Coptic SCRIPTORIUM Diplomatic Transcription Guidelines

Version: 1.2_2016.8.26

Caroline T. Schroeder[1] & Amir Zeldes[2]

1. University of the Pacific
2. Georgetown University

1. Preamble

This document details guidelines for transcribing a diplomatic edition of a manuscript in Sahidic Coptic according to the Coptic SCRIPTORIUM project scheme. The diplomatic transcription currently requires extensive manual annotation, due to the complexities of processing a diplomatic text in which no word breaks exist in the original and yet words and even morphemes span across line, column, and page breaks.

The transcription procedure assumes familiarity with basic paleography and traditional manuscript transcription following the Leiden conventions. (http://www.stoa.org/epidoc/gl/latest/app-glossary.html#leiden)

The diplomatic transcription also utilizes XML (eXtensible Markup Language) -like tagsets, including some of the TEI (Text Encoding Initiative) XML markup language, although the resulting document is **not** a valid XML document. Wherever possible, the EpiDoc subset of TEI XML is utilized for element nomenclature. EpiDoc TEI conventions were created by and for epigraphers and have come to be a standard in markup of ancient texts, epigraphic or otherwise. (http://sourceforge.net/p/epidoc/wiki/Home/) In contrast to TEI, SCRIPTORIUM utilizes no milestone XML tags (e.g., <cb/>). Instead, all tags are span annotations (e.g., <cb>This is a column of Coptic text.</cb>).

We recommend using an XML editor such as Oxygen to ensure the encoding is well-formed and well-structured.

The aim is twofold: 1) to achieve a transcription that documents the text and visualization of the manuscript as closely as possible to the original; 2) to provide a text file that can be processed by various digital tools and software, such as a tokenizer, a part-of-speech tagger, or the ANNIS database infrastructure (http://www.sfb632.uni-potsdam.de/annis/; Zeldes et al. 2009). Coptic SCRIPTORIUM has bundled some of these tools in a Natural Language Processing web service.

The resulting transcription itself does not resemble a traditional text of a diplomatic edition. The markup ensures optimization for processing and search using such tools and software. For examples of the diplomatic editions visualized in HTML generated from the post-ANNIS transformations, see corpora at data.copticscriptorium.org. Valid EpiDoc TEI XML versions of the documents are also provided from this site.

2. Character Encoding

Texts are encoded using the UTF-8 (Unicode) Coptic language character set. The freely available Antinoou font and Coptic-English keyboard created by Michael Everson in cooperation with the International Association of Coptic Studies is the standard (http://www.evertype.com/fonts/coptic/). Unicode characters in the private use area are not recommended.

2.1 Alphanumeric Characters

Characters follow the orthography of the manuscript.

Mark oversize characters with XML tagging. Do not use uppercase version of the character.

2.2 Punctuation and Decoration

Punctuation and decoration follows the manuscript as closely as possible within the Unicode character set. Not all decoration and punctuation can be encoded using characters; deviations or documentation that can't be keyed in is instead typically indicated in a note element.

Notes on individual specific punctuation characters:

> For the character ` that occasionally appears at the end of words in some manuscripts, use U+2CFF. Example:
>
> ⲡⲉⲙⲙⲟⲛ̀ⲧⲉ

2.3 Accentuation and Supralinear Strokes

Accentuation and supralinear strokes follow the orthography of the manuscript. Some manuscripts have binding strokes between letters (e.g. ϩⲛ̄) whereas others in the case of the same word might only provide a stroke over a single letter (e.g., ϩⲛ̄). The diplomatic transcription follows the conventions of the manuscript, even if the manuscript is internally inconsistent or contains what seem to be errors.

