Interim Report

PW-51672-14

Coptic SCRIPTORIUM: Digitizing a Corpus for Interdisciplinary Research in Ancient Egyptian

Project Directors:
Caroline T. Schroeder, University of the Pacific
Amir Zeldes, Georgetown University

Institution: University of the Pacific

May 31, 2015

This report outlines the activities and accomplishments originally planned for the major phases of work described in the original work plan for the grant. It then narrates our progress on meeting these objectives during during December 1 2014 through May 2015 (the period since our last interim report from November 2014). Additional activities are described at the end of the report. As noted in the previous interim report, co-Director Dr. Amir Zeldes has taken a position at Georgetown University; the grant activities take place primarily at the University of the Pacific in Stockton, California, and Georgetown University in Washington, D.C. Coptic SCRIPTORIUM has met most of the projected milestones for the grant period thus far. Two objectives have been postponed or delayed, either because the process took longer than anticipated or another more urgent task became apparent during our work.

Phase 2: Fall 2014-early Spring 2015:
Some of this work was begun or completed in the fall and was described in our November 2014 interim report. We provide here information on activities since November.

Continued transcription and digitization of expanded corpus (Shenoute, Sayings). Complete formatting of the select biblical text and begin pilot import, tokenizing, and tagging of papyri from Papyri.info.
Editors/Annotators/Digitization Contributors continue to digitize and annotate Biblical texts, Sayings of the Desert Fathers, and Shenoute material. Although we have not completed manual annotation of two full books of the Bible as originally planned, we are on target to complete those by the end of the summer, and for the chapters of Mark and 1 Corinthians we have published with manual annotation, we have also provided the aligned Greek New Testament. (Manually annotated documents are first annotated with our tools but are then subjected to additional layers of human editorial review at each stage of the process.) The Greek text is the digital edition from the Society of Biblical Literature (SBL); we aligned their apparatus, as well, allowing concurrent search, of Coptic, Greek, English translation and the apparatus to the SBL Greek text. We have also used the tools and technologies developed in our NEH ODH grant HD-51907 to publish a purely machine-annotated version of the entire Coptic New Testament online (tokenized, normalized, and tagged for part of speech and language of origin).

Update metadata based on data curation models and other guidance provided by consultant.
As noted in our November report, Bridget Almas of the Perseus Digital Library completed her modeling of data curation standards, and we engaged another contractor to implement her plan (Luke Hollis of Archimedes Digital) in a web application that resolves stable URNs for our data to the most recent versions of visualizations, files, and searchable text plus annotations for our corpora. We restructured our metadata model and updated our entire corpora with this model. The technical implementation as a web application is being funded by our NEH ODH grant HD-51907 and is being deployed at http://data.copticscriptorium.org this summer. This will allow scholars to cite our data while relying on identifiers that remain stable through possible changes in URLs and Web architecture in the future.

Refinement of methods and documentation.
We have added further documentation to our website and have begun a wiki for the project, which has facilitated communication between annotators and increased the consistency of our pipeline between different participants and newcomers to the project.

Phase Three (Winter/Spring-Summer 2015):
Since we are still in the middle of Phase Three, we outline below only those milestones on which we have begun work.

Continued transcription and digitization of expanded corpus (Shenoute, Sayings) and completion of import, tokenizing, and tagging of papyri.
In May, we published test case documentary papyri from papyri.info. Incorporating these into our corpora was successful, although our tools had higher error rates because of the different vocabulary and less regularized grammar and orthography. Two documents were manually corrected in full following automatic pre-processing. Team member Rebecca Krawiec completed annotations several documents for the Shenoute corpus, including digital manuscript transcriptions donated to the project by David Brakke (the Ohio State University). We also published additional *Sayings* annotated by Amir Zeldes' students from Humboldt University as a result of a course on Coptic incorporating DH technology into the curriculum last year.

Consultation with Almas to create models for linked data, such as RDFs (Resource Description Frameworks).
Some of these linked data conversations have been postponed as we have been focused on modeling our corpora according to the data curation standards recommended by Almas for reference, citation, and sustainability. The data modeling and implementation of the web application to provide stable citations (in the form of URNs) and access to our data has taken longer than anticipated, but it is time well spent, since we now have a model for referencing digital Coptic literary texts and visualizations. Some aspects of this goal have been accomplished, though. The inventory of model URNs developed by Almas is available as text, RDF, turtle, and XML files. We also have a JSON manifest as an API for the work in progress of the web application being developed to provide stable citations, etc., as part of the implementation of Almas' data model. This API, available at http://data.copticscriptorium.org/api/manifest, will allow machine readable and interoperable recovery of our inventory by external projects, present and future.

Discussion and planning regarding creating an online interface for researchers to contribute text and/or annotations to the corpus; research the potential of adapting LAUDATIO, SoSOL, or creation of our own.
We have incorporated more participants into the project who are contributing text and annotations through our GitHub repositories. At the *Digital Coptic 2 Symposium and Workshop* (described below), we trained attendees in our annotation process, and two of them have begun annotating *Sayings of the Desert Fathers*. Our experience expanding the

team and training the attendees has led us to begin discussing what features we would want in an online interface: separate interfaces for paleographic annotations (like SoSOL) and further linguistic or analytical annotations (as provided in a system such as PERSEIDs). These discussions are still in the preliminary phases.

