NATIONAL ENDOWMENT FOR THE
Humanities

OFFICE OF DIGITAL HUMANITIES

Narrative Section of a Successful Application

The attached document contains the grant narrative and selected portions of a previously funded grant application. It is not intended to serve as a model, but to give you a sense of how a successful application may be crafted. Every successful application is different, and each applicant is urged to prepare a proposal that reflects its unique project and aspirations. Prospective applicants should consult the Office of Digital Humanities program application guidelines at http://www.neh.gov/grants/odh/digital-humanities-start-grants for instructions. Applicants are also strongly encouraged to consult with the NEH Office of Digital Humanities staff well before a grant deadline.

Note: The attachment only contains the grant narrative and selected portions, not the entire funded application. In addition, certain portions may have been redacted to protect the privacy interests of an individual and/or to protect confidential commercial and financial information and/or to protect copyrighted materials.

Project Title:	Coptic SCRIPTORIUM: A Corpus, Tools, and Methods for Corpus Linguistics and Computational Historical Research in Ancient Egyptian
Institution:	University of the Pacific
Project Directors:	Caroline T. Schroeder
Grant Program:	Digital Humanities Start-Up Grants, Level 2

Coptic SCRIPTORIUM: A Corpus, Tools, and Methods for Corpus Linguistics and Computational Historical Research in Ancient Egyptian

2. List of Participants

Project Director
Schroeder, Caroline T. Associate Professor, Religious and Classical Studies, and Director of the Humanities Center, the University of the Pacific (Stockton, California).

Associate Director and Technical Lead
Zeldes, Amir. Researcher, the Institute for German Language and Linguistics and the Collaborative Research Center on Information Structure, Humboldt University (Berlin, Germany).

Consultant
Almas, Bridget. Lead Software Developer and Architect, Perseus Digital Library, Tufts University (Medford, Massachusetts)

ANNIS Development Team
The ANNIS development team is comprised of graduate researchers at the Institute for German Language and Linguistics. The team varies from year to year and is supervised by Dr. Zeldes. For a current staff list, visit http://www.sfb632.uni-potsdam.de/annis/contact.html.

Advisory Board
Delattre, Alain. Assistant Professor, Department of Languages and Literatures, Université libre de Bruxelles; Papryi.info.

Grossman, Eitan. Assistant Professor, Department of Linguistics and the School of Language Sciences, Hebrew University.

Imhof, Robin. Humanities Librarian and Associate Professor, University Library, the University of the Pacific.

Note: The Coptic SCRIPTORIUM team is applying for funding from a variety of other sources to expand the text base through digitization in 2014-2015; very few digitized sources in Coptic exist aside from scanned PDFs (without optical character recognition) of old, often problematic editions. The funding from these other sources will support 3 Digitization Contributors on staff and additional time for Prof. Schroeder to supervise the digitization process. The NEH Digital Humanities Start-Up funding will not be used to support digitization, though this work will occur at the same time. Current and future digitization contributors on staff and their dates of employment on the project are:

Platte, Elizabeth. Postdoctoral Fellow, Kalamazoo College. (2013-14; 2014 pending funding)

Krawiec, Rebecca S. Associate Professor, Religious Studies and Theology, Canisius College. (2014-15 pending funding)

McDermott, Lauren. Undergraduate Student, University of the Pacific. (2013)

Digitization Contributor to be hired for 2014-15 pending funding.

3. Abstract and statements of innovation and humanities significance

Abstract
Coptic, having evolved from the language of the hieroglyphs of the pharaonic era, represents the last phase of the Egyptian language and is pivotal for a wide range of disciplines, such as linguistics, biblical studies, the history of Christianity, Egyptology, and ancient history. Coptic SCRIPTORIUM provides the first open-source technologies for computational and digital research across the disciplines as applied to Egyptian texts. The project is developing a digitized corpus of Coptic texts available in multiple formats and visualizations (including TEI XML), tools to analyze and process the language (e.g., the first Coptic part-of-speech tagger), a database with search and visualization capabilities, and a collaborative platform for scholars to contribute texts and annotations and to conduct research. The technologies and corpus will function as a collaborative environment for digital research by any scholars working in Coptic.

Statement of Innovation
SCRIPTORIUM provides the first open-source, digital environment for computational research in the Egyptian language. A unique interdisciplinary collaboration, it designs tools and methodologies and applies them for the first time to Coptic literary texts to create a rich open-access corpus using state-of-the-art digital standards. The project will enable new collaborative, cross-disciplinary digital humanities work and help promote digital humanist methods in multiple fields.

Statement of Humanities Significance
Coptic has proven essential for the decipherment and continued study of Ancient Egyptian and is of major interest for Afro-Asiatic and Coptic linguistics in their own right. Coptic manuscripts preserve biblical texts and document ancient and Christian history. They are the cultural heritage of the current Christian minority in Egypt and its diaspora. SCRIPTORIUM increases access to and understanding of material of historical, religious, cultural, and linguistic significance in a collaborative environment.

4. Project Narrative

Enhancing the Humanities through Innovation

The past decade has witnessed research advances into language and literature through the computer-aided and quantitative analysis of large samples of language data, owing in large part to advances in corpus linguistics and computational methods. Collections (known as corpora) of texts, oral histories, newspapers, and other sources enable research into real-world language expression. They also allow for data-mining and analysis on thematic or topical research questions. SCRIPTORIUM (**S**ahidic **C**orpus **R**esearch: **I**nternet **P**la**t**fo**r**m for **I**nterdisciplinary m**u**ltilayer **M**ethods) is developing and providing open-source technologies and methodologies for interdisciplinary research in linguistics, ancient history, religious studies, and literature as applied to a corpus of texts in Coptic, the last phase of the world's longest historically documented language. Whereas languages like Classical Greek and Latin have enjoyed advances made in digital humanities with fully-fledged online research environments accessible to students and scholars, no computational tools for Coptic exist. Nor is a digital research corpus available. The project will design new tools and adapt existing technologies for the analysis of Coptic, including facilities to search and visualize linguistic and historical data mined from the digital corpus.

The Coptic language, and particularly the Sahidic dialect, is significant for linguistics as well as several other academic disciplines. A direct descendent of the language of ancient Egypt, Coptic developed during the Roman period of Egyptian history and consists of Egyptian grammar, vocabulary, and syntax written primarily in the Greek alphabet; some Egyptian letters were retained, and some Greek vocabulary was incorporated into the language.[1] Coptic was essential for the deciphering of Egyptian hieroglyphs in the nineteenth century and continues to inform research into ancient Egyptian languages.[2] Additionally, Coptic is fundamental to biblical studies and the history of Christianity. A substantial number of gospels and other early Christian texts that were not included in the Christian Bible survive in Coptic. The most famous is the "Nag Hammadi" library, second through fourth century Gnostic texts that have transformed our understanding of Christian origins. Important Coptic biblical manuscripts also survive. The earliest Christian monasteries developed in Egypt, and many of their texts were written in Coptic, not Greek. These texts form an important part of the cultural heritage of the contemporary Coptic Orthodox Church (the Christian minority community in Egypt and the diaspora).

