Advanced Collaborative Support (ACS) Awards

Advanced Collaborative Support (ACS) is a scholarly service at 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.

Calls for proposals to participate in the ACS program go out approximately once per year. For questions, please send an email to acs@hathitrust.org.

2021 Awardees

Projects funded by the Andrew W. Mellon Foundation through the Scholar-Curated Worksets for Analysis, Reuse & Dissemination (SCWAReD) grant project.

Mining the Native American Authored Works in HathiTrust for Insights

Kun Lu, Raina Heaton, and Raymond Orr (University of Oklahoma)

This project seeks to compile a collection of Native American authored works in HathiTrust and apply various text mining methods to the collection to reveal the coverage, subjects, perspectives, and writing styles of Native authors. A list of Native authors and their works will be compiled from an existing database created by a member of the project team and from other online resources. This list will be aligned with the HathiTrust digital library to create a workset of Native American authored works in HathiTrust for further analysis. Then, a variety of text mining methods will be used to analyze the subjects, topics, language use, and writing styles of Native American authors. Comparative analysis will be carried out to understand the characteristics of this textual community. The project is expected to develop a database of Native American authors and the bibliographic information of their works, create a reusable workset of Native American authored works in HathiTrust, identify potential gaps in the HathiTrust corpus on this textual community, and provide insights into the characteristics of the community by text mining their works.

The Black Fantastic: Curated Vocabularies, Artifact Analysis and Identification

Clarissa West-White (Bethune Cookman University) and Seretha Williams (Augusta University)

This project focuses on identifying Black Fantastic texts in the HathiTrust Digital Library. The project proposes that characteristics of the Black Fantastic—the cultural production of African Diasporic artists and creators who engage with the intersections of race and technology in their work—exist in historical and current cultural artifacts, including those created by and about future-forward personalities, such as Dr. Mary McLeod Bethune. It builds on previous and ongoing work to create a bibliography of the Black Fantastic that is featured in *Third Stone Journal*. Works in HathiTrust will be analyzed along with Black Fantastic artifacts from other collections, such as the Dr. Mary McLeod Bethune collection in the Bethune-Cookman University archives. By working across collections, the project will test methods for locating Black Fantastic texts and lives.

Creating Period-Specific Worksets for Latin American Fiction

José Eduardo González (University of Nebraska, Lincoln)

This project seeks to create large datasets to research the history of Latin American fiction and question traditional periodization of this literature by attempting to detect the boundaries between literary periods and subgenre distinctions in Latin American fiction. It will look critically at the techniques for detecting genre distinctions that have developed over the last few years and evaluate how they apply to the particular development of Latin American literary system. While many of the subgenres in the English-speaking literary market such as detective fiction, the Gothic novel, and speculative fiction have followers in Latin America, the genres that have traditionally been considered important for the changes in the literary history of the region are less formulaic and more closely linked to national and regional historical and/or social developments. Instead of attempting to identify canonical documents that typify a genre, this project will examine how documents diverge from a particular canon in order to explore the social and cultural reasons an author might accept or deviate from a dominant style.

The National Negro Health Digital Project: Recovering and Restoring a Black Public Health Corpus

Kim Gallon (Purdue University)

This project draws on HathiTrust's collection of public health documents on Black health to explore how early twentieth Black public health officials communicated and addressed health disparities that impacted African American communities. The major goal of the project is to create a series of worksets and visualizations that scholars and students of African American health and medicine along with public health experts and physicians can use to deepen historical narratives about Black health that might offer insight into the development of contemporary health communications targeted toward African American communities. The project also establishes some of the research for *Technologies of Recovery: Black DH Theory and Praxis*, a book in-progress. Finally, the work will fill a gap in the history of African American public health.

