KELLIA Final White Paper

National Endowment for the Humanities Grant HG-229371

Deutsche Forschungsgemeinschaft (BE 4172/1-1)

Koptische/Coptic Electronic Language and Literature International Alliance (KELLIA)

Project Directors:

Caroline T. Schroeder, University of the Pacific (American PI)

Amir Zeldes, Georgetown University (co-PI)

Heike Behlmer, Georg-August University, Göttingen; Göttingen Academy of Sciences and Humanities (German PI)

Institutional Grantees:

University of the Pacific (NEH)

Georgetown University (NEH)

Georg-August University (DFG)

Report Authors: Caroline T. Schroeder, Heike Behlmer, Elizabeth Platte, Ulrich Schmid, Amir Zeldes, So Miyagawa, Frank Feder

27 April 2019

Project Activities

The KELLIA project is a collaboration between multiple institutions. At the time of the grant application, the primary institutions were the Seminar for Egyptology and Coptology at Georg-August University (Göttingen, Germany), the Department of Religious Studies at the University of the Pacific (Stockton, California, USA), and the Department of Linguistics at Georgetown University (Washington, DC, USA) with additional partners at the University of Münster and the Berlin-Brandenburg Academy of Sciences and Humanities. Prior to the funding of the project proposal, the Digital Edition of the Coptic Old Testament Project was established at the Göttingen Academy of Sciences and Humanities, and they became a primary partner. During the course of the grant, activities involved the following projects:

- Coptic SCRIPTORIUM (University of the Pacific, Stockton CA; Georgetown University, Washington DC)
- Digital Edition of the Coptic Old Testament (Göttingen Academy of Sciences and Humanities)
- Seminar for Egyptology and Coptic Studies (Georg August University Göttingen)
- Institute for New Testament Textual Research (INTF) (University of Münster)
- *Thesaurus Linguae Aegyptiae* (TLA) (Berlin-Brandenburg Academy of Sciences and Humanities (BBAW), Berlin):
- Database and Dictionary of Greek Loan Words in Coptic (DDGLC) (Leipzig University, Freie Universität Berlin, BBAW)
- Electronic Text Reuse Acquisition Project (eTRAP) (Georg-August University, Göttingen)
- SFB 1136 “Bildung und Religion in Kulturen des Mittelmeerraums und seiner Umwelt von der Antike bis zum Mittelalter und zum Klassischen Islam” (Georg-August University, Göttingen)
- PAThs An Archaeological Atlas of Coptic Literature (Sapienza University of Rome)

The KELLIA project hosted multiple project meetings and workshops to facilitate collaborations that resulted in the Accomplishments and Grant Products described below. Project meetings and workshops were held in Goettingen (September 2015), Claremont (July 2016), Goettingen (July 2017), Washington DC (June 2018), Goettingen (June 2018), and Washington DC (September 2018). In addition, collaborators met informally at external conferences they mutually attended. Between physical meetings, individuals worked at participants’ home institutions in the US and Germany and communicated electronically. Agendas are in Appendix 1. The following individuals participated in KELLIA Activities:

KELLIA Participants (Individuals)

Heike Behlmer, Georg-August University, Göttingen, and Göttingen Academy of Sciences and Humanities
Caroline T. Schroeder, University of the Pacific
Amir Zeldes, Georgetown University
Frank Feder, Georg-August University, Göttingen, and Göttingen Academy of Sciences and Humanities
Tonio Sebastian Richter, Freie Universität, Berlin; Berlin-Brandenburg Academy of Sciences and Humanities
Siegfried Richter, University of Münster
Ulrich Schmid, Georg-August University, Göttingen, and Göttingen Academy of Sciences and Humanities
Elizabeth Platte, Reed College
Maxim Kupreyev, Berlin-Brandenburg Academy of Sciences and Humanities
Troy Griffitts, Georg-August University, Göttingen, and Göttingen Academy of Sciences and Humanities
Christine Luckritz Marquis, Union Presbyterian Seminary
Rebecca Krawiec, Canisius College
So Miyagawa, Georg-August University, Göttingen
Malte Rosenau, Georg-August University, Göttingen, and Göttingen Academy of Sciences and Humanities
Uwe Sikora, Georg-August University, Göttingen
Julien Delhez, Georg-August University, Göttingen
Emma Manning, Georgetown University
Shuo Zhang, Georgetown University

Additional Contributors and Participants

Mitchell Abrams, Georgetown University
Diliana Atanassova, Göttingen Academy of Sciences and Humanities
Julian Bogdani, Sapienza University of Rome
Dave Bricetti, independent programmer and software consultant
Dylan M. Burns, Freie Universität Berlin
Paola Buzi, Sapienza University of Rome
Elizabeth Davidson, The Westminster Schools
Paul Dilley, University of Iowa
Luke Gessler, Georgetown University
Katrin John
Theresa Kohl, Georg-August University, Göttingen

Emma Manning, Georgetown University
Lance Martin, Catholic University of America
Simon Schweitzer, Berlin-Brandenburg Academy of Sciences and Humanities
Melissa Harl Sellew, University of Minnesota
Laura Slaughter, University of Oslo
Agostino Soldati, Sapienza University of Rome
Alin Suciu, Georg-August University, Göttingen, and Göttingen Academy of Sciences and Humanities
Nathan Thomas, independent programmer
Tiffany Ziegler, Georg-August University, Göttingen

Coptic Scriptorium Advisory Board

Alain Delattre, Assistant Professor, Department of Languages and Literatures, Université libre de Bruxelles; Papyri.info.
Eitan Grossman, Assistant Professor, Department of Linguistics and the School of Language Sciences, Hebrew University.
Marco Büchler, Head of the eTRAP Research Group, Institute of Computer Science, Georg-August-Universität Göttingen.
Robin Imhoff, Humanities Librarian, University of the Pacific.
Maged S. A. Mikhail, Professor of History, California State University, Fullerton, and Managing Editor of Coptica, the Journal for Coptic Studies.
Ellen Muehlberger, Associate Professor of Near Eastern Studies and History, University of Michigan.
Joshua Salyers, University of the Pacific Digital Collections Librarian.
Laura Slaughter, Associate Professor, Centre for Scalable Data Access, Department of Informatics, University of Oslo.
Janet A. Timbie, Adjunct Associate Professor, Department of Semitic and Egyptian Languages and Literatures, Catholic University of America.

Digital Edition of the Coptic Old Testament Advisory Board

Anne Boud'hors, Director of Research, Institut de recherche et d'histoire des textes
Jan Dochhorn, Associate Professor of New Testament, Durham University
Reinhard Gregor Kratz, University Professor for Old Testament, Georg August University, Göttingen
Gerhard Lauer, Chair for Digital Humanities, University of Basel
Tonio Sebastian Richter, University Professor in the Egyptology Seminar, Freie Universität Berlin

Project activities were publicized via the Coptic Scriptorium blog, the Coptic Old Testament blog, the Digital Coptic email list, conference presentations, institutional publicity offices, and social media. Grant Products are disseminated on the main KELLIA project website (http://kellia.uni-goettingen.de/) in addition to the Coptic Scriptorium and Coptic Old Testament project websites (http://copticscriptorium.org, http://coptot.manuscriptroom.com/).

Accomplishments

In addition to the tools, publications, and standards outlined in the Grant Products section below, KELLIA partners successfully:

- built partnerships with other digital projects in the ancient world (Pelagios and Pleiades for linked geographic data, PATHs for text-bearing objects from Coptic Egypt, eTRAP for text-reuse, among others);
- secured additional funding for a Coptic OCR (optical character recognition) project in the form of a 12,000 Euro fellowship ("Digitalisierung und computergestützte Analyse“) from the University of Göttingen for Dr. Eliese-Sophia Lincke to train and use OCRopy for Coptic (supervised by PIs Heike Behlmer, Marco Büchler, Frank Feder, and Camilla Di Biase-Dyson);
- Produced publications (completed and in progress) and conference papers presented (see Appendix 2)

All proposed grant outcomes were completed

- Outcome 1: Milestones for data standards
- Outcomes 2 and 4: Converter and Annotation tools
- Outcomes 3 and 5: Integration of Corpus Linguistics Methods with Digital Editions Methods

Audiences

When the grant began, the Virtual Manuscript Room for Coptic did not exist; now 83 VMR accounts exist for the Coptic Old Testament project. Improvements made to the VMR during the KELLIA project also applied to the Greek New Testament project, which has over 1500 users. The users include researchers, students, clergy, and lay/non-academic users with training in Greek. The Coptic Old Testament project is now working to expand collaborations with working on Coptic and Arabic bilingual

manuscripts (Biblia Arabica Project, Prof. Ronny Vollandt)[1] and people working in the Bohairic dialect of Coptic.

Since the Coptic Dictionary Online (http://coptic-dictionary.org) went online two years ago, it has become the most widely used electronic dictionary for the language. According to Google Analytics, the site receives 895 distinct visitors per month (6 month average), with each visitor returning for an average of 2.4 sessions monthly, viewing an average of 9.7 pages in a single session, or approximately 3-4 dictionary queries, in an average 10 minutes of usage time. An anonymized browser Search Engine Result Page (SERP) analysis suggests that the dictionary is robustly represented, ranking first for the Google queries “Coptic Dictionary”, “Coptic Lexicon” and “Coptic online dictionary”.

Coptic Scriptorium’s ANNIS search engine, which allows users to perform advanced lexical and linguistic searches in Coptic corpora, receives an average of 1,111 search queries each month (4 month average), in addition to an average 30 requests for complex frequency analyses, enabling quantitative research on Coptic data. The queries cover a wide range of categories, including common words and names, which are searched for often, and a long tail of complex grammatical queries, which are usually unique. On average, the interface receives 641 unique queries of the Coptic corpora each month. About 47% of the queries target a single corpus, while the rest target at least two works, with just over 40% targeting all Coptic corpora, suggesting aggregate studies or users looking for attestations of dictionary headwords from the online dictionary.

