HathiTrust
research center

How to "Read" Millions of In-Copyright Books: The HathiTrust Digital Library and Research Center

Glen Layne-Worthey & Ryan Dubnicek
HathiTrust Research Center (via the University of Illinois iSchool)

Lowering the Barriers to Computational Access for Digital Archivists
Digital Preservation Coalition
July 6, 2022

HathiTrust origin stories (1/2)

- **December 2004**: Google & five large research libraries announce massive book scanning project.

- **September 2005**: Authors Guild sues over “massive copyright infringement.”

The New York Times **Technology**

A FREE e-mail with Theater seats at great prices

NYTimes: Home - Site Index - Archive - Help — Welcome, - Member Center - Log Out

Go to a Section — Go — Search: All of Technology — Go

Technology Home | Circuits | Product Reviews | How To's | Deals

Advertisement

Google Is Adding Major Libraries to Its Database

By JOHN MARKOFF
and EDWARD WYATT
Published: December 14, 2004

Google, the operator of the world's most popular Internet search service, announced today that it had entered into agreements with some of the nation's leading research libraries and Oxford University to begin converting their holdings into digital files that would be freely searchable over the Web.

It may be only a step on a long road toward the long-predicted global virtual library. But the collaboration of Google and research institutions that also include Harvard, the University of Michigan, Stanford and the New York Public Library is a major stride in an ambitious Internet effort by various parties. The goal is to expand the Web beyond its current valuable, if eclectic, body of material and create a digital card catalog and searchable library for the world's books, scholarly papers and special collections.

Google - newly wealthy from its stock offering last summer - has agreed to underwrite the projects while also adding its own technical abilities to the task of scanning and digitizing tens of thousands of pages a day at each library.

Enlarge This Image

Thor Swift

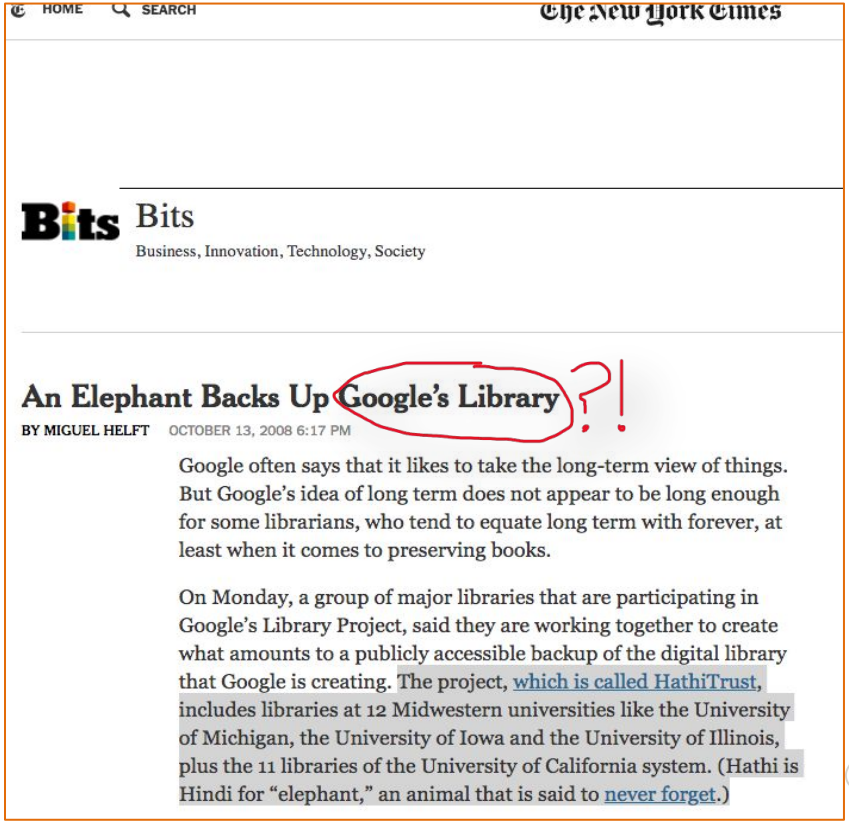
A book is scanned at Stanford University. Google's plans for digital files include the University of Michigan and the New York Public Library.

ARTICLE TOOLS

E-Mail This Article

HathiTrust origin stories (2/2)

- **October 2008**: The ***HathiTrust Consortium*** is announced jointly by 23 research libraries in the US Midwest and California.
  - ***Hathi*** is Hindi for ***elephant***
  - Currently more than **200** member libraries
- **2009-2011**: Authors Guild suit expanded to include HathiTrust & its member libraries



HOME SEARCH

The New York Times

Bits

Business, Innovation, Technology, Society

An Elephant Backs Up Google's Library ?!

