The HathiTrust+Bookworm Project as a Model for Collaborative Research at Large Scale

Abstract:

We present the HathiTrust+Bookworm (HT+BW) project, a two-year (Sep 2014 — August 2016) multi-disciplinary, multi-institutional endeavor funded by the National Endowment for Humanities (NEH) and involving collaborators from the fields of scholarship, librarianship and education in the humanities, and from culturomics, computer science, and applied mathematics.

The HathiTrust Digital Library (HTDL), of which the HathiTrust Research Center (HTRC) is the research arm, comprises over 13 million volumes of digitized library materials, as well as their accompanying bibliographic metadata. The HTRC team includes personnel who combine an interest in humanistic research based on texts and their interpretation with openness towards the possibility of scientific and statistical analyses of corpora. The Bookworm tool was created at the Cultural Observatory (originally at Harvard, now at Rice), which previously collaborated with Google to create the Google Ngram Viewer; the lead developer of Bookworm, Benjamin Schmidt, is now on the faculty at Northeastern. This team has a similar interest but complementary expertise: HTRC creates corpora, which Bookworm is designed to analyze.

Our collaboration is characterized by several shared values, such as a commitment to open-source development. We hope to reach out to humanists at large by presenting a model of how scientific approaches can complement and enhance traditional modes of humanistic inquiry. We also hope to reach a scientific audience, enabling and encouraging engagement with individual texts and humanistic methods. Our presentation will discuss our collaboration, including lessons learned and challenges faced along the way that may be useful for others engaging in similar interdisciplinary endeavors.