Evaluation

The two main partners in the grant application, Coptic SCRIPTORIUM and the Seminar for Egyptology and Coptology, established a plan to have have KELLIA activities evaluated by the advisory boards and agencies of Coptic SCRIPTORIUM and the Coptic Old Testament Project. Both projects also write regular self-reports, annually for the Coptic Old Testament Project, and semi-annually for Coptic SCRIPTORIUM. CS’s reports are posted publicly online.[2]

Coptic SCRIPTORIUM consulted with its advisory board en masse 3 times a year and with individual board members as needed on particular issues. Evaluations were

[1] https://biblia-arabica.com/

[2] http://copticscriptorium.org/reports

generally positive, and recommendations were incorporated into future work. A draft of the White Paper was reviewed by all members of the Advisory Board.

The Coptic Old Testament Project created a pre-evaluation self-report in April, 2017. Evaluators chosen by the Scientific Commission (Wissenschaftliche Kommission) of the Union of German Academies read the report and conducted a site visit; subsequently the Commission created an evaluation that is on file at the Academy. The letter from the Commission to the president of the Göttingen Academy gave the project a positive evaluation and particularly emphasized as a strength the networks with other projects, which include the KELLIA collaboration. The next evaluation will be in 2022.

Continuation of the Project

The Coptic Dictionary Online, ANNIS search engine of Coptic Scriptorium, and Coptic Natural Language Processing API are set to continue offering the same services, with additional developments depending on the project's ability to secure funding for future resources. A follow up grant for the development of new Coptic Scriptorium tools was awarded by the NEH in 2018.

Specific collaborations between KELLIA partners will continue. The Coptic OT project will use the NLP API to annotate their digital editions in the future. We continue to share textual data, as Coptic Scriptorium publishes additional Coptic documents digitized by German partners. Support and ongoing development of the Coptic Dictionary will continue.

Long Term Impact

As a result of the project, our participating universities are known for digital research in Coptic, generating interest from researchers, visiting scholars, and prospective students. Many of the tools and texts are being used in the classroom (the dictionary, the NLP pipeline, the treebanking annotations, digital corpora and annotations) and for research. The materials are used in the Coptic Summer Schools offered by Goettingen and other partners.

The Coptic Old Testament project has the potential for funding until 2036; the tools and collaborations developed here are the foundation for this long term project. The tools and methods are also foundational for multi-language support being developed (e.g., Greek, other Coptic dialects, Arabic, Armenian).

The Coptic Dictionary will continue to be used widely. Coptic NLP tools are being and will continue to be used by Coptic scholars for studying stylometry, linguistics, and text reuse.

Grant Products

Online Coptic Dictionary

The online Coptic dictionary (http://coptic-dictionary.org) was created by members of the KELLIA collaboration using data from KELLIA partner Thesaurus Linguae Aegyptae (TLA) at the Berlin-Brandenburg Academy of Sciences (BBAW). Bidirectional links have been integrated into the digital text in Coptic SCRIPTORIUM's corpora. Entries in the online dictionary link to Coptic SCRIPTORIUM corpora by lemma; a link in each dictionary entry directs the user to results of a query for all instances of the lemma in the corpora published in ANNIS. In the Coptic. Likewise clicking on links of words in the Coptic SCRIPTORIUM corpora published in ANNIS direct the user to a dictionary search for that word's lemma. Normalized digital editions of Coptic SCRIPTORIUM texts also link individual words to linked queries in the online dictionary. The integration of data from the BBAW lexicon and Coptic SCRIPTORIUM's annotated digital editions increases the scholarly utility and discoverability of both projects' data. Any project can use these linking features by linking digital Coptic texts to the dictionary via a query for lemmas.

Coptic Treebank

The Coptic treebank includes Coptic SCRIPTORIUM texts with full syntactic annotations.[3] This treebank forms the basis for training stochastic parsers for automatic analysis of Coptic syntax, which is still in very early steps. The syntactic analyses in the treebank are foundational to further work on automatic entity tagging in the corpora, since syntactic annotation allows us to recognize the word borders covered by entity mentions and to resolve their discourse relations to each other (e.g. apposition, pronominalization and subsequent mention). This automatic syntactic annotation lays the groundwork for linking data about entities within Coptic texts, as outlined in the entity recognition section, below. In 2017, the Coptic Treebank joined the Universal Dependencies project, offering over 100 syntactically annotated corpora in more than

[3] http://copticscriptorium.org/treebank.htm

60 languages, all annotated using uniform guidelines.[4] As of the latest release, the treebank is included in the public UD dataset (Nivre et al. 2017), which is used to train automatic parsers cross-linguistically.

Entity recognition

During the KELLIA grant period, US partner Amir Zeldes has completed preliminary work on Named Entity Recognition (NER) on Coptic SCRIPTORIUM texts. First entity tagging tests have been done on select *Apophthegmata* as well as Shenoute's "Not Because a Fox Barks" as the gold standard for these tools. Using entity lists taken from Coptic SCRIPTORIUM corpora, Zeldes used xrenner to recognize classes of entities.[5] NER annotations are possible due to the dependency annotations provided by the treebank, and we expect future advances for NER in the Coptic SCRIPTORIUM corpora and in Coptic literature in general as the treebank data improves. This work in NER will allow for entity disambiguation and linking opportunities for Coptic literature, which will in turn provide data to scholars working on, for instance, mapping or social networking in Coptic literature.

VMR-Coptic SCRIPTORIUM Converter

The VMR (German partners) generates digital manuscript transcription data. The Coptic SCRIPTORIUM (US partner) produced tools to semi-automatically analyze Coptic text data. The idea was to build a conversion routine to transform VMR-data to the Coptic SCRIPTORIUM EpiDoc-Format and push it to the NLP-Pipeline (natural language processing) in order to integrate the digital text corpora developing in both projects. The final result is a VMR-API call based on some scripts using the JAVA version of the Saxon XSLT-Processor as a subprocess and converting data retrieved from the VMR to the format used by Coptic SCRIPTORIUM.[6] This VMR-Coptic SCRIPTORIUM Converter can be used independently from the VMR-API and is available on GitHub.[7]

[4] http://universaldependencies.org/

[5] https://corpling.uis.georgetown.edu/xrenner/#

[6] http://coptot.manuscriptroom.com/community/vmr/api/projects/kellia/

[7] https://github.com/KELLIA/vmr_converter

GitDox Online Transcription and Annotation Tool

GitDox is a light-weight transcription and annotation tool customizable for individual projects and for multiple languages.[8] Created by researchers at Georgetown University during the KELLIA project, it currently contains a transcription/text editor with customizable encoding validation options and a spreadsheet editor for collaborative editing of a multi-layer annotated document. The tool can be adapted to different languages and is currently also in use for other annotation projects: it has been used for building the English GUM corpus (http://corpling.uis.georgetown.edu/gum/) and for annotating texts from reddit forum discussions as part of a course on Corpus Linguistics at the 2017 Linguistic Institute of the Linguistic Society of America (http://lsa2017.uky.edu/). the Coptic Scriptorium version is linked to our Coptic natural language processing tools. After researchers transcribe a Coptic text with light XML markup for structural information (*i.e.* page breaks, missing text, etc.), they can click a button to run the text through the NLP tool pipeline; this annotated text is presented to the researcher in a multilayer format in spreadsheet mode. Researchers commit the data and subsequent edits to our repositories on GitHub. The tool includes space for document metadata and customizable validation mechanisms. Access to the Coptic Scriptorium instance requires a username and password, managed by Dr. Zeldes at Georgetown University. The tool is open-source (Apache 2.0 license[9]) and available for download and installation.

Digital Corpora: Richly Annotated Coptic Literary Texts

During the grant period, KELLIA members have published several literary texts with linguistic annotations. These texts are available in ANNIS, the search and visualization tool used by Coptic SCRIPTORIUM, and are available for download in TEI XML, PAULA XML, and relANNIS formats from the Coptic SCRIPTORIUM GitHub repository. Texts published as part of the KELLIA grant include sections of Shenoute's Acephalous Work 22 and Some Kinds of People Sift Dirt; large portions of his *I See Your Eagerness*; over two dozen sayings in the *Apophthegmata Patrum*; letters of Besa; Pseudo-Theophilus' *On the Cross and the Thief*; and sections of the *Martyrdom of Saint Victor the General* and the *Canons* of Apa Johannes. We have also published updates to Shenoute's *Not Because the Fox Barks* and some biblical books. These texts include

[8] https://github.com/gucorpling/gitdox/

[9] https://github.com/gucorpling/gitdox/blob/master/LICENSE

material previously transcribed prior to the KELLIA project, contributed by Diliana Atanassova, David Brakke, Alin Suciu, and the Marcion project.

Digital Corpora: Coptic Old Testament Texts

The Sahidic Coptic Old Testament consists of some 50+ books basically covering the entire Old Testament available from the Greek translation of the Jewish scriptures. During the grant period, the CoptOT team has assembled, augmented and revised authoritative text files for said body of literature. These files are used as base texts throughout the vmr in order to facilitate transcription work. Moreover, these files are also used to successively replace and enhance the files that are currently serving a search tool created by Christian Askeland and Matthias Schulz for the Chrome and Firefox web browsers.[10] In addition said files will also be exposed under a Creative Commons license (BY-SA) via our api to potential users outside of the mentioned use cases. Finally, the entire Sahidic Old Testament files - be it authoritative base text files or transcriptions of extant manuscripts created in the course of the ongoing CoptOT project - is destined to be fed via the VMR-Coptic SCRIPTORIUM Converter into the NLP-pipeline.