Coptic SCRIPTORIUM will have four major components: 1) a digitized corpus of Coptic texts available in multiple formats and visualizations (including EpiDoc TEI XML)[3], 2) digital and computational tools to analyze, process, and visualize the language (e.g., the first Coptic part-of-speech tagger, converters between digital formats and fonts, etc.), 3) a database created from the texts and tools using the ANNIS search and visualization infrastructure (described on p. 5), and 4) a collaborative platform for scholars to contribute texts and annotations as well as conduct research using the corpus, tools, and database. The project will also serve as a model for future digital humanities projects utilizing historical corpora or corpora in languages outside of the Indo-European and Semitic language families. A proof-of-concept (Components 1-3) is at http://coptic.pacific.edu (backup: http://www.carrieschroeder.com/scriptorium). We seek Level II funding to expand into a full-fledged Pilot. Funding will support additional tool development, advancements in technologies in the search and visualization infrastructure, and the planning and initial development of the collaborative aspects of the online platform.

---

[1] Bentley Layton, *A Coptic Grammar*, 3rd Edition, Rev., Porta Linguarum Orientalium Neue Serie 20 (Wiesbaden: Harrassowitz, 2011), 5.

[2] E.g., Barbara Egedi, "Possessive Constructions in Egyptian and Coptic. Distribution, Definiteness, and the Construct State Phenomenon," *Zeitschrift für Ägyptische Sprache und Altertumskunde* 137 (June 2010): 1–12; John B. Callender, *Studies in the Nominal Sentence in Egyptian and Coptic* (Berkeley: U. of California Press, 1984).

[3] XML is the standard markup language for annotating digitized texts. The Text Encoding Initiative has developed guidelines for encoding, which have been widely adopted in the digital humanities (http://www.tei-c.org/). EpiDoc is the widely used TEI subset for ancient inscriptions and manuscripts (http://epidoc.sourceforge.net/).

The project will develop tools, technologies, and methods to produce the first digital repository of richly-annotated Coptic literary texts as well as methodologies to conduct research using this corpus. (The Start-up grant will not fund digitization of the texts. Digitization is ongoing [2012-14], and the team is seeking funds from other sources to expand the text base.) The Pilot corpus will be comprised of a) writings of the Egyptian monk Shenoute of Atripe; b) select additional monastic texts; c) select sayings of the Desert Fathers in Coptic translation with parallel Greek originals; d) select Coptic papyri; e) select New Testament texts in Coptic translation. (a)-(c) are the core historical corpus and focus of the grant period.

The writings of the Egyptian monk Shenoute of Atripe (ca. 347-465 CE) are the heart of the historical text corpus. Shenoute is universally acknowledged as the most important and influential Coptic author. His texts also form an important source for Biblical studies, since his writings contain some of our earliest Coptic biblical citations. Additionally, the scriptorium at his monastery copied and distributed biblical, monastic, and theological texts throughout southern Egypt. Most of the other texts listed above (a-c, e) survive in manuscripts found at Shenoute's monastery in the 18th and 19th centuries. Shenoute's writings are also instrumental for understanding the history of monasticism, since they are our largest collection of historical, non-hagiographical texts documenting life in a monastery from the very earliest period of the monastic movement: the fourth- to fifth-centuries.

Corpora in corpus and computational linguistics research typically contain multiple layers of data: one base layer of the text being analyzed, followed by additional layers of information about the text. The base layer is often created by applying automated or semi-automated tools known as tokenizers to a digitized text; these tokenizers break up a corpus into searchable and annotatable minimal units or "tokens" of text. In Indo-European languages, the tokens are usually orthographic words between spaces or punctuation. Each token can then be tagged in multiple ways (in layers known as "stand-off markup"): for parts of speech, syntax and grammatical function, named entities, geographic categories, or other research categories. This method produces what is termed a "richly-annotated corpus." See Appendix E of a screenshot of the multiple annotation layers visualized in ANNIS.

Coptic, however, poses challenges for the most basic stage of producing a searchable corpus. It is not an Indo-European language. In fact it is an agglutanitive language, which means that many parts of speech combine into one word using a complex system of prefixes, converters, roots, and suffixes. Existing tokenizers and other tools for creating and formatting an online corpus are not immediately applicable for Coptic. Lemmatizing tools identify the dictionary forms of words or morphemes together with a part-of speech tagging scheme are crucial instruments for syntactic analysis, quantitative vocabulary studies, analysis of text reuse (such as allusions to or quotations from the Bible and other ancient writings), and computational research into themes and topics in the corpus. Such technologies have never before been developed for Coptic. Documentation (including methodologies and best practices as well as linguistic and technical issues) will accompany the tools. (See Appendix D for a list and description of tools).

The project will apply these methodologies and tools to digitized Coptic texts in order to produce a Pilot richly-annotated corpus (minimum size 20,000-30,000 tokens, likely larger with the inclusion of biblical texts). Users will be able to search and visualize data for linguistic and historical research using the open-source tool ANNIS.[4] ANNIS was developed for computational and corpus linguistic research at Potsdam University and Humboldt University; Associate Director Amir Zeldes supervises the ANNIS development team. The SCRIPTORIUM team will work with the ANNIS team on customization of the infrastructure for the Coptic language (enabling complex search in Coptic, visualizing editions of manuscripts, visualizing linguistic and syntactic relationships, visualizing the corpus architecture, etc.)

[4] Amir Zeldes et al., "ANNIS: A Search Tool for Multi-Layer Annotated Corpora," *Proceedings of Corpus Linguistics 2009* (2009), http://ucrel.lancs.ac.uk/publications/cl2009/. See also http://www.sfb632.uni-potsdam.de/d1/annis/.

We will provide documentation in online text and video screen captures. The data will be freely available for search and visualization online or for download. The texts will also be available in TEI XML, for online browsing and for download for further annotation by interested scholars.

The SCRIPTORIUM technologies and corpus will enable new digital research into critical questions in history, biblical studies, and linguistics. For example, researchers will be able to investigate whether Shenoute (one of the earliest sources for Coptic biblical text via his biblical quotations) provides evidence for the early standardization of biblical texts. They can also investigate the relationship between his citations and later developments of the scriptural canon. Computational methods will help scholars probe research questions about intertextuality and citations. Researchers can use statistical methods to conduct a philological study of the circumstances in which monastic vocabulary is used, to develop a quantitatively-informed understanding of the usage of terms. As a prolific Coptic author, Shenoute is a prime test case for research into Greek-Egyptian bilingualism and the influence of Greek philosophy and education in Egypt.[5] We will be able to compare the usage of native Egyptian Coptic vocabulary and Greek loanwords, and even investigate whether Shenoute and other authors were bilingual or received some level of classical education.

The first version was published online in May 2013 as a proof-of-concept. Screenshots of the site, a sample TEI XML file, and queries and visualizations in ANNIS are in Appendix E. The Pilot platform to be developed during the grant period will expand to include a collaborative component to enable users to more easily contribute their own digitized text or layers of annotations to the corpus. The team will research creating our own interface and/or the potential for adapting technologies from LAUDATIO (**L**ong-term **A**ccess and **U**sage of **D**eeply **A**nnotated Informa**tio**n project at Humboldt University) or SoSOL (the content editor for Papyri.info).[6]

This project will be of use to students and advanced scholars in linguistics, the history of Christianity, ancient history, and biblical studies. Researchers beginning their studies of Coptic can use the corpus and tools to aid in their understanding of the language. Experienced scholars will be able to search the corpus or add to it. The Project Director and team will continue to develop and publish best practices guidelines for collaborators, and will vet contributions to ensure compatibility and quality, while simultaneously maintaining a commitment to the most open and collaborative of research principles possible.