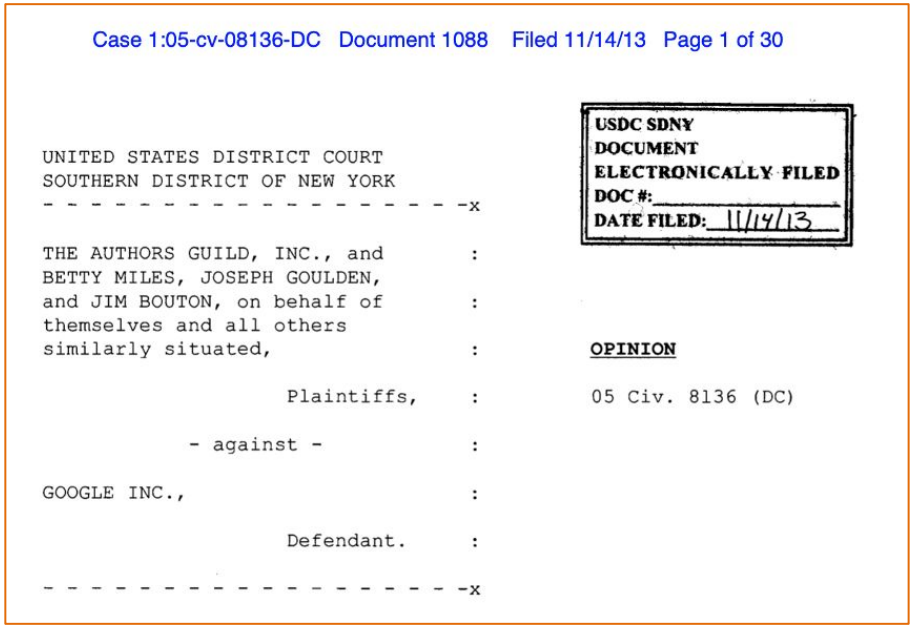
BY MIGUEL HELFT OCTOBER 13, 2008 6:17 PM

Google often says that it likes to take the long-term view of things. But Google's idea of long term does not appear to be long enough for some librarians, who tend to equate long term with forever, at least when it comes to preserving books.

On Monday, a group of major libraries that are participating in Google's Library Project, said they are working together to create what amounts to a publicly accessible backup of the digital library that Google is creating. The project, which is called HathiTrust, includes libraries at 12 Midwestern universities like the University of Michigan, the University of Iowa and the University of Illinois, plus the 11 libraries of the University of California system. (Hathi is Hindi for "elephant," an animal that is said to never forget.)

A consequential legal victory for HathiTrust, *et al.*

- **November 2013**: Authors Guild v Google, HathiTrust, and member libraries resolved in favor of Google, *et al.*
- Consequences:
  - Google (and other) scanning of in-copyright allowed to continue unimpeded
  - HathiTrust free from legal jeopardy
  - Digitization of in-copyright works, *without permission,* for purposes of *digital scholarship*, is found to be a Fair Use
  - The principle of *non-consumptive use* is elaborated



Case 1:05-cv-08136-DC Document 1088 Filed 11/14/13 Page 1 of 30

USDC SDNY
DOCUMENT
ELECTRONICALLY FILED
DOC #:
DATE FILED: 11/14/13

UNITED STATES DISTRICT COURT
SOUTHERN DISTRICT OF NEW YORK
- - - - - - - - - - - - - - - - - - - -x

THE AUTHORS GUILD, INC., and BETTY MILES, JOSEPH GOULDEN, and JIM BOUTON, on behalf of themselves and all others similarly situated,

Plaintiffs,

- against -

GOOGLE INC.,

Defendant.