Optical Character Recognition

Optical character recognition for Coptic printed texts except those of Bohairic liturgical texts has not been available up to this point. KELLIA, represented by So Miyagawa, and the eTRAP research team at the Institute for Computer Science (University Göttingen) tested existing OCR programs applicable to Coptic texts (OCROPY and TESSERACT) and found out that OCROPY is the better choice, especially for texts with diacritical marks. The team provides trained data for Sahidic and Bohairic on GitHub.[11] This (sub-)project was considered so successful, that KELLIA partners secured funding for a 6-month fellowship to train more data and add to the corpus of OCRed Coptic printed texts (see above, Accomplishments).

[10] https://chrome.google.com/webstore/detail/sahidic-bible-askeland-sc/mbhdolnomjodfmgihfajipihojajgdjk, https://addons.mozilla.org/de/firefox/addon/sahidic-bible/

[11] https://github.com/KELLIA/CopticOCR

Digital Humanities Field Survey

As part of the work on metadata standards the German KELLIA team produced a survey of DH projects in Coptology/Egyptology and related fields (Historic Philologies, Ancient History). The survey was conducted in 2015 and 2016 and aggregated data for 39 such DH projects. These data have been compiled in an eXIST_db application and published in html. The html version has been published on the KELLIA website.[12]

Metadata Standards

Recommended standards and practices regarding metadata were developed as a result of the field survey and can be found in Appendix 4 (Metadata) and Appendix 5 (Linked Data).

Virtual Manuscript Room Enhancements

The following user-facing improvements were made to the VMR during the grant period:

- New Quire metadata capture facility to record codicological data
- New Biblia Coptica manuscript display option to show manuscript information in a format familiar to all Coptologists
- New Transcription Importer which allows transcriptions to be uploaded from multiple sources and imported to selected user account and publish level
- Greatly improved transcription display with better support for corrector popups, page/folio number, multiple project and translation language support with additional tabs, better inscriptio/subscriptio and lection metadata support
- Better information and improved user experience for image management, including new folio number generation tool
- Better handling of manuscript pages which include both Old and New Testament
- All new interactive Published Apparatus display (dECM display)
- Improved Zotero integration for bibliography data(worked around the 150 entry restriction)
- New Lection Index metadata facility to capture the intricate details for lectionary entries
- Improved Institution registry service to better support sharing Holding Institution collection with partner projects

[12] http://kellia.uni-goettingen.de/editions/

- Added the beginnings of support for page fragments by allowing individual images assigned to a page to designate to which shelf number they belong. This mechanism will be used when a page is divided into many fragments and those fragments live in various different institutions. Now one image can represent each individual fragment and associated with the appropriate shelf instance so no ambiguity remains about which fragments are at which institutions labeled as what shelf number.
- Integrated new release of Transcription Editor from Trier, which includes many bug fixes and improved editing experience
- Improved export options to include page ranges

Linked Data Standards

Please see Appendix 5

Appendix 1: Agendas from Project Workshops and Meetings

1st KELLIA Workshop in Göttingen, September 7-11, 2015

Participants

KELLIA participants: Heike Behlmer, Troy Griffitts, So Miyagawa, Ulrich Schmid, Caroline T. Schroeder, Uwe Sikora, Amir Zeldes, Tonio Sebastian Richter, Julien Delhez, Siegfried Richter, Frank Feder

Coptic Old Testament partners beyond the KELLIA group: Diliana Atanassova, Malte Rosenau

Göttingen Center for the Digital Humanities: Marco Büchler

Thesaurus Linguae Aegyptiae: Maxim Kupreyev, Simon Schweitzer

Dictionary and Database of Greek Loan Words in Coptic: Katrin John

Agenda

Monday, September 7

Overview

- Present NLP pipeline
    - API for communication with VMR?
- Discuss segmentation (all)
    - Subword morphemes (mnt-, r- & co)
    - Bound groups (Till vs. Layton)
    - Word boundaries (esp. handling of compounds)
    - VMR and annotation tools (Carrie, Amir, Ulrich, Troy, So)
    - TEI subsets for annotation
    - VMR and Scriptorium teams tag subsets / standards
    - Switch scriptorium annotators to use VMR? Separate instance?
    - Other annotation interfaces (EtherCalc prototype or similar to replace Excel)
    - Github connection (commit VMR text?)
- Lemmatization

    - Show new lemmatizer
    - Discuss use of TLA lemma list
    - technical integration
- Metadata scan - Uwe (+Carrie, Amir, Heike, So)
    - incl linked data (below)
    - incl persistent urns -- show data.copticscriptorium.org
    - Trismegistos and CTS URN
    - authority files and vocabularies
- Linked data - parsing, entities (people, geo, …), coreference
- versification and indexation
- Collaboration with TLA
    - Crum in XML
    - TLA in JSON, couchDB and their software which runs on Java
    - Funk lemmata
    - Funk corpus was already lemmatized.
    - license problem
    - Hyper-lemmatization; good for diachronic research

5:45 pm Meeting with Marco Büchler

7:30 pm Dinner at APEX, Burgstraße

Tuesday, September 8

Simon, Maxim (TLA); CoptOT; Kellia

09:00 TLA presentation (lemmatization)

Afternoon: metadata

Wednesday, September 9

10:00 Segmentation of Coptic texts

Use of the VMR, Coordination and Cooperation of CoptOT and INTF

Thursday, September 10

Continued General discussion of above topics

Skype presentation from Paul Dilley (Iowa) on Coptic Text Analysis

Friday, September 11

General discussion/Summary

Decisions

Tasks/Next steps

Dissemination

2016 KELLIA meeting

Saturday, July 23 and Sunday, July 24

Burkle Building, Room 12

Claremont Graduate University

Each portion of the program will consist of a short, informal presentation by the listed speaker(s) followed by discussion.

Participants:

Heike Behlmer, Georg-August-Universität Göttingen

Paul Dilley, University of Iowa

Frank Feder, Akademie der Wissenschaften zu Göttingen

Troy Griffitts, Akademie der Wissenschaften zu Göttingen

Christine Luckritz Marquis, Union Presbyterian Seminary

Rebecca Krawiec, Canisius College

Maxim Kupreyev, Berlin-Brandenburg Academy of Sciences and Humanities

So Miyagawa, Georg-August-Universität Göttingen

Beth Platte, Coptic SCRIPTORIUM

Tonio Sebastian Richter, Freie Universität Berlin

Caroline Schroeder, University of the Pacific

Melissa [Philip] Harl Sellew, University of Minnesota

Uwe Sikora, SUB Göttingen

Alin Suciu, Akademie der Wissenschaften zu Göttingen

Amir Zeldes, Georgetown University

Program

Saturday, July 23

9:00 Welcome

9:15 Uwe Sikora: Data model

So Miyagawa: Coptic OCR with a neural network modelled OCR engine and high-performance computing

10:15 Break

10:30 Maxim Kupreyev: TLA Lexicon update

11:30 Amir Zeldes: Web Application for the TLA lexicon

12:30 lunch

1:45 Amir Zeldes: NLP Pipeline and spreadsheet editor

2:45 Troy Griffitts: VMR and satellite sites

3:45 Break

4:00 Frank Feder: Transcription guidelines (developed with C. Askeland)

Sunday, July 24

9:00 Welcome

9:15 Frank Feder: Report on OT Base text

10:15 Break

10:30 Beth Platte/Amir Zeldes: Entities

11:30 Sebastian Richter: Database and Dictionary of Greek Loanwords in Coptic

12:30 lunch

1:45 Paul Dilley: Big Ancient Mediterranean/Iowa Canon of Coptic Authors and Works project report

2:45 Break

3:00 Melissa Harl Sellew: Ancient Lives project report

4:00 Wrap up/plan for next year

KELLIA Meeting in Washington, DC, December 2017

Participants

Elizabeth Davidson

Rebecca Krawiec

Christine Luckritz Marquis

Caroline T. Schroeder

Amir Zeldes

Via Skype: Elizabeth Platte

Agenda

Coptic Treebanking

Annotating and preparing corpora for publication

Linked Data (geographical)

Data Standards

Documentation

KELLIA Meeting in Göttingen, June 19-23, 2017

Participants

Agostino Soldati (PAThs)

Alberto Winterberg (DDGLC)

Alin Suciu (University of Göttingen/CoptOT)

Amir Zeldes (SCRIPTORIUM)

Becky (Rebecca) Krawiec (SCRIPTORIUM)

Beth (Elizabeth) Platte (SCRIPTORIUM)

Carrie (Caroline T.) Schroeder (SCRIPTORIUM)

Diliana Atanassova (University of Göttingen/CoptOT)

Frank Feder (University of Göttingen/CoptOT)

Heike Behlmer (University of Göttingen/CoptOT)

Julian Bogdani 20-23 (PAThs)

Julien Delhez (University of Göttingen/CoptOT)

Katrin John 21 via (DDGLC)

Laura Slaughter (University of Oslo)

Matt (Matthew) Munson (GCDH)

Marco Büchler (GCDH)

Malte Rosenau (University of Göttingen/CoptOT)

Maxim Kupreyev (BBAW)

Paola Buzi (PAThs)

Paul Dilley (University of Iowa) via Videolink

Sebastian Richter (BBAW/DDGLC)

Siegfried Richter (INTF, Münster)

So Miyagawa (University of Göttingen/CoptOT/Kyoto University)

Theresa Kohl (University of Göttingen/CoptOT)

Troy Griffitts (University of Göttingen/CoptOT)

Ulrich Schmid (University of Göttingen/CoptOT)

Wolf-Peter Funk (Guest)

Student Assistants: Lina Elhage-Mensching, Joanna Hypszer, Eva Kremer-Brinkmann

Schedule

Monday, June 19

10:00-12:00

Friedländer Weg: guided visit for KELLIA participants by CoptOT staff; Internal Scriptorium meeting;(simultaneously: CoptOT Steering Committee Meeting in Heyne-Haus)

13:30-15:00

Agenda; Discussion of progress from last meeting

15:30 Discussion with eTrap members (Text re-use) OCR of Coptic Texts (So, Kirill, and Marco)