Notes on encoding individual specific accents, strokes, etc, using the Coptic-English keyboard for Antinoou (for MacIntosh):

> ◌̄ (as in ⲙ̄) the supralinear stroke above only one letter: type the letter followed by Unicode U+0304 (; on keyboard)
>
> ◌͞◌ (as in ⲙ͞ⲛ) the binding stroke between two letters: type first letter then U+FE24 (< in the Coptic-English keyboard) then second letter then U+FE25 (> in the Coptic-English keyboard), i.e. m<n> on a Mac using the Coptic-English keyboard
>
> ◌͞◌͞◌ (as in ⲙ͞ⲛ͞ⲧ) binding stroke over three letters: type the first letter then U+FE24 (< on a Mac using the Coptic-English keyboard) then second letter then U+FE26 (: [i.e. shift+;] on a Mac using the Coptic-English keyboard) then third letter then U+FE25 (> on a Mmac using the Coptic-English keyboard), i.e. m<n:t>)

◌͂ (as in o͂ⲩ) circumflex combining two letters: U+1DCD (keystroke shift+option+/ on a Mac using the Coptic-English keyboard) typed between the letters, so oi͂u (o then shift+option+/ then u)

For squiggly curved or jagged strokes over etas, use a regular circumflex rather than a dot or line or trema (◌̂): type the letter followed by U+0302 (option+3 on the keyboard)

Tremas (ï, ⲏ̈): type the letter followed by U+0308 (option+7 on the keyboard)

3. Text Divisions

A diplomatic transcription aims to preserve the formatting of the original text. Line breaks, column breaks, and page breaks as they appear in the manuscript are all documented.

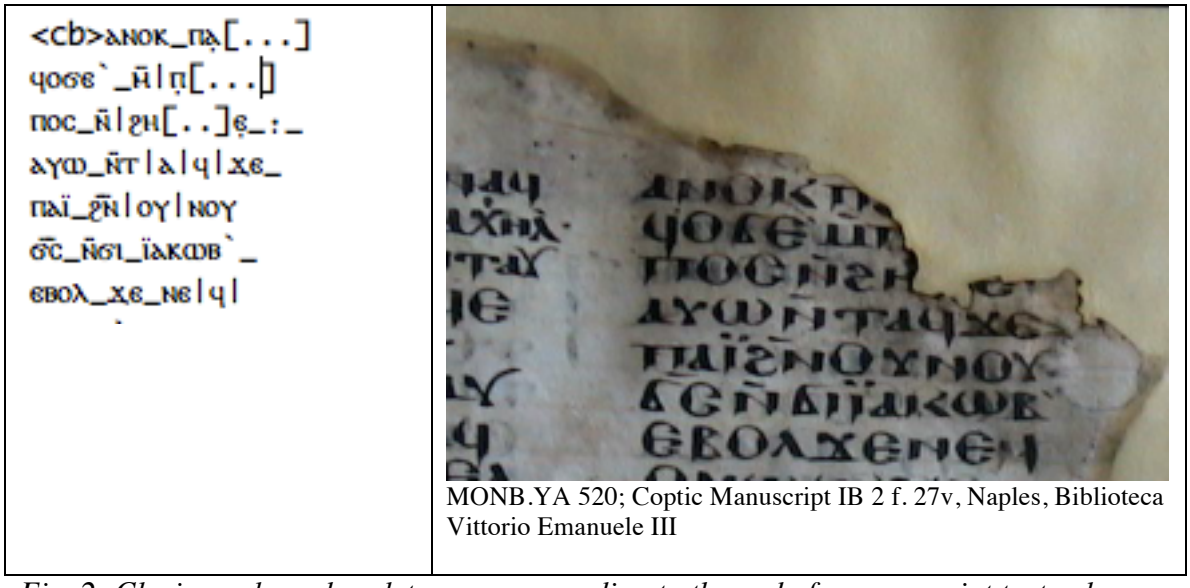
3.1 Line Breaks

All line breaks in the transcription should follow the line breaks of the manuscript. Editors may manually encode line breaks using the tags <lb></lb>. However, if you plan to use the Coptic NLP web service to further annotate your text, you may use the "Enter" or "Return" key to produce a line break in the text file of the transcription. Selecting the option for "meaningful line breaks" in the NLP web service will insert encoding for the line breaks.