Environmental Scan

Current corpus linguistics projects provide large corpora and tools for computational research, especially in modern languages; online repositories digitizing ancient texts also exist, but often without linguistics and historical markups to enable advanced computational research. No digital corpus with tools exists for literary texts in Coptic. The Perseus Digital Library provides digitized ancient texts in Greek, Latin, Old English, Arabic and a few other languages.[7] In collaboration with the Alpheios Project, it provides tools for digital research on Greek, Latin, and Arabic,[8] as well as searchable text using ANNIS.[9] It does not currently support Coptic, but looks forward to doing so in collaboration with Coptic SCRIPTORIUM (see

[5] Raffaella Cribiore, *Gymnastics of the Mind: Greek Education in Hellenistic and Roman Egypt* (Princeton: Princeton University Press, 2001); Arietta Papaconstantinou, ed., *The Multilingual Experience in Egypt from the Ptolemies to the Abassids* (Burlington: Ashgate, 2010).

[6] "LAUDATIO," http://www.laudatio-repository.org/; "The Son of Suda On Line," https://github.com/papyri/sosol. Schroeder has tested T-PEN, an NEH-funded manuscript transcription project; it has incompatibilities with Coptic Unicode characters and do not provide the same kind automated annotation as LAUDATIO and SoSOL. Future releases may be more applicable. ("T-PEN: Transcription for Paleographical and Editorial Notation," http://t-pen.org/TPEN/.)

[7] "Perseus Digital Library", http://www.perseus.tufts.edu/hopper/.

[8] "Alpheios Texts | Alpheios", http://alpheios.net/.

[9] "Perseus Latin and Ancient Greek Treebank - Annis Query Tool", http://annis.perseus.tufts.edu/.

letter from Greg Crane). An advanced project that models a multi-format, richly-annotated corpus is the RIDGES German corpus of herbology texts, created by Dr. Zeldes and colleagues as part of a Google Digital Humanities award. RIDGES is freely available in TEI XML next to a variety of formats.[10] Aside from our proof-of-concept, no such corpus exists for Egyptian language corpora.

Complementary but not overlapping Coptic and Egyptian language projects are in early stages of development. We are in close dialogue with them. The Egyptology Department at the University of Leipzig has begun a Database and Dictionary of Greek Loanwords.[11] The database is not yet public, is not in XML, and applies only to vocabulary that originated in Greek. The Corpus dei Manoscritti Copti Letterari (Rome, Hamburg) hosts a subscription site, which contains black-and-white photographs of some manuscripts and transcriptions.[12] Transcriptions are in SGML (an early markup language that has been supplanted by XML standards), Coptic characters are not in Unicode. The texts are not organized as a relational database, and the site is not open-access. Compared to these projects, Coptic SCRIPTORIUM provides tools as well as a searchable, richly annotated XML corpus in Unicode. A Sahidic biblical manuscript database (Digitale Gesamtedition des koptisch-sahidischen Alten Testamentes) is being planned at Georg-August University, Göttingen, Germany, but has not begun. We anticipate eventual collaborations. (See the letter of support from Project Director Heike Behlmer. Dr. Behlmer and her colleague Dr. Juan Garcés participated in the May 2013 Workshop.) The *Thesaurus Linguae Aegyptiae*, a lexical database of Ancient Egyptian, plans to incorporate Coptic. SCRIPTORIUM is more than lexical and is launching before the *TLA*'s Coptic repository (anticipated to begin in 2018). *TLA* director Dr. Frank Feder attended our May 2013 Workshop and expressed interest in incorporating our corpus. Projet Ramsès in Liège, Belgium, will produce an annotated corpus in Ancient Egyptian.[13] Ramsès currently is not open access. It will build on technologies developed by Dr. Zeldes and collaborators.[14] We were invited to the August 2013 Ramsès board meeting but due to scheduling conflicts could not attend.

Papyri.info contains a repository of searchable, digitized papyri (including some Coptic) in Unicode using the EpiDoc subset of TEI XML standards. They are primarily documentary texts (wills, contracts, private letters, etc.), not literary or monastic. Using only the EpiDoc format alone is not optimal for linguistic research, which requires additional linguistic standoff markup. SCRIPTORIUM texts are formatted for both linguistic and historical research in XML standards including, but not limited to, TEI EpiDoc. One of our Advisory Board members, Alain Delattre, is a Coptic editor for them and attended our May 2013 Workshop. His involvement with Coptic SCRIPTORIUM will ensure important collaborations, including possibly eventually importing Papyri.info texts into our corpus at a later stage.

Even print editions of our core historical corpus are often lacking. No critical edition of a majority of Shenoute's writings exists. Texts are scattered across different publications, with folios from a single text or codex in multiple books and journals, typically with unpublished components, as well. A print critical edition is in progress under the direction of Dr. Stephen Emmel, but no timetable for publication has been released. (Dr. Emmel also attended our May workshop.)

History and Duration of the Project

A detailed History and Work Plan is in Appendix A. Coptic SCRIPTORIUM was born at the 2012 NEH Institute for Advanced Topics in Digital Humanities "Working with Text in a Digital Age," hosted by the Perseus Digital Library, Tufts University. Dr. Zeldes was an instructor teaching Statistical Methods for

[10] "RIDGES - Register in Diachronic German Science", http://korpling.german.hu-berlin.de/ridges/.
[11] "DDGLC - Home", http://www.uni-leipzig.de/~ddglc/index.html.
[12] "CMCL - Studies in Coptic Civilization", http://cmcl.aai.uni-hamburg.de/.
[13] "Le Projet Ramses," www.egypto.ulg.ac.be/Ramses.htm.
[14] See the reference to ANNIS in Stéphane Polis and Jean Winand, "The Ramses Project: Methodology and Practices in the Annotation of Late Egyptian Texts," (unpublished) 9, n. 7. Ramsès is now listed on the website for projects employing ANNIS (http://www.sfb632.uni-potsdam.de/annis/cooperations.html).

the Digital Humanities. Prof. Schroeder attended in order to develop her long-standing interests in Coptic and early Christian history using emerging digital technologies. At the Institute, Schroeder and Zeldes met and began exploring the current challenges of producing digital tools and a digital Coptic corpus.