Tuesday, June 20

9:00–9:30 NLP pipeline and online spreadsheet (Amir)

9:30-10:00 Coptic Treebank (Amir)

10:00-12:00 Exchange formats and text segmentation in digital editions: discussion

13:30-14:00 Metadata standards and exchange (So, Beth, Ulrich, Troy)

14:00-15:00 News from CoptOT: Infrastructure, Base-Texts, and Metadata

(Frank, Malte, Ulrich, Troy) & Discussion

15:30 Topic modelling in the Greek NT (Paul, via Videolink)

Wednesday, June 21

9:00–10:00

Discussion about the Coptic Dictionary Online

10:00-10:30 Presentation of the actual status of the Coptic Lemma List and BTS (Maxim)

10:30 DDGLC (Katrin via Skype)

13:30-14:00 PAThs presentation (Paola)

14:00-14:30 Digital Infrastructure of PAThs (Julian) & Discussion

15:30 Colophons of Biblical and other Mss (Agostino) & Discussion of future

cooperation

Thursday, June 22

9:00-10:30 C(anonical)T(ext)S(ervice) and CapiTainS Suite (Matt) & Discussion

10:30-12:00 Discussion about Unicode

13:30-15:00 How to construct a Coptic wordnet? (Laura) & Discussion

15:30 Scriptorium's internal meeting

Friday, June 23

Discussion of results; further steps; final report for KELLIA

KELLIA Concluding Meetings in Washington DC (11-14 June 2018), Göttingen (18-19 June 2018), Washington DC (13-14 September 2018)

Participants (Washington DC, June 2018)

So Miyagawa

Amir Zeldes

Via Skype: Caroline T. Schroeder, Rebecca Krawiec, Christine Luckritz Marquis

Agenda

Corpus editing and publication

Text data exchanges across projects

Treebanking and linguistic annotations

Participants (Göttingen, June 2018)

Diliana Atanassova

Heike Behlmer

Frank Feder

So Miyagawa

Ulrich Schmid

Caroline T. Schroeder

Agenda

Detecting text-reuse in Coptic (using eTrap's Tracer)

Documentation, White Paper

Text data exchanges

Future collaborations

Participants (Washington DC 13-14 September 2018)

Rebecca Krawiec

Christine Luckritz Marquis

Elizabeth Platte

Caroline T. Schroeder

Amir Zeldes

Agenda

Updating NLP tools

Annotating and publishing corpora

Linked data

Documentation, White Paper

Future, post-KELLIA collaborations

Appendix 2: Publications and Presentations

Publications

Behlmer, Heike. (2016) “Digitale Gesamtedtion und Übersetzung des koptisch-sahidischen Alten Testamentes” (annual report) In *Jahrbuch der Akademie der Wissenschaften zu Göttingen*. Berlin: De Gruyter, 277-282.

Behlmer, Heike and Ulrich Schmid. “Towards a New Digital Edition of the Coptic Old Testament,” In *Studies in Coptic Culture: Ordinary Lives, Changing Times* (ed. M. Ayad; Cairo: American University Press), forthcoming.

Behlmer, Heike. (2017) “Digitale Gesamtedition und Übersetzung des koptisch-sahidischen Alten Testamentes” (annual report) In *Jahrbuch der Akademie der Wissenschaften zu Göttingen* Berlin: De Gruyter, forthcoming.

Behlmer, Heike and Frank Feder. (2017) "The Complete Digital Edition and Translation of the Coptic Sahidic Old Testament. A New Research Project at the Göttingen Academy of Sciences and Humanities", *Early Christianity* 8, 97–107

Feder, Frank, Kupreyev, Maxim, Manning, Emma, Schroeder, Caroline T. and Zeldes, Amir (2018) "A Linked Coptic Dictionary Online". *Proceedings of LaTeCH 2018 - The 11th SIGHUM Workshop at COLING2018*. Santa Fe, NM, 12-21.

Krawiec, Rebecca and Caroline T. Schroeder. (2019, forthcoming) “Digital approaches to Studying Authorial Style and Monastic Subjectivity in Early Christian Egypt.” In *Digital Humanities and Religious Studies*, ed. Christopher Cantwell and Kristian Peterson.

Miyagawa, So, Kirill Bulert, Marco Büchler, and Heike Behlmer. (forthcoming). “Optical character recognition of typeset Coptic text with neural networks”. *Digital Scholarship in the Humanities.*

Miyagawa, So, Amir Zeldes, Marco Büchler, Heike Behlmer and Troy Griffitts (2018) Building Linguistically and Intertextually Tagged Coptic Corpora with Open Source Tools. In: Chikahiko Suzuki (ed.), *Proceedings of the 8th Conference of Japanese Association for Digital Humanities.* 139-41. Tokyo: Center for Open Data in the Humanities.

Miyagawa, So, Marco Büchler and Heike Behlmer. (forthcoming) “Computational Analysis of Text Reuse/Intertextuality: The Example of Shenoute Canon 6”. In Hany N. Takla, Stephen Emmel, and Maged S. A. Mikhail (eds.), *Proceedings of the*

Eleventh International Congress of Coptic Studies. Orientalia Lovaniensia Analecta. Leuven.

Miyagawa, So. (2018) Koputogo Saīdo Hōgen no Gengo Shiryō to Bunpō Chūshaku: Napori Kokuritsu Vittōrio Emanuēre 3 Sei Toshokan-zō Bēsa ni yoru Tekusuto no Danpen [Text Corpus and Grammatical Annotation of Sahidic Coptic: Fragments of a Text by Besa Preserved at Biblioteca Nazionale Vittorio Emanuele III, Naples]. *Journal of KIJUTSUKEN (Descriptive Linguistics Study Group)* 10, 271-320.

Schroeder, Caroline T., Amir Zeldes, *et al.,* Coptic SCRIPTORIUM corpora, 2015-2018, versions 1.5.0-2.6.0, http://copticscriptorium.org.

Zeldes, Amir and Caroline T. Schroeder. (2016) "An NLP Pipeline for Coptic". In: *Proceedings of LaTeCH 2016 - The 10th SIGHUM Workshop at the Annual Meeting of the ACL*. Berlin, 146-155.

Zhang, Shuo and Amir Zeldes. (2017) "GitDOX: A Linked Version Controlled Online XML Editor for Manuscript Transcription". In: Proceedings of FLAIRS 2017, Special Track on Natural Language Processing of Ancient and other Low-resource Languages. Marco Island, FL.

Presentations

Albrecht, Felix and Frank Feder. (2017) “Editing the Coptic Old Testament,” Society of Biblical Literature Annual Meeting, Boston.

Albrecht, Felix and Malte Rosenau. (2017), “Digital Old Testament,” 33. Deutscher Orientalistentag, Jena.

Behlmer, Heike. (2018) “Digitale Edition des koptischen Alten Testaments: Herausforderungen und Perspektiven,” Basel University Theology Faculty.

Behlmer, Heike. (2017), “Digitale Gesamtedition und Übersetzung des koptisch-sahidischen Alten Testaments (Keynote), Leibniz-Projekt “Polyphonie des spätantiken Christentums, Frankfurt

Behlmer, Heike. (2017) “Chancen und Herausforderungen der Digitalisierung des koptischen Kulturgutes (Keynote)” Altorientalisch-evangelischer theologischer Dialog, Berlin Evangelische Kirche in Deutschland

Behlmer, Heike. (2017) “‘Ich verbitte mir alle ungezogenen Bemerkungen darüber, dass ich Handschriften des Auslands nicht benutzt habe…’ Der Zugang zu Manuskripten und Sammlungen und die Erforschung der koptischen Bibel,” Berlin-Brandenburg Academy of Arts and Sciences

Behlmer, Heike and Frank Feder. (2017) “Digital Edition and Translation of the Coptic Old Testament” (poster), Ständigen Ägyptologenkonferenz, Goettingen.

Behlmer, Heike and So Miyagawa. (2017) “KELLIA” (poster), Ständigen Ägyptologenkonferenz, Göttingen.

Bulert, Kirill and So Miyagawa. (2017) “Optical Character Recognition with a Neural Network Model for Coptic.” Digital Humanities 2017, McGill University. Montreal, Canada.

Bulert, Kirill and So Miyagawa. (2016). “Optical Character Recognition with Neural Network Model and High-performance Computing on Printed and Handwritten Texts in Coptic.” Shenoute and the Bible International Conference. Georg-August-Universität Göttingen.

Dilley, Paul. (2016) “Coptic Scriptorium beyond the Manuscript: Towards a Distant Reading of Coptic Texts,” Digital Humanities Panel at the International Association of Coptic Studies Congress, Claremont Graduate University

Feder, Frank. (2017) “The Digital Coptic Old Testament: Towards the Virtual Reconstruction of a Unique Linguistic, Literary, and Religious Monument,” Coptic Studies in the Digital World Conference, University of Toronto.

Feder, Frank. (2017) “Editing the Coptic Old Testament,”Society of Biblical Literature Annual Meeting, Boston.

Griffitts, Troy. (2017) “The NT VMR as a Gateway to Manuscripts,” Society of Biblical Literature Annual Meeting, Boston.

Griffitts, Troy. (2016) “The Virtual Manuscript Room Collaborative Research Environment,” Quinto convegno annuale Associazione per l'Informatica Umanistica e le Culture Digitali, Ca’ Foscari University of Venice.

Krawiec, Rebecca. (2018) “Studying Ancient Egyptian Christianity in a Modern Digital World,” Digital Humanities Speaker Series, Canisius College

Krawiec, Rebecca. (2016) “Charting Rhetorical Choices in Shenoute: Abraham our Father and I See Your Eagerness as Case-Studies,” KELLIA Digital Coptic Studies Panel at the International Association of Coptic Studies Congress, Claremont Graduate University

Lincke, Eliese-Sophia. (2016) “Optical Character Recogition (OCR) for Coptic. Testing Automated Digitization of Texts with OCRopy” Digital Humanities Panel at the International Association of Coptic Studies Congress, Claremont Graduate University

Luckritz Marquis, Christine. (2016) “Reimagining the Apophthegmata Patrum in a Digital Culture,” KELLIA Digital Coptic Studies Panel at the International Association of Coptic Studies Congress, Claremont Graduate University

Miyagawa, So. 2018. Quotation from the Psalms and Its Authority in Shenoute’s Monastic Education. “As It Is Written"? Uses of Sources in Ancient Mediterranean Texts. Georg-August-Universität Göttingen.

