Coptic SCRIPTORIUM Cheat sheet for commonly used EpiDoc XML tags for manuscript transcription

Transcription for a digital corpus combines traditional transcription methods with standardized coding, called “tagsets.” This will produce a document that the Natural Language Processor (NLP) on CopticScriptorium.org will be able to process and annotate to normalize spelling, diacritics and tag for part of speech, loan words, and lemmas.

When you transcribe, you will be adding your own annotations for manuscript information (column breaks, page breaks and page numbers, ornamented characters

A full description of our diplomatic transcription guidelines is available on the website. For the purposes of this workshop, we have prioritized the most common and useful tagsets to include in our on-line transcription editor.

IMPORTANT: All tagsets must have at least one letter or punctuation mark between the tags. All tagsets should be nested (e.g., <pb><cb>columnofCoptictext</cb></pb> not <cb><pb>columnofCoptictext</cb></pb>)

Page breaks — <pb> at the beginning of the page and </pb> to close the page

Column breaks — <cb n=”1”> to begin the first column and </cb> to end the column; use <cb n=”2”></cb> for the second column

Character ornaments:

Use <hi> </hi> (for “highlighting”) with an attribute added to specify the kind of highlighting, usually rend (for rendering). E.g. type <hi rend=”ekthetic”></hi> wrapped around the ekthetic charact(ers). Transcriptions can combine two attributes which should be entered with a space, no punctuation, between them.

<hi rend=”[attribute]”></hi>

1. Outside the left margin: rend=”ekethetic”
2. Large letters: hi rend=”large”
2. Ekthetic & large: hi rend=”ekthetic large”
3. Letters that are above the line: rend=”superscript”
4. Letters that are below the line: rend=”subscript”
5. Letters that stretch tall above the line: rend=”tall”
6. Letters that stretch long below the line: rend=”long”
7. Letters that appear in red ink: rend=”red” (or brown, green, etc.)

Additional notations: Use a note for decorations in the margins or other information that might be difficult to tag. Wrap the tags around the nearest text.

<note note="there is a drawing of a bird below the column"></note>

Lacunae, damage:

If you find lacunae or damaged characters, transcribe them using the usual conventions (underdot for partially visible/ambiguous character, square brackets for missing and reconstructed characters, etc.) Here are some common tags (taken from the papyri.info cheat sheet):

Letters ambiguous outside of their context:	ⲁⲭⲥⲫⲧⲙ (no tags, underdots only)
Vestiges of letters visible but illegible	<gap reason="illegible" quantity="3" unit="character"/>[...]</gap>
Characters lost but restored	<supplied reason="lost">[ⲁⲭⲥ]</supplied>ⲱⲧⲙ
Characters lost (lacuna such as ms tear)	<gap reason="lost" quantity="3" unit="character">[...]</gap>

Note: for papyri.info editors, we use very similar tags and conventions, but there are some differences due to our multilayer annotation format.