Schroeder began work in Fall 2012 and through Spring 2013 supervised the development of an encoding converter, digitized text, supervised text digitization by two contributors (Drs. Janet Timbie and Rebecca Krawiec), and supervised manual encoding of EpiDoc TEI XML metadata by a student (Alex Dickerson). (Other Shenoute scholars, including Dr. Timbie, have also volunteered digitized texts for the project in the future, as documented in the letters of support for this application.) Dr. Zeldes began work on the project in Winter-Spring 2013, when he created prototypes of a tokenizer and a part-of-speech tagger, both the first of their kind for the Egyptian language. Dr. Zeldes and the ANNIS development team began adapting the ANNIS infrastructure for Coptic. We used the tokenizer, tagger, and manual annotation to produce the multi-layered proof-of-concept corpus for ANNIS. Using ANNIS's export and visualization capabilities and other tools, we produced TEI XML files, HTML files, and visualizations. The team also made presentations at conferences. (Listed in Appendix B) The proof-of-concept with these tools, the multi-format corpus, visualizations, and documentation was released online in May. We also hosted a Workshop on Digital and Computational Research in the Coptic Language at Humboldt University in May 2013. Over 20 scholars from six countries participated. (See Appendix C for the program.) During 2013-14, we will produce an article about the part-of-speech tagger; continue text digitization and documentation; update and debug the proof-of-concept; plan for the 2014-15 grant period.

Coptic SCRIPTORIUM is a multi-year project conducted in the United States and Germany. Schroeder will seek additional funding from internal sources at Pacific, the ACLS Digital Innovation Fellowships, the NEH, and fellowships for release time at research institutions (e.g., the Stanford Humanities Center, the NYU Institute for the Study of the Ancient World, etc.). We are also seeking funding from European sources. Finally, Schroeder has the ability to renew her 2011-12 Alexander von Humboldt Foundation research fellowship in 2015-16. Long-term storage of the corpus will also be provided by LAUDATIO (Humboldt University) and Perseus (see letters from Prof. Anke Lüdeling and Prof. Greg Crane).

Work Plan
A detailed History and Future Work Plan (including Advisory Board meeting agendas) is in Appendix A.

Phase One (May – Fall 2014): Corpus architecture and data curation development; tool development (tokenizer, part-of-speech tagger, lemmatizer); planning for the collaborative online platform. One week working session in California to address linguistic questions, assess tool development and technologies, update documentation, plan for the next phase of the grant period.

Phase Two (Fall 2014-Winter/Spring 2015): Continued evaluation of technologies and methodologies by testing and applying tools developed in Phase One; beginning development on additional tools; Advisory Board meeting; refine data curation and metadata standards; second one week joint working session for Zeldes and Schroeder in Berlin or California.

Phase Three (Winter/Spring-Summer/Early Fall 2015): complete development of tools begun in Phase Two; begin development of collaborative interface; compile and release new version of corpus; write and disseminate documentation; develop and begin investigation of linguistic and historical research questions to be conducted using the corpus; third working session in California or Berlin; draft white paper.

Although this work plan may seem ambitious, the pace of work we accomplished between November 2012 and May 2013 while working with even less time per week dedicated to the project indicates that this is a reasonable timeline.

Post-Start Up: continued development and support of the online platform; expand the corpus; create and

investigate additional research questions, such as detecting and analyzing textual reuse and thematic and lexical research; publicize platform through papers at digital humanities and disciplinary conferences.

Technical Resources Required: all staff will work on existing computers and laptops. Server space is provided by the Institute for German Language and Linguistics and the University of the Pacific. (See existing hosting at http://coptic.pacific.edu). Existing software used will be open-source and/or open-access and/or standard installs on university computers (i.e., Excel).

Staff
Project Director Caroline T. Schroeder, Associate Professor of Religious and Classical Studies and Director of the Humanities Center at the University of the Pacific, will be responsible for: overall project management; formatting corpus through annotation and markup using SCRIPTORIUM tools and methodologies; supervision of consultants; supporting Zeldes in the design and production of tools; design and documentation of methodologies and best practices in historical corpus research; managing and updating the online site and platform; writing presentations and articles about project results. Schroeder's previous related work in Coptic consists of a published monograph and several peer-reviewed articles and conference papers, and a book in progress on children in early Egyptian monasteries. Schroeder's technical expertise includes XML, HTML, and CSS.

Associate Director and technical lead Amir Zeldes, Researcher at the Institute for German Language and Linguistics at Humboldt University in Berlin, will be responsible for: supervising and directing software design and development; formatting corpus through annotation and markup using SCRIPTORIUM tools and methodologies; design and documentation of methodologies and best practices in historical corpus research; writing presentations and articles about project results. Zeldes is an internationally recognized expert in corpus linguistics digital humanities work as well as morphology and syntax. At the Collaborative Research Center on Information Structure, he has been responsible for the ANNIS search and visualization tool since 2007 and conducts research on text repositories in a variety of languages using this and many other tools, including the statistics software R. Other relevant previous technical and linguistics work includes: being on the development team for <tiger2/>, an XML format used in computational linguistics;[15] creating tools and training data for corpus linguistics annotations.

The Project Directors' unique combination of technical skills, knowledge of Coptic, and multidisciplinary expertise form a team singularly positioned to produce this project.

Other collaborators: Bridget Almas, Senior Software Developer at Perseus Digital Library, will consult on data curation. She will assist in developing models for stable identifiers, uniform references, and linked data standards. Pending separate funding, three Digitization Contributors will continue digitizing texts. *ODH funds will not be used for digitization*. The Advisory Board will provide feedback on all aspects of the project; see the description and members in sec. 6 *Biographies*, and Appendix A for meeting agendas.

Final Product and Dissemination
Technologies, documentation, and project updates as well as the richly-annotated corpus are disseminated on the SCRIPTORIUM website under open-source licenses. The site provides links to and documentation for the digital Coptic corpus, the tools used to develop and annotate the texts, the search and visualization tool, and documentation. Results of research questions investigated using search, analysis, and visualization of data from the corpus will be presented at conferences and published in journals and conference proceedings. The project's white paper describing our procedures, technologies and best practices will be distributed on the SCRIPTORIUM site.

[15] "<tiger2/> - An XML serialization of the SynAF syntactic annotation model", http://korpling.german.hu-berlin.de/tiger2/homepage/index.html.

6. Biographies

Project Director Caroline T. Schroeder is an Associate Professor of Religious and Classical Studies and Director of the Humanities Center at the University of the Pacific, where she has taught since 2007. She received her Ph.D. from Duke University in 2002 and A.B. from Brown University in 1993 (*summa cum laude*, with honors, and Phi Beta Kappa). Her book *Monastic Bodies: Discipline and Salvation in Shenoute of Atripe*, was published by the University of Pennsylvania Press in 2007. It examines the ascetic practices and theologies of the fourth- and fifth-century Egyptian monk Shenoute who led a monastery of approximately 4000 women and men. It integrates the traditional methodologies of textual history, historical theology, and social history with literary theory and anthropology in an examination of material culture, published and unpublished Coptic texts and other early Christian writings. Other publications include articles in the *Journal of the American Academy of Religion*, the *Journal of Early Christian Studies*, *Church History*, the *Journal of Ancient Near Eastern Studies*, and various collected volumes. She is completing a monograph in progress entitled *Monks and Their Children: Family and Childhood in Early Egyptian Monasticism*. She is co-editing with Catherine Chin (University of California, Davis) the book in progress *Melania: Early Christianity through the Life of One Family*. Her research has received financing from the National Endowment for the Humanities, the American Academy of Religion, the Alexander von Humboldt Foundation, the Woodrow Wilson National Fellowship Foundation, and other institutions. She has experience with several web and digital technologies. She has used XML for 1) creating an annotated repository of research notes for her current book project using the Eclipse IDE; 2) annotating ancient texts for online editions using oXygen; 3) marking up ancient texts in TEI XML compliant code; 4) creating a prototype of an interactive parallel Coptic text and English translation using the Alpheios Alignment Editor tool. Schroeder also programs in HTML and CSS and has created maps of monastic sites in Egypt visualizing data from literary sources in the open-source QGIS software.

