Maria Bonn

Computation, corpus, community

The HathiTrust Research Center today

Editors' note: One of our recurring columnists, Maria Bonn, returns this month to report on her conversations with the people behind the HathiTrust Research Center. She is going to fill us in on the center's background, accomplishments, and future directions.

The HathiTrust Research Center (HTRC), launched in 2011, in its own words:

> . . . provides [computational] research access to the public domain corpus of the HathiTrust [HT] Digital Library. The HTRC is a collaborative research center launched jointly by Indiana University [IU] and the University of Illinois [UIUC], along with the HT Digital Library, to help meet the technical challenges of dealing with massive amounts of digital text that researchers face, by developing cutting-edge software tools and cyber-infrastructure to enable advanced [computational] access to the growing digital record of human knowledge. The HTRC provides an infrastructure to search, collect, analyze, and visualize the full text of more than [5] million public domain works and is intended for nonprofit and educational researchers.[1]

Although HTRC has been on my radar since its inception, and I have attended one of the periodic "uncamps" designed to educate and engage potential and actual users, it was not until the occasion of this column that I educated *myself* about its past, present state, and future plans.[2]

HTRC was in gestation several years before its launch and, in fact, years before there was an HT. It was a twinkle in the eye of the original academic library partners in the Google Books project from the moment scanning negotiations began. After scanning commenced and prior to HT becoming a reality in 2008, what I will call the "Hathi Brain Trust" (a group of key figures in managing contractual and production relationships with Google), mostly based at the University of Michigan and IU, was considering the research potential of the extensive digital corpus that would emerge from the scanning partnership. In the course of contractual discussions with Google, this brain trust assured the inclusion of a clause that permitted large-scale, nonconsumptive[3] (a phrase coined by Dan Clancy, who was the chief engineer for the Google Books scanning effort) computational research from the outset of the project.

As an early initiative, the HT executive

Maria Bonn is senior lecture at the Graduate School of Library and Information Science at the University of Illinois at Urbana-Champaign, email: mbonn@illinois.edu

Contact series editors Adrian Ho, director of digital scholarship at the University of Kentucky Libraries, and Patricia Hswe, digital content strategist at Penn State University, at crlnscholcomm@gmail.com with article ideas.

committee designed a competitive process to assign responsibility for the research center to institutions wishing to lead in its creation. A task force of representatives from HT partner institutions created a call for proposals "to develop a HathiTrust Research Center."[4] The task force and the HT executive committee encouraged multi-institutional proposals. The successful proposal, awarded by the HT executive committee, came forward from IU and UIUC.

In my interview with him, John Wilkin, founding executive director of HT and a key player in the negotiations that made HTRC possible, said he recognized from the outset that the Google Books effort would result in the largest and most comprehensive digital body of research library collections anywhere. This corpus, he asserted, makes possible lines of investigation that aren't bounded by discipline. While he noted that its uses so far tend toward humanities research, he highlighted the fact that there are many kinds of questions that we can ask of this corpus. While in his own initial vision he imagined something like the census data enclaves (physical locations for qualified researchers to have access with a direct network connection to data that is typically sensitive), he was impressed and pleased that the reality has exceeded that vision, and that the research community now has a richer functional model in which HTRC supports remote application of large-scale computational resources to the corpus.

Almost five years into its life, HTRC has become, as HT's current director, Mike Furlough, said to me in conversation, the "first large-scale, service-focused research center for nonconsumptive research on a text corpora this diverse and extensive."

In my discussion with Ted Underwood, UIUC faculty member, scholar of textual corpora, and widely acknowledged HTRC power user, he called out HTRC as having the potential "to be the primary on-ramp for researchers who are trying to understand and access the enormous resources of HathiTrust." He went on to say that HTRC can help those researchers "understand what's in there, how to select a subset of works, and how to wade through complex metadata."

In these early years at HTRC, considerable effort has gone into developing tools and support around corpus building for the purposes of individual research efforts. HTRC refers to such an individualized corpus as a "work set, an aggregation of materials brought together for the purpose of analysis." As one HTRC staffer points out, scholars are excited by the potential of access to the entire corpus, but for research purposes those scholars need and want a subset. This January, the Andrew W. Mellon Foundation awarded researchers at UIUC a grant[5] to support further implementation of work-set models and to focus on making them more actionable.

In addition to its emphasis on technology and tools, HTRC has also partnered with the libraries of its home institutions to develop a complementary program for user training and instruction. When I spoke with Beth Namachchivaya, associate university librarian for research at the UIUC Library, she noted the importance of the program as many "scholars are interested in the possibilities, but aren't technically there yet." The libraries step in to help bridge this gap between interest and expertise and develop frameworks of support for scholars. As Namachchivaya pointed out, many support needs draw upon traditional library skills such as constructing frameworks for query, evaluation of retrieval, and (re)mediation of metadata. The UIUC Library is dedicated to making these frameworks more robust.

"Digging Deeper, Reaching Further: Libraries Empowering Users to Mine the HathiTrust Digital Library Resources," a three-year project funded by a Laura Bush 21st Century Librarian grant award from the Institute of Museum and Library Services, is developing curricular materials to be disseminated through a "Train the Trainer" program. The program aims to provide librarians with new content for instructional services, empower librarians to become

active research partners on digital projects at their institutions, and provide the foundation to transform academic libraries' digital scholarship and digital humanities centers into more data-intensive collaborative learning spaces.

Furlough, in considering the achievements of HTRC to date, argues, "[T]he biggest success story is what kinds of work have been enabled by use of the corpus through the Center."

