The HathiTrust Research Center: Exploring the Full-Text Frontier

Many readers of EDUCAUSE Review are familiar with the HathiTrust Digital Library (HTDL), which was featured in an E-Content column by Jeremy York and Brian E.C. Schottlaender a couple years ago. However, readers may not yet be aware of the research counterpart to this mass-scale digital library: the HathiTrust Research Center (HTRC). In this column, we want to help readers better understand the HTRC mission—a mission that supports new knowledge creation through novel computational uses of the HTDL.

Birth of a Research Center
In the original Google Books Settlement Agreement in 2008, funds were to be set aside to create a research center that would enable researchers worldwide to accomplish data-mining and analysis on texts in the public domain and under copyright in a manner that was secure and compliant with appropriate U.S. copyright law. This did not happen, because the court rejected the agreement in 2011. However, the HathiTrust Board of Governors believed in this public good and announced a request for proposals to build such a center in 2009. In 2011, the HTDL announced that Indiana University Bloomington and the University of Illinois at Urbana-Champaign would run the HTRC under an agreement with the HathiTrust Board of Governors and the University of Michigan. The HTRC has been an active production service since 2014, with tools available to analyze a set of out-of-copyright content equaling around 4.4 million volumes. In 2016, the HTRC plans to enable analysis of the entirety of the 14 million volumes held currently by the HTDL.

Understanding Non-Consumptive Research
To understand the tools and methodologies of the HTRC, one needs to understand the conditions under which scholars are permitted to access and use the collection. Use of the HathiTrust collection is governed in part by agreements among Google and others who digitized the works. Much of the collection is in copyright, so any use of those materials must carefully follow copyright law. Thus the HTRC has focused on non-consumptive research using the digital surrogates. The rejected settlement defines this as “research on one or more books, but not research in which a researcher reads or displays.” Operationally, the HTRC defines non-consumptive research as that in which no action or set of actions on the part of users, either acting alone or in cooperation with other users over the duration of one or multiple sessions, can result in sufficient information gathered from a collection of copyrighted works to reassemble pages from the collection.

Current Tools and Methods
Currently the tools of the HTRC can be broken into four distinct categories that build off each other in terms of usefulness to the scholar:

1. The HTRC Workset Builder (HTRC-WB) enables scholars to utilize a faceted browsing interface to build unique collections from within the HTDL collection. This is part of the overall HTRC Portal.
2. The HTRC Data Capsule (HTRC-DC) is the heart of the platform, providing scholars with a unique Virtual Machine (VM) and data enclave that is supported by cyberinfrastructure at Indiana University to run analysis on the content of the individual worksets created by scholars.
3. The HTRC Solr/Lucene API allows scholars to work directly with the Solr/Lucene index of the HTDL corpus in novel ways.
4. The HTRC Extracted Features Dataset (HTRC-EFD) was...
published in May 2015 and has since been used to create a further refined extracted features dataset based on volumes published between 1700 and 1922 and including specific information for genre.

**Bridging the Scholarship Gap**

The work of the HTRC has enabled a contextualized usage of the metadata plus the full text of the entire corpus—an approach described by many in the digital humanities fields as *distant reading.* This type of distant reading scholarship meshes well with the one-to-many uses of the completely digitized content and metadata of the HTDL. Key concepts for future work around distant reading scholarship need to include options for enhanced metadata improvement and new models for workset creation within the HTRC platform. This work will be useful when exploring how to enhance discovery within the HathiTrust production platform, perhaps by enhancing metadata with new data on genre, date, or gender of the author.

**Advanced Collaborative Support (ACS) Program**

A critical aspect of having a multiyear funding model for the HTRC includes the opportunities that have been created through the HTRC Advanced Collaborative Support (ACS) program. This micro-grant program offers scholars in-kind development support supplied by the HTRC staff. The development support is specific to the project that is proposed to the ACS program by the scholar and includes assistance in migrating code to work on the HTRC compute platforms as well as in developing advanced worksets and other types of data enhancements. The first round of ACS awards was announced in January 2015. As was evident at the third annual HTRC UnCamp in 2015, the ACS program has encouraged many strides ahead in use of the HTRC platform in disciplines as diverse as English literature, language topic modeling, literary theory, and economics.

**A Scholars Commons Outreach Approach**

The HTRC is working to extend the pedagogy of its tools and services through an innovative partnership between the Scholars’ Commons at Indiana University Libraries and the Scholarly Commons at the University Library of the University of Illinois. These two libraries are partnering to develop instruction and models for conducting research consultations to support the integration of text data-mining concepts into scholarly research and the classroom. This partnership, in conjunction with three other libraries, is developing and testing a set of curricular material for use as “train the trainer” library workshops that will focus on teaching key concepts in text data-mining research. This effort, supported by funding from the Institute of Museum and Library Services (IMLS) Laura Bush 21st Century Librarian Program, will enable the delivery of effective text data-mining instruction, based on freely accessible training materials, to graduate students and other groups of scholars. The HTRC aims to enable more productive research for scholars as well as integrating its tools and services within a variety of disciplines.

**Conclusion**

Today’s digital scholars are embracing new opportunities to explore their disciplines in ways that are enhanced through the types of computational analysis that the HTRC provides. Nonetheless, the heightened level of integration and collaboration required to offer these new scholarly services requires a transformed working relationship between the academy, the digital scholar, and research support units as described in this article. For this reason, the HTRC is pooling resources that operate above any one campus, in order to create a shared service that can be used by all.

**Notes**

5. See the HathiTrust Research Center Documentation.
7. This concept was coined by scholar Franco Moretti in his book of the same name: *Distant Reading* (New York: Verso, 2013).

© 2016 J. Stephen Downie, Mike Furlough, Robert H. McDonald, Beth Namachchivaya, Beth A. Plale, and John Unsworth.