3.2 Column Breaks

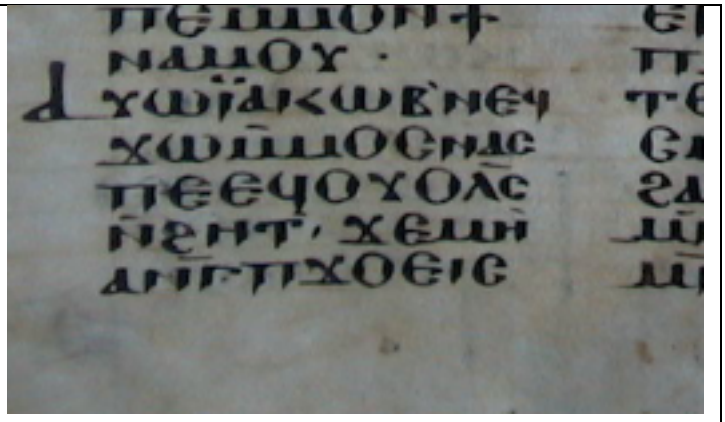
All column breaks in the transcription should follow the column divisions in the manuscript. Columns are wrapped in span annotations using the <cb></cb> tagset.

Fig. 1: Opening column break tag, corresponding to beginning of manuscript column



MONB.YA 520; Coptic Manuscript IB 2 f. 27v, Naples, Biblioteca Vittorio Emanuele III

Fig. 2: Closing column break tag, corresponding to the end of a manuscript text column

ⲁⲙⲏⲣⲉ̀_ⲉ⧸ⲁⲙⲱ ⲡⲉ_ⲛ̄ⲛⲟ⧸ⲛ̀_☥⧸ ⲛⲁ⧸ⲛⲟⲩ_·_ ⲁⲩⲱ_ⲓⲁⲕⲱⲃ̀_ⲛⲉ⧸ⲩ⧸ ⲭⲱ_ⲛ̄ⲛⲟ⧸ⲥ_ⲛⲁ⧸ⲥ_ ⲡⲉ_ⲉ⧸ⲩ⧸ⲟⲩⲟⲗ̄ⲥ_ ⲛ̄⧸ⲯⲏⲧ_·_ⲭⲉ_ⲙⲏ̀_ ⲁⲛ̄ⲧ⧸ⲡ⧸ⲭⲟⲉⲓⲥ_</cb>	 MONB.YA 520; Coptic Manuscript IB 2 f. 27v, Naples, Biblioteca Vittorio Emanuele III,"

3.3 Page Breaks and Numbering

All page breaks in the transcription should follow the page divisions in the manuscript. Page numbering in the transcription reflects the page numbering in the original manuscript codex. Codex sigla in the example below are two-letter codes following the White Monastery codex siglum list created by Tito Orlandi (Orlandi 2002; also http://www.cmcl.it/). Page breaks are wrapped in TEI compatible span annotations using the <pb></pb> tagset with the xml:id element. The entire page of text (including the relevant column tags) should be wrapped with these tags. Thus <pb xml:id="YA518"> is the opening tag for page 518 in White Monastery codex YA (MONB.YA). The xml:id should not contain spaces. (Thus, xml:id="YA518" not xml:id="YA 518">.)

Fig. 3: Closing and opening page break tags indicating the end of one page and beginning of the next. (Note: the opening tag for the first page and closing tag for the second page are not visible here but are required.)

```
ⲑⲛⲟⲥ̀_ⲉ⧸ⲛⲁ⧸ⲁⲙⲱ⧸ⲩ_·_
ⲁⲩⲱ_ⲡ⧸ⲉⲧ⧸ⲛⲏⲩ_
ⲉⲃⲟⲗ_ⲉ̄ⲛ_ⲛⲉⲕ̀_·_</cb></note></pb>
<pb xml:id="YA519"><note note="page number φιθ
barely visible in upper right"><cb>[....]ⲉ_·_ⲓⲥⲁⲁⲕ̀
[...]ⲩ_ⲡ⧸ⲉⲧ⧸ⲛⲁ⧸
[ⲕⲗⲏ]ⲣⲟⲛⲟⲏⲉⲓ_
ⲛ̄ⲛⲟ⧸ⲕ_·⎣
ⲑⲧⲃⲉ_ⲡⲁⲓ̈_ⲣⲱ̄_ⲁ⧸ⲥ⧸
ⲭⲟⲟ⧸ⲥ_ⲉ̄ⲛ⧸ⲟⲩ⧸
```

The location and Coptic numeration of the page number is currently documented in a note element. (See Figure 3 above).

