Digital Humanities from Scratch: A Pedagogy-Driven Investigation of an In-Copyright Corpus

Brian Croxall
brian.croxall@brown.edu
Brown University, United States of America

Following the publication of Franco Moretti's *Graphs, Maps, Trees*, scholars looking to apply digital humanities methods to literature have increasingly been drawn to "distant reading." The influence of distant reading in digital humanities is apparent not only in the work it has inspired (see, among others, Cordell and Smith; Elson, Dames, and McKeown; Jockers; Long and So; Rhody; and Underwood) but also for its regular inclusion as a method in courses introducing DH. "Teaching digital humanities," it turns out, often means "teaching distant reading."

Teaching students the techniques of distant reading can be challenging as it depends on re-framing the familiar object of study. But another difficulty altogether is that this approach depends on a digitized corpus; and such a corpus, in turn, depends on someone, somewhere doing the difficult labor of digitization. One might ask, then: if "teaching digital humanities" means "teaching distant reading," shouldn't it also mean "teaching digitization"?

In this paper, I will discuss a collaborative, multi-year assignment that I conducted in two of my "Introduction to Digital Humanities" courses at Emory University: the digitization and analysis of the complete works of Ernest Hemingway (Croxall). With the goal of teaching my students not only how to do distant reading but also about the intense labor that goes into corpus preparation, we digitized the whole of Hemingway's work in just two weeks. Working from newly purchased copies of the texts, the students and I rapidly scanned hundreds of pages, performed and corrected optical character recognition, and assembled a corpus—with each of us spending no more than 4 hours on the task. Our from-scratch corpus was composed expressly so we could draw important distinctions among Hemingway's works: individual works vs the whole collection; fiction vs non-fiction; and works published before while Hemingway was alive vs those that appeared after his death in 1961. I will detail what we learned from rapid digitization and how those lessons affected the second iteration of the assignment.

After preparing the corpus, students worked in groups to analyze the many works of Hemingway that they had not had time to read. Making use of Voyant Tools, they identified themes in the corpus and charted patterns that could never have been observed through regular, close reading methods. For example, the class confirmed that while Hemingway insists on writing about "men," the women to whom they are attached are inevitably just "girls." In an attempt to chart the patterns of Hemingway's diction, another group of students investigated the terms he uses to introduce dialogue. Unsurprisingly, the students discovered that "said" is by far the most frequent such term across the entire corpus. What was more surprising, however, was to observe that in late and posthumous writings, the frequency of "said" suddenly drops by 50%. In short, by building our own corpus from scratch, the students were able to conduct original research, something that is relatively rare for many undergraduates in humanities programs.

Building our collection of texts from scratch had two critical advantages. First, we were able to create a small, relatively clean corpus whose provenance we knew. This provided a sense of confidence in the data as we began to distant read. Furthermore, while our analysis of Hemingway's works was "distant" compared to traditional close reading of a single novel or story, it was not nearly as distant as projects that deal with several thousand texts. We became engaged, in short, in close-distant reading. Second, digitizing the texts ourselves allowed us to skirt a problem that frequently plagues distant reading texts from the twentieth century: copyright. As an educational endeavor focused on teaching the students how to prepare their research materials, this guerilla digitization project fell under the regime of fair use in the United States.

To close, I will discuss how students at Brown University and I have taken further steps with the Hemingway corpus and with their digital humanities education as we have used it as a means to explore the methods and utility of topic modeling. Topic modeling is frequently deployed to come to terms with large and unwieldy corpora (see Jockers; Nelson; Nelson, Mimno, and Brown; Underwood and Goldstone). But working with a small, relatively clean corpus that is created from scratch allows students to better

understand what takes place via unsupervised machine learning. At the same time, topic modeling allows us to ask in a new way some of the same questions that my former students had already uncovered: how does Hemingway's dialog differ from his prose? how different are the topics in Hemingway's fiction from those of his non-fiction? to what degree does his late—or even posthumous—work differ from what he wrote three decades earlier?

In the end, the process of modeling Hemingway becomes a means by which we can model all of digital humanities—both analysis and corpus creation—in a student-focused environment (see also Brier; Croxall and Singer; Harris; Hirsch; Jewell and Lorang; and Swafford). By doing digital humanities from scratch, students can be engaged in original research and see for themselves, from start to finish, how digital humanities gets done.

Bibliography

Brier, S. (2012). "Where's the Pedagogy? The Role of Teaching and Learning in the Digital Humanities." In Gold, M. K. (ed), *Debates in the Digital Humanities*. Minnesota University Press, pp. 350-367.

Cordell, R. and Smith, D. A. (2017). *Viral Texts: Mapping Networks of Reprinting in 19th-Century Newspapers and Magazines*. http://viraltexts.org/ (accessed 7 April 2017).

Croxall, B. (2015). "How to NOT Read Hemingway." *Intro to DH*. http://www.briancroxall.net/s15dh/assignments/how-to-not-read-hemingway/ (accessed 7 April 2017).

Croxall, B. and Singer, K. (2013). "The Future of Undergraduate Digital Humanities." Digital Humanities 2013, Lincoln, NE, July 2013.

Elson, D. K., Dames, N. and McKeown, K. R. (2010). "Extracting Social Networks from Literary Fiction." *Proceedings of the 48th Annual Meeting of the Association for Computational Linguisticsi,* Uppsala, Sweden. http://www.cs.columbia.edu/~delson/pubs/ACL2010-ElsonDamesMcKeown.pdf (accessed 7 April 2017).

Goldstone, A. and Underwood, T. (2014). "The Quiet Transformations of Literary Studies: What Thirteen Thousand Scholars Could Tell Us." *New Literary History* 45.3: 359-384.

Harris, K. D. (2011). "Pedagogy & Play: Revising Learning through Digital Humanities." Digital Humanities 2011, Stanford, CA, June 2011.

Hirsch, B. D. (2012). *Digital Humanities Pedagogy: Practices, Principles and Politics*. Open Book Publishers.

Jewell, A. and Lorang, E. (2016). "Teaching Digital Humanities Through a Community-Engaged, Team-Based Pedagogy." Digital Humanities 2016, Kraków, Poland, July 2016.

Jockers, M. L. (2013). *Macroanalysis: Digital Methods and Literary History*. Urbana Champaign, IL: University of Illinois Press.

Long, H. and So, R. J. (2016). "Literary Pattern Recognition: Modernism between Close Reading and Machine Learning." *Critical Inquiry* 42.2: 235-267.

Moretti, F. (2013). *Distant Reading*. London: Verso.

Moretti, F. (2007). *Graphs, Maps, Trees*. London: Verso.

Nelson, R. K. (2011). *Mining the Dispatch*. http://dsl.richmond.edu/dispatch/ (accessed 7 April 2017).

Nelson, R. K., Mimno, D. and Brown, T. (2012) "Topic Modeling the Past." Digital Humanities 2012, Hamburg, Germany, July 2012.

Rhody, L. M. (2013). "Revising Ekphrasis: Methods and Models." *The Association for Computers and the Humanities.* http://ach.org/2013/12/30/revising-ekphrasis-methods-and-models/ (accessed 7 April 2017).

Sinclair, S. and Rockwell, G. (2017). *Voyant Tools*. http://voyant-tools.org/ (accessed 7 April 2017).

Swafford, J. E. (2016). "Read, Play, Build: Teaching Sherlock Holmes through Digital Humanities." Digital Humanities 2016, Kraków, Poland, July 2016.

Underwood, T. (2013). *Why Literary Periods Mattered: Historical Contrast and the Prestige of English Studies*. Stanford: Stanford University Press.