Well, I'm very sorry for Lady Diana, but I must tell her father to-day.

Links directly to texts in the HTDL

HTRC Analytics for HT Members

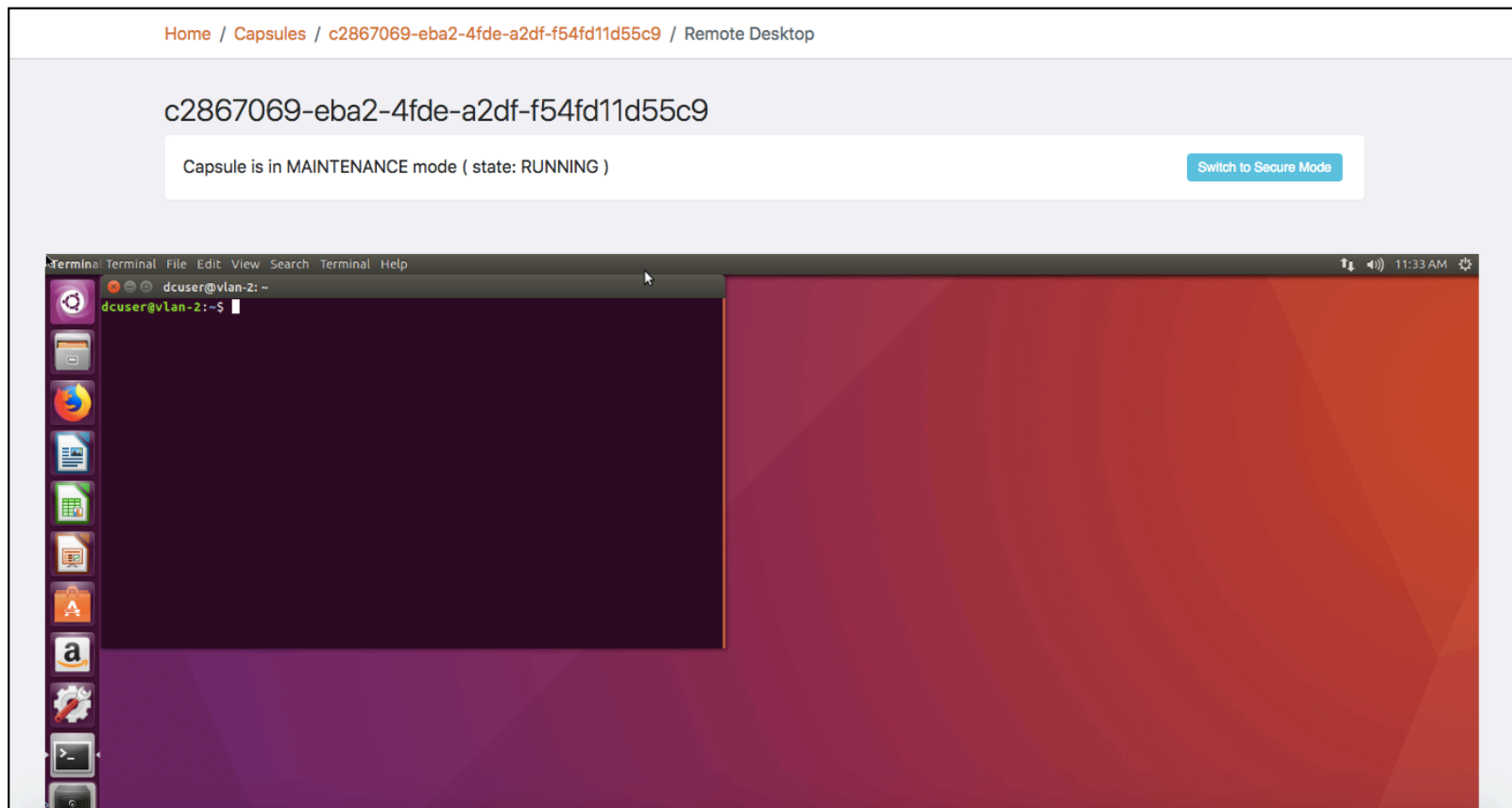
Tool	Function	Data access
HTRC Data Capsule	A secure computing environment for text analysis on the HathiTrust corpus, using the researcher's tools of choice	Access to copyrighted items using an HTRC Data Capsule is available ONLY to HathiTrust member-affiliated researchers

Why? Because we anticipate significant demand for this service and HTRC has finite resources to support it.

See [HathiTrust Member Institutions](#)



HTRC Data Capsule



How Is This Possible?

- HathiTrust exists to enable lawful research and educational uses of its collection.
- In recent years, US courts have recognized that there is a legal basis for non-consumptive research on copyrighted materials.
- In 2016, HathiTrust established a Non-Consumptive Use Research Policy.
- That policy is now embodied in the HTRC Analytics services.



How to get started

- Create an account for [HTRC Analytics](#)
 - HT-member institution affiliation NOT required
- Identify HathiTrust volumes to analyze
 - [Search & create collection](#)
 - [Make use of metadata services](#)
 - Ask for help!
- Create a Data Capsule, if desired



How to get started

- Download Extracted Features for the volumes
-

- Import HT collection/volume list as a workset
and
 - Run an algorithm against your workset
-

- Import volumes into your Data Capsule
and
- Analyze using your preferred tools



Any questions later

Joseph.Hafner@mcgill.ca

or

htrc-help@hathitrust.org



Acknowledgements

- At Indiana University, HTRC is affiliated with and supported by the [IU Pervasive Technology Institute](#), the [School of Informatics, Computing, and Engineering](#), and the [IU Bloomington Libraries](#). Additional financial support comes from the [Office of the Vice Provost for Research](#). Computational resources are provided by the Pervasive Technology Institute.
- At the University of Illinois Urbana-Champaign HTRC is hosted and supported by the [School of Information Sciences](#) in collaboration with the [University of Illinois Library](#). Financial support is provided by the [Office of the Provost](#) and the [Office of the Vice-Chancellor for Research](#). Additional resources to advance the mission of HTRC are supplied by the [National Center for Supercomputing Applications](#).



Acknowledgements

- Thanks to: Eleanor Dickson Koehl, HTRC Digital Humanities Specialist for sharing information for this presentation with me.



Questions?

Thanks!
Merci beaucoup!

