Advanced Collaborative Support (ACS) Awards

Advanced Collaborative Support (ACS) is a scholarly service at the HTRC offering collaboration between external scholars and HTRC staff to solve challenging problems related to HTRC tools and services. By working together with scholars, we facilitate computational access to HathiTrust Research Center digital tools (HTRC) as well as the HathiTrust Digital Library (HTDL) based on individual scholarly need. ACS will drive innovation at the scholar's digital workbench for enhancing and developing new techniques for use within the HTRC platform. For questions, please send an email to acs@hathitrust.org.

2017 Awardees:

Computational Support for Reading Chicago Reading

Robin Burke, John Shanahan, Ana Lucic, DePaul University

The Reading Chicago Reading team will seek to extend their own research on the “One Book, One Chicago” city-wide reading program by incorporating textual analysis on books chosen for the OBOC program, as well as comparison texts. Further, the resulting textual analysis—including toponym extraction, sentiment analysis, and story arc detection—will be paired with library patron, circulation and demographic data to present a fuller picture about the OBOC program, and the books chosen for inclusion.

Project report: Computational Support for ‘Reading Chicago Reading’

Modeling the History of Book Design

David Bamman and Bjorn Hartmann, University of California, Berkeley

This project will utilize the HTRC Data Capsule to conduct feature extraction on page images from 10,000 in-copyright books in the HathiTrust repository, extracting features such as page construction, line justification, leading between baselines, kerning between letter pairs/combinations, line density per page, characters per line, position of images, typeface (serif, sans-serif) and font size. Beyond the analysis and utility of the extracted feature set, this project also seeks to serve as a use case for engagement with HathiTrust/HTRC beyond books-as-strings-of-words analysis.

The Power of Place: Structure, Culture, and Continuities in U.S. Women’s Movements

Laura Nelson, Northeastern University

Dr. Nelson’s project will study the women’s movement in the United States from 1848-1975 in two cities, New York City and Chicago, using new advances in network analysis and computational text analysis to identify structural and cultural diversity. This approach is three-pronged: building a workset of writing by individuals and organizations within the movement in New York and Chicago, using network analysis to measure the structure of this movement, and conducting computational text analysis to measure the underlying culture and ideas within the movement, including lexical analyses to identify distinctive words and topic modeling to identify dominant themes.


Richard Jean So, McGill University

Dr. So’s project seeks to write a new history of the American novel by examining a series of large textual datasets focused on the full cycle of the U.S. literary field from production to reception to canonization. The major goal is to identify the emergence of new patterns of language, style, discourse and themes in American novels as they appear at different moments in the cycle of literary production and reception, including publication via large publishing houses such as Random House, and book reviews in major U.S. periodicals. This will be achieved through using the HTRC Data Capsule environment to undertake text analysis of full texts, including using various methods in Machine Learning and Natural Language Processing, such as topic models, word embeddings, and specialized tools such as BookNLP, which allows for the extraction of grammatical dependencies and characters.


Measuring Literary Novelty

Laura McGrath, Devin Higgins, Arend Hintze, Michigan State University

This work draws on ongoing collaborative efforts to develop a method for applying genetic sequencing tools to the study of literature in order to identify and measure literary novelty, and address questions of literary history, canonicity, and prestige. Previous results have been suggestive of a prominent connection between the purely information-based novelty of the sequences of characters that comprise literary texts, and the experimental newness we associate with modernist literary texts. Leveraging the HTRC Data Capsule will offer the potential to apply this theory at scale for the first time, and potentially lead into new research into modernism and the literary history of the 20th century.
A Writer’s Workshop Workset with the Program Era Project (PEP)

Nicholas Kelly, Loren Glass, Nikki White, University of Iowa

The PEP team will compile a proof-of-concept workset with, at first, prominent individuals (faculty, staff, students) who were involved with the Iowa Writers’ Workshop (IWW), then produce “style cards” for each author’s works (by volume), based on stylometric data gathered through text analysis of the IWW workset within the HTRC Data Capsule. It is the goal of the project to also create a living workset that can be continually updated for scholars who wish to engage with IWW authors and their writing.

