EMORY UNIVERSITY
Department of English

November 29, 2020

To the Register of Copyrights,

I am an assistant professor of English at Emory University, with a courtesy appointment in Quantitative Theory and Methods. I am writing a book on the conglomeration of the United States publishing industry, under contract with Columbia University Press, for which I require computational analyses of thousands of novels published since 1945. I also teach computational analysis; presently, I am teaching Practical Approaches to Data Science with Text. I am writing in my individual capacity in support of an exemption to Section 1201 of the DMCA for the purposes of text and data mining research (TDM).

My book-in-progress, *The Conglomerate Era*, asks how the conglomeration of US publishing changed fiction. In the 1950s, almost every publisher in the US was independent. By 2000, only six multinational media conglomerates controlled a large majority of the sector. How can I make arguments about change at such scale? I cannot read enough fiction to make judgments myself. Instead, to detect patterns of change across thousands of novels across decades, I use TDM. TDM methods are exciting; they promise to expand considerably our understanding of literary history. But, at present, scholars in my field (post-1945 literature) are severely limited by Section 1201 of the DMCA.

The only option I and fellow scholars have is to use HathiTrust to build sufficient datasets. 1201 makes it otherwise impossible. I am grateful that Hathi exists, because otherwise I would be unable to pursue my research at all, but Hathi has considerable limitations. I can access Hathi's collection because Emory is a partner institution. Colleagues whose employers are not partner institutions, or who are independent scholars, lack access. To access works under copyright through Hathi, I need my own data capsule, a secure virtual computing environment. Hathi's data capsules are cumbersome. Navigating them takes much more time than does navigating a standard computer today. Opening and closing windows, accessing files, and other basic tasks require patience. One must also navigate between data capsules' "secure" and "maintenance" modes. In secure mode, internet is disabled. So if I need the internet for any reason while working in secure mode—for example, if I'm working with code, as I often do, and must do in secure mode to work with text under copyright, but need to debug a bit that's not working, as is common, by searching the web—I need to switch to maintenance mode. Switching between modes can take a few minutes. While, individually, these delays might sound minor, in aggregate they make my work two, three, or four times slower than it would be otherwise. Further, the challenges of working in data capsules are enough to inhibit most scholars from even attempting TDM in my field of study.

Hathi data capsules, further, have limited computing power. When I initially launched my capsule, I ran into limits with fairly basic analyses; for example, I had to cut some very long novels from my corpora because I did not have enough computing power to process them in my models, adding an artificial bias in my corpus. I won one of Hathi's Advanced Collaborative Support awards for the 2019-2020 year, granting me enhanced computing power. Even still, I do not have enough to run the most advanced and demanding models, like certain neural networks and transformers.

Beyond the limits of data capsules, Hathi's holdings themselves limit my research. Hathi is not comprehensive. Its holdings are the holdings of select university libraries, which do not acquire all fiction equally. There are, thus, vast gaps in Hathi. Worse, scholars do not yet know the contours of the gaps. This means that my findings based on Hathi's holdings are necessarily provisional and partial.

In my capacity as a professor, 1201 inhibits my ability to teach TDM. For my students to use TDM in our field of post-1945 literature, they need proficiency with Hathi's data capsules. In most cases, this is too larger a barrier to overcome. It takes too much time to teach students of literary studies, whether undergrads or grads, to use Hathi over the course of a semester. In practice, this means students turn to older periods where literature is not under copyright or to text they can acquire from the internet. So long as students do not pursue TDM in the field, the field will be stunted.

If I were exempt from 1201, I would be able to write a better, truer book about conglomeration. Maybe more profoundly, I would be able to teach TDM to the next generation of scholars who would transform our field of study. I am working to build the foundation for this future work. Laura B. McGrath and I have co-founded the Post45 Data Collective, which will launch in coming months. We are building a system of peer review and a single home for metadata such as author gender and race, MFA site, thesis advisor, and titles' publisher, prizes, literary agent. As of now, we have to cross-reference this metadata with Hathi IDs for scholars to study the metadata with the text. Exemption from DMCA 1201 would allow researchers far greater ease to do research with the data in the collective, which we believe will be transformative for our collective knowledge of literature and literary history.

Sincerely,

Dan Sinykin
Assistant Professor of English; Courtesy Appointment in Quantitative Theory and Methods
Emory University