Associate Director and Technical Lead Amir Zeldes is a researcher at the Institute for German Language and Linguistics at Humboldt University in Berlin. He received his doctorate in General Linguistics from Humboldt University in Berlin in 2012 (*summa cum laude*), Magister Artium from Humboldt University and Potsdam University in 2007 in Computational, Historical and German Linguistics (1,0, mit Auszeichnung) and his B.A. from the Hebrew University of Jerusalem in 2005 (*summa cum laude*) in General Linguistics and Cognitive Science. At the Hebrew University, he studied linguistics, Late Egyptian, and Sahidic and Bohairic Coptic with leading Egyptian linguists Prof. Ariel Shisha-Halevy and Prof. Eitan Grossman. Zeldes's book, *Productivity in Argument Selection: From Morphology to Syntax*, was published by De Gruyter press in 2012. His research applies quantitative methodologies in linguistics to theoretical questions and pursues new methods in computational linguistics. As part of his work at Humboldt University he has been involved in multiple digital humanities projects, most recently the compilation of the RIDGES: Register in Diachronic German Science corpus. He has been a primary contributor to several collaborative, computational and corpus linguistics projects, including the ANNIS open-source architecture for multi-layer corpus search and visualization. He is a member of CLARIN working group 5.7 on Interoperability and Standards and has worked on the standardization of XML formats for syntactic annotation in the <tiger2/> project. He possesses computer skills in PHP, PERL, R, Java/JavaScript, VB, Python, HTML, and XSLT, among other technologies. He also has extensive experience with database management, in particular with the PostgreSQL and SQL Server infrastructures under Windows and Linux platforms.

Advisory Board
The Advisory Board will meet two-three times per year in a virtual, online setting. The international, interdisciplinary group of scholars will assess and evaluate the project. They will advise the team on

linguistic issues, editorial issues, project management, data curation, and potential collaborations with other Coptic and digital humanities projects. They will also provide feedback on effectiveness of tools and technologies and the user interface of the project. They will provide feedback on these issues as needed via email, phone calls, online meetings, and personal meetings. The Advisory Board consists of:

Delattre, Alain. Assistant Professor, Department of Languages and Literatures, Université libre de Bruxelles; Papryi.info.

Grossman, Eitan. Assistant Professor, Department of Linguistics and the School of Language Sciences, Hebrew University.

Imhof, Robin. Humanities Librarian and Associate Professor, University Library, the University of the Pacific.

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 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.

Appendix A: Detailed Project History and Future Work Plan

Planning phase (Fall 2012)
Coptic SCRIPTORIUM was born at the 2012 NEH Institute for Advanced Topics in Digital Humanities "Working with Text in a Digital Age," hosted by the Perseus Digital Library at Tufts University. Dr. Zeldes was an instructor teaching Statistical Methods for the Digital Humanities. Prof. Schroeder attended in order to develop her long-standing interests in Coptic and early Christian history using emerging digital technologies. At the Institute, Schroeder and Zeldes met and began exploring the current challenges of producing digital tools and a digital Coptic corpus.

During Fall 2012, Dr. Schroeder supervised the beginning development of a character encoding converter by Eric E. Johnson (of Tibco Software, Palo Alto, Ca.) in collaboration with Dr. Stephen Emmel (University of Münster), one of the most prominent scholars of Coptic internationally. (Dr. Emmel also later attended the May 2013 Digital Coptic Workshop.) Dr. Schroeder also began organizing and acquiring the texts to be digitized for the proof-of-concept planned for release in May 2013.

Development of the Proof-of-Concept (Winter-Spring 2013)
Dr. Schroeder supervised the completion of the character encoding converter. She also acquired manuscript facsimilies not already in her possession necessary for the proof-of-concept (namely, the letter *Abraham Our Father* and select Coptic *Sayings of the Desert Fathers*). Having worked on Shenoute and other Coptic texts since the late 1990s, Schroeder has amassed photographs and microfilms of manuscripts and developed the necessary relationships with other scholars working on Shenoute as well as with museums, libraries, and universities that hold manuscripts. She digitized text and supervised text digitization by two contributors (Drs. Janet Timbie and Rebecca Krawiec). This involved transcribing the text directly from manuscript facsimilies. Simultaneously, she developed the transcription policies and best practices for the digital processing tools also developed for the project. These policies and best practices are documented on the project site as the Diplomatic Transcription Guidelines. Krawiec also provided the project with an English translation of *Abraham Our Father*. Schroeder and Zeldes manually aligned the translation in standoff markup.

During Winter and Spring 2013, Dr. Zeldes with editorial assistance from Dr. Janet Timbie (the Catholic University of America) and supervised by Schroeder produced a small corpus of select passages from the *Sayings of the Desert Fathers*. Parallel to the text digitization, Zeldes developed two open-source tools for processing Coptic text to enable sophisticated computational search. First is a tokenizer that processes digitized Coptic text and converts it intoa multi-layered standoff-markup file in which the original text (with its word breaks and/or line, column and page breaks) is aligned with the text broken into morphemes. This is the first known Egyptian language tokenizer. Second was a part-of-speech tagger (developed from the open-source TreeTagger tool). This is the first known Egyptian part-of-speech tagger. These tools were made possible thanks to the donation of a digital lexicon by Dr. Tito Orlandi (Sapienza University, Rome, and the University of Hamburg).

Dr. Zeldes and the ANNIS team in close consultation with Dr. Schroeder began adapting the ANNIS search and visualization infrastructure for Coptic and for interdisciplinary research. They embedded the Unicode-compatible font Antinoou and a Coptic keyboard into ANNIS. They also developed visualizations of the manuscript pages, texts, and translations within the infrastructure and also for export as HTML files, browsable on the web.

In Spring 2013, Dr. Schroeder also supervised an undergraduate student (Alex Dickerson, the School of Engineering, the University of the Pacific) who developed scripts to help process the language and encoded the TEI XML metadata in document headers compliant with the EpiDoc subset of TEI standards.

In April and May, the team produced and then published the proof-of-concept with a multiformat corpus, tools, and a database using the ANNIS search and visualization infrastructure. The digitized text with multilayer stand-off annotations was imported into the ANNIS infrastructure. TEI XML metadata headers (manually encoded) were combined with TEI-XML texts (automatically exported from ANNIS). Multiple visualizations in HTML were also produced through automatic exporting functions from ANNIs. See http://coptic.pacific.edu (backup: http://www.carrieschroeder.com/scriptorium). A working session in May in Berlin where Schroeder and Zeldes facilitated this work.

In May, the project also organized and hosted a Workshop on Digital and Computational Research in the Coptic Language at Humboldt University. Over 20 scholars and librarians from six countries participated. All provided their own travel expenses in order to come, indicating the importance of this event for Coptic scholars. (See Appendix ## for the agenda and outcomes of the Workshop. A link to the agenda and presentations is also online on the Coptic Scriptorium site.) One of the most significant outcomes was a discussion of standards for universal references for Coptic authors, texts, and digital artifacts. Dr. Schroeder chaired the workshop and facilitated the discussion.