- - - - - - - - - - - - - - - - - - - -x

OPINION

05 Civ. 8136 (DC)

https://clearinghouse.net/doc/74833/

Non-consumptive and other *fair* uses in U.S. law

- Descriptive metadata*
  - The "telephone directory" principle: there is no copyright in *facts*
  - This extends to *facts about a text*, including:
    - Book-level catalog data
    - Page counts, word counts, snippets
    - Part-of-speech and other grammatical facts
- Non-consumptive use*
  - Also called "Non-Expressive Use"
  - "Research where **computational analysis** is performed on one or more volumes in the HathiTrust collection, but does **not include reading or displaying of substantial portions of** a volume in order to understand its **expressive content**." – https://www.hathitrust.org/htrc_ncup

* ...inclusive of *in-copyright* texts

Extracted Features: What & How (to Use)

HTRC Extracted Features (EF) Dataset

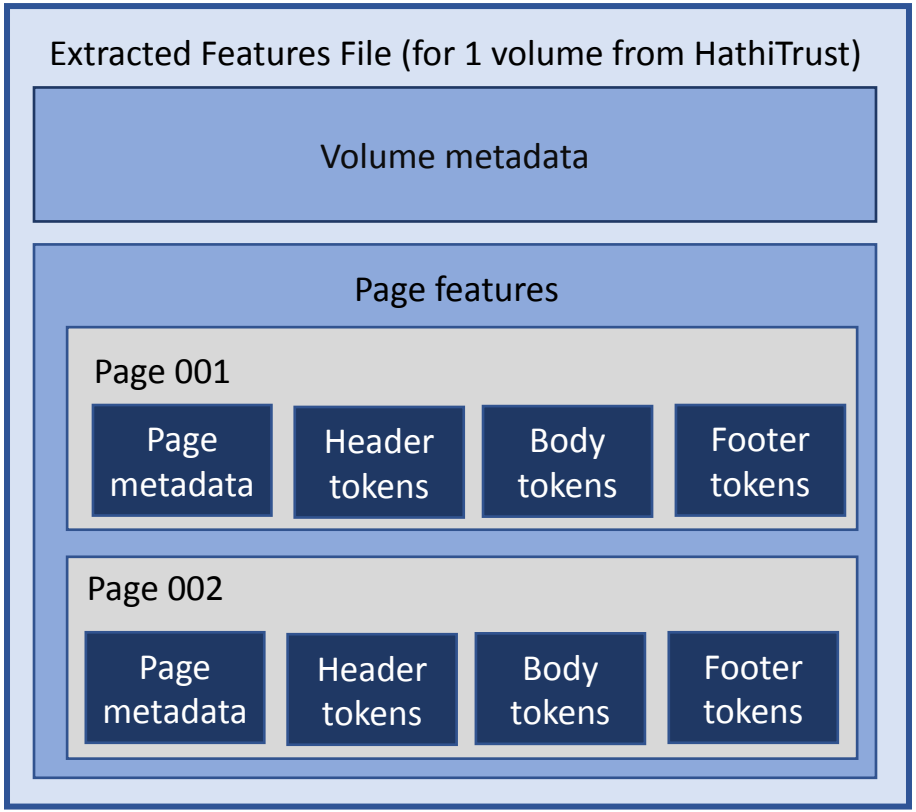
- Public domain, fully downloadable
- Structured data consisting of human-supplied metadata and algorithmically-derived features
- From 17.1 million volumes
- Form of non-consumptive research access

HTRC EF Dataset

- The features are:
  - Extracted from raw text
  - Volume- and page-level
  - Selected data and metadata
  - Extracted from raw text
- Position the researcher to begin analysis
  - Some common preprocessing is already done