Project report: Program Era Project

Off-Cycle project:

The Life of Words

David-Antoine Williams, The University of Waterloo

This project aimed to match Oxford English Dictionary (OED) references to HathiTrust volume IDs, in order then to draw down associated metadata using the heterogeneous and fragmented bibliographical data in OED2 and OED3. It furthered the work of The Life of Words (LOW), a research project in its third year, led by Dr. Williams at St Jerome’s University in the University of Waterloo in Canada. The aim of the project is to enhance the OED with metadata concerning its corpus of 3.5 million quotations.

Project report: The Life of Words

2016 Awardees:

Fighting Fever in the Caribbean: Medicine and Empire, 1650-1902

Mariola Espinosa, University of Iowa

This project seeks to explore the history of yellow fever in the Caribbean by comparing how the disease was described by residents of the Caribbean to the European perspective, including through sentiment analysis of text referencing yellow fever. Her work will be visualized spatially in a map generated with support from the University of Iowa’s Digital Scholarship and Publishing Studio. She will build a corpus of text from the HathiTrust Digital Library related to yellow fever and filth in the Caribbean to track the use of the terms “filth” and “filthiness” from 1650 to 1902.

Project report: Fighting Fever in the Caribbean

Inside the Creativity Boom

Samuel Franklin, Brown University

This project will map the increasing use and shifting meanings of the words “creative” and “creativity,” with a particular focus on the twentieth century. A custom “creativity corpus” will be assembled and processed to identify linguistic patterns via a number of text analysis and natural language processing techniques. Brown’s project will make use of the functionality developed for HathiTrust + Bookworm.

Project report: Inside the Creativity Boom

The Chicago School: Wikification as the First Step in Text Mining in Architectural History

Dan Baciu, Illinois Institute of Technology

This project will look at the Chicago School of architecture and examine its history of reception over the last 75 years, as well as identify patterns in its international spread and influence. Baciu will use named entity recognition in his analysis, notably deploying the Wikifier tool on a large sample corpus of HathiTrust data for the first time.

Project report: The Chicago School: Evolving Systems of Value

Signal and Noise and Pride and Prejudice: Toward an Information History of Romantic Fiction

Dallas Liddle, Augsburg College

This project will test two hypotheses about information theory and information density as they relate to a digital humanities approach to studying Romantic-era British fiction. The concept of “information” used in mathematical information theory may help digital humanists evaluate the information density of textual forms. This project tests a theory that the popular and critical success of the novel in British print culture after 1815
may be related to increased information density and sophistication of information encoding in those years, especially via innovations introduced by Jane Austen and Walter Scott.

**Project report:** Signal and Noise and Pride and Prejudice

### 2015 Awardees:

#### The Trace of Theory

*Geoffrey Rockwell, Laura Mandell, Stefan Sinclair, Matthew Wilkens, Susan Brown, University of Alberta, Texas A&M University, University of Notre Dame*

Rockwell, Mandell, Sinclair, Wilkens, and Brown aim to subset theoretical subsets from the HT public corpus and apply large-scale topic modeling on the subsets. The researchers will develop tools and computational methods for tracking the concept of "theory".

**Project report:** The Trace of Theory project

#### Detecting Literary Plagiarisms: The Case of Oliver Goldsmith

*Douglas Duhaime, University of Notre Dame*

Duhaime will work on developing tools for detecting plagiarisms. He will focus on the case of Oliver Goldsmith, to detect the literary thefts of Goldsmith by using machine learning techniques.

**Project report:** Coming soon!

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

*Colin Allen, Jaimie Murdock, Indiana University Bloomington*

Allen and Murdock will carry out a cultural-scale investigation and topic modeling on HT public-domain full text through random sampling to select collections according to the Library of Congress Subject Headings (LCSH).

**Project report:** Towards Cultural-Scale Models of Full-Text project