Throughout 2012-13, the team has made presentations at conferences and workshops. (Appendix ##)

Corpus Expansion and Planning for the Grant Period (2013-14)
Primary activities during the 2013-14 academic year are a conference paper and article about the part-of-speech tagger; continued text digitization and documentation; updating and debugging the current proof-of-concept; planning for new tools and technologies to be developed during the 2014-15 grant period.

Schroeder and Zeldes will continue with text digitization efforts and documentation. Schroeder is supervising two contract staff members (Dr. Elizabeth Platte and Ms. Lauren McDermott) who are digitizing additional texts for the corpus. They have an additional working session in Berlin in August 2013 (funded by the University of the Pacific and Humboldt University).

An Advisory Board Meeting is planned for Spring 2014.

Work Plan During the Grant Period
Phase One (May – Fall 2014):

- Preliminary consultation will occur with Bridget Almas about corpus architecture, uniform references, and data curation.
- Tool development of existing tools: tokenizer, part-of-speech tagger
- Dvelopment of new tool(s): lemmatizer.
- Based on digitization of texts that occurred prior to the grant period, Schroeder will review and make note of needs and requirements for an online platform for users to add text and/or annotations (Component 4 of the project).
- A one week working session in California is planned to address linguistic issues that have arisen; update documentation (or plan for documentation updates); plan for the next phase of the grant period; review the progress of tools and technology development and their effectiveness in processing the digitized text; continue discussions on data curation and corpus architecture; plan for next phase of grant period.

Phase Two (Fall 2014-Winter/Spring 2015):

- Internal evaluation of technologies and methodologies by testing and applying tools to digitized.
- A test expanded corpus will be produced and imported into the ANNIS search and visualization structure.

- Update metadata, linked data and corpus architecture based on data curation models developed in consultation with Almas.
- Meeting of Advisory Board in Fall 2014: board members will assess quality of corpus; provide feedback on decisions regarding corpus architecture, data curation, and linguistic questions; and give guidance on improving processes, standards, and documentation.
- Refinement of existing tools and methods based on internal evaluation and Board evaluations.
- Continued development of lemmatizer and begin development of other tools (e.g., named-entity tagger, collation navigation and visualization tool).
- Discussion and planning regarding creating an online interface for researchers to contribute text and/or annotations to the corpus.
- Work will be conducted at individual institutions and in a second one week joint working session for Zeldes and Schroeder in Berlin or California.

Phase Three (Winter/Spring-Summer/Early Fall 2015):

- Complete development of tools begun in Phase Two.
- Implement development of collaborative interface.
- Edit documentation (drafts composed in process in earlier phases) for the tools.
- Advisory Board meeting to provide feedback on current state of tools, multi-format corpus, and platform.
- Develop and begin investigation of linguistic and historical research questions to be conducted using the corpus—vocabulary studies, morphological studies, and syntax studies are the obvious first investigatory steps.
- Publicly release expanded multi-platform corpus, tools, and first version of SCRIPTORIUM platform at end of period.
- Draft white paper.
- Work will be conducted at individual institutions and in third working session in California or Berlin

Post-Start Up:

- continued development and support of the online platform
- expand the corpus to include more monastic texts, select biblical texts (including texts in collaboration with the Digitale Gesamtedition des koptisch-sahidischen Alten Testamentes), and papyri
- Create and investigate additional research questions, such as detecting and analyzing textual reuse and thematic and lexical research.
- Publicize and expand the user base of the platform through conference papers and workshops at digital humanities and disciplinary society conferences, such as the international annual Digital Humanities conference, the Society of Biblical Literature annual meeting, the North American Patristics Society annual meeting, the International Association of Coptic Studies Congress, the Corpus Linguistics conference, the Language Resources and Evaluation Conference (LREC), and the conferences of the Association for Computational Linguistics.

Appendix B: List of Project Presentations and Important Links

Site: http://coptic.pacific.edu

Backup site: http://www.carrieschroeder.com/scriptorium

Video description of the project on YouTube:
http://www.youtube.com/watch?v=rt_Wr5CvHFU&feature=youtu.be

“Searching for Scripture: Digital Tools for Detecting and Studying the Re-use of Biblical Texts in Coptic Literature,” Forthcoming: Society for Biblical Literature Annual Meeting 2013, Rhetoric and New Testament Section, Baltimore, Maryland, November 2013

“Raiders of the Lost Corpus: Digitizing Fragmented Manuscripts for Linguistic and Historical Research,” Digital Humanities Summer Institute Colloquium, Victoria, Canada, June 7-10, 2013

"Raiders of the Lost Corpus," Institute for German Language and Linguistics, LAUDATIO Workshop on Historical Corpora, Humboldt University, Berlin, May 8, 2013

“Cracking the Code: A Coptic Digital Corpus for Interdisciplinary Research,” Word, Space, Time: Digital Perspectives on the Classical World (Digital Classics Association Conference), Buffalo, April 5-6 2013; http://www.youtube.com/watch?v=m9iak8ot20k

"Coptic Studies on the Digital Frontier: Creative Approaches to Manuscript Publication," Society of Biblical Literature Annual Meeting, November 2012

Appendix C: Program from Workshop on Digital and Computational Scholarship in the Coptic Language

Hosted by the University of the Pacific (Stockton, California) and Humboldt University (Berlin)
Location: Institute for German Language and Linguistics at Humboldt University
Dorotheenstrasse 24, 3.246
May 14, 2013
A link to this program at http://coptic.pacific.edu (backup http://www.carrieschroeder.com/scriptorium).

9:00-9:15 Arrivals, coffee

9:15 Introductions

9:30-10 Vincent Walter (Leipzig): Database and Dictionary of Greek Loanwords in Coptic (DDGLC)
http://www.uni-leipzig.de/~ddglc/

10-10:30 Frank Feder (Berlin): The Thesaurus Linguae Aegyptiae and Its Planned Coptic Components
http://aaew.bbaw.de/tla/

10:30-11 coffee break

11:00 Tito Orlandi/Alin Suciu (Hamburg): CMCL
http://cmcl.let.uniroma1.it

11:30-12 Carrie Schroeder (Pacific) /Amir Zeldes (Humboldt University): Coptic Scriptorium
http://go.pacific.edu/copticscriptorium, Presentation Slides

12-12:30 discussion

12:30-2 pm lunch hosted by the University of the Pacific

2:00 Heike Behlmer/Juan Garces (Göttingen): Digitale Gesamtedition des koptisch-sahidischen Alten Testamentes
Presentation Slides

2:30 Alain Delattre (Université Libre de Bruxelles) EpiDoc/Duke Database of Papyri
http://sourceforge.net/p/epidoc/wiki/Home/, Presentation Slides

3-5 pm Discussion of technologies and research questions (Schroeder will moderate; coffee available)

Also attending:
Stephen Emmel (Münster)
Slavomír Čéplö (Charles University, Prague)
Eliese-Sophia Lincke (Humboldt University)
Matthias Müller (Basel)
Frank Kammerzell (Humboldt University)
Sandra Hodecek (ONB, Vienna)
Doris Topman and Simon Schweitzer (TLA)
Tonio Sebastian Richter, Frederic Krueger,
Katrin John (DDGLC at Leipzig)
Daniel Werning (Humboldt University, Institute for Archaeology)

A report on the discussion is at http://earlymonasticism.org/2013/06/02/workshopmay2013/ .