4. Word Segmentation, Spacing, and Tokenization

Sahidic Coptic bound groups are formed by several words and/or morphemes attaching together. A word refers to one noun, preposition, article, etc. One complex word can be comprised of multiple morphemes, including affixes such as ⲁⲧ, ⲙⲛⲧ, or ⲣⲉⲩ, or compound words, such as complex numbers (e.g. -teens) and verbs formed with ⲣ. One bound group may include multiple prepositions and objects, or a verbal auxiliary + subject + infinitive, or even more words and morphemes strung together (generally speaking clitics). The copula, which some might consider a clitic, remains unbound. Coptic SCRIPTORIUM follows the practices in Bentley Layton's grammar (Layton 2011) for word, morpheme, and bound group segmentation.

> *Examples of individual words comprised of one morpheme:*
> ⲥⲱⲧⲙ
> ⲛⲟⲃⲉ
> ϩⲏⲧ
>
> *Examples of individual words comprised of multiple morphemes:*
> ⲙⲛⲧⲁⲧⲥⲱⲧⲙ
> ⲣⲉⲩⲣⲛⲟⲃⲉ
>
> *Examples of bound groups comprised of words with multiple morphemes:*
> ⲧⲙⲛⲧⲁⲧⲥⲱⲧⲙ
> ⲙⲡⲣⲉⲩⲣⲛⲟⲃⲉ
> ⲡⲣⲙⲛϩⲏⲧ
> ϫⲉⲛⲧⲁⲩϫⲓⲧⲥ
>
> Note: if a project wishes to annotate on the morpheme level (i.e. internal analysis of units like ⲙⲛⲧ) and not just on the *word* level, the morphemes need to be tokenized. Coptic SCRIPTORIUM annotates on the word level and then provides additional annotation on the morpheme level for compound words and words with affixes. (See section 4.4 for more information.)

In most manuscripts, no spaces between words or bound groups are provided. Sometimes a diacritical mark, such as ̀ does appear, but word segmentation following diacritics and punctuation does not always correspond with contemporary segmentation practices (such as Layton or Till (1960)). More study of this marking is required.

4.1 Word Segmentation

SCRIPTORIUM diplomatic transcription marks word segmentations according to Layton's conventions (Layton 2011). The transcriber inserts a unique character, such as an underscore ("_"), after each Coptic bound group, even when the end of the bound group falls at the end of a line.

Likewise, all punctuation is followed by an underscore.

(1) ⲉⲧⲉⲓⲥⲙⲁⲛⲗ_ (word ends at end of line)
(2) ⲡⲉ_ⲛⲩⲛⲁⲕⲗⲏ (two words, in which the second bound group flows into line 3)
(3) ⲣⲟⲛⲟⲙⲉⲓ_ⲙ (the bound group continues from line 2, is followed by an underscore)
(4) ⲙⲟⲕ_ⲁⲛ_·_ (punctuation followed by an underscore)

These underscores are not and do not need to be visualized in HTML transformations of the diplomatic editions; they are nonetheless essential for processing the text, since they demark breaks between bound groups and will enable searches and visualizations of a word-segmented text.

We do not recommend using spaces to demarcate bound groups and punctuation, since spaces may occur elsewhere in the document (such as inside XML tags), and lead to confusion during automatic processing.

