HATHI
TRUST
RESEARCH
CENTER

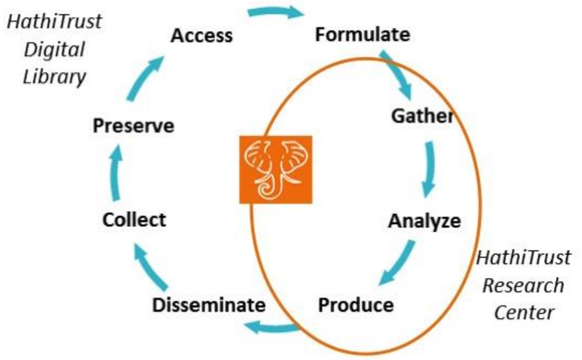
CNI Fall 2016 Membership Meeting
December 12, 2016

The HathiTrust Research Center (HTRC) has achieved a level of growth over the period 2014 – 2016 that suggests that its initiatives are having value to researchers and educators:

New Users	Registered Users	Data Capsule Users	Total No of Institutions
923	1127	130	257

HTRC's place in the ecosystem of HathiTrust (HT) services to its members is cementing as an important complement to existing HT services (see Fig right). HTRC has firmly established itself as a research center within the ecosystem of HathiTrust member services, complementing the existing HT discovery and access platform.

In the Spring of 2016 the HT and HTRC teams successfully ingested the full HT corpus of nearly 15,000,000 digitized volumes into the HTRC research environment. In the Fall of 2016 HTRC provided early access to trusted researchers while we made infrastructure adjustments, including the purchase of a data storage cluster, to enable broader access and faster compute capability



Looking Forward

HTRC strategy for future success is multi-pronged:

Grow demand:

Target new communities, esp. in the social sciences. Continue outreach to train librarians in text mining. Adapt HTRC tools, services and documentation for instructional use. Incorporate lessons learned from outreach events for continuous improvement of HTRC.

Lower barriers to use:

- Reduce technical and process barriers for research and classroom use, for uses both small (<1,000 volumes) and large (1 million volumes), for HTRC's three major use modes:
    - HTRC web tools – SEASR analysis tools, Bookworm analytics, workset builder. See https://analytics.hathitrust.org
    - HTRC Extracted Features – download extracted features (word counts, parts of speech). See https://analytics.hathitrust.org/features
    - HTRC Data Capsule – employ own analysis tools in secure environment. See https://wiki.htrc.illinois.edu/display/COM/HTRC+Data+Capsule

Partnering opportunities:

Pilot for shared access across collections with publishers as path to common platform and HTRC sustainability.

Develop cost model for in-kind contributions to HTRC.

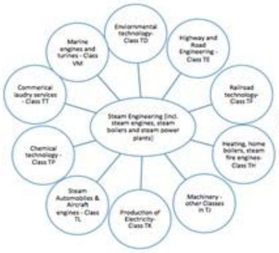
HATHI TRUST RESEARCH CENTER

CNI Fall 2016 Membership Meeting
December 12, 2016

Tracking Technology Diffusion Through Time in the HathiTrust Corpus

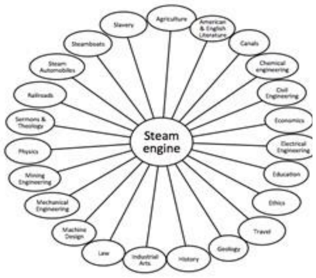
Michelle Alexopoulos, University of Toronto

Dr. Alexopoulos, an economist, is using the vast historical record contained in the HathiTrust to study the diffusion of various technologies over time. By tracking word usage trends of 1,214 technology-related terms identified by Alexopoulos, such as the steam engine, her research based on HathiTrust book content has the potential to overturn accepted theories about the economic and societal impacts of a technology.


Steam Engineering (incl. steam engines, steam boilers and steam power plants)
Environmental technology-Class TD
Highway and Road Engineering - Class TE
Railroad technology-Class TF
Heating, home boilers, steam fire engines-Class TH
Machinery - other Classes in TJ
Production of Electricity-Class TK
Steam Automobiles & Aircraft engines - Class TL
Chemical technology - Class TP
Commerical laudry services - Class TT
Marine engines and turines - Class VM

Linkages to "Steam Engines" implied by the Library of Congress Classification

vs.



From HT text: Selected subject terms linked to "Steam engine" n-gram by 1910

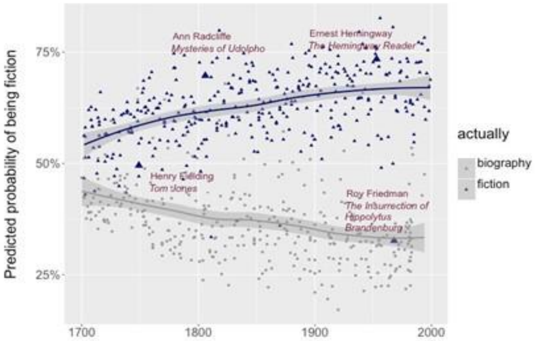
- 1,012,633 volumes analyzed.
- Over 22 hours of processing using a 32-node cluster on Indiana University's high-performance supercomputer, Big Red II.
- Each node had 32 cores and 64 GB of RAM.

