

The HathiTrust Project

Analysing the New Zealand Corpus

Robert Marchman

Dartmouth College

Dr James Smithies

Senior Lecturer in Digital Humanities

University of Canterbury

<https://github.com/rlmv/HathiTrust-SoeR>

“The mission of the HathiTrust is to contribute to the common good by collecting, organizing, preserving, communicating, and sharing the record of human knowledge.”¹

The HathiTrust Digital Library

Founded in 2008, the HathiTrust is a joint venture between 70 partners – including Google, the Internet Archive, and major US research institutions.



Vision and Goals

- Be the first of it's kind in providing a comprehensive and accessible digital repository of human culture.
 - Meet the archival needs of its partner institutions.
 - Develop tools and best practices for large scale and long term digital curation.
-

The Collection

- 72 contributing organizations
 - 10,641,980 volumes
 - 3,725,089,550 pages
 - 477 terabytes of data
 - 1/3 in the public domain
 - Hosted at the University of Michigan.
-

HathiTrust Research Center (HTRC)

The HTRC is a collaboration between Indiana University, the University of Illinois, and the HathiTrust digital library.

HTRC is committed to:

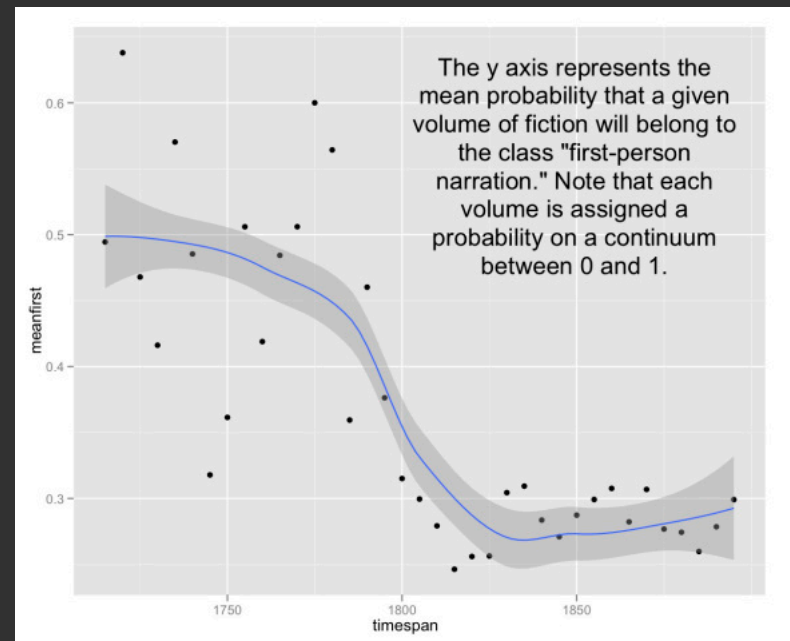
- Enabling research within the HathiTrust collection.
 - Pioneering a paradigm of 'non-consumptive' research tools allowing full access to the HathiTrust without infringing academic copyright laws.
-

The Digital Humanities

- The term Digital Humanities encompasses a diverse and rapidly evolving set of fields seeking to apply modern computational techniques to research questions in the arts and humanities.
 - Acquiring high-quality datasets can be challenging, depending on the desired content. As such, the size of HathiTrust collection is an invaluable resource for digital humanities scholars.
-

An example...

Ted Underwood, Associate Professor at the University of Illinois, has been working with a collection of 18th and 19th century texts drawn from the HathiTrust. His thesis: 'We don't already know the broad outlines of literary history', but computational tools can provide the answer.²



Summer of eResearch Project Goals

- Explore and leverage the HathiTrust's APIs, tools and datasets, and attempt to assess the usefulness of the HathiTrust for New Zealand researchers.
 - Attempt to identify a 'New Zealand corpus'; a subset of the HathiTrust collection pertaining to NZ and relevant to NZ researchers.
 - Lay groundwork for future access to HathiTrust resources by NZ researchers.
-

Accessing the collection...

The Bibliographic API

- An access point for programmatic retrieval of metadata and bibliographic records.
 - Possible use cases:
 - Building a database of bibliographic information for a small standalone collection.
 - Replacing lost metadata.
 - Validating or improving metadata in an existing collection.
 - Specification at http://www.hathitrust.org/bib_api
-

The Data API

- For small-scale retrieval of textual information.
 - Protected by OAuth certification.
 - Returns:
 - OCR (Optical Character Recognition) text
 - Raw document images
 - Structural metadata (METS)
 - http://www.hathitrust.org/data_api
-

Example OCR and raw image source

8 PRELIMINARY ARRANGEMENTS OF THE

" It would be easy to swell the list of those whom circumstances have predisposed to emigration, by describing the benefits which it holds out to the struggling yeoman, the small capitalist, the enterprising trader ; to these the prosperity promised by good colonization cannot fail to render our Settlement specially attractive ; but its peculiar feature consists in the benefits which it is intended to hold out to persons of refined habits and cultivated tastes, whom the moral evils inherent in our present modes of emigration have prevented from availing themselves of its material advantages."