4.2 Spacing

Encoding of blank space is preferred to using the space key. The encoding should match spaces in the manuscript. Consequently, if the manuscript provides no spaces between words or punctuation, the diplomatic transcription contains no spaces. Where there are significant spaces in the manuscript that the transcriber wishes to draw attention to, the transcription should encode a space using TEI XML tags in order to visualize the white space in the manuscript. Encode the word, morpheme, or punctuation next to the white space, as in these examples:

(1) <hi rend="1_space_right">·</hi> will visualize one space to the right of the ·
(2) n̄<hi rend="1_space_right">ⲃⲟⲛⲑⲟⲥ</hi> will visualize one space to the right of he n t
(3) ϫⲱ<hi rend="2_space_right">ⲓ̈</hi>_<hi rend="1_space_right">·_</hi>_ⲉⲧⲃⲉ_ will visualize two spaces to the right of ï and one space to the right of the ·

It is important to make sure that attributes are surrounded by straight, not curly quotes (i.e. " on both sides).

4.3 Tokenization of Words

If one wishes to manually segment bound groups into words, one can do so using the pipe character ("|").

(1) ϩⲓⲧⲙ|ⲡ|ⲛⲟⲩⲧⲉ_ (preposition|article|noun)
(2) ⲉⲧⲉ|ⲓ̈ⲥⲙⲁⲛⲗ_ (converter|noun)
(3) ⲡⲉ̀_ⲛ|ⲩ|ⲛⲁ|ⲕⲗⲏ (word_auxiliary|subject pronoun|future marker|verb (verb continues to line 4)
(4) ⲣⲟⲛⲟⲙⲉⲓ̀_ⲙ̄

The NLP web service contains a tokenizer that will take as input bound groups and provide as output word segmentation with pipes. Coptic SCRIPTORIUM's standalone tokenizer tool will do the same.

4.4 Tokenizing and Annotating Morphemes below the Word Level

To conduct research on the morpheme level in compound words or other words that contain multiple morphemes, the words will need to be tokenized and annotated below

the word level and on the morpheme level. In Coptic SCRIPTORIUM, text is annotated on the word level for the part of speech (see SCRIPTORIUM Part-of-Speech Tagsets for Sahidic Coptic) and other characteristics, such as language of origin. Tokenizing and annotating on the morpheme level allows for additional search, visualization, and research capabilities.

Examples of individual words comprised of multiple morphemes, tokenized on the morpheme level:

word	ⲙⲛⲧⲁⲧⲥⲱⲧⲙ		
morpheme	ⲙⲛⲧ	ⲁⲧ	ⲥⲱⲧⲙ

word	ⲣⲉⲩⲣⲛⲟⲃⲉ		
morpheme	ⲣⲉⲩ	ⲣ	ⲛⲟⲃⲉ

Examples of bound groups comprised of words with multiple morphemes:

bound group	ⲧⲙⲛⲧⲁⲧⲥⲱⲧⲙ			
word	ⲧ	ⲙⲛⲧⲁⲧⲥⲱⲧⲙ		
morpheme	ⲧ	ⲙⲛⲧ	ⲁⲧ	ⲥⲱⲧⲙ

bound group	ⲙⲡⲣⲉⲩⲣⲛⲟⲃⲉ				
word	ⲙ	ⲡ	ⲣⲉⲩⲣⲛⲟⲃⲉ		
morpheme	ⲙ	ⲡ	ⲣⲉⲩ	ⲣ	ⲛⲟⲃⲉ

bound group	ⲡⲣⲙⲛϩⲏⲧ			
word	ⲡ	ⲣⲙⲛϩⲏⲧ		
morpheme	ⲡ	ⲣⲙ	ⲛ	ϩⲏⲧ

Compound words that involve an article or affixed personal pronoun to the second item of the compound typically are tokenized as bound groups comprised of multiple words, not as one word comprised of multiple morphemes.

