6. Data Management Plan

Roles and Responsibilities

The implementation of the data management plan during the course of the grant will be the responsibility of Hannah Alpert-Abrams. Three parties are responsible for maintaining the data. Taylor Berg-Kirkpatrick at UC Berkeley will maintain an open-access version of the modified Ocular code. The Early Modern OCR Project (eMOP) at Texas A&M University, under the direction of Laura Mandell, Anton duPlessis, and Matt Christy, will be responsible for maintaining a second open-access copy of the Ocular code. The University of Texas Libraries will provide preservation services, as well as web access, for the transcriptions. This will be mediated by Ladd Hanson, Associate Director for Information Technology Architecture and Strategy, and Aaron Choate, Head of Technology Integration Services at UT Libraries.

Expected Data

Our project will produce data in three forms: software code, transcriptions, and language models. All data and transcriptions will be stored and backed up daily on GitHub, where they will be made publicly available via the web. They may be viewed as individual files, or downloaded as a zipped archive.

In addition, the software code will be preserved and made available on the personal website of Taylor Berg-Kirkpatrick, the developer of Ocular. It will also be made available to anyone using eMOP, the Early Modern OCR Project hosted by Texas A&M University. eMOP will maintain a copy of the software on its servers and provide access via the eMOP website and Github code respository. The transcriptions will be produced in the form of XML files using the ALTO standard. These files will be preserved by UT Libraries.

This project draws on data in the form of PDF scans of books from *Primeros Libros*, currently preserved and made available at the UT Libraries-hosted project website. Our OCR system uses language data collected from the open-access text repository Project Gutenberg, and from private collections. All language data is preserved as plain-text files and backed up daily on GitHub. Though some of the data in these models is proprietary, language models based on the data will be made available for out-of-the-box use with the OCR system.

Data Formats and Dissemination

Format: All transcriptions will be made available in their final format as text files, with markup and metadata encoded using ALTO, the Library of Congress XML format for OCR transcriptions. This will include language tags for all transcribed words, along with document-level metadata drawn from the *Primeros Libros* website. Transcriptions will also be published on the *Primeros Libros* website.

All data produced during the course of this project will be made freely and publicly available. The OCR code will be held under the GNU General Public License version 3, in accordance with the original Ocular code. The transcriptions, like all content associated with the *Primeros Libros* project, will be public domain.

Dissemination: All data stored on GitHub is made publicly available to anyone with web access, and does not require any form of registration or user account. The same is true for the *Primeros Libros* website.

Likewise, all code developed for and by eMOP and its workflow is available via Github for download and use.

The only proprietary data in this project are transcriptions of early modern Nahuatl documents made by scholars in the field. Though we cannot make this data available, our project will produce new digital corpora of Nahuatl that will be made freely available. We will also be able to provide a Nahuatl language model based on the proprietary data, which can be used to run our OCR system on Nahuatl documents.

Data Storage and Preservation of Access

All data stored on GitHub will be made available indefinitely, and can be maintained over time; it will also be available for user collaboration in the form of bug fixes and other features. The University of Texas Libraries has a long-term commitment to the preservation of all *Primeros Libros* content,, including the transcriptions produced as part of this project. The Initiative for Digital Humanities, Media, and Culture at Texas A&M University is committed to the long-term storage, maintenance, and upkeep of data related to the Early Modern OCR Project.

Following consultation with the University of Texas Libraries staff members, we plan on depositing the research data in the University of Texas Digital Repository (UTDR). We will submit the necessary metadata and other resources to make the data accessible for future users. The UTDR will preserve the data indefinitely and is committed to responsible and sustainable management of submitted works as well as associated descriptive and administrative metadata, by employing a strategy combining the following: nightly secure backups, storage media refreshment, file format migration (including possible migration to standard formats during submission), and assignment of a unique and persistent URL.