Miyagawa, So, Amir Zeldes,Marco Büchler, Heike Behlmer and Troy Griffitts. (2018) Building Linguistically and Intertextually Tagged Coptic Corpora with Open Source Tools. Eighth Conference of Japanese Association for Digital Humanities (JADH2018) "Leveraging Open data". Hitotsubashi Hall, Tokyo.

Miyagawa, So and Marco Büchler. (2018) “The Use of Digital Tools in Studies of Biblical Intertextuality in Early Christian Authors: The Case of Shenoute and Besa, Coptic Abbots in the 4-5th Centuries Digital Humanities.” Biblical Studies, Early Jewish and Christian Studies (EABS). SBL 2018 International Meeting. University of Helsinki, Helsinki, Finland.

Miyagawa, So and Amir Zeldes. (2018) "A Semantic Map of the Coptic Complementizer če Based on Corpus Analysis: Grammaticalization and Areal Typology in Africa," International Workshop on Semantic maps: Where do we stand and where are we going? Liège, Belgium.

Miyagawa, So. (2018) “Evaluation of the Digital Text Reuse Detection on Coptic Texts by TRACER A Case Study on Besa.” eTRAP Sponsor Meeting, Georg-August-Universität Göttingen.

Miyagawa, So. (2018) “Digitisation of Coptic Manuscripts and Digital Humanities Initiatives in Germany, the U.S., Japan, and Israel.” The 21st Century Curatorship Seminar Series. The British Library, London, the United Kingdom.

Miyagawa, So. (2017) “Shenoute and the Bible: Digital Text Re-Use Analysis of Selected Monastic Writings from Egypt.” Tag der GSGG 2017. Georg-August-Universität Göttingen.

Miyagawa, So. (2017) “Word Segmentation and Polysynthesis in Coptic.” International Summer School on Typology and Lexicon (TyLex) National Research University Higher School of Economics, jointly with Hebrew University of Jerusalem and Stockholm University, Moscow. (poster session)

Miyagawa, So, Julien Delhez and Heike Behlmer. (2017) “Schriftauslegung und Bildungstraditionen im koptischsprachigen ägyptischen Christentum der Spätantike:

Schenute, Kanon 6.” Jahrestagung "Was ist Bildung in der Vormoderne?"Georg-August-Universität Göttingen. (poster session)

Miyagawa, So, Kirill Bulert and Marco Büchler. (2017) “Utilization of Common OCR Tools for Typeset Coptic Texts” DATeCH (Digital Access to Textual Cultural Heritage), Georg-August-Universität Göttingen. (poster session).

Miyagawa, So. (2017) “The Integration of Existing Digital Text Corpora and Corpus-Linguistic Tools into Research on Transitivity and Valency in Coptic.” Inaugural Workshop of the GIF Project „Transitivity and Valency in Contact: The Case of Coptic“ Berlin-Brandenburgische Akademie der Wissenschaften Berlin, Germany.

Miyagawa, So. (2016) “The Use of the Psalms in Educational Monastic Literature in Byzantine Egypt: A Case Study of Shenoute, an Egyptian Monastic Leader of the Fourth and Fifth Centuries.” Third Parekbolai Symposium on Byzantine Literature and Philology National and Kapodistrian University of Athens & Byzantine and Christian Museum Athens, Greece.

Miyagawa, So. (2016) Coptic Studies and Digital Humanities Digitale Unterstützung in den Geistes- und Gesellschaftswissenschaften Georg-August-Universität Göttingen Göttingen. Germany.

Miyagawa, So. (2016) “Biblical Quotations and Allusions in Coptic Literature.” International Conference ‘The World in Motion: Language, Culture, and Intercultural Communication in Asian and African Studies’ University of Warsaw Warsaw, Poland

Miyagawa, So, and Marco Büchler. (2016) “Computational Analysis of Text Reuse in Shenoute and Besa,” Digital Humanities Panel at the International Association of Coptic Studies Congress, Claremont Graduate University

Miyagawa, So and Behlmer, Heike. (2016) “Processing Non-Biblical Texts (e.g. Besa and Shenoute) for the Old Testament Virtual Manuscript Room and Using Text Re-Use Software for Detecting and Describing (Biblical) Intertextuality.” Shenoute and the Bible International Conference. Georg-August-Universität Göttingen.

Miyagawa, So. (2016) “Koputo Ejiputogo Saīdo Hōgen Tekusuto no Deijitaruka ni okeru Shomondai: Unicode, OCR, Denshikōpasu wo Chūshin ni [Problems in Digitization of the Texts in the Sahidic Dialect of Coptic Egyptian: with a Focus on Unicode, OCR, and Electronic Corpus].” Tōzaibunka no Tayō to Kyōzon Moderu: “Tōhō Kirisutokyōken” wo Takakuteki ni Kangaeru Gakusaiteki Kokoromi, 2015 Kyōto Daigaku Bun’ya Ōdan Purattofōmu Kōchiku Kikaku (Kenkyū Daigaku Kyōka Sokushin Jigyō “Hyakka Sōmei” Puroguramu) [Diversity and Coexistance Model in the Cultures of the West and East: Interdisciplinary Attempt to Think on

"Eastern/Oriental Christian Area," 2015 Kyoto University Interdiciplinary Platform Building Plan (Research Univeristy Intensification "Hyakka Sōmei" Program)], Kyoto University.

Miyagawa, So and Kirill Bulert. (2016). "Coptic OCR", Workshop on Supercomputing in the Humanities Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen, Göttingen.

Platte, Elizabeth and Caroline T. Schroeder. (2016) "Coptic Scriptorium: Data from the Desert." Linking the Big Ancient Mediterranean Conference, University of Iowa.

Rosenau, Malte. (2017) "Digital Humanities and Coptology," Colloque Census—Recenser et identifier les manuscrits par langue et par pays, Bibliothèque Nationale Paris.

Schroeder, Caroline T. (2018) "The Materiality of Digital Apocryphal Studies," The Material of Christian Apocrypha Conference (of the North American Society for the Study of Christian Apocryphal Literature), University of Virginia.

Schroeder, Caroline T. (2018) "Annotating Heresies (Keynote)." Future Philologies. Institute for the Study of the Ancient World. New York University.

Schroeder, Caroline T. (2018) "A Homily is a Homily is a Homily is a Corpus: Digital Approaches to Shenoute," The Transmission of Early Christian Homilies from Late Antiquity to the Middle Ages Conference, Goethe-Universität Frankfurt am Main.

Schroeder, Caroline T. (2017) "Web-based Natural Language Processing and Annotation Technologies for Ancient Languages," Leipzig University.

Schroeder, Caroline T. (2016) "Coptic SCRIPTORIUM: A Digital Platform Research in Coptic Language and Literature," KELLIA Digital Coptic Studies Panel at the International Association of Coptic Studies Congress, Claremont Graduate University

Schmid, Ulrich. (2018) "Towards a joint digital catalogue of Sahidic biblical manuscripts," Manuscript Cataloguing in a Comparative Perspective: State of the Art, Common Challenges, Future Directions Conference, Hamburg University.

Schmid, Ulrich. (2016) "The Virtual Manuscript Room," Global Philology Planning Meeting, Freie Universität Berlin.

Schmid, Ulrich. (2016) "Eine digitale Edition der koptischen (sahidischen) LXX - Probleme und Lösungen" Kirchliche Hochschule Wuppertal, 6. Internationale Septuaginta-Konferenz " Die Septuaginta, Geschichte, Wirkung, Relevanz"

Schmid, Ulrich. (2016), “Der New Testament Virtual Manuscript Room (NTVMR) und das koptisch-sahidische Alte Testament”, Düsseldorf, Nordrhein-Westfälische Akademie der Wissenschaften, Workshop AG “eHumanities”

Sikora, Uwe. (2016) “Text Encoding - Opportunities and Challenges,” Digital Humanities Panel at the International Association of Coptic Studies Congress, Claremont Graduate University

Zeldes, Amir, Rebecca Krawiec, Elizabeth Platte, Caroline T. Schroeder. (2018) “A Linked Digital Environment for Coptic Studies,” Friday Speaker Series, Georgetown University Department of Linguistics

Zeldes, Amir. (2016) “A Quantitative Approach to Syntactic Alternations in Sahidic,” KELLIA Digital Coptic Studies Panel at the International Association of Coptic Studies Congress, Claremont Graduate University

Training Workshops on Tools and Methods

Coptic Fonts & Coptic Bible. International Association of Coptic Studies Congress, July 2016, Claremont CA. Led by Christian Askeland and Frank Feder, Caroline T. Schroeder participant.

Digital Tools for Beginners (Workshop on Coptic SCRIPTORIUM). International Association of Coptic Studies Congress, July 2016, Claremont CA. Caroline T. Schroeder, Amir Zeldes, and Rebecca Krawiec.

Digital Editions and Text Analysis: Coptic as Case Study. North American Patristics Society Annual Meeting, May 2017, Chicago IL. Caroline T. Schroeder, Rebecca Krawiec.

Digital Humanities Tools and Methods for Studying Antiquity. Department of Ancient History, University of Basel. 11-15 December 2017. Caroline T. Schroeder

Einführung in digitale Forschungsmethoden der Digital Humanities (DH) in der Ägyptologie und Koptologie. Georg-August-Universität Göttingen. 20 Juli 2018. So Miyagawa and Uwe Sikora.