Examples of bound groups containing compound words with articles or pronouns on the second unit of the compound:

bound group/compound	ⲣⲯⲛⲁⲩ		
word	ⲣ	ⲯⲛⲁ	ⲩ
no tokenization & annotation on the morpheme level below the word level			

bound group	ⲙⲡⲉⲧⲛⲣⲡⲙⲉⲉⲩⲉ				
word	ⲙⲡⲉ	ⲧⲛ	ⲣ	ⲡ	ⲙⲉⲉⲩⲉ
no tokenization & annotation on the morpheme level below the word level					

(where ⲣⲡⲙⲉⲉⲩⲉ *is considered to contain multiple words, not morphemes below one word level)*

[Note: the part-of-speech tagger developed by Coptic SCRIPTORIUM operates on the *word* level, not the sub-word morpheme level. So, ⲣⲯⲟⲧⲉ is tagged as one V, ⲙⲛⲧⲁⲧⲥⲱⲧⲙ as one N, etc.]

Transcription conventions for segmenting morphs should utilize a unique character, such as a dash or hyphen. E.g.:

ⲧ|ⲙⲛⲧ-ⲁⲧ-ⲥⲱⲧⲙ

ⲙ|ⲡ|ⲣⲉⲩ-ⲣ-ⲛⲟⲃⲉ

If you plan to use Coptic SCRIPTORIUM's NLP web service, you may transcribe the Coptic in bound groups with no pipes or morphemes. The NLP web service's tokenizer can provide as output segmentation with pipes between words and dashes between morphs. Likewise, Coptic SCRIPTORIUM's stand-alone tokenizer can output words with segmented morphs. The webservice can further automatically annotate the segmented words and morphs for part of speech, language of origin, and lemma.

5. Rendering and Leiden Transcription Conventions

Coptic SCRIPTORIUM uses Leiden and Leiden+ conventions for transcribing manuscripts. The encoding follows the EpiDoc guidelines. Not all Leiden documentation is currently XML encoded as Leiden+, however.

5.1 Characters Highlighted, Raised, Lowered, or Set Apart in Some Way

Characters that are raised, lowered, or printed in different colors or styles are encoded using the TEI XML element <hi> with the rend attribute. Letters written above the line are encoded: <hi rend="superscript">. Characters written below the line are encoded: <hi rend="subscript">. Letters in a different color ink are encoded with the color ink, e.g., <hi rend="red">. It is possible to combine these annotations, e.g. <hi rend="red subscript">. Coptic SCRIPTORIUM currently encodes large, tall (the letter stretches above the line), long (letter stretches below the line), thin, superscript, subscript, and colors. Any additional information can be provided in a note element. To encode two attributes, use a space (not a comma) between the two attributes.

Example

ⲉⲓⲧⲙ̄ⲡⲛ<hi rend="superscript"><note note="o is directly above the ⲩ">o</note></hi>ⲩ ⲠⲠⲑⲦⲚⲁⲚⲞ<hi rend="large">ⲩ</hi><hi rend= "long thin">ⲯ</hi>_._

Diplomatic Visualization

ⲉⲓⲧⲙ̄ⲡⲛ^{o}ⲩ (ANNIS) or ⲉⲓⲧⲙ̄ⲡⲛ\o/ⲩ (EpiDoc XSLT)

ⲠⲠⲑⲦⲚⲁⲚⲞⲩ ⲯ.

Other encodings are colors (red, brown, green, etc.) "ekthetic" should be used for characters that are part of the ongoing text but written to the left of the margin line. See below, in which the Ⲡ is encoded <hi rend="red large ekthetic">Ⲡ</hi>

ⲉ̀ⲧⲁϩⲏⲕⲉ̀ⲃⲟⲗ·¬ⲛ̀ⲧⲉ

5 ⲡⲉⲑⲃⲃⲓⲟⲛ̀ ⲉⲏⲧ·¬

Ⲡⲁⲓⲇⲉⲁⲭⲁⲱⲛⲟⲩⲱⲁ

hi@rend cannot contain more than five words as per Epidoc guidelines and may contain only alphanumeric characters. (No punctuation. So <hi rend= "long, thin">ⲯ</hi> is invalid.)

