

Data Management Plan

The proposed work detailed in the narrative and project management documents comprises two inter-related but parallel elements – the collection and computational analysis of crowd-sourced transcriptions of ancient texts and the qualitative analysis of data for the classics research thrust. This document describes the overall data management plan for the activities in both of these elements.

Types of Data

Table 1 shows examples of the key types of data products used in the proposed work over the main project activities. It also shows the main code-base utilized for each of the four main activities.

	Project Activity	Data Types (format)	Code-base
Project Element 1	Crowd-sourcing data collection (Ancient Lives interface)	Fragment images (.jpg) Database records (SQL)	Zooniverse Ouroboros (Rails 3)
	Computational analysis of crowd-sourced data (Greek and Coptic processing pipeline)	Database records (SQL) Unicode text files Image files (.jpg)	Matlab and Python scripts, C++ executables
	Curation of crowd-sourced data (Curation interface)	Unicode text files Database records (SQL)	Datomic/Clojure C++, Javascript, Python
Project Element 2	Collection of research data (Manuscript identification and Universal Search Engine)	Text files (.doc, .xls, Unicode .txt) XML files	C++, Javascript, Python

Table 1: Examples of the data types, formats and code-base for the main activities in the project.

Metadata standards, archiving, access/use rights, security and dissemination plans

We now provide more descriptive details of the data and their associated services and products for each of the main activities. Where relevant we address the issues of metadata standards, archiving, access and use rights, security and dissemination plans.

The data and code-base associated with the Ancient Lives crowd-sourcing project (ancientlives.org) and the Zooniverse.org platform. While the proposing team does not oversee the technical aspects of the Zooniverse, we felt it was appropriate to describe the Zooniverse cyber-infrastructure and how our work depends on it. As a Zooniverse project, the Ancient Lives project uses Amazon Web Services (AWS) to deliver images to public volunteers via web-technologies and to then record their clicks. The driving engine behind this project is Ouroboros – the currently proprietary software infrastructure (api.zooniverse.org) developed by the Zooniverse development team.¹ This software serves images and the javascript-based classification software to users. As they input clicks and character identifications, the system stores their results in a simple SQL database. We record the individual character classifications and session information, including the user id, time of the classification, click position, and the selected character. Access to “real-world” user information is strictly controlled and Institutional Review Board procedures are followed if that data is requested by any team. When the user load increases, the Amazon Cloud automatically adds new virtual machines to scale with the load. Use of this system also allows us to take advantage of automated hourly-backups of the databases which are stored for one month together

¹ A recent blog post with information about the Zooniverse code release policy is here:
<http://blog.zooniverse.org/2013/02/18/making-the-zooniverse-open-source/>

with daily database exports that are persisted to their object store ‘S3’ for three calendar years. The reliability of the Amazon Cloud is extremely high, however if there was a failure, the data for this phase of the project is eventually reproducible.

The data and code-base associated with the processing and analysis of the crowd-sourced data including curation. The data we collect will represent tens of thousands of fragments with a total data volume of less than ten Gigabytes. The image data for the project will be in the range of fifty Gigabytes. These data are also stored on AWS. We have developed extensive tools that we use to process and visualize the data in the existing Greek texts (see narrative section 3f and Appendices C & D). The code-base for these tools will be made publicly available on Github through an Apache 2.0 license. The Coptic dictionary compiled for the development of the Coptic pipeline and searchable database will be backed up locally and also stored in the Amazon Cloud. This project adheres to the guidelines set forth by the Text Encoding Initiative (TEI). As an ancient language intensive project, Unicode and all XML and XSLT methods comply with TEI standards, so that all coding and markup can be migrated and iterated upon in the future. All data files will follow standard naming conventions based on the fragment identification from their collections and data constructions in the file. For example, each fragment in Ancient Lives has a “subject” identifier (eg. AAL139406) that forms the basis of the filename; each step of the pipeline process tacks the step name to the subject identifier (eg. the output of the consensus step is: fragment_AAL139406_consensus_8.txt). The Datomic database engine supporting the webapp automatically captures all transactions taken by its users; the architecture is designed to separate storage from the application enabling a variety of storage location options such as local storage services at UMN, Oxford or AWS.

Data collected from external resources relevant to the scholarly research proposed in the project.

One of the main sources of data from the ancient texts used in this project will be taken from the Oxyrhynchus collection housed at the University of Oxford as well as from other archival sites around the world (detailed in section 3f). Some images and specific published editions of text in this project may be proprietary, in which case specific arrangements are made with the collection owners allowing public access to the images through the interfaces described in the proposal. Much of this data is controlled by the Universities which house the original documents, including the University of Oxford’s Oxyrhynchus collection. However all research results will be open access with no embargo period. There are no specific privacy requirements for the data. The security of our on-line systems are adequate for the protection of the proprietary collections (through AWS). There are no specific sharing requirements from other agencies or universities for this project.

Overall Access and Dissemination

The audience for the Classics research results will be primarily scholars of early Christian history. However, the broader impact and interest in the data will be across a much wider community. Papyri data have been studied for hundreds of years, so the creation of a large number of new transcriptions and identifications will have an impact for decades to come. Initially the edited manuscripts will be available through the Zooniverse Ancient Lives website. We believe the software used for the analysis will also be of use to a broad scholarly community. This software will be released under an open source license and made available on Github for free use. The long-term access to the collected data will be curated by the University of Oxford. These organizations, including University of Oxford, have a proven record at preserving documents for hundreds of years. The results synthesized from this work, including the edited fragment transcripts and statistical information about the fragment layouts, will be published in scholarly journals and made available to the public through the Zooniverse website housed in the Amazon Cloud. The algorithms used and the processing pipeline will be published in appropriate journals. Scholarly presentation of the research results as well as the computational efforts will take place at national and international professional meetings. The first presentations of the results will be within six months after the start of the project.