9. Data Management Plan

Responsibilities

PI Zeldes will oversee the data management plan. He will oversee the development of digital tools, manage the servers described below, and manage the GitHub repositories. Co-PI Schroeder will oversee the project website and digital document and collection management. At the end of year 1, the DH Specialist will take over the maintenance of the project website (including documentation) and GitHub repositories. Staff from Georgetown's Institutional Repository (IR) and University of the Pacific IR will assist with long-term data retention (see *Sustainability*).

Expected Data

Our expected data fall into the following four general classifications: digitized texts, tools, corpus database, documentation. The data types within each classification are described in more detail below in *Data Formats and Dissemination* (see also *Final Product and Dissemination* for more details on data content).

Data Formats and Dissemination

The project participants embrace the principles of timely, rapid, and open-source data distribution. Open source and open access practices assist in data management and dissemination by creating resources that can be used and disseminated beyond the life of the project. For example, Coptic SCRIPTORIUM published a pilot Coptic language treebank for linguistic research as part of the Universal Dependencies project (http://universaldependencies.org/), which coordinates universal, cross-language standards to facilitate multi-lingual research using the same annotations. Our work now appears alongside those of over 50 other languages, although the resource is still in development and only at a pilot stage.

Tools and Technologies: The digital tools described in this proposal will be written in Python, Java, JavaScript, and other languages. We will also pursue the adaptation of existing open-source tools. The digital tools will be developed and distributed as open-source software and free public downloads under open-source licenses, normally the Apache 2.0 license or similar (depending on usage of imported library licenses, when those are incompatible with Apache 2.0). The software will be developed and distributed on the project's GitHub repositories, with links to these resources provided from the Coptic SCRIPTORIUM website and from Georgetown University Computational Linguistics webpages.

Digitized Text: The digitized text will be raw data in the form of text files, TEI XML files, PAULA XML, SGML files in English, Greek Unicode and Coptic Unicode (displayed in the Antinoou font created by the International Association of Coptic Studies.) We anticipate no changes to Unicode standards for Coptic characters that would affect our work. The digitized text files will be stored on version-controlled servers at Georgetown University (e.g., the Corpus Linguistics server at http://corpling.uis.georgetown.edu/) using Git or Subversion, as well as the Coptic SCRIPTORIUM GitHub site at http://www.github.com/CopticScriptorium.

Some of the project text data will be drawn from ancient and medieval manuscripts or scans of books out of copyright. Under intellectual property law in the United States, the text from the manuscripts is in the public domain; editorial work can be under copyright. The project will not be publishing online editorial work that has been published in print or in digital formats under existing copyright. Prepublication working files will be stored on the above-named servers. Images of manuscript pages will not be circulated outside of the direct project participants unless authorization has been provided by the image providers.

Digital publication of textual data will occur on Coptic SCRIPTORIUM's site and the Georgetown Corpus Linguistics server and site. The subdomain of Coptic SCRIPTORIUM's site, data.copticscriptorium.org, publishes annotated digitized text and is hosted on an Amazon Web Services server. Published text, corpora, and textual annotations will be distributed under the

Creative Commons attribution (CC-BY 4.0) license whenever possible (https://creativecommons.org/licenses/by/4.0/legalcode).

Documentation: The project will provide documentation of the tools, technologies, methods, standards, and data models developed and used during the grant period as text files and pdfs. Video tutorials with screenshots may be made available on YouTube under public licenses. Documentation will be labeled with date and version information and disseminated under open-source licenses, such the Apache license, GNU Free Documentation License, and the Creative Commons Attribution (CC-BY 4.0) License. News and information about project progress (with links to the tools or more detailed formal documentation) will appear on the project blog; text of blog posts will be licensed Creative Commons Attribution (CC-BY 4.0).

Period of Data Retention

Tools, documentation, and database files will be released as they are created on our public GitHub repositories, with links provided on the project website and the Georgetown University Corpus Linguistics website. Digitized text files which contain edited, annotated text will also be released as soon as the editorial process is completed and made available indefinitely (see below).

Data Storage and Preservation of Access

Long-term storage and access for tools and published annotated digital text will be provided by the Georgetown University and University of the Pacific Institutional Repositories (IRs). Both of these IRs are open access. Georgetown's repository at Lauinger Library, called DigitalGeorgetown (http://www.library.georgetown.edu/digitalgeorgetown), is powered by DSpace, an open source software solution, and hosts the scholarly output from faculty members, institutes, centers, publishers, and round tables across all of Georgetown's campuses, with the goal of providing long term open access to this material. The repository includes working papers; journal articles; theses and dissertations; data sets; citation and image databases; and much more.

In addition to helping disseminate the scholarly output of the University, DigitalGeorgetown also serves as a robust digital preservation platform. Each item within the repository is assigned a handle, which acts as a persistent identifier and permanent URL to the resource. This ensures that links to items within DigitalGeorgetown will always resolve, even if the software and underlying architecture of the repository change over time. In addition to the use of the Handle System (https://www.handle.net/), DigitalGeorgetown also utilizes the Academic Preservation Trust repository as a back-end dark archive for digital preservation. APTrust (http://academicpreservationtrust.org/) is a consortium of higher education institutions, including Georgetown, which are committed to providing a preservation repository for digital content. Georgetown University Library's Digital Preservation Policy ensures that there is a commitment to preserving all content from DigitalGeorgetown into APTrust, irrespective of the bitstream and data format and storage requirements. Metadata is also harvested by the Digital Public Library of America (DPLA), thereby increasing the findability of resources produced in the project. The University of the Pacific IR can hold additional copies of our files in all formats as well. Repository Staff will assist Faculty in ensuring proper documentation and metadata for the IR.

The final White Paper will be disseminated on the NEH website, the two university IRs, and the project website (under Reports at http://copticscriptorium.org/reports, along with prior NEH reports and White Papers). Project participants will also disseminate their results in journal articles and at professional conferences and symposia. The most important of the latter events are the Society of Biblical Literature annual meetings, the quadrennial international congress for the International Association of Coptic Studies, and the annual international Digital Humanities conference. In Computational Linguistics, technical advances will be presented at regional and international conferences of the Association for Computational Linguistics (ACL) and Corpus Linguistics conferences (for example the Language Resources and Evaluation Conference, Treebanking and Linguistic Theories, or the international Corpus Linguistics conference).