

7. Data Management Plan

Responsibilities

Project Director Caroline T. Schroeder will oversee the data management plan, ensuring that all processes are implemented. Schroeder will also supervise the creation and management of the SCRIPTORIUM public web platform. Associate Director Amir Zeldes, as the technical director of the project, will co-direct the creation of digital tools, manage the server for the public ANNIS repository, and manage the versioning software for the ANNIS repository and the pre-publication data on separate server space.

Expected Data, Collection Methods, Data Formats, and Data Dissemination

Digitized Text: The digitized text will be raw data in the form of text files, XML files, and Microsoft Excel files in English and the free, Unicode Coptic Antinoou font. (We anticipate no changes to the Unicode standards that would affect our work.) The digitized text files and Excel files will be stored internally on a version-controlled server but not published. Some of the text data will be drawn from ancient and medieval manuscripts. Under intellectual property law in Germany and the United States, the text from the manuscripts is in the public domain; editorial work can be under copyright. SCRIPTORIUM will use data that has appeared in publication only if the copyright has expired, if permission has been granted, or if we have also consulted with the original manuscripts to produce our own original editorial work. Other text data will be drawn from online texts that are licensed for use and are of sufficient scholarly quality. Examples may include Papyri.info TEI XML files (with the open access, open source CC-BY license) and biblical texts from the Sahidica.org project (which are derived from the standard digitized Coptic New Testament used in the field and are licensed for academic use with attribution).

Tools and Technologies: The digital tools to annotate and format the text files will be written in Java, Python, or other scripting languages. We will pursue the adaptation of existing open-source and open-access tools (such as TreeTagger, LAUDATIO, SoSOL) and the development of our own tools. We will distribute the tools via links on the SCRIPTORIUM web platform as free public downloads under open-source licenses, such as the Apache 2.0 license. The software will also be distributed on GitHub, which is already used for ANNIS development. Version control for the tools in development will be managed through a standard version control software, such as Subversion.

Richly-Annotated Corpus Database: The files will be created using digital tools and manual annotations. These processes will produce files in formats such as text files, .csv files, Excel files, and XML files. The final output for the database will be the relANNIS format for the ANNIS infrastructure,³¹ as well as human readable formats generated by ANNIS and the open-source converter framework SaltNPepper.³² SaltNPepper enables conversion to and from a variety of formats for easy standoff markup of tokenized text. Metadata for the ANNIS files will include metadata conforming to TEI XML standards (the EpiDoc subset) and TEI versions of the documents will be downloadable. The corpus comprised of these files will be publicly available and accessible via the links on the SCRIPTORIUM platform under an open-source license, such as the Creative Commons license. (Open-source ANNIS code is currently distributed on the ANNIS project site. SaltNPepper is distributed on its project site. SCRIPTORIUM Associate Director Amir Zeldes is a primary contributor for both of these projects, which are funded by the Collaborative Research Center on Information Structure in Berlin and Potsdam.) Visualizations of the text in HTML will also be available for viewing and download online.

Documentation: SCRIPTORIUM will provide internal documentation of developing versions of the tools, the corpus database, and methodologies and best practices. When the corpus and tools are disseminated at

³¹ "Download ANNIS - A Tool for Searching in Multilevel Linguistic Corpora", n.d., <http://www.sfb632.uni-potsdam.de/d1/annis/download.html>.

³² "SaltNPepper - SaltNPepper - Korpling Projects", n.d., <https://korpling.german.hu-berlin.de/p/projects/saltnpepper/wiki/>.

the end of the Start-Up phase, public documentation for each element will be created and disseminated freely on the SCRIPTORIUM web platform. Documentation will be labeled with date and version information and disseminated under open-source licenses, such the GNU Free Documentation License, and the Creative Commons Attribution (CC-BY) License.

Research data generated from searches, queries, statistical analyses, and visualizations of the corpus: This data will be disseminated at conferences and published in journals and conference proceedings. Selected aggregate datasets (e.g. frequency lists) may be offered directly on the website depending on feedback from interested users.

We expect no legal or ethical restrictions on our data. Our digital textual data is based on transcriptions of ancient texts, which are no longer under copyright restrictions. We are not reproducing images of the objects themselves without permission from their repositories, nor are we reproducing editorial work under copyright from published editions still under copyright.

Period of Data Retention

The SCRIPTORIUM team embraces the principles of timely, rapid, and open-source data distribution. SCRIPTORIUM is a multi-year project. Tools, methodologies, documentation, and the digital corpus will be released through the SCRIPTORIUM platform as they are created. New versions of the digital corpus will be released with documentation about additions as new data is generated. The SCRIPTORIUM staff will maintain the tools, corpus, and documentation on the SCRIPTORIUM platform for a period of no less than five years.

Storage of the corpus will be provided by LAUDATIO (Humboldt University) and the Perseus Digital Library (see letters from Prof. Anke Lüdeling and Prof. Greg Crane).

Data Formats and Dissemination

Data for the duration of the project and the public SCRIPTORIUM web platform will be stored, hosted, and managed on servers provisioned by Schroeder and the University of the Pacific and by Zeldes and the Institute for German Language and Linguistics, Humboldt University. The SCRIPTORIUM team will support and manage the web platform and corpus database for a period of no less than five years. The LAUDATIO project at the at Humboldt University and the Perseus Digital Library at Tufts University have also offered to host the corpus long-term.