Appendix 3: Transcription and Encoding Standards

1. Introduction

This document provides recommended guidelines for anyone working in digital Coptic textual studies, including manuscript studies and paleography, linguistics and natural language processing, philology, and other relevant fields. These guidelines were produced as an outcome of a collaborative, international exchange. They should be considered minimal recommendations applicable to most projects. Each researcher or research group will have particular research questions, and therefore project-specific needs for transcription and encoding. ***We recommend each individual researcher or project produce their own guidelines and practices specific to the needs of their particular research in consultation with these guidelines.***

These guidelines assume either manual transcription of text or the encoding and annotation of previously digitized text. We anticipate they can be easily adapted for text digitized via Optical Character Recognition (OCR) for Coptic as that technology improves.

These guidelines cover transcription of text, the encoding of Coptic characters, metadata (information about the textual object), and annotations of the text itself.

2. Transcription environments

Many projects find it useful to transcribe plain text before annotating it with further information. When transcribing Coptic on one's personal computer, transcription with a simple text editor in a plain text file (.txt format) is recommended. Scholars of Coptic have reported multiple occasions when proprietary word-processing software such as Microsoft Word does not visualize Coptic characters properly.

Transcriptions with annotations can be composed in a variety of programs installed on one's computer or with web-based tools. Installed programs include simple text editors (which require manual typing of annotations or tags) or more robust programs, such as the Oxygen editor (which can be customized for a project's particular annotation schema). A variety of open-source web-based tools for transcription also exist, many of which would need customization for Coptic (e.g., T-Pen) or for a particular project's needs (e.g., papyri.info's Papyrological Editor). **We strongly recommend projects**

working in Coptic contact KELLIA partners about adapting the following two tools developed for transcription and annotation in Coptic:

- GitDox: a light-weight transcription and annotation tool customizable for individual projects and for multiple languages, linked to Coptic natural language processing tools (supported by Coptic SCRIPTORIUM)
- Virtual Manuscript Room: a transcription and annotation tool used for biblical and literary manuscript transcription (supported by the Coptic Old Testament project)

Researchers will likely need to customize the tool or their own workflow.

3. Character Encoding

We recommend using the official Unicode (UTF-8) Coptic character set. Transcribing in ASCII legacy fonts leads to a mismatch between the digital characters in the digital file and the visualization of those characters using a font; thus digital texts using ASCII characters and legacy fonts are neither sustainable nor easily shared.

A table of the Unicode Coptic Characters can be found on the Coptic SCRIPTORIUM wiki and the Pennsylvania State University languages site.[13]

We invite researchers in digital Coptic to contribute their expertise in the use and application of these characters to the Coptic SCRIPTORIUM wiki.

Use of unofficial characters or character encodings in the “Private Usage Area” are *not* recommended, due to potential problems with sustainability and exchange of data. Researchers should be aware that the Coptic character set for manuscript and paleographical symbols apart from the alphabet is incomplete. As of this writing, two email list-serves exist to discuss digital Coptic and Coptic Unicode. Researchers interested in discussing these issues in more detail are encouraged to contact a member of the KELLIA group about joining one or both of these list-serves.

The Antinoou font is recommended for properly visualizing the Coptic characters.[14] It also visualizes combining characters (such as supra-linear strokes). The font created by

---

[13] http://wiki.copticscriptorium.org/doku.php?id=kellia:unicode:coptic_unicode_standards_and_guidelines_for_coptologists; http://sites.psu.edu/symbolcodes/languages/ancient/coptic/copticchart/

[14] https://www.evertype.com/fonts/coptic/

the Institute français d'archéologie orientale uses Private Usage Area character encodings and is not recommended.[15] The New Athena Unicode font is an alternative to Antinoou.[16] When typing in Unicode characters, one must install both a font to visualize the characters on screen and a digital keyboard to map your computer's keystrokes on to the Coptic character set.

Many scholars have digitized text in legacy (pre-Unicode) fonts. We recommend all projects convert or re-transcribe texts in these legacy fonts. Coptic SCRIPTORIUM and PATHs both provide tools to convert legacy fonts into Unicode characters.[17]

4. Transcription and Digitization

The following guidelines apply to text transcription and digitization, whether using a web-based tool or simple text editor.

- Preserve the original source text spellings and orthography, whether that source is a manuscript, a print edition, or previously digitized edition. Resist the temptation to "correct" spelling in the source text; instead use annotations for normalization, lemmas, etc.
- We recommend against using capitalization, since Coptic does not have the same concept of "capital" letters as modern languages. We instead recommend using annotation to mark oversize characters.
- Projects may wish to use a previously digitized text (e.g., Warren Wells' Sahidica New Testament or OCR of a print edition) as a "base text" which then is modified in consultation with a manuscript or another source; using a base text can save time in transcribing a manuscript, for example. Exercise care in proofreading the work, as with all transcriptions.
- Projects that wish to use the **Natural Language Processing services of Coptic Scriptorium**[18] should use word segmentation that follows principles of binding and segmentation that conform to the linguistic principles of Bentley Layton's

[15] http://www.ifao.egnet.net/publications/publier/outils-ed/polices/

[16] https://apagreekkeys.org/NAUdownload.html

[17] https://github.com/CopticScriptorium/converters, https://github.com/paths-erc/cmcl2unicode with demo at http://paths.uniroma1.it/cmcl2unicode/index.html.

[18] Amir Zeldes and Caroline T. Schroeder, "An NLP Pipeline for Coptic," *Proceedings of the 10th ACL SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH2016)* Berlin, 2016. https://doi.org/10.18653/v1/W16-2119. The NLP pipeline can be accessed at https://corpling.uis.georgetown.edu/coptic-nlp/.

Coptic Grammar.[19] Place a **unique character** (such as a space or underscore) between Coptic bound groups. If a project transcribes using Walter Till's principles of transcribing Coptic[20] and wishes to use Coptic Scriptorium's NLP service, we recommend transcribers place a unique character between morphemes or words that are bound in Coptic Scriptorium's protocols[21] but typically remain unbound using Till's principles (e.g., long prepositions followed by articles or nouns). This character can be removed prior to running the text through NLP; it can be replaced by a space prior to the project visualizing the text in other forms.

5. Digital Annotation and Encoding for Textual Structure and Metadata

We urge all projects provide rich metadata as documentation of analog and digital information about their texts. Research into Coptic text increasingly is addressing the historical context of these documents and their circulation in the ancient and modern worlds. Additionally, Coptic Studies as a field highly values the ability to gauge "authenticity" and "validity" of textual and philological research. Information about the source, curatorial, and editorial histories of the digitized text enables further research and confers upon the project a greater likelihood of recognized scholarly legitimacy by the field.

Depending on the specific research questions of the project, you may seek to annotate digital text with information (paleographic information about columns, page breaks, marginalia; linguistic information such as part of speech; citations and references of "text reuse" such as quotes and allusions to biblical passages or other ancient literature; etc.). Plain text (in UTF-8 character encoding) is often quite a useful format for sharing source material and for some forms of research; other projects may need further annotation.

---

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

[20] Walter C. Till, "La Séparation des Mots en Copte," *Bulletin de l'Institut Français d'archéologie Orientale* 60 (1960): 151–70.

[21] See Section 4 of Schroeder and Zeldes, "Coptic SCRIPTORIUM Diplomatic Transcription Guidelines," v. 1.3 (2018), online, accessed 18 June 2018; the most up-to-date version of the guidelines can be found at http://copticscriptorium.org/documentation.

For encoding both textual data and metadata, we recommend the following considerations:

1. Although it is possible to specify one's own schemas we recommend consulting existing standards and models, such as the specifications of the Text Encoding Initiative (TEI)[22]. TEI XML counts as the most used and broadly accepted standard to describe textual phenomena.[23] The TEI-Standard consists of several modules focusing on different aspects of textuality and thus represents a highly customisable Schema to create TEI-valid but project-specific Sub-Schemas.

2. Existing standards may not always fulfill all the needs of a project and do not guarantee interoperability across projects. Application of standards involves project-specific interpretation and modification.

3. Nonetheless, using or adapting existing standards may help a project with data-modeling, even if the project ultimately does not use fully-compliant TEI XML. For example, a project may use spreadsheets or databases to record metadata. The categories and data-modeling provided by TEI XML may inform a project's data model, even if the project doesn't use XML.

4. One sub-schema broadly used and tested by projects describing epigraphical data is the EpiDoc (Epigraphic Documents in TEI XML) Schema.[24] Since Coptology is confronted with handwritten manuscripts not conforming the modern concept of typographic textuality and its implications, EpiDoc[25] is recommended to encode Coptic text-bearing objects or as a data model.

6. Additional textual annotation considerations

Best practices include not only using TEI tags but documenting the usage of the tags and structures from the TEI set. The structure and tagset of XML-Documents can be specified by DTDs (Document Type Definition) or Schema-languages like XML-Schema [26], Relax NG (RNG)[27] or Schematron[28] to guarantee standardised and valid XML-data.

---

[22] http://www.tei-c.org/

[23] http://www.tei-c.org/Guidelines/Customization/

[24] https://sourceforge.net/p/epidoc/wiki/Home/

[25] http://www.tei-c.org/Guidelines/P5/

[26] https://www.w3.org/XML/Schema

An ODD Format (“One Document does it all”)[29] can be used. It brings together the Documentation of Tags and the formal declaration which can be compiled[30] into different schema-languages.

7. Tools

Encoding textual data is typically accomplished by using so called markup-languages. XML (Extensible Markup Language) is used as today's de facto standard to represent Texts as hierarchically structured, machine readable data. Furthermore XML markup and related processing scripts are non-proprietary and thus free to use while specified, refined and documented by the W3C (World Wide Web Consortium) (e.g., XSLT, a mechanism to query XML encoded texts or to transform them into different visualizations such as HTML web pages; and XQuery, a language to query XML encoded texts).