Project Directors' meeting
Project Directors Schroeder and Zeldes met and reviewed the web application's progress and data modeling in March 2015, just prior to the *Digital Coptic 2 Symposium and Workshop* (described below). All the annotators/encoders involved in the project also attended a working meeting with the project directors that week, as well.

Additional Activities
In addition to the originally planned goals and milestones, we have several further achievements:

- Schroeder and Zeldes' co-authored article on part of speech tagging Coptic has been accepted for publication in a special issue of the *Journal of Digital Scholarship in the Humanities*
- Zeldes presented the paper "Tagging the Desert Fathers: part of speech analysis in Sahidic Coptic corpora" to the North American Conference on Afroasiatic Linguistics.
- We organized a two-day conference: *Digital Coptic 2 Symposium and Workshop* that brought together scholars from around the world working at the intersection of Digital Humanities and Coptic Studies or related disciplines (such as Syriac). The schedule and program are in the Appendix to this report. The conference was highly successful, leading to conversations about digital standards and linguistic issues, and to cross-pollination among projects with similar interests (such as crowd-sourcing, digitizing fragmentary text, etc.).
- Revised our guidelines for digitally transcribing and tokenizing Coptic (http://copticscriptorium.org/download/tools/SCRIPTORIUMDiplTranscriptionGuidelines.pdf)

Challenges and Future Plans
The human annotation of the biblical text has taken longer than anticipated, in part due to inconsistencies in the original data and in part due to the time required for annotators to learn and become comfortable with our tools and technologies. As a result, we are planning to provide even more detailed documentation and more training sessions for potential contributors. The Workshop day of *Digital Coptic 2* was quite successful; though participants were new to the process and experienced occasional technical "hiccups", they reported that it was a valuable experience that should be repeated. Two attendees have volunteered to digitize and annotate more documents for the project. Linked data planning has also been postponed, because we needed more time to further refine our data models. Due to these factors, we anticipate that we may request an extension to the grant period in order to be able to offer more workshops in person or online, to provide more detailed documentation, and to map out the linked data models.

Appendix: Digital Coptic 2 Workshop and Symposium Schedule

Thursday, March 12: Symposium

9 am-6 pm
Georgetown University
Poulton Hall, Room 230
1421 37th St. NW
Washington, DC 20057

Presentations are each 20 minutes long followed by 10 minutes for questions and discussion.
Preliminary program:

9:00	**Coffee and Arrivals**
9:15	**Welcome** **Amir Zeldes, Georgetown University**
9:30-12:00	**New and Expanding Digital Projects in Coptic Studies** **Chair: Caroline T. Schroeder, the University of the Pacific** LIVE STREAM at Google Hangouts 9:15-10:30 and 10:40-12:00 *Adventures in Crowd-Sourcing Papyri - The Resurrecting Early Christian Lives DH Project* Philip Sellew, the University of Minnesota *Website Galleries of the White Monastery Candle Room Manuscript Fragments: Challenges of Digitization and Classification* Mary K. Farag, Yale University *The Digital Edition of the Coptic-Sahidic Old Testament and its planned Virtual Manuscript Room (VMR)* Frank Feder, Akademie der Wissenschaften zu Göttingen *Digitizing Language Contact: Lexicography and Technological Perspectives at the Database and Dictionary of Greek Loanwords in Coptic (DDGLC)* Frederic Krueger and Katrin John, Universität Leipzig Discussion (30 minutes)

12:00-1:00 **Lunch**

1:15-3:15 **Digital and Computational Research in Coptic Language and Literature**
Chair: Elizabeth Platte, Valparaiso University
LIVE STREAM at Google Hangouts

Coptic SCRIPTORIUM: Current Possibilities and Future Directions
Amir Zeldes, Georgetown University, and Rebecca S. Krawiec, Canisius College

Coptic Scriptorium beyond the Manuscript: Tokenization and Corpus Analysis
Paul Dilley, the University of Iowa

Synthesis, Boundness, and Clitics in Sahidic Coptic
So Miyagawa, Kyoto University

Discussion (30 minutes)

3:15-3:30 **Coffee Break**

3:30-5:30 **Digital Humanities and Eastern Christian Traditions beyond Coptic Studies**
Chair: Christine Luckritz Marquis, Union Presbyterian Seminary
LIVE STREAM at Google Hangouts

Digital Preservation and Oral History of Displaced Syriac Speakers in the Middle East
Robin Darling Young, the Catholic University of America

A New XML Exchange Format for Aligning Translations, Quotations, and Other Versions of Texts
Joel Kalvesmaki, Dumbarton Oaks

Ex uno pro pluribus: Digitization, cataloging, and study of Eastern Christian manuscript collections at the Hill Museum & Manuscript Library

	Adam Carter McCollum, Hill Museum and Manuscript Library
	Discussion (30 minutes)
5:30-6:00	**Concluding Remarks and Wrap-Up**

Friday March 13

9 am-6 pm
Georgetown University
Poulton Hall, Room 230
1421 37th St. NW
Washington, DC 20057

A day-long workshop on Coptic SCRIPTORIUM for the SCRIPTORIUM team, collaborators, contributors, and those interested in becoming collaborators or contributors. We will discuss SCRIPTORIUM technologies, how to contribute and annotate text corpora for the project, future directions, and possible collaborations. A final agenda will be distributed in March. Space is limited. Please indicate on the registration form if you wish to join us, and we will confirm your attendance.