5.2 Damaged Characters

Characters that are damaged but restored based on context are marked with an underdot. Coptic SCRIPTORIUM uses the diacritical character ̣ (Unicode U+0323). These characters are not currently encoded in TEI XML using the EpiDoc tagset for Leiden+. Coptic SCRIPTORIUM uses the underdot character rather than annotation to designate this information.

5.3 Lacunae and Lost Characters

Lost lines and characters (lacunae) are indicated using square brackets, as in the Leiden conventions. They may be encoded using the EpiDoc tagset, but it is not required. See EpiDoc guidelines for more details ("EpiDoc Guidelines: Lost Characters, Quantity Unknown"; "EpiDoc Guidelines: Editorial Restoration: Characters Lost but Restored by Modern Editor"; "EpiDoc Guidelines: Lost Characters, Quantity Approximate"; "EpiDoc Guidelines: Lost Characters, Quantity Known"; "EpiDoc Guidelines: Erased and Lost"; "EpiDoc Guidelines: Lacunas, Other Units").

(1) Example encoded using the gap element:
<gap reason="lost">
[ ]_
[ ]_
[ ]_
</gap>
(2) Unencoded gaps (no XML elements):
[.....]ⲡⲣ[..]

5.4 Other

Other rendering information is encoded either according to EpiDoc conventions or recorded as information within a note element. See the cheatsheet for Leiden+ conventions in EpiDoc at https://sourceforge.net/p/epidoc/code/HEAD/tree/trunk/guidelines/msword/cheatsheet.doc?format=raw and http://papyri.info/docs/leiden_plus. See also the full list of text transcription guidelines here http://www.stoa.org/epidoc/gl/latest/app-alltrans.html.

Transcribing in Oxygen or a similar XML editor is recommended, to ensure tags are well-structured.

6.0 File Format and Document Preferences

Documents are transcribed in a text editor such as TextEdit. Document preferences are set to UTF-8 encoding without byte-order Mark (BOM). (E.g., in TextEdit 1.7.1 for MacIntosh, in the File-->Preferences menu, click on "Open and Save," and select "Unicode (UTF-8)" for Opening files and Saving files.)

Bibliography

An up-to-date bibliography can be found at the project's Zotero page: https://www.zotero.org/groups/coptic_SCRIPTORIUM/items/collectionKey/8IHTW3NZ

Bodard, Gabriel. "EpiDoc Appendix: Glossary: Leiden, Leiden-plus." *Appendix: Glossary*. 18 Jun. 2013. <http://www.stoa.org/epidoc/gl/latest/app-glossary.html#leiden>.

"Corpus Dei Manoscritti Copti Letterari." *CMCL - Studies in Coptic Civilization*. 11 Sep. 2012. <http://cmcl.aai.uni-hamburg.de/>.

"EpiDoc Guidelines." *EpiDoc Guidelines*. 25 May 2013. <http://www.stoa.org/epidoc/gl/dev/>.

---. *EpiDoc: Epigraphic Documents in TEI XML*. 25 May 2013. <http://sourceforge.net/p/epidoc/wiki/Home/>.

"Evertype: Antinoou." *Evertype: Antinoou - A Standard Font for Coptic* 2012. 29 May 2013. <http://www.evertype.com/fonts/coptic/>.

Layton, Bentley. *A Coptic Grammar*. 3rd Edition, Revised. Wiesbaden: Harrassowitz, 2011. Print.

Orlandi, Tito. "The Library of the Monastery of Saint Shenute at Atripe." *Perspectives on Panopolis: An Egyptian Town from Alexander the Great to the Arab Conquest*. Leiden: Brill, 2002. 211–231. Print.

Till, Walter C. "La séparation des mots en Copte." *Bulletin de l'Institut français d'archéologie orientale* 60 (1960): 151–70.

Zeldes, Amir, Ritz, Julia, Lüdeling, Anke & Chiarcos, Christian "ANNIS: A Search Tool for Multi-Layer Annotated Corpora." *Proceedings of Corpus Linguistics 2009* (2009) : n. pag. 10 Sep. 2012. <http://ucrel.lancs.ac.uk/publications/cl2009/>.