2020 Awardees

Read project updates

Surveying Applicability of Energy Recovery Technology for Waste Treatment

Aduramo Lasode (University of Minnesota)

This project focuses on data collection and analysis regarding six *prime movers*: gas turbines, steam turbines, microturbines, reciprocating internal combustion engines, solid oxide fuel cells and Stirling engines. Prime movers are part of an energy recovery effort that reuses heat and power generated from combustible gases produced in waste treatment. In order to aid renewable energy, prime mover technology needs to be optimally implemented within the waste treatment industry. The project will address the difficulty of making optimal choices for waste treatment applications and aid a growing distributed waste treatment industry. First, data collection from the HathiTrust Digital Library will extract values that influence use of these prime movers, mainly power output, efficiency, capital cost, and fuel composition. Then, data analysis will be performed using methodology from a preliminary study evaluating efficiency-based applicability for five of the six aforementioned prime movers. The major output of this project is a dataset containing power, efficiency, cost and fuel-related information for six prime movers, with a goal to publish data and resulting analysis in addition to being incorporated into the researcher's dissertation. The project outcomes will impact relevant fields in sustainability, combustion, and energy policy through a practical decision guide for choosing prime movers in waste treatment facilities, as well as by highlighting the need for innovation, with a special focus on prime mover applicability in the growing distributed waste treatment industry.

Detecting and Transcribing Arabographic Texts

David Smith (Northeastern University), Matthew Thomas Miller (University of Maryland), Maxim Romanov (University of Vienna), and Sarah Bowen Savant (Aga Khan University, London)

While not predominant, significant material is available in Arabographic languages, such as Arabic and Persian, in HathiTrust. Transcription accuracy in these languages is lower, however, than in Latin-script languages. Perhaps due to the use of synthetic data from digital fonts to train optical character recognition (OCR) systems, OCR models perform poorly on the large numbers of Arabographic printed books set in historical fonts or lithographed from manuscripts. When Arabic or Persian text is embedded in books in other languages, such as English, the non-Latin text is often transcribed as if it were English in a very strange font, resulting in near-zero accuracy. This project will work toward solutions to both problems. First, the research team will improve baseline Arabic and Persian OCR by finding editions of canonical texts in the HathiTrust collection and aligning images of those editions with existing digital transcriptions. Then they will use this aligned data to train new generalized models for a wide variety of typefaces and lithographed book hands. Second, the researchers will build classification models to detect pages with embedded Arabic and Persian, using textual features of the matrix language alone using the HTRC Extracted Features dataset. They will align existing digital transcripts of texts for 30 works in Arabic and 24 works in Persian in HathiTrust to run their baseline OCR system, and then extract page images of lines matched to spans of text in the digital editions.

Tracing the shifting rhetoric of ethnoracial difference in federal responses to education, 1958-2018

Andrés Castro Samayoa (Boston College)

This project leverages HathiTrust's U.S. Federal Documents Collection to investigate how materials produced by the U.S. federal government document shifts in terminologies of ethnoracial difference. The project will focus on the documents and materials published by the Department of Education (formerly United States Department of Health, Education, and Welfare) and related congressional documents from hearings in specialized subcommittees from 1958 until the present. It will explore how the rhetorics of ethnoracial difference overlapped with the growing allocation of federal resources to postsecondary institutions, particularly Minority Serving Institutions, in the latter half of the 20th century. The start of the National Defense Education Act in 1958 was a watershed moment that signaled the greater engagement of the federal government in higher education.The subsequent passing of the Higher Education Act in 1965, alongside amendments through the 1990s and 2000s, allocated specific federal appropriations to support colleges and universities, including Historically Black Colleges & Universities, Tribal Colleges & Universities, Hispanic Serving Institutions, and Asian American & Native American Pacific Islander Serving Institutions. The project contributes to current work focusing on the history of federal responses to higher education in the United States, and the growing visibility of Minority Serving Institutions as a valuable sector of the postsecondary sector in the United States' higher education.

2019 Awardees

Read project updates

Building Large-Scale Collections of Genre Fiction

Laure Thompson and David Mimno (Cornell University)

This project will develop methods for automatically constructing large-scale collections of genre fiction from HathiTrust. Even, and especially, in digital libraries as large as HathiTrust, it can prove challenging to understand whether the library contains suitable representations of a chosen genre. The researchers plan to focus on collections of speculative fiction novels as a case study, but they intend their solutions to be generalizable. They will identify robust methods for correlating author-title pairs to matching volume sets in HathiTrust. Using these methods in conjunction with lists of novels that were curated by hand, they will build their collections and investigate which works are (over)represented and which are missing. They expect their project will enable scholars to better understand the suitability of studying genre fiction through HathiTrust and highlight underserved author and genre groups. Moreover, the project will result in collections of genre fiction which can be readily reused and reorganized for different lines of humanistic inquiry.

