

KELLIA White Paper on Transcription and Encoding Standards for Digital Coptic


By Caroline T. Schroeder, Ulrich Schmid, So Miyagawa, Elizabeth Platte, Amir Zeldes

From the Koptische/Coptic Electronic Language and Literature International Alliance (KELLIA) Project (<https://kellia.uni-goettingen.de>)

Product of a joint Grant from the National Endowment for the Humanities (HG-229371) and Deutsche Forschungsgemeinschaft (BE 4172/1-1)

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)

The KELLIA White Paper on Transcription and Encoding Standards for Digital Coptic by [Caroline T. Schroeder, Ulrich Schmid, So Miyagawa, Elizabeth Platte, Amir Zeldes](#) is licensed under a [Creative Commons Attribution-NonCommercial 4.0 International License](#). 

27 March 2019

Transcription and Encoding Standards for Digital Coptic

Members of the Advisory Boards for the Coptic Scriptorium project and the Digital Edition of the Coptic Old Testament project as well as project participants have reviewed this document prior to publication. This White Paper is also attached to the main KELLIA project White Paper

(<https://kellia.uni-goettingen.de/downloads/KELLIA-white-paper.pdf>) as Appendix 3.

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.¹

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

1

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/>

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.² It also visualizes combining characters (such as supra-linear strokes). The font created by the Institut français d'archéologie orientale uses Private Usage Area character encodings and is not recommended.³ The New Athena Unicode font is an alternative to Antinoou.⁴ 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.⁵

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

² <https://www.evertype.com/fonts/coptic/>

³ <http://www.ifao.egnet.net/publications/publier/outils-ed/polices/>

⁴ <https://apagreekkeys.org/NAUdownload.html>

⁵ <https://github.com/CopticScriptorium/converters>, <https://github.com/paths-erc/cmcl2unicode> with demo at <http://paths.uniroma1.it/cmcl2unicode/index.html>.

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**⁶ should use word segmentation that follows principles of binding and segmentation that conform to the linguistic principles of Bentley Layton's *Coptic Grammar*.⁷ 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 Coptic⁸ 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 protocols⁹ 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.

⁶ 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/>.

⁷ Bentley Layton, *A Coptic Grammar*, 3rd Edition, *Porta Linguarum Orientalium Neue Serie* 20 (Wiesbaden: Harrassowitz, 2011).

⁸ Walter C. Till, “La Séparation des Mots en Copte,” *Bulletin de l'Institut Français d'archéologie Orientale* 60 (1960): 151–70.

⁹ 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://copticcriptorium.org/documentation>.

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.

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)¹⁰. TEI XML counts as the most used and broadly accepted standard to describe textual phenomena.¹¹ 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.¹² Since Coptology is confronted with handwritten manuscripts not conforming the modern concept of typographic textuality and its implications, EpiDoc¹³ is recommended to encode Coptic text-bearing objects or as a data model.

¹⁰ <http://www.tei-c.org/>

¹¹ <http://www.tei-c.org/Guidelines/Customization/>

¹² <https://sourceforge.net/p/epidoc/wiki/Home/>

¹³ <http://www.tei-c.org/Guidelines/P5/>

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-Schema¹⁴, Relax NG (RNG)¹⁵ or Schematron¹⁶ to guarantee standardised and valid XML-data. An ODD Format (“One Document does it all”)¹⁷ can be used. It brings together the Documentation of Tags and the formal declaration which can be compiled¹⁸ 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.¹⁹

¹⁴ <https://www.w3.org/XML/Schema>

¹⁵ <http://relaxng.org/>

¹⁶ <http://schematron.com/>

¹⁷ 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>

¹⁸ Using the Tool Roma provided by TEI on <http://www.tei-c.org/Roma/>

¹⁹ <https://kellia.uni-goettingen.de/downloads/KELLIA-metadata-white-paper.pdf>, <https://kellia.uni-goettingen.de/downloads/KELLIA-white-paper.pdf>

9. Visualizing and Publishing Encoded Text Data

TEI-C provides several stylesheets (XSL) to convert xml files into various file formats including html.²⁰ It is recommended that projects use existing stylesheets and amend them where necessary to ensure the proper display of all of their encoded data.

For visualisation of linguistic text data, ANNIS is recommended.²¹ 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 Salt N Pepper or Exmeralda, one can convert various XML file formats including TEI XML into PAULA XML.

²⁰ <https://github.com/TEIC/Stylesheets/tree/dev/html>

²¹ <http://corpus-tools.org/annis/>