8. Additional metadata considerations

See the KELLIA White Paper on Metadata Standards, also published as Appendix 4 of the main KELLIA project White Paper.[31]

9. Visualizing and Publishing Encoded Text Data

TEI-C provides several stylesheets (XSL) to convert xml files into various file formats including html.[32] It is recommended that projects use existing stylesheets and amend them where necessary to ensure the proper display of all of their encoded data.

---

[27] http://relaxng.org/

[28] http://schematron.com/

[29] See http://www.tei-c.org/release/doc/tei-p5-doc/en/html/TD.html and http://www.tei-c.org/release/doc/tei-p5-doc/en/html/USE.html

[30] Using the Tool Roma provided by TEI on http://www.tei-c.org/Roma/

[31] https://kellia.uni-goettingen.de/downloads/KELLIA-metadata-white-paper.pdf, https://kellia.uni-goettingen.de/downloads/KELLIA-white-paper.pdf

[32] https://github.com/TEIC/Stylesheets/tree/dev/html

For visualisation of linguistic text data, ANNIS is recommended.[33] ANNIS is a highly multifunctional visualization platform of XML data with linguistic annotation. One can add various things such as syntactic tree, morphological information, part-of-speech, translation, audio, video as well as philological information such as page, column, and quire numbers and identification number of manuscripts. ANNIS can visualise the data written in PAULA XML. Using SaltNPepper or Exmaralda, one can convert various XML file formats including TEI XML into PAULA XML.

[33] http://corpus-tools.org/annis/

Appendix 4: Metadata Standards

1. Introduction

Metadata records basic descriptive and administrative metadata like license statements and call numbers from holding institutions, as well as much more detailed descriptions, e.g., of the materiality of information carriers or more abstract aspects of the encoded source and its cultural implications are possible.

German partners completed a survey of Metadata Standards and formats used in the field of Coptic studies and neighboring disciplines. The survey results are in a database published as KELLIA E-ditions.[34] This appendix contains a summary of findings and recommendations based on the survey and work in KELLIA. The survey was conducted *prior* to the establishment of the PAThs project in Rome.[35] PAThs will be providing unique identifiers to Coptic text-bearing objects; we encourage projects to follow PAThs ongoing work.

2. Encoding

Most projects use TEI-P5 or a specialised subset of TEI as EpiDoc to ensure interoperability in theory. TEI XML (and the EpiDoc subset) includes the msdescription-module with elements for describing manuscripts.[36]

Since every project has its own perspective, in reality many tags are interpreted in different ways due to the very general TEI-Definitions. Projects often develop their own metadata categories and map a set of metadata on to TEI-P5. The Coptic Old Testament project's metadata model is available online as an example.[37]

[34] http://kellia.uni-goettingen.de/editions/

[35] http://paths.uniroma1.it/

[36] http://www.tei-c.org/release/doc/tei-p5-doc/en/html/MS.html

[37] https://docs.google.com/spreadsheets/d/189qBxOylUyo0rgUSP20kdkE5SBaAAHg7tAerIL1D5Yg

3. Authority files

Recommended is also the use of authority files or similar data to link identical entities to a general information resource. For Coptic Studies, currently the best solution is to link to Trismegistos People[38] for a historic person or authority data for proper names. (Note: the Trismegistos People database *is not* a complete or precise prosopography.)

Pelagios[39] and Pleiades[40] are the best authorities for linking ancient geographical entities. To find names to specify a geographic location one may use GeoNames[41] or Getty.

PAThs is developing identifiers for place names in Egypt; projects should contact PAThs regarding their metadata and identifiers for provenance.

A general tendency is to link personal, geographic or other kinds of entities to Wikidata[42] to use structured linked open data-sets (also known as "Wikification"). One benefit is that one can generate data-sets that can be used by anyone.

4. Standards and Controlled Vocabularies

There are no cross-project controlled vocabularies for metadata, and existing controlled vocabularies (Getty, Dublin Core, EAGLE) do not suffice for Coptic Studies. We recommend that each project publish their own controlled vocabularies to ensure data integrity and consistency within the project; new projects should survey existing projects so they do not need to "reinvent the wheel." For more on controlled vocabularies, see KELLIA White Paper on Linked Data Standards and Practices, also published as Appendix 5 of the main KELLIA project White Paper.[43]

---

[38] www.trismegistos.org

[39] http://commons.pelagios.org/

[40] https://pleiades.stoa.org/

[41] http://www.geonames.org/

[42] www.wikidata.org

[43] https://kellia.uni-goettingen.de/downloads/KELLIA-linked-data-white-paper.pdf, https://kellia.uni-goettingen.de/downloads/KELLIA-white-paper.pdf

The PAThs project in Rome will be publishing stable identifiers for every Coptic manuscript and for place names in Egypt; it will also Clavis Coptica entries for each known Coptic text work.[44] ***KELLIA partners encourage all Coptic projects to include these identifiers in their metadata.***

The Coptic Old Testament project explored mapping VMR-Data to METS/MODS to provide an internationally approved metadata exchange format. This task could not be accomplished:

- Mapping the complex VMR structure to the even more complex METS/MODS was time consuming; we aborted the trial during the conception-phase after realising that we do not have the time to dig deep into METS/MODS to achieve a proper result.
- Due to the fragmentation of Coptic manuscripts, one “document” would be split in the VMR into different items with different rightholders and holding institutions. Mapping multiple items to a single METS/MODS representation was difficult, since METS/MODS is designed to present a single dataset for a single legal Resource.

5. Data-Access

To make data accessible for future research the metadata should be saved and provided in a machine-readable form and under special licence agreements that allow re-usage.

Many Coptological projects are accessible via self published or institutionally hosted websites.[45] But it depends on the project holders themselves if and how they publish their data on this platform. On some websites one can easily download the required information because free access to the data is provided. Just to name a few examples, this is the case for the following projects: Coptic Scriptorium[46], Inscriptions of Israel / Palestine[47], U.S. Epigraphy Project[48], epidat - epigraphische Datenbank[49], Monasterium

[44] http://paths.uniroma1.it/

[45] This White Paper’s scope is digital editions projects in a narrower sense (see above). As a side note, the blog of Alin Suciu should nevertheless be mentioned, where resources regarding coptological (mainly philologically centered) publications are provided. (http://alinsuciu.com)

[46] http://copticscriptorium.org/

[47] http://cds.library.brown.edu/projects/Inscriptions/index.shtml

[50], Epigraphische Datenbank Heidelberg[51], Inscriptiones Graecae[52], digilibLT - Biblioteca digitale di testi latini tardoantichi[53], Bibliotheca Palatina digital[54], Papyri.info[55] and Papyrus und Ostraka Projekt[56]. Most of them provide the data via XML-files. The Deutsches Textarchiv[57] and Germania Sacra. Die Kirche des Alten Reiches und ihre Institutionen[58] even offer more data formats like TCF, Turtle or json-ld). Sometimes a free registration is required to download the information (Papsturkunden des frühen und hohen Mittelalters[59]) and on a few occasions one can only register by paying a fee (Corpus dei Manoscritti Copti Letterari (CMCL)[60]).

Several homepages provide a section in which a reference to similar or related projects such as Epigraphische Datenbank Heidelberg[61], U.S. Epigraphy Project[62], Inscriptions of Israel / Palestine[63], Papyri.info[64], and Germania Sacra: Die Kirche des Alten Reiches

---

[48] http://usepigraphy.brown.edu/projects/usep/collections/

[49] http://www.steinheim-institut.de/cgi-bin/epidat

[50] http://monasterium.net/mom/home?_lang=deu

[51] http://edh-www.adw.uni-heidelberg.de/home?lang=de

[52] http://telota.bbaw.de/ig/

[53] http://digiliblt.lett.unipmn.it/index.php

[54] http://digi.ub.uni-heidelberg.de/de/bpd/index.html

[55] http://papyri.info/

[56] https://papyri.uni-leipzig.de/content/start.xml?XSL.lastPage.SESSION=/content/start.xml

[57] http://www.deutschestextarchiv.de/

[58] https://adw-goe.de/fr/forschung/forschungsprojekte-akademienprogramm/germania-sacra/

[59] http://www.papsturkunden.de/EditMOM/home.do

[60] http://www.cmcl.it/

[61] http://edh-www.adw.uni-heidelberg.de/links

[62] http://usepigraphy.brown.edu/projects/usep/links/

[63] http://cds.library.brown.edu/projects/Inscriptions/related.shtml

[64] http://papyri.info/docs/resources

und ihre Institutionen[65] can be found. In doing so, it is easier for the user to get to know and browse already existing and theme related undertakings. But this is a far cry from an extensive catalogue in which projects using digital methods are listed as it is just rudimentary linking from project-website to project-website. A central coptological platform aggregating discipline-related digital editions or similar data does not exist at the moment.

Some projects provide access to content data via defined APIs like OAI-PMH. (See Das Altägyptische Totenbuch. Ein Digitales Textzeugenarchiv[66] OAI-PMH interface.)

6. Recommendations

Data created by computer aided Coptological research should be provided as machine-readable and thus digital data. In this manner encoded data including metadata should be archived by institutionally or disciplinary bound repositories which allow long-term digital preservation. This comprises not just storage space or its support but also persistent identification of digital resources via persistent identifiers (PIDs) like DOI[67] or other kinds of Uniform Resource Identifiers (URI).[68] As an institutional Repository, for example, TextGrid-Repository[69] can be mentioned which created a consistent long-term preservation policy and infrastructure: Each digital resource [there] is identified by a PID and accessible via an URL.[70] Furthermore the Repository and its data is fulltext searchable and thus provides direct access to all digital data produced. Whereas recommending a specific Repository is not in the scope of this whitepaper's responsibility, repositories that are qualified by a **DINI-Certificate**[71] or a similar certification mark may be more desirable.