Project report: Building Large-Scale Collections of Genre Fiction: Final Report

Mapping scientific names to the HathiTrust Digital Library

Matthew J. Yoder and Dmitry Mozzherin (University of Illinois)

This project will create an index of all the scientific names of the Earth's species found within the HathiTrust corpus. The index, which will likely measure in the hundreds of millions to billions of entries, will consist of a simple link between the scientific name and the volume and page location of that name within HathiTrust. The index will assist in identifying volumes that may be medically relevant, for example by identifying all of the volumes containing the scientific name for the mosquito that carries illnesses such as Zika virus ('Aedes aegypti'). The index will also allow volumes to be grouped into clusters based on which scientific names they contain to show which taxon (e.g. "mammals") are most common. This team of researchers has completed similar work across the data of the Biodiversity Heritage Library. Their ACS project will allow them to do cross-corpora comparisons.

Project report: **Global Names and the HathiTrust: Towards comprehensive indexing of taxon names in real time**

Supporting The Conglomerate Era Project

Dan Sinykin (Emory University)

This project furthers the researcher's investigation into how the conglomeration of the publishing industry changed literature. The results will be included the researcher's in-progress book titled The Conglomerate Era: A Computational History of Literature in the Age of the Agent. The project explores a set of publisher-based corpora to see whether there are distinctions in what is published by large publishing houses versus independent presses. It will make use of predictive modeling to further the researcher's existing work to build a computational model of genre that aids in identifying latent patterns in the publishers' editorial practices.The project will utilize methods such as genre detection through unsupervised modeling; stylistic differentiation through text classification and supervised learning via logistic regressions; and social network analysis with metadata to determine latent literary connections, especially with regard to gender and race of the author.

Project report: **The Conglomerate Era Project**

Deriving Basic Illustration Metadata

Stephen Krewson (Yale University)

This project aims at identifying all pictorial elements in educational texts from 1800-1850 to explore the interplay between progressive education and print media in the early nineteenth century. The resulting research will characterize the extent to which wood engravings and other reprographic materials were shared among educational publishers. The researcher will extract specific features from page images, such as illustration location, using advances in machine learning. The project intends to make use of the process developed to identify pictorial elements to motivate a new metadata field that describes the location and type of illustrations on the page. An ultimate goal of the project is to move toward "machine-read" texts where the data generated by classifiers and dimensionality reduction techniques are bundled as metadata with the corresponding volumes and made available to future research. ("Machine-read" is a term is borrowed from researcher Ben Schmidt.)

Project report: **Derived Metadata for Early 19C Illustrations: ACS Grant Final Report**

Semantic Phasor Embeddings

Molly Des Jardin, Scott Enderle, and Katie Rawson (University of Pennsylvania)

This project intends to explore a novel way of abstracting and representing textual data that could aid in new ways of discovering and deduplicating items in HathiTrust, detecting and analyzing genre, or analyzing narrative analogies. The project team will investigate the utility of a certain kind of mathematical representation of text documents, called semantic phasor embeddings, that combine a mathematical structure called phasors with data from standard word embeddings (strings of numbers that represent an item). If successful, the vectors could represent documents with a tunable degree of granularity, which could provide an opportunity to share vectors representing copyright-protected without concerns about wholesale text reproduction. The vectors would also carry valuable information about the global ordinal structure of the volumes, so that the items could be queried, clustered, and visualized in a robust way that recognizes similarity not just in the content of the items, but also their structure.

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.

Project report: **Modeling the History of Book Design, HTRC Whitepaper: Summary of Activities**

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 movements 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.

Project report: **The Power of Place: Structure, Culture, and Continuity in U.S. Women's Movements**

A Computational History of the U.S. Novel, 1950-2000

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.

Project report: **A Computational History of the U.S. Novel, 1950-2000**

Measuring Literary Novelty

Laura McGrath, Devin Higgins, and 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, and 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 heterogenous 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 (University of Alberta), Laura Mandell (Texas A&M University), Stefan Sinclair (University of Alberta), Matthew Wilkens (University of Notre Dame), and Susan Brown (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.

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

Colin Allen and 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