

# Data Management Plan

## 1. Roles and responsibilities

This data management plan will be implemented and managed by J. Stephen Downie. Sayan Bhattacharyya and the HTRC PhD Research Assistant will assist with project documentation and data deposit throughout the grant period.

In clarifying roles and responsibilities, it is also critical to distinguish between data generated as part of project activities, and the data against which the tools being implemented are run. As such, we must clearly delineate: 1) HathiTrust data; 2) HathiTrust Research Center data; and 3) project data. The HathiTrust data refers to the 11 million (and counting!) digitized volumes of the HathiTrust Corpus. The HathiTrust Research Center currently holds over 3 million digitized volumes that are out of copyright and in the public domain. The HathiTrust Research Center also includes HTRC worksets and other data that are available to, but not necessarily generated by, this project. Project data refers specifically to data generated during the course of project activities.

In all cases, the HathiTrust, the HathiTrust Research Center, and the project team are committed to providing long-term preservation and access to data wherever permitted by law. For more information on data availability through the HathiTrust, see <http://www.hathitrust.org/data>. For more information on data availability through the HathiTrust Research Center, see [http://www.hathitrust.org/htrc/technical\\_documentation](http://www.hathitrust.org/htrc/technical_documentation).

## 2. Expected data

The HathiTrust data is under the purview of the HathiTrust and not available for release. However, the project team expects to generate new data derived from analysis of the HathiTrust corpus using the Bookworm instance, which will be made publicly accessible. The project team will also generate a suite of technical documentation, user documentation, and classroom training materials. The project's source code and documentation for both HTRC and our contributions to Bookworm will be available publicly and freely online through GitHub for download, distribution, and continued contributions and modifications.<sup>4</sup> The final project white paper and all project disseminations (posters, papers, etc.) will be deposited in IDEALS (the University of Illinois' digital repository for research and scholarship)<sup>5</sup> linked from the HTRC website.<sup>6</sup>

Our goal is to facilitate use of HTRC worksets in Bookworm and to use Bookworm to save new HTRC worksets or modify existing HTRC worksets. Worksets fall under the auspices of HathiTrust Research Center data, and each workset is generated by an individual scholar who is given the option whether their worksets are public or private.

## 3. Period of data retention

---

<sup>4</sup> <https://www.github.com/bmschmidt/BookwormAPI>

<sup>5</sup> <https://ideals.illinois.edu>

<sup>6</sup> <http://www.hathitrust.org/htrc>

The HathiTrust Research Center and the project team agree to provide scholars with ongoing access to their HTRC worksets and data derived from Bookworm analysis for a minimum of 5 years, but the HTRC envisions making this permanent feature of the Center. Worksets that are not flagged as private will also be available to the general public. Project code and technical documentation will be retained in GitHub indefinitely. Project disseminations and training materials will be retained in IDEALS indefinitely.

#### **4. Data formats and dissemination**

For each digitized volume in the HathiTrust Research Center, available data includes bibliographic metadata as a METS XML file and a zip file containing a separate txt file for each page of text. These data utilized by the project team but fall under the purview of the HathiTrust Research Center. HTRC workset are currently exportable as CSV files, and the Center intends to continue workset development to include additional features and allow for greater interoperability. The data underlying individual charts derived from analysis of HTRC's data using Bookworm will be accessible to users as a CSV file from the web site.

#### **5. Data storage and preservation of access**

HathiTrust Digital Library is a digital preservation repository and highly functional access platform. It provides long-term preservation and access services for public domain and in copyright content from a variety of sources, including Google, the Internet Archive, Microsoft, and in-house partner institution initiatives.<sup>7</sup> The HathiTrust Research Center enables computational access for nonprofit and educational users to published works in the public domain and, in the future, on limited terms to works in copyright from the HathiTrust using a non-consumptive paradigm that protects the security of sensitive in-copyright data.

Bookworm provides yet another access point for computed data derived from analysis of the HathiTrust corpus. We will provide ongoing access to HTRC worksets and data derived from Bookworm analysis for a minimum of 5 years, but the HTRC envisions making this permanent. Source code will be deposited and openly available in GitHub. All project documentation, working papers, technical reports, and other research material will be deposited in the IDEALS system. We will also provide access to these materials from the HTRC and HT websites.

#### **6. Post-Award Monitoring**

The project team is committed to ongoing evaluation of our data management practices. We will continue monitor community standards for data formats, data storage, data preservation, and data sharing. If we alter our original plan due to changes in perceived best practices in the data curation community or data are generated during the course of the project that were not previously described in the data management plan, we will outline project decisions for ongoing management in interim and annual reports to NEH as well as in our final project white paper.

---

<sup>7</sup> The HathiTrust is a TRAC-certified digital repository. For information about its digital preservation policies, see: <http://www.hathitrust.org/preservation>