Full EF documentation: https://analytics.hathitrust.org/datasets#ef

Extracted Features model



Bag-of-words text data

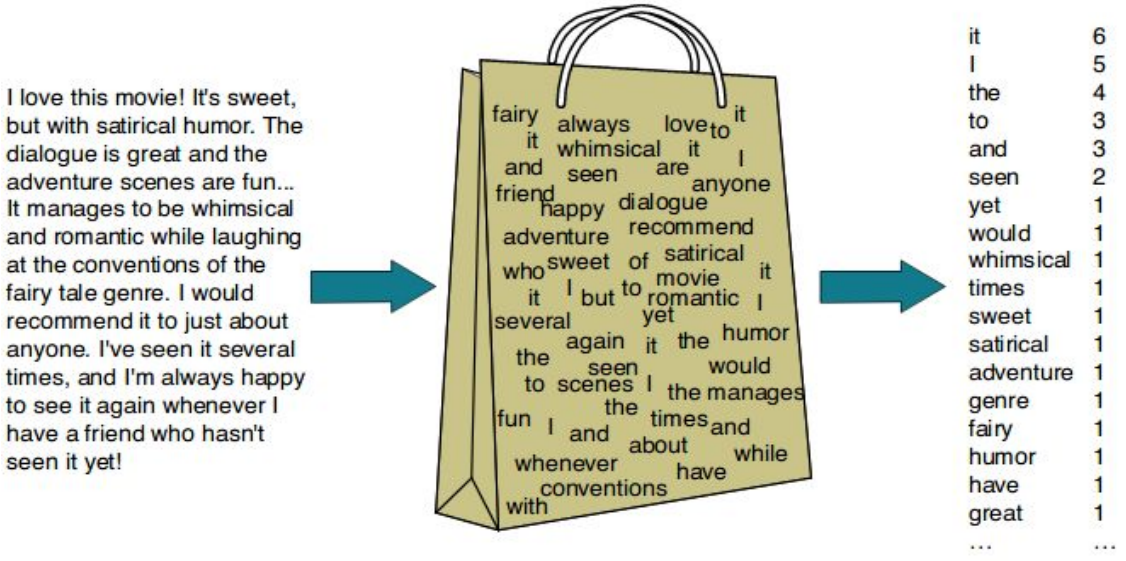
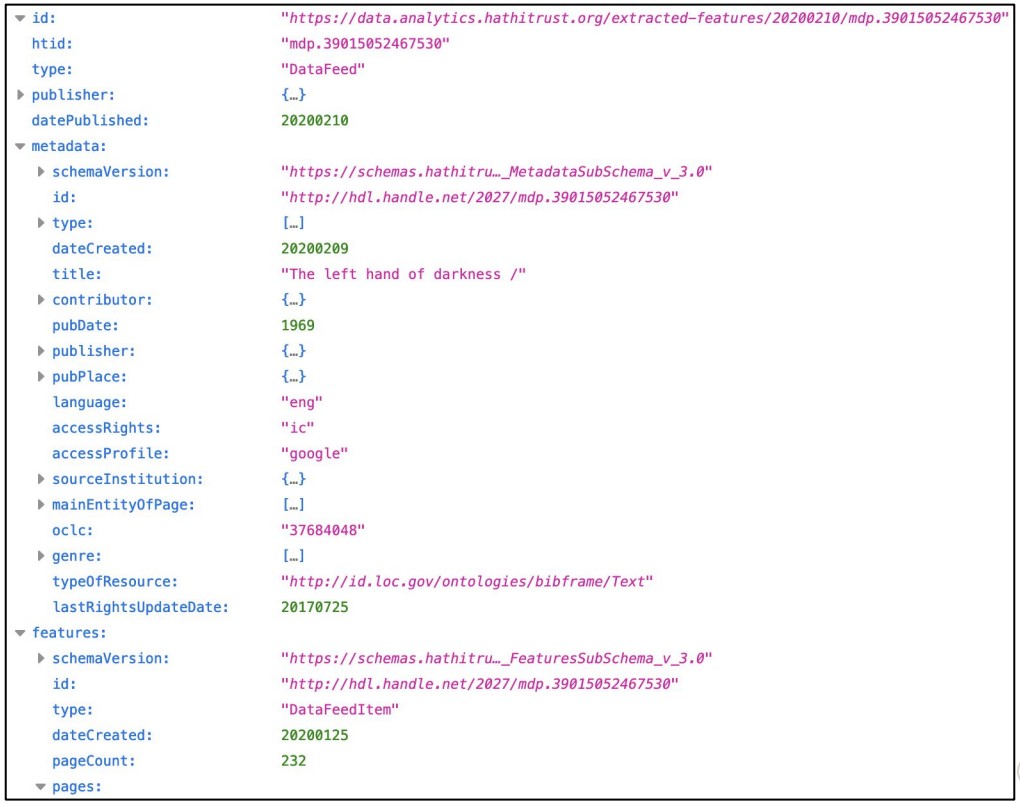


Image source: https://dudeperf3ct.github.io/lstm/gru/nlp/2019/01/28/Force-of-LSTM-and-GRU/

Per-volume features

Sourced from bibliographic metadata:

