Data Management Plan

Roles and Responsibilities

This data management plan will be implemented and managed by Matt Shoemaker, under the project supervision of Peter Logan. Mr. Shoemaker will oversee maintenance, backup, and archiving of data generated by the project. And he will manage the transfer of data for project research to Drexel's Metadata Research Center. He will also be responsible for final project data transfers to the OTA and Humanities CORE repositories. All repository data will be publically accessible. If Mr. Shoemaker leaves Temple University during the course of the grant, his role will be taken over by his successor as Coordinator of the Digital Scholarship Center.

Data Storage Hardware

During the production phase, all data is stored locally in the DSC on its internal server, consisting of a networked pair of 6TB hard drives in a RAID1 configuration. Access to the DSC server is limited to project participants authorized by Dr. Logan. Project files are automatically archived daily to a separate 4TB external hard drive. Files are also synchronized daily with an online Box service provided by Temple University, with remote access limited to team members.

Data Formats

1. Image files. These are duplicates of external image files downloaded from the Hathi Trust or the Internet Archive (see Appendix D).
2. AFR Project Files. These are file folders generated by ABBYY FineReader to organize large amounts of page images for scanning.
3. AFR User Files and Dictionary Files. These are proprietary files that store custom information used to fine-tune the OCR process.
4. HTML files. There are two types:
    a. Those output by AFR, with one file for each print page of the Encyclopedia.
    b. Final files generated by XSLT from the project's TEI-XML data files with full metadata, for uploading to the OTA at the completion of the project.
5. XSLT scripts.
6. TEI-XML files. These contain the core textual data of the project and the generated metadata.
7. Oxygen Project Files. These are small data files for use by Oxygen XML Editor to organize large numbers of XML files.
8. TXT files, generated by XSLT from the project's TEI-XML data files for internal use in testing the viability of using online topic-modeling to identity concept drift.
9. Project spreadsheets documenting the creation of all files and any modifications made to them.

Data Organization Plan

Core Data and Metadata Organization

In order to output consistent metadata with the textual data, our 110,000 individual entry files are organized within a hierarchy of file dependencies, with entry files at the bottom level (see Appendix F). Above them are container, or "wrapper" files for each volume, followed by wrappers for each edition, and finally a corpus wrapper containing basic encoding to all files. Files at lower levels of the hierarchy inherit the attributes of those above them. They also have unique metadata describing the content of each entry. This structure allows us to keep the encoding of entry files as simple as possible, while the series of dependencies means that each entry will have a comprehensive set of metadata associated when output into its final format for repository storage.

Production File Organization

A comprehensive data organization plan for use by project participants is spelled out in the "Data Organization" section of the online Project Manual, specifying the storage location for all forms of project data (https://tu-plogan.github.io/#source/data_organization.html).

Expected Data for Preservation

The project will generate approximately 110,000 different files of textual data. Each file represents one entry in the four Encyclopedia editions. These files serve as "master files" that are used to generate output in other file formats for end-users, including researchers. From the files, we will generate final "digital edition" files TEI-XML format the include all of their critical metadata within each file, and so serve the needs of researchers working from file metadata, rather than the textual data alone.

Permanent Preservation and Access

One copy of these "digital editions" files will by uploaded to Humanities CORE in bulk form, with all files for each print volume contained within a single ZIP or RAR archive. The four editions contain a total of 89 volumes, so the archive will consist of 89 ZIP or RAR files. The textual data will also be output in two other formats: HTML and TXT, the most useful formats for researchers who want to work with the complete data set. The data set in these alternate formats will be uploaded to Humanities CORE in the same ZIP or RAR archive form. Humanities CORE promises permanent storage and open access for data deposited with them.

A second copy of the data will be made publically accessible in perpetuity by the Oxford Text Archive, when they are uploaded at the end of the project. OTA will make them publically available for free as HTML files readable online. Additional details will be negotiated with them, such as whether or not they wish to supply alternate formats, like EPUB or TXT, and the DSC can provide them with those formats.

Five-Year Preservation

All AFR Project, User, and Dictionary files, plus the XSLT and Python scripts used in the production process, will be preserved internally in the DSC for a minimum of five years, to allow us to regenerate the raw textual data at any time. The XSLT and Python scripts used to modify that data and convert it to TEI-XML will also be uploaded to the project GitHub site, for free download by anyone interested (https://github.com/TU-plogan/encyclopedia-project) during the same time period.

Test Data

In the final stage of the process, Dr. Logan and Dr. Greenberg will trial two different methods for identifying concept drift in the data set and write a journal article explaining their results. The data for those tests will be preserved and posted on Humanities CORE, for use by readers of the journal article or others interested in the outcomes. Some of it will also be shared at the DH2019 conference during our presentation.

Continuing Research

The full textual data set and master files will also be retained by Dr. Logan for continuing research on nineteenth-century knowledge. Dr. Greenberg also will retain a copy of the TEI-XML digital edition files for use in her future teaching.