Appendix D: Coptic SCRIPTORIUM Tool List

Tools Developed in 2013 to Be Refined and Advanced During the Grant Period
Tokenizer: The first known tokenizer for the Egyptian language; version 1.0 developed Spring 2013. Currently applicable only for text files encoded in specific formats, which involve a significant amount of manual annotation. (For more details, please see the online documentation entitled "Diplomatic Transcription Guidelines.") We will work on developing new versions of the tokenizer so that we can digitally process text in a more "raw" form as well as text in different formats. This will enable our project to accept text contributions from other scholars without having to provide a lot of manual encoding to make them ready for inclusion in the corpus, and it will be a tool useful for other linguists and Coptologists to use in their own projects.

Part-of-speech tagger: first known part-of-speech tagger for the Egyptian language family; developed in spring 2013. Needs to be tested and developed on a broader set of texts. Its technologies also need more development to make it a "smarter" tool (e.g., taking into account more of the context surrounding the word or morpheme being tagged to determine the accurate part of speech).

Character encoding converter: transforms Coptic digitized texts encoded in an older ASCII Coptic font into fully-compliant Unicode characters; prototypes developed winter-spring 2013. More testing and debugging required.

New Tools Planned for Development During the Grant Period
Lemmatizer: will automatically annotate words or morphemes with their lemmas (the dictionary entry of the word).

Named-entity tagger: to automate the annotation of named entities, such as people and places.

Collation navigation and visualization tool: a tool to visualize and navigate the collation of multiple manuscript witnesses of an individual text in the corpus. Many Coptic texts are fragmented, often requiring the reader to piece together fragments from multiple ancient copies of a text, which are now housed in different modern repositories or print editions in order to read one text. This tool will visualize the fragments and aid the reader in navigating the text online. Collation and annotation of fragmented texts has been a recent topic on the TEI-XML email list; we anticipate this tool to be a useful model for other projects with similar collation issues.

Appendix E: Screenshots of the Project Site and Multi-Format Corpus

Note: full materials available at http://coptic.pacific.edu and backup site http://www.carrieschroeder.com/scriptorium.

Coptic SCRIPtorIuM (Sahidic Corpus Research: Internet Platform for Interdisciplinary multilayer Methods) is a collaborative, digital project created by Caroline T. Schroeder (University of the Pacific) and Amir Zeldes (Humboldt University).

Coptic SCRIPtorIuM provides a platform for interdisciplinary and computational research in texts in the Coptic language, particularly the Sahidic dialect. As an open-source, open-access initiative, the SCRIPtorIuM technologies and corpus function as a collaborative environment for digital research by any scholars working in Coptic. It provides:

- tools to process Coptic texts
- a searchable, richly-annotated corpus of texts using the ANNIS search and visualization architecture
- visualizations of Coptic texts
- a collaborative platform for scholars to use and contribute to the project
- research results generated from the tools and corpus

We hope SCRIPtorIuM will serve as a model for future digital humanities projects utilizing historical corpora or corpora in languages outside of the Indo-European and Semitic language families.

Please read our Frequently Asked Questions for more information on the project, methodologies, and terminology.

We hosted a workshop on digital research and scholarship in Coptic at Humboldt University on May 14, 2013. The program and presentations are available.

A video introduction to the project, including how to use ANNIS, is available.

Corpora

The corpora below offer some examples of mark-up for diplomatic transcription and normalization. Most data is available in TEI XML, PAULA XML and

relANNIS for use with the ANNIS corpus search software. Links are provided to search the corpus online in ANNIS. Individual documents can also be viewed in HTML for reading purposes in either diplomatic or normalized transcriptions with English translations. [For more information on TEI, PAULA, and ANNIS, check out our FAQ.]

Abraham our Father

- About the corpus
- Search the corpus in ANNIS (click here if the Coptic characters do not appear properly)
- Download entire corpus in the following formats: 
- Individual documents:
  - Abraham.YA518-20: download TEI XML, HTML: diplomatic, normalized
  - Abraham.YA525-30: *download TEI XML*, HTML: diplomatic, normalized
  - Abraham.YA535-40: *download TEI XML*, HTML: diplomatic, normalized
  - Abraham.YA547-50: *download TEI XML*, HTML: diplomatic, normalized
  - Abraham.YA551-54: *download TEI XML*, HTML: diplomatic, normalized
  - Abraham.ZHfrgmts1a_d: *download TEI XML*, HTML: diplomatic, normalized
  - Abraham.XL93-94: *download TEI XML*, HTML: diplomatic, normalized

Apophthegmata Patrum

- Search the corpus in ANNIS
- Download entire corpus in the following formats: 
- Individual documents:
  - AP.42.sarah: download TEI XML, HTML: diplomatic, normalized

- AP.90.olympius: download TEI XML, HTML: diplomatic, normalized
- AP.157.papnoute: download TEI XML, HTML: diplomatic, normalized
- AP.172.antonius: download TEI XML, HTML: diplomatic, normalized
- AP.216.besarion: download TEI XML, HTML: diplomatic, normalized

Tools

Some of the tools below use a Sahidic Coptic lexicon based on data kindly provided by Prof. Tito Orlandi and the CMCL project. When using the part-of-speech tagging models or the tokenization script and its lexicon please make sure to refer back to the CMCL project.

Part-of-Speech Tagging

- Scripts and models
  - Tokenization script and lexicon (assumes normalized Coptic, see tokenization guidelines and tagging guidelines)
  - TreeTagger - an open source part-of-speech tagger (additional Windows interface WinTreeTagger)
  - Coptic TreeTagger training models - for the fine and coarse grained tagsets (see documentation)
- Documentation
  - Diplomatic Transcription Guidelines
  - Tokenization Guidelines
  - Part-of-Speech Tagging Guidelines

Converters

- Coptic encoding converter (converts older text character systems used for fonts such as Coptic and Laser Coptic into standards-compliant Coptic Unicode characters)
  - Simple recoding script in Perl (supports CMCL, Laser Coptic and UTF-8 encoding conversion)
  - Converter for ASCII encoding / UTF-8 of Dirk Van Damme and Gregor Wurst
- SaltNPepper - a metamodel based Java framework for multi-format conversion

- Excel-Plugin for importing and exporting EXMARaLDA XML, SGML, PAULA XML and subsets of TEI XML