One example of the research endeavors enabled by HTRC is the project entitled "Literary Geography At Scale,"[6] conducted by Matthew Wilkens of the University of Notre Dame. He uses natural language processing algorithms and automated geocoding to extract geographic information from nearly 11 million digitized volumes held by the HT Digital Library. The project extends existing computationally assisted work on U.S. and international literary geography to new regions, new historical periods (including the present day) and a vastly larger collection of texts. It also provides scholars in the humanities and social sciences with an enormous, yet accessible, trove of geographic information.

A different approach to realizing the potential of the magnitude of this digital textual corpus can be seen in "Taxonomizing the Texts: Towards Cultural-Scale Models of Full Text," a project led by Colin Allen and Jaimie Murdock of IU-Bloomington. Allen and Murdock are carrying 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. It is intended to show the relationship between a topic model created on a random sample of volumes and the entire category from which it is drawn.

Michelle Alexopoulos of the University of Toronto is engaged in "Tracking Technology Diffusion Through Time"[7] using the HT corpus. Alexopoulos, an economist, is using the vast historical record available in HT to study the diffusion of various technologies over time. By tracking the usage trends of 1,214 technology-related terms identified by Alexopoulos, such as *steam engine,* her research based on HT book content has the potential to overturn accepted theories about the economic and societal impacts of a technology.

While a burgeoning number of scholars are actively engaged with HTRC and deploying its tools, resulting in an impressive amount of computational textual research, it would be disingenuous to ignore the fact that HTRC has challenges to overcome now and in the years ahead. Both the former and the current executive directors of HT point out that there is still significant work to do in stabilizing the code, putting in place some real change management controls, and otherwise making HTRC a robust, well-defined, and consistently supported production service. Wilkin says that "based on what the leadership of HTRC has presented to its stakeholders, it is well on the way to being both robust and well defined." Furlough notes, "It's taking significant effort to develop secure, virtual environments for non-consumptive research." That effort is underway, and funding has been successfully secured for the development of the "Data Capsule" project, which will provide just such environments. This requires not only the technical development but also an investigation into needs and a clear understanding of what kinds of questions researchers want to ask of the corpus.

The need to understand and engage the scholars and researchers who are the target user base for HTRC was a recurrent theme in my conversations with the HTRC stakeholders.

Underwood, for example, argued for uses of the HT corpus that move beyond the portal that has been the focus of so much of the HTRC work so far. While he continued to emphasize that the portal is extremely important for many users for many types of analysis, he also believed that the most advanced analysis will be done by researchers downloading and manipulating

large data sets in their local environment. HTRC understands and appreciates this kind of need. The data capsule model currently under development, intended to provide "virtual and secure environments for scholars to conduct research using HathiTrust digital library texts,"[8] will afford that sort of high-level direct engagement.

To fully realize its aspirations, HTRC needs to connect as fully as possible with its actual and potential users. In our conversation, Namachchivaya stressed the need for "assertive outreach to scholars," and that HTRC "would benefit from infiltration into more scholarly and technical communities." She underlined the importance of maintaining momentum in outreach and bringing new people into process, and concluded on a hopeful note by pointing to the opportunity of reaching more scholars through the growing HT membership and its constituent communities.

Furlough rightly observes that "we have never had this kind of detailed data about our intellectual heritage." Through the work of HTRC, we see scholars enabled and supported in collecting and analyzing the material needed to create substantive narratives that advance our understanding of the world documented by that data.

I both entered and concluded my investigation into the work of HTRC with optimism about its possibilities and the contribution it will make to scholarship. Mindful of my audience in this venue and my own enduring interests in both emergent and traditional forms of scholarly communication, I often found myself wondering whether HTRC had a role to play in enriching the ways in which scholars can share their work and extend its reach and impact.

My greatest insight upon this topic came when Namachchivaya pointed out that while the work sets created by scholars working in HTRC are not in themselves sharable, the scripts constructed to create those work sets are freely sharable and reproducible through the HTRC portal. Particularly for humanities scholars, who are the most active users of HTRC to date, sharing the process of scholarly inquiry as well as the results constitutes a new practice of scholarly communication. In exposing and sharing this scholarly method and offering it up for the possibility of reuse, I see promise of a significant impact upon inquiry and the resultant body of knowledge.

Notes

1. See The HathiTrust Research Center at https://sharc.hathitrust.org/.

2. Throughout the course of my investigation, I benefited from generous gifts of time and conversation from Stephen Downie, Megan Senseney, John Wilkin, Beth Namachchivaya, Mike Furlough, and Ted Underwood. I extend heartfelt thanks for their part in making this column possible. The limits of my space here constrain me from fully sharing the insights that they all shared with me. I hope to do them more justice in the future.

3.This call for proposals is still accessible at https://www.hathitrust.org/documents/hathitrust-research-center-rfp.pdf.

4. Visit https://mellon.org/grants/grants-database/grants/university-of-illinois-at-urbana-champaign/41500672/.

5. See https://al.nd.edu/news/54897-breaking-new-ground-in-the-digital-humanities/.

6. See preliminary discussion at https://www.researchgate.net/publication/287250030_Towards_Cultural-Scale_Models_of_Full_Text.

7. Visit http://ineteconomics.org/grants-research-programs/grants/digitally-tracking-technologies-and-their-effects-across-time-and-space.

8. See https://wiki.htrc.illinois.edu/display/COM/HTRC+Data+Capsule.

ACRL Scholarly Communication Toolkit

Learn more about scholarly communication issues and trends in the ACRL Scholarly Communication Toolkit. The toolkit was designed by the ACRL Research and Scholarly Environment Committee to support advocacy efforts designed to transform the scholarly communication landscape. The toolkit is freely available online at http://acrl.ala.org/scholcomm/.