" Our settlement will be provided with a good College, good schools, Churches, a Bishop, clergy, all those moral necessities, in short, which promiscuous emigration of all sects, though of one class, makes it utterly impossible to provide adequately."

8 PRELIMINARY ARRANGEMENTS OF THE

" It would be easy to swell the list of those whom circumstances have predisposed to emigration, by describing the benefits which it holds out to the struggling yeoman, the small capitalist, the enterprising trader ; to these the prosperity promised by good colonization cannot fail to render our Settlement specially attractive ; but its peculiar feature consists in the benefits which it is intended to hold out to persons of refined habits and cultivated tastes, whom the moral evils inherent in our present modes of emigration have prevented from availing themselves of its material advantages."

" Our settlement will be provided with a good College, good schools, Churches, a Bishop, clergy, all those moral necessities, in short, which promiscuous emigration of all sects, though of one class, makes it utterly impossible to provide adequately."

The Solr Proxy

The HathiTrust Research Center exposes a Apache Solr search index through a web API.

<http://wiki.htrc.illinois.edu/display/COM/2.+Solr+API+User+Guide>

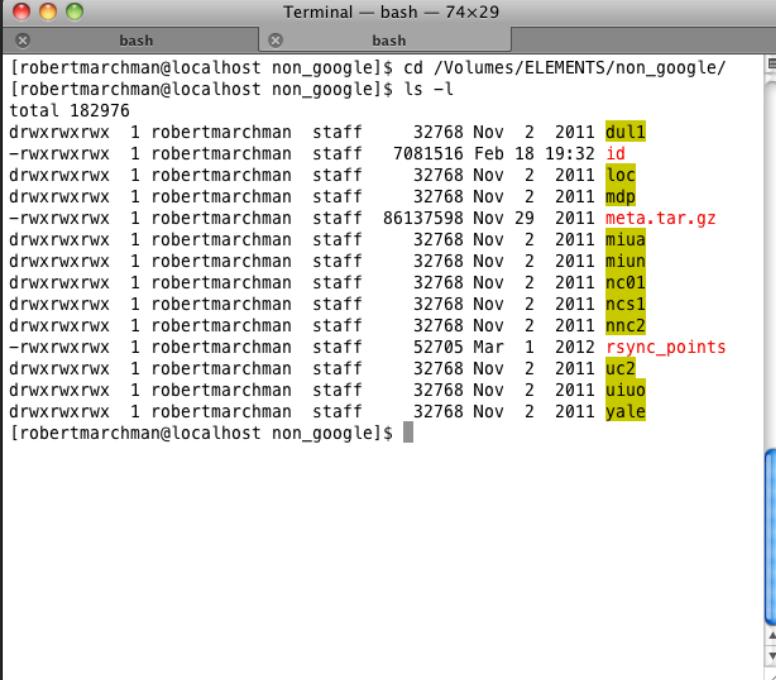
Identifying a NZ corpus

Using the Solr proxy to search metadata for New Zealand content (using keywords such as 'New Zealand', 'Maori', and names of NZ cities and universities) delivers ~2500 documents.