Acknowledgments

- Thank you
- Bibliography

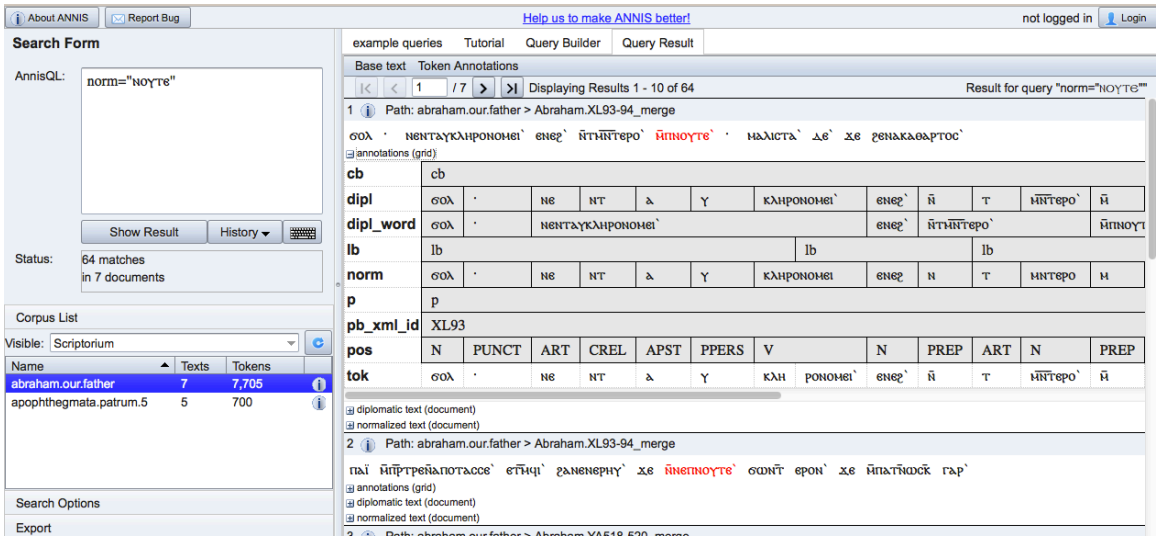
Page last updated 24 June 2013

Screenshots of excerpts of sample TEI-XML file from corpus:

```
− <msIdentifier>
    <msName>MONB.YA 518-520</msName>
  − <altIdentifier>
      <repository>Biblioteca Nazionale di Napoli</repository>
      <idno>Borgia Collection IB2 ff. 26-27</idno>
    </altIdentifier>
  </msIdentifier>
− <history>
  − <origin>
      <objectType>Codex</objectType>
      <country>Egypt</country>
      <placeName>Atripe</placeName>
      <origPlace>The White Monastery</origPlace>
      <origDate notBefore-custom="0900" notAfter-custom="1200" precision="medium">Between 900 and 1200 C.E.</origDate>
    </origin>
  </history>
</msDesc>
</sourceDesc>
</fileDesc>
− <encodingDesc>
  − <p>
      This data encoded to comply with EpiDoc Guidelines and Schema.
      <ref>http://www.stoa.org/epidoc/gl/5</ref>
    </p>
</encodingDesc>
− <profileDesc>
  − <langUsage>
      <language ident="cop">Coptic</language>
```

```
</teiHeader>
- <text>
  - <body>
    - <div>
      - <p>
          ⲥⲟⲛ[.] ⲛ ⲯⲟⲥⲟ̄
          <lb/>
          ⲉ ⲛ ⲯ[ⲟ]ⲟⲛ : ⸗
          <lb/>
          ⲥⲉⲛⲟⲩ ⲑⲓⲟⲩ ⲉⲡⲓⲥⲧⲟⲗⲏ :
          <lb/>
```

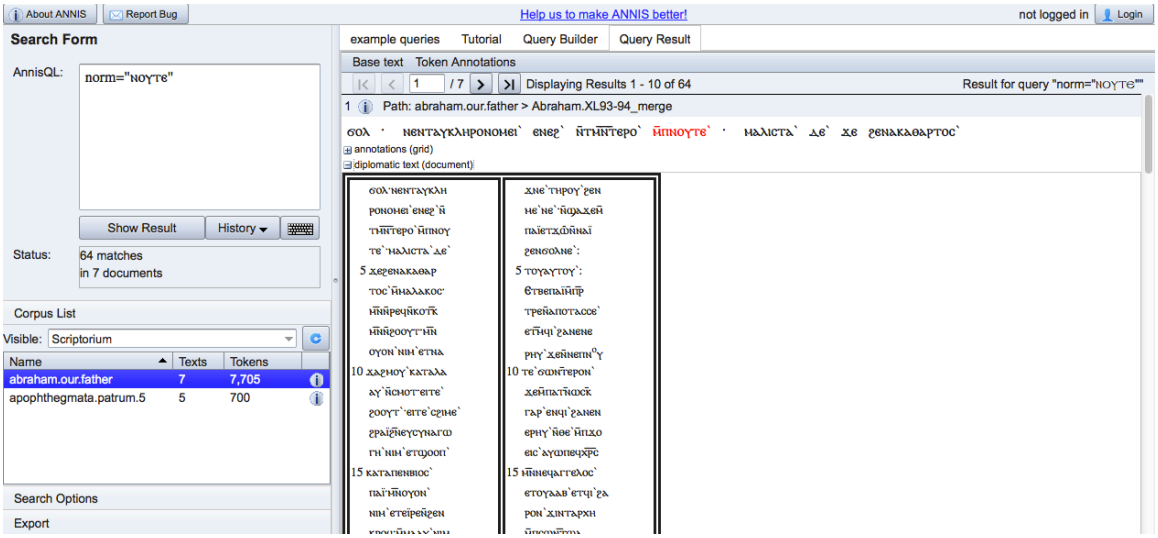
Screenshot of the multi-layered proof-of-concept corpus in ANNIS search and visualization tool:



Annotation layers visible include:

- cb: Column breaks in the original manuscript
- dipl: Diplomatic edition of the text (conforming exactly to spelling and punctuation in the original manuscript); text segmented into morphemes rather than words
- dipl_word: Diplomatic edition of the text (conforming exactly to spelling and punctuation in the original manuscript); text segmented into words
- lb: line breaks in the original manuscript
- norm: Normalized text: spelling and punctuation all normalized; text segmented into morphemes; optimal layer for most searches involving vocabulary or word forms
- p: paragraph segmentation aligned with the English translation (English translation not visible here)
- pb_xml_id: Manuscript ID and page number (i.e., XL93 is White Monastery codex XL, page 93)
- pos: Part of speech; automatically annotated using the part-of-speech tagging tool
- tok: base token layer, which reflects segments of morphemes as well as text segmentation cutting across column, line, and page breaks. Tokens are the smallest pieces into which the text is broken; since Coptic manuscripts have words and morphemes that break across lines, the token layer is NOT the same as the morpheme or word layer.

Screenshot of diplomatic edition of the manuscript visualization generated in and by ANNIS from the XML:


About ANNIS
Report Bug
Help us to make ANNIS better!
not logged in
Login
Search Form
AnnisQL:
norm="ΝΟΥΤΕ"
Show Result
History
Status:
64 matches
in 7 documents
Corpus List
Visible:
Scriptorium
Name
Texts
Tokens
abraham.our.father
7
7,705
apophthegmata.patrum.5
5
700
Search Options
Export
example queries
Tutorial
Query Builder
Query Result
Base text
Token Annotations
Displaying Results 1 - 10 of 64
Result for query "norm="ΝΟΥΤΕ""
Path: abraham.our.father > Abraham.XL93-94_merge
annotations (grid)
diplomatic text (document)