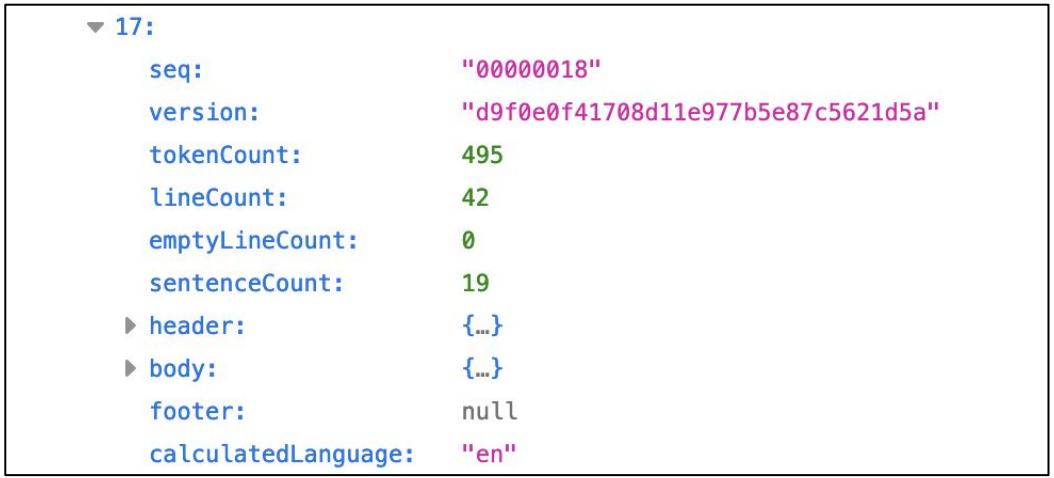
- Title
- Author
- Language
- Identifiers



Per-page features

Sourced from EF algorithm, organized by page sequence:

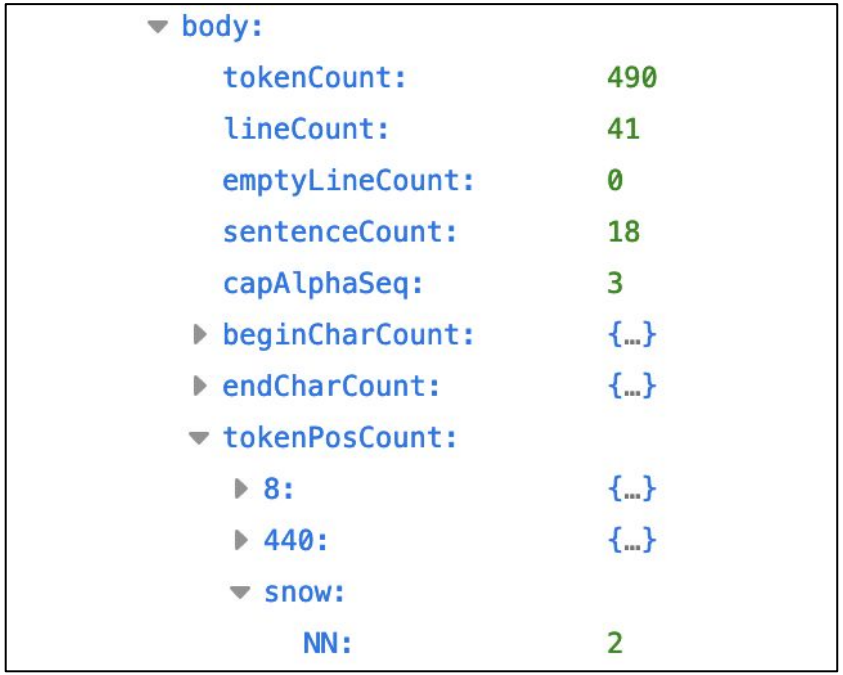
- Word count
- Line count
- Empty line count
- Sentence count
- Language



Per-page-section features

For each header, body, footer of each page:

- Line, empty line, and sentence count
- Counts of beginning- and end-line characters
- Tokens, POS tags, token counts



(Some) ways to use the EF Dataset

Explore tokens

- Identify parts of a book
  - Using descriptive metadata
- Analyze a volume's text
  - Topic modeling
  - Lexical analysis
- Modeling volumes or language
  - Train and use a genre/language/period-specific classifier
  - Generate a large model of vectorized text for a given language
  - Supervise a model using bibliographic metadata as ground truth

Explore metadata

- Visualize the HTDL
  - Year, Place, Language, Contributor, Subject*
- Analyze or improve volume metadata
  - Evaluate/augment volume-level metadata:
    - Language, author, publisher
  - Compare metadata records for "identical" volumes
- Compare token analysis to metadata:
  - e.g. evaluate OCR or lexical analysis over time, by publisher, etc.

HATHITRUST
research center

Thanks, questions, and discussion

Glen Layne-Worthey | gworthey@illinois.edu
Ryan Dubnicek | rdubnic2@illinois.edu
htrc-help@hathitrust.org

HathiTrust Research Center
School of Information Sciences
University of Illinois Urbana-Champaign

School of Information Sciences
UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN