Overview: Enhancing the Humanities

The most transformative development in humanities research in the past two decades has been the development of large corpora of digital texts, whether those that are optically scanned, such as Cengage-Gale’s Eighteenth Century Collections Online (ECCO), or those that are born digital, like much of JSTOR. These online archives make hundreds of thousands of texts available to researchers around the world. They promise researchers the ability to easily find material that had previously been locked away in physical archives. While the availability of these corpora has prompted scholars to imagine the possibilities of opening up what Franco Moretti has called “the great unread” to scholarly attention, relatively little work has been done to develop tools to help scholars to meaningfully navigate these immense archives. For the most part, the search engines that scholars use to access these archives were developed more than a decade ago, and they are unhelpful in answering any but the most rudimentary research questions.

Simply put, the search engines provided by database vendors such as Cengage-Gale and ProQuest make little use of the rich textual material they purport to open for research. Many of the existing tools are predicated, like Google, on a simple form of keyword search, enriched, perhaps, with the possibility of including a few Boolean modifiers. While the simplicity of this approach makes these tools easy to use, it does little to enable the more nuanced, qualitative judgements that scholars tend to make when they are deciding what material is most likely to be useful to a research project. Complicating this matter further, these search platforms are often exceedingly opaque, slow to navigate, and poorly maintained. In short, using the search tools that are currently available, it is exceedingly hard for a researcher to find what he or she wants. One particularly vivid example of this comes from Peter de Bolla, a professor at Cambridge University. In the introduction to his recent book, *The Architecture of Concepts*, he has described how he spent hundreds of hours entering countless permutations of the phrase “human rights” into a Gale search platform. This predicament is, we believe, all too common.

Our team has developed the proof-of-concept of a search platform that answers these problems. Using a technique called *concept search* that was pioneered in legal practice for the purpose of expediting the process of document review, our search engine allows scholars in the humanities to search, not for single keywords, but for clusters of terms that define a concept. The algorithm is much more robust than those that underlie currently available search platforms. Ideally, it involves a researcher reading widely in her field, compiling a loading set of passages that she judges to be relevant to her concept, evaluating that loading set for a collection of up to thirty frequent search terms that she judges to define that concept, and searching for that collection of terms across millions of passages. This approach is superior to conventional keyword-search algorithms in a number of ways. It answers the problem of polysemy that plagues keyword search, because it searches for the aggregate signal of a collection of terms. It is also much more transparent than conventional search algorithms, because its results display the dirty OCR of the optically read file, rather than the apparently clean image of the source document. Furthermore, because our search platform allows the researcher to elect to receive a full set of results via csv spreadsheet, she can quickly sort and filter her data, allowing her to more quickly zero in on material that is likely to be relevant to her research.

Now that we have developed the underlying code for this search engine, it is relatively trivial to acquire and prepare new corpora. In addition to ECCO, we have already acquired and prepared the first module of ProQuest’s Early English Books Online Text Creation Partnership, as well as HathiTrust’s non-Google collection of historical texts. Together, these databases constitute millions of pre-twentieth century, non-copyright documents. Harvard Library is very invested in this project, and they are currently working to acquire the second module of EEBO TCP, Google Books, and JSTOR, along with other databases regularly used by humanists. Together these will bring the digital holdings available for our search platform to over 7 million books.

History of the Project

This project grew out of a research problem that I faced in beginning my current book project, *On Principle: Newton to Coleridge*, which concerns what I call “the culture of the principle” in the eighteenth century, the obsession with encapsulating books or even whole fields of knowledge in single propositions. Its starting point is the realization that many of the most famous statements of the period – Newton’s inverse-square law of gravity, Kant’s categorical imperative, and Bentham’s utility principle – constituted a genre to which intellectuals felt an imperative to contribute, and that readers expected to find whenever they opened a work of intellectual significance. David Hume put it succinctly. A principle, he said, offers “a whole science in a single theorem.” By promising to encapsulate and disseminate an author’s most fundamental ideas, the principle became the preeminent intellectual device of the Enlightenment.

During the early stages of this project, I encountered a problem that almost all humanist scholars face when they set out to write a book: what should I read? Specifically, how could I judge which texts contained these propositions, as well as passages in which authors discussed their contemporaries’ enthusiasm for this device. Sometimes, they point to this special sort of sentence in the very titles of their works, as in Newton’s *Principia* or Bentham’s *Principles of Morals and Legislation*. In other cases, the propositions in question are famous enough that they are trivial to locate: Malthus’s principle of population, for example. Early on, however, I began to suspect that less famous principles might lie buried in books, some obscure, and that the most common aids of research – long reading, critical bibliographies, colleagues’ advice, and Library of Congress subject headings – would be of only limited help. How then to proceed?

My initial research led me to search for various keywords in vendor-provided search platforms. I quickly realized how unreliable and inefficient this process was. If I searched for a particular keyword such as *principle*, might I not risk missing those instances where an author referred instead to his encapsulating proposition as an *axiom* or a *maxim*? Worse still, the platforms’ HTML results were exceedingly slow to sort through, especially because the relevance ranking function that Gale provided with its ECCO search engine failed to work if a search returned a certain number of hits. Compounding this problem, once I decided to click through to a particular volume, there was no way to judge quickly which part of it might be most relevant to my research.

Frustrated with these problems, I acquired the 207,000 underlying .txt files of ECCO from Gale and set out with my team to make a search engine of that would answer my needs. Working with Ista Zahn of Harvard’s Institute for Quantitative Social Science, I developed a rudimentary proof-of-concept search algorithm in R. To complete this phase of the project, we received a $12,000 grant from Harvard’s Lasky-Barajas Dean’s Innovation Fund for Research Computing, which funded the cost of storage and processing our data. We completed the proof-of-concept in July 2017 and presented the tool, along with some of the research findings that we have derived from its use, at an October 2017 colloquium meeting through Harvard’s Humanities Center. Subsequent to this presentation, we have performed searches for a number of graduate students and faculty including Jess Keiser of Tufts’s English Department and Andrew Warren of Harvard’s English Department. The use of the search engine was also mandated as part of an assignment in Stephen Osadetz’s Fall 2017 undergraduate lecture on The Eighteenth-Century Novel. His students were asked to use the tool to discover new contextual material relevant to their particular paper topics.

Although a search engine written in R is simple to use and adequate for our own research needs, it quickly became apparent that it would be unsuitable if we planned to make our tool more broadly available on a web platform. Our data was stored in immense document-term matrixes that had to be read into working memory with each new search, significantly increasing the processing time. Initially, a single search could take as long as an hour. In order to remedy this problem, we enlisted the help of Rashmi Singhal and Cole Crawford of Harvard’s Arts and Humanities Research Computing office to restructure our data, transferring it into relational databases. In November 2017, we completed a proof-of-concept of this new data structure on 1,000 texts from the ECCO corpus. One advantage of this approach is that many of the statistical measures we use have been pre-processed, so that, once the restructuration and processing are complete, searches will be able to be done in real time.