HathiTrust UnCamp Research Plan

My object of inquiry for this research trip is the HathiTrust Digital Library itself, which I seek to examine for both the eighteenth-century and Digital Humanities elements of my dissertation. I am optimistic that its holdings will be invaluable to by study of popular works published in the 1790s, and certain that its unusual technological structure will form a key example for my study of the interpretive assumptions that underly scholarly databases.

The "UnCamp" being offered by HathiTrust is a combination of a conference and workshop, similar to the International Image Interoperability Framework (IIIF) event in New York at which I presented in 2016 with Alex Gillespie. At the HathiTrust event, I will have the opportunity to hear about a wide range of work being undertaken with the tools, meet and discuss future directions with the individuals actually writing the code for the archive, and receive hands-on instruction in the archive's underpinnings.

I have been interested in HathiTrust since my Masters Essay research, when I discovered that, of the 208 Gothic works which I wished to consider, only 7 were available as high-quality scholarly digital texts, but 115 were included in the HathiTrust Digital Library. At that time, I disregarded text-mining entirely, but the HathiTrust Digital Library has, in the past few years, emerged as an increasingly popular alternative corpus to the Eighteenth Century Collections Online / Early English Books Online (ECCO/EEBO) Text Creation Partnership (TCP) corpus. At the most recent Digital Humanities conference (DH2017, in Montreal), HathiTrust projects made up nearly half of the research presented on the eighteenth century.

The HathiTrust Digital Library is in many ways a problematic archive to incorporate into a multifaceted project. The page-image scans in its collection are largely provided by Google Books, and the plain-text files have been generated by Optical Character Recognition rather than scholarly editing. Thus, although the HathiTrust Digital Library is able to include many texts which are not available elsewhere, it is difficult to integrate into scholarly systems, and the quality of each individual holding is less reliable. A key objective of my trip to the HathiTrust UnCamp is to gain a detailed insight into the archive's technological underpinnings, and Google's impact as a private for-profit company integrating with a large number of university libraries.

The potential of HathiTrust lies in its provision of plain-text files for works well outside the scholarly canon. Non-TCP ECCO texts have, like HathiTrust texts, also been processed with OCR to produce plain-text files, and even in many cases double-typed, but ECCO uses this plaintext exclusively for its search engine and does not release these text files for scholarly use. HathiTrust does not make its text files easy to access, but it does allow scholars to work with them. Its catalog lists 4,385 works published in the United Kingdom between 1789 and 1799, of which 4,378 are available as full texts — an order of magnitude more than the 466 plain-text transcripts in the ECCO-TCP corpus published during that decade. Moreover, my research to date suggests that many of HathiTrust's holdings are not duplicated in other archives, providing unique access to often-ephemeral popular works.

I remain, in some ways, skeptical of HathiTrust's technological assumptions, but this skepticism increases my interest in the archive. My dissertation seeks to provide useful new knowledge both about literary production in Britain 1789-99 and about database-based literary study. To fully address the current state of literary corpora, it seems crucial to properly account for what the HathiTrust Digital Library archive can and cannot be used to discover.