HTRC Use Case: Collaboration between Scholars and the HTRC

Predicting the Past: How digital libraries speak to literary questions about genre

Ted Underwood, University of Illinois

This project seeks to use digital libraries, and the large scale data analysis of their volumes, to create predictive models for determining genre of both a subset of volumes, and an individual volume. Using frequency of individual indicator words, such as action verbs, references to body parts or political terms, Dr. Underwood's work has been able to reliably determine genre in either case for volumes in the HathiTrust Digital Library (HTDL).



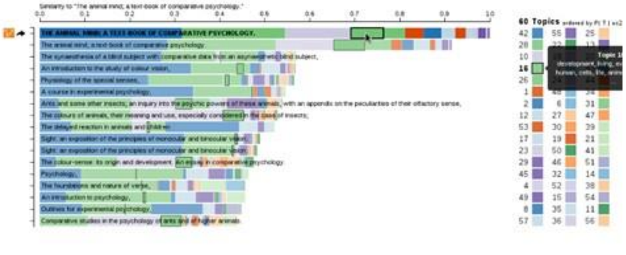
- Processed and released genre-specific word counts for English language literature in HTDL, 178,381 volumes (available from HTRC).
- Has conducted similar analysis to predict grammatical gender.
- Process to ID fiction over entire PD corpus took over 24 hours on over 48 cores.

HTRC Use Case: Collaborating with Scholars

Taxonomizing the Texts: Towards Cultural-Scale Models of Full Text

Colin Allen & Jaimie Murdock, Indiana University

This project from Indiana University leverages the HTRC Data Capsule framework to test and visualize topic models. It will show the relationship between a topic model created on a random sample of volumes and the entire category, from Library of Congress, from it is drawn. Topics generated from LoC categories are visualized in an online service called "Topic Explorer".



Similarity to "The animal mind; a text-book of comparative psychology."

THE ANIMAL MIND; A TEXT-BOOK OF COMPARATIVE PSYCHOLOGY.
The animal mind, a text-book of comparative psychology.
The synaesthesia of a blind subject with comparative data from an asynaesthetic blind subject,
An introduction to the study of colour vision,
Physiology of the special senses,
A course in experimental psychology,
Ants and some other insects; an inquiry into the psychic powers of these animals, with an appendix on the peculiarities of their olfactory sense,
The colours of animals, their meaning and use, especially considered in the case of insects;
The delayed reaction in animals and children
Sight; an exposition of the principles of monocular and binocular vision,
Sight; an exposition of the principles of monocular and binocular vision,
The colour-sense: its origin and development. An essay in comparative psychology
Psychology,
The foundations and nature of verse,
An introduction to psychology,
Outlines for experimental psychology,
Comparative studies in the psychology of ants and of higher animals.

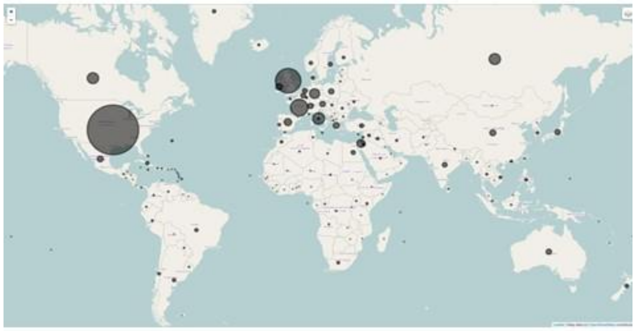
- Extracted Library of Congress Subject Headings for 1,606,302 volumes for building topic models by subject headings.
- Utilized 6 virtual machines in the HTRC data capsule
- Topic Explorer readily available in the HTRC Data Capsule

HTRC Use Case: Collaborating and Supporting Community Scholars

Literary Geography at Scale

Matthew Wilkens, University of Notre Dame

With the help of natural language processing, Dr. Wilkens will extract and geocode place names nearly eleven million volumes, from the HathiTrust, including twentieth and twenty-first century texts. This project to geolocate world literature, supported by a 2014-2015 American Council for Learned Societies Digital Innovation Fellowship, is one of the largest humanities text-mining projects to date.



- Pilot analysis completed. Named entity extraction run on 10,000 randomly sampled volumes.
- Pilot took 55 minutes of processing time on Indiana University's high-performance supercomputer, Big Red II, using an 8-node cluster.
- Next step: Process the entire corpus!

HTRC Use Case: Collaboration between Scholars and the HTRC