---

[65] https://adw-goe.de/fr/forschung/forschungsprojekte-akademienprogramm/germania-sacra/links/

[66] http://totenbuch.awk.nrw.de/

[67] https://www.doi.org/

[68] Persistent identifiers are not just to identify resources but also to cite content from digital resources.

[69] https://textgridrep.org/

[70] See for example the resource textgrid:123rw that may be accessed via https://textgridrep.org/browse/-/browse/123rw_0

[71] http://www.dini.de/dini-zertifikat/

Machine-readable data created during the process of digital editing should be explicitly bound to a license which allows free access by means of scientific reuse. Therefore, the Creative Commons Licenses CC-BY[72] and CC-BY-SA[73] are recommended.

Data access does not just concern physical access to analog or digital resources but also legal aspects with regard to creation and usage of resources. In that manner in the metadata, legal technicalities should be clarified for both the source and the digital encoded data regarding authorship, data privacy and personality rights if applicable.

For linking and access, well-resourced projects should consider providing access to metadata via APIs (REST, OAI-PMH). Smaller projects can provide metadata downloadable as csv files for further manipulation and research.

One desiderata is a central Coptological platform that catalogues projects with descriptive and administrative metadata that can be searched (possibly drawn from linked data API's from the projects). Desirable attributes include the following features:

- in English and including projects world wide
- Searchable metadata and documentation:
    - the link to the according project website (maintained to avoid dead links)
    - a brief outline of the project
    - the responsible persons and institutions
    - metadata with regard to geographically, chronologically, thematically, institutionally, data-access (free, registration, fee-based), status (ongoing, completed, discontinued) information
- a defined scope, eg. "digital editions", "linguistically encoded corpora" etc.
- interactive: the user can contribute to the list of digital projects or correct an entry
- list can be downloaded in various forms
- license for use of data is stated

[72] https://creativecommons.org/licenses/by/3.0/

[73] https://creativecommons.org/licenses/by-sa/3.0/

- connection with other networks / platforms

Appendix 5: Linked Data Standards and Practices

1. Introduction

As part of the overall KELLIA collaboration goal of establishing data standards for related and future projects, KELLIA set out to create standards that ensure collaboration, data-exchange, and compatibility in linked-data initiatives across digital Coptic projects. To this end, we have met with members of other projects working on linked data initiatives in the ancient world, created products that are a result of data exchange and link data across project in the KELLIA collaboration, and considered current practices in linked data when establishing standards, especially for metadata. KELLIA partners and other digital projects working in Coptic need robust linked data infrastructure that allows linking between KELLIA projects and to other projects, even when standards between projects vary.

Several projects currently present opportunities to link geographical data from the ancient world. In particular, US KELLIA partners have begun to work with Pleiades (https://pleiades.stoa.org/), a digital gazetteer of the ancient world. In doing so, they have also worked with Pelagios (http://commons.pelagios.org/), which provides infrastructure for linking geographic data. Pelagios works largely with projects focused on the ancient and medieval periods. Both US and German KELLIA projects have also used Trismegistos (http://www.trismegistos.org/) identifiers in metadata. Trismegistos is a group of metadata databases of information about texts, collections, authors, people, and places. It originally began with information from papyrological and epigraphic information from Egypt, but has since expanded to information from the ancient Mediterranean world more generally.

Finally, KELLIA partners also look forward to working with PATHs (http://paths.uniroma1.it/), a project focused on putting Coptic literary texts into their geographic context. PATHs plans to produce an "Archaeological Atlas of Coptic Literature," and, in doing so, collect and classify information about Coptic literature, collections, manuscripts, colophons, authors, copyists, donors, and institutions. Unlike Trismegistos, which does not contain a comprehensive catalogue of Coptic manuscripts and manuscript fragments, the PATHs project plans to provide permanent identifiers for Coptic text-bearing objects. Ideally, future work will entail collaborating with Trismegistos to update their catalogues. Such work will be very valuable for KELLIA members, and controlled vocabularies and metadata standards in Coptic digital humanities will certainly be influenced by the work of the PATHs project. Because

PATHs is a relatively new project, KELLIA partners are awaiting PATHs outcomes to take advantage of linked data opportunities more fully. Sharing and linking data makes the work of the individual projects and initiatives that are part of KELLIA more valuable, as other projects use and expand it. This has already been demonstrated by outcomes of sharing data noted below. Linking data to projects both within and outside of KELLIA also makes our individual work more discoverable.

2. KELLIA Linked Data Products

Online Coptic Dictionary: Entries in the online dictionary (http://coptic-dictionary.org) are linked to Coptic SCRIPTORIUM digital textual data (described in more detail in the Grant Products section). The same linking opportunities are available to KELLIA partners and other Coptic-language projects, who can freely annotate their digital texts with links to online dictionary lemma searches.

Coptic Treebank: The automatic syntactic annotation lays the groundwork for linking data about entities in Coptic texts, as described below. The treebank standards themselves are described in more detail in the Grant Products section.

Entity recognition: Proof-of-concept machine-processed entity-recognition (described in more detail in the Grant Products section) will allow for entity disambiguation and linking opportunities for Coptic literature to named entities projects such as Pelagios, Pleiades, Trismegistos, PATHs.

VMR-Coptic SCRIPTORIUM Converter: This converter (described in more detail in the Grant Products section) transforms digital text data produced by the Coptic Old Testament project's Virtual Manuscript Room to the EpiDoc TEI XML format used by Coptic SCRIPTORIUM and pushes this XML text to Coptic SCRIPTORIUM's natural language processing pipeline. This converter facilitates the integration of digital text corpora developing in both projects.

3. Recommended standards

The KELLIA project recommends the following linked data standards, described in more detail below.

- Projects should adopt and publicly document controlled vocabularies for use in metadata, thus ensuring data integrity and the possibility of linking via API (section 3a)

- Projects should provide automated metadata validation whenever possible (section 3a)
- Projects should include metadata fields for collaborators' unique identifiers (section 3a)
- Projects should assign persistent identifiers to the digital objects they produce (section 3b)
- Projects should include clear licensing information to facilitate data exchange (section 3b)

3a Standards for controlled vocabulary

In order to facilitate retrieving, linking, and exchanging data, KELLIA partners have agreed upon recommended metadata and data structures, as outlined in the metadata standards appendix of this grant report. We recommend all digital Coptic projects establish internal controlled vocabularies for metadata. Such vocabularies provide internal standards for populating metadata fields to ensure data integrity. Additionally, in the absence of a robust hub for linked data (akin to Pelagios or Trismegistos), controlled vocabularies lay the groundwork for linking via API's or queries rendered in http URI syntax. (E.g., in Coptic Scriptorium's environment, the query http://data.copticscriptorium.org/filter/author=Shenoute retrieves all documents where the "author" metadatum is "Shenoute", and http://data.copticscriptorium.org/filter/collection=Borgia%20Collection retrieves all documents where the "collection metadatum is "Borgia Collection"). Controlled vocabularies accessed through URI based queries can enable a project to achieve 4-star level linked data. Well-documented internal standards can also help to facilitate the exchange of data, as conversions can be applied to communicate across different project standards. Standards thus lay the groundwork for a more robust linking environment for Coptic data in the future (such as Pelagios is for ancient geography).

To date, standard controlled vocabularies (such as Dublin Core, Getty, or the Europeana EAGLE Vocabularies[74]) do not provide the coverage needed for Coptic literary materials, and there is no comprehensive standard for controlled vocabulary for fields such as places, names, document identification, etc. in digital Coptic. Despite this lack of standard, we recommend projects develop internal standards, and the metadata guidelines in this White Paper provide models. We also recommend an automated process for metadata validation to ensure adherence to internal standards whenever possible. Our current annotation editor, the open source online XML editor GitDox, provides automatic checking of metadata field names and values, ensuring that required

[74] https://www.eagle-network.eu/resources/vocabularies/

metadata appears and matches regular expression patterns specifying valid values across our documents. Likewise, whenever possible, projects should include unique identifiers assigned to metadata fields by other projects. KELLIA partners should therefore include a field for CoptOT/Coptic SCRIPTORIUM identifiers in metadata.

The Coptic Old Testament project has created and shared a list of standard names for institutions that have Coptic manuscripts in order to standardize their metadata. Their list can be found here: (https://docs.google.com/spreadsheets/d/1jDeRLhM9tlG9pzkUWbyoi7CD2_L4yLdDM1yJ18ToT9A/edit#gid=0). It includes references to the respective entries in the Trismegistos "Collections" table http://www.trismegistos.org/coll/index.php Such standards will make the discovery and exchange of data across projects easier.

3b Standards for unique identifiers

As mentioned in the recommendations for data curation, digital objects created by each projects should be assigned persistent identifiers, such as Digital Object Identifiers or other Unique Resource Identifiers that are accessible via a URL (1.3.3). Section on KELLIA current practices. Coptic SCRIPTORIUM assigns a unique document title to each document as well as CTS (Canonical Text Service) URNs (http://www.homermultitext.org/hmt-doc/cite/cts-urn-overview.html) for the digital data in their corpora. The URN schema includes two namespaces created specifically for Coptic materials: copticLit for literary materials and copticDoc for documentary materials. Coptic SCRIPTORIUM provides a service (data.copticscriptorium.org) to resolve these URNs, and they can also be used in the ANNIS search and visualization interface.

Finally, as mentioned in the recommendations for data curation standards, all projects should have clear licensing attached to all products in order to facilitate data sharing and linking (1.3.4). Ideally, this information should both be included in the metadata associated with each digital object and be evident or easily findable through any web-based interface used to display the object.

4 Future plans in Coptic linked data

KELLIA partners look forward to continuing to share data and standards, as well as collaborating with other Coptic and ancient studies projects. We have already begun to work with Pelagios and Pleiades to link our geographic data and make it discoverable to a wider scholarly and popular audience. We are especially excited about working with the PAThs project in the future, as their proposed interlinked databases will provide an

opportunity for other Coptic projects to review data and metadata standards in relation to their work.