6 Data management plan

6.1 Roles and responsibilities

The project director and the postdoctoral fellow will have primary responsibility for data management during the grant period. The project director will be responsible for data management after the expiration of the grant. The director and postdoctoral fellow will be assisted by members of the cyberinfrastructure team headed by Chris Sweet within the Notre Dame Center for Research Computing for issues related to data management during the grant. For long-term preservation issues, the CurateND team based in the Hesburgh Libraries will assist with repository issues.

6.2 Expected data

As noted in the project narrative, we will produce data of two types: geographic data associated with digitized texts and code used both to produce that data and to implement the project interface site. The anticipated total size of the data generated by the project is estimated to be several terabytes, almost all of which is accounted for by the dataset itself. Processing and interface code occupies no more than a few gigabytes.

For purposes of compliance with OMB Circular A-16 and Executive Order 12906, we certify that the geospatial data products we propose will be produced in compliance with applicable guidance from the Federal Geographic Data Committee.

6.3 Period of data retention

Data will be retained on all platforms through at least 2023 and in Notre Dame's institutional repository indefinitely.

6.4 Data formats, storage, and preservation of access

Data will be made available in multiple formats as appropriate for its intended uses. Geographic data will be accessible through the project interface and downloadable in CSV and JSON formats. Project source code will be managed on GitHub or a comparable service in the event that GitHub becomes unavailable or unsuitable at a future date.

We already own the hardware on which the project is hosted. The server and disk array have substantial overhead for increased use; as noted in the project narrative, we have 64 cores, 64 GB RAM, and 34 TB of RAID storage collocated in Notre Dame's high-availability data center. Network connectivity and data backup are provided on an ongoing basis as indirect costs.

We see two aspects of the project that will require resources beyond the end of the grant period, and have made provisions for both.

First, the geographic dataset will need to be stored and made available to users in minimally processed form. This need will be met via three independent channels: the project site, Notre Dame's institutional repository (CurateND; https://curate.nd.edu), and the HathiTrust Research Center. The project site allows us the greatest flexibility in formatting, subsetting, and interactivity. Its costs after the grant period are minimal and will be covered

for at least five years by Notre Dame funds already allocated for that purpose. CurateND is specifically designed for ongoing, long-term preservation of research products and is provided to Notre Dame-originated projects on a no-cost basis. Finally, our collaboration with the HTRC is designed to make both our data and our computational methods available to the largest possible community of users. The HTRC is committed not only to hosting our full dataset in a form compatible with their existing products, but also to implementing our ingest pipeline so that it can be applied to volumes that are added to the HathiTrust Digital Library in the future. This ensures the preservation of grant-produced data, extension of that data to newly acquired materials, and additional documentation of the processing and ingest methods.

The second element of the project that will require support beyond the end of the grant period is the user interface site. This site is hosted on the production server described above and consists of code written primarily in Python and JavaScript. We have three channels of long-term sustenance for this aspect of the project. As in the case of the underlying dataset, the server that hosts the project will be supported for a period of at least five years through funds granted by Notre Dame. We will deposit the full source code for the project in CurateND for long-term storage, with plans to provide updated, versioned releases at major milestones. Finally, we will make all source code available on GitHub (which currently hosts the development code) or a comparable platform under an open-source license for public use. Using GitHub helps to ensure easy practical access to the project's code by interested users and facilitates user contributions back to us, thereby extending intellectual collaboration around the project.