The non-Google dataset

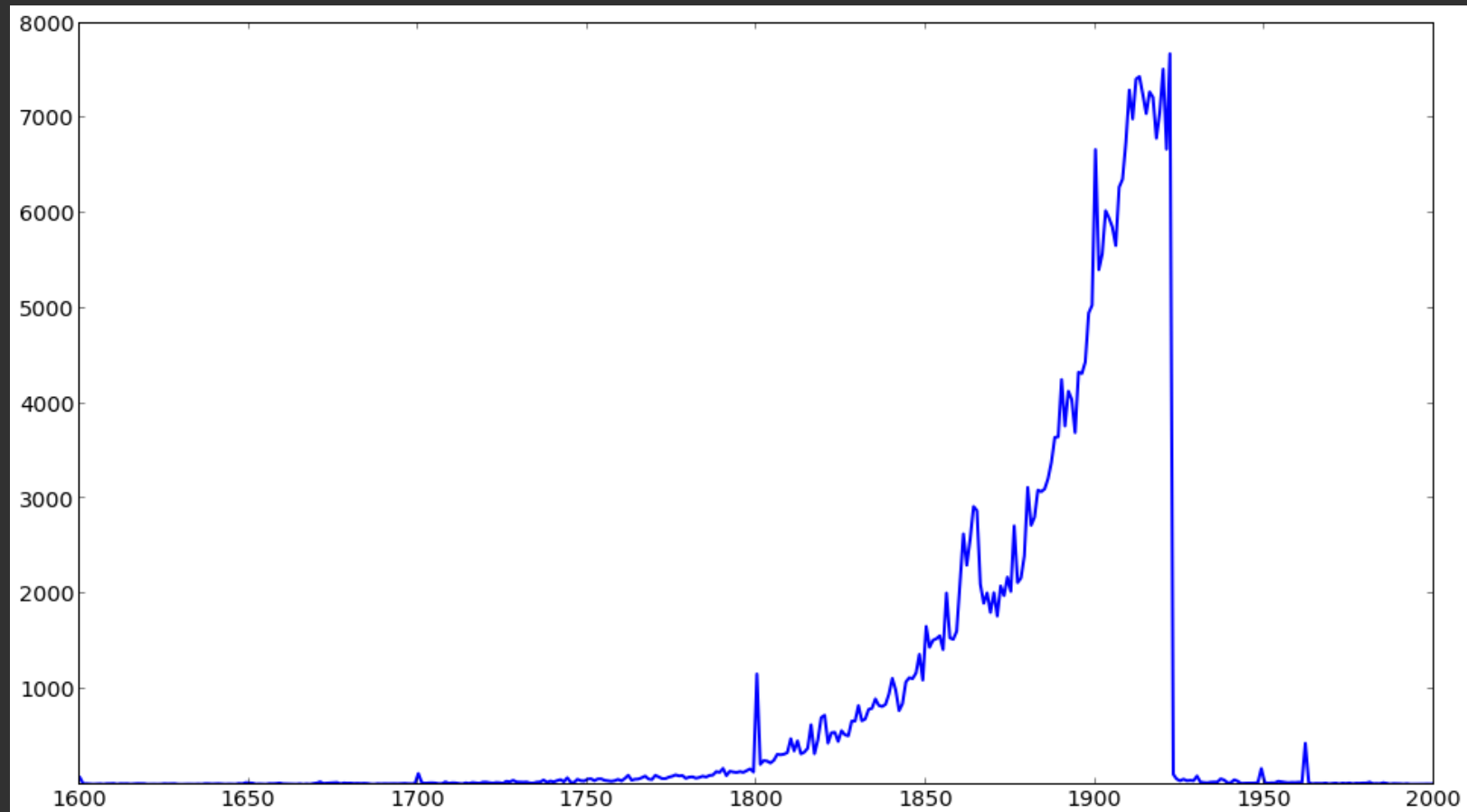
We acquired the non-Google public domain dataset for local use.

The collection consists of 290,000 volumes.



```
Terminal — bash — 74x29
[robertmarchman@localhost non_google]$ cd /Volumes/ELEMENTS/non_google/
[robertmarchman@localhost non_google]$ ls -l
total 182976
drwxrwxrwx 1 robertmarchman staff 32768 Nov 2 2011 dul1
-rwxrwxrwx 1 robertmarchman staff 7081516 Feb 18 19:32 id
drwxrwxrwx 1 robertmarchman staff 32768 Nov 2 2011 loc
drwxrwxrwx 1 robertmarchman staff 32768 Nov 2 2011 mdp
-rwxrwxrwx 1 robertmarchman staff 86137598 Nov 29 2011 meta.tar.gz
drwxrwxrwx 1 robertmarchman staff 32768 Nov 2 2011 miua
drwxrwxrwx 1 robertmarchman staff 32768 Nov 2 2011 miun
drwxrwxrwx 1 robertmarchman staff 32768 Nov 2 2011 nc01
drwxrwxrwx 1 robertmarchman staff 32768 Nov 2 2011 ncs1
drwxrwxrwx 1 robertmarchman staff 32768 Nov 2 2011 nnc2
-rwxrwxrwx 1 robertmarchman staff 52705 Mar 1 2012 rsync_points
drwxrwxrwx 1 robertmarchman staff 32768 Nov 2 2011 uc2
drwxrwxrwx 1 robertmarchman staff 32768 Nov 2 2011 uiuo
drwxrwxrwx 1 robertmarchman staff 32768 Nov 2 2011 yale
[robertmarchman@localhost non_google]$
```


volumes by year



Metadata

The dataset metadata is distributed in MARC (MAchine-Readable Cataloging) XML format, a library standard schema for metadata encoding and transmission.

Normalizing Metadata

MARC quality varies greatly. Formatting is inconsistent, requiring extensive cleaning for any sort of metadata analysis.

For example, here are years found in the metadata for the non-Google collection:

- 1891
- 184[5?]
- 186-?]
- 18--
- c19
- 18 -19
- MDCCLXXIX.
- 5682 [1921]

Using the MARC records, we identified 180 documents in the non-Google collection that contain metadata references to New Zealand.

Open Questions

Classifying by metadata is accurate, but limited – how can we accurately identify relevant documents?

The future

Perform non-consumptive research over the entire collection.

Integrate NZ libraries and organizations such as digitalNZ.org with HathiTrust resources.

Contribute content from NZ research libraries to the HathiTrust.

Thanks to...

- James Smithies
- Yiming Sun, Stephen Downie, and the HTRC
- Tim McNamara, Richard Hosking, Sina Masoud-Ansari, Nick Jones, and NeSI



-
- ¹ www.hathitrust.org/mission_goals
 - ² <http://tedunderwood.com/2013/02/08/we-dont-already-know-the-broad-outlines-of-literary-history/>
-