Specifying XML source for Codex Sinaiticus

Revised for website release of XML transcription 11.11.2011.

Header and file structure

As the file is not fully TEI-compatible, the transcription is enclosed within <egXML> tags, in order to enable it to be viewed in web-browsers and other automatically-parsing software without the generation of error messages.

A standard TEI header was supplied following the agreement of the Project Board on 28.10.2010.

*Layout: a) by *

** ... ** encloses the whole transcript file.
** ... ** encloses each individual book.
** ... ** encloses each individual chapter.
<ab id="V-...">...</ab> encloses each verse.
<w>...</w> encloses each word.
All of these are numbered apart from .

Layout: b) by page

<pb>...</pb> identifies each page.
Pages are identified by quire number, then page number, so 34-8r is Quire 34, Page 8 recto.
Each <pb> includes the following identifiers:

- **scribeid="n"** for the copyist of the page. This may also include information about overwriting of the page by a later scribe.
- **archive="n"** for the holding library (BL / LUL / SC / NLR)
- **localfol="n"** for the folio number assigned to the page by the holding library

<cb>...</cb> identifies each column.
Columns are identified by quire number, page number and column number, e.g. 34-8r-4
<lb>...</lb> identifies each line.
Lines are identified using the cumulative system described, e.g. 34-8r-4-23.
Lines may be positioned in the following ways:

- **<lb rend="indent">...</lb>** line indented to the right
- **<lb rend="indentextra">...</lb>** line indented to the right by twice the usual width (in books with two columns)
- **<lb rend="hang">...</lb>** line overhanging to the left
- **<lb rend="center">...</lb>** line centre-justified

Each page has four (most commonly), two (quite common) or one (rarely) columns.
Some pages are fragmentary and blank columns may be added for display purposes.

The start and end of each page, column and line are marked by break elements linked by **id** and **corres** attributes. Thus
<cb id="S-34-8r-1" corres="E-34-8r-1" n="1"/>
indicates the start (S) of the first column of page 8r of quire 34, and
<cb id="E-34-8r-1" corres="S-34-8r-1" n="1"/>
indicates the end (E).

Margins

Each page may have one or more of the following margins:
<margin type="topmargin">...</margin>
<margin type="bottommargin">...</margin>
<margin type="rightmargin">...</margin>
<margin type="leftmargin">...</margin>
These appear between the opening and closing page break markers.

- Within <margin type="topmargin"> and <margin type="bottommargin"> there is a further set of margins: <margin type="marginright"> for right-aligned text, <margin type="marginleft"> for left-aligned text, <margin type="margincenter"> for centred text.
- Within <margin type="leftmargin"> and <margin type="rightmargin"> there is: <margin type="middle"> for vertically-centred text (this applies to the binding marks).

Each column may have one or more of the following margins:
<margin type="coltopmargin">...</margin>
<margin type="colbottommargin">...</margin>
These appear between the opening and closing column break markers.

- Within these margins, there is a further set of margins: <margin type="right"> for right-aligned text, <margin type="left"> for left-aligned text, <margin type="center"> for centred text.

Each line may have the following margins:
<margin>...</margin>
<margin type="GL">...</margin>
If this appears within the opening line break (<lb id="S-...">...</lb>) it is to the left of the text. If this appears within the closing line break (<lb id="E-...">...</lb>) it is to the right of the text.

Margins may include text and/or graphic elements. Words are not numbered within margins.

Corrections

Where text has been altered, it is enclosed within an **<app>...</app>** element (for apparatus).
Within each <app> tag, each reading is enclosed by **<rdg>...</rdg>**.

The identity of the hand is included in the <rdg> element:

- The original reading is identified as **<rdg type="main-corr">...</rdg>**
- Correctors are identified by type="corr", with the name of the corrector: Thus **<rdg type="corr" n="S1">...</rdg>** is a correction by the corrector known as S1.

Words within a correction are enclosed within **<w>...</w>** tags.
The same text may have been altered by more than one corrector, so there may be numerous <rdg> elements within an <app> element.
Where a reading is blank, the text is either omitted (if type="main-corr") or deleted (if type="corr").
The default <rdg> displayed in the online transcription is always that which occurs **first** in the <app> element. In 95% of cases this is the original reading.
Line breaks and other formatting information are only included in this first element.

In the cases where an alteration consists of the addition of a block of text in the margin, this has been indicated by use of the **<ptr>** element. The <app> element is used as normal in the course of the text, but the correction in the margin has an extra id in the <rdg> element, e.g. **<rdg n="D" type="corr" id="AM-B7K11V18-07-1CHR-1">**.
The <ptr> element is placed at the appropriate point on the page in order to be able to display this material where it appears on the page as well as in the course of the text.
There are two types of margin:

- When the <ptr> appears in <margin type="coltopmargin"> or <margin type="colbottommargin">, it is encoded as <ptr type="appmargin" n="AM-B7K11V18-07-1CHR-1" />, horizontally aligned by <margin type="left"> or <margin type="right">
- When the <ptr> appears on the far left or far right margin of any opening (<margin type="leftmargin"> or <margin type="rightmargin">, again horizontally aligned by <margin type="left"> or <margin type="right">) the <ptr> must also be vertically aligned with the number of the line to which it corresponds. This is encoded by putting the line number in the 'rend' attribute. Thus <ptr type="appmargin" n="AM-B15K7V23-15-JER-1" rend="27"/> should appear next to line 27 in the neighbouring column.

Graphics

These are characters which are not present in the Unicode character set and so must be represented by images.
The following graphic elements are included in the transcriptions:
<g ref="#libstamp"/> *for Leipzig library stamp on many Leipzig leaves*
<g ref="#libstamp2"/> *for St Petersburg library sticker on 3-4r*

Notes

There are the following types of note:
1. **<note type="editorial">...</note>** for comments added by the editors of the project.

2. **<note type="gloss">...</note>** for non-biblical material added to the text by later hands. The following attributes may also be present:

attribute scribe="x": Scribe: x (optional)
attribute reading="x": Reading: x (optional - usually for Arabic glosses)
attribute translation="x": Translation: x (optional - usually for Arabic glosses)
attribute comment="x": Comment: x (optional)

3. **<note type="colophon" scribe="n">...</note>** for colophons to biblical books.
4. **<note type="ECN">...</note>** for Eusebian canon numbers.
Eusebian canon numbers appear in the Gospels only and are split over two lines (indicated by a <lb />). They are usually in red (<hi rend="red">).
The Greek numbering system is rendered in the following way:

- **Ammonian="n"** for the Ammonian section (top number)
- **Canon="n"** for the canon it belongs to (bottom number)
- **comment="..."** for any Comments (optional)

5. **<note type="folionum">...</note>** for folio numbers physically written on the page.
6. **<note type="quireSig" n="n">...</note>** for quire signatures.
The attribute comment="..." is optional, and is used to record corrections.
7. **<note type="rt">...</note>** for running titles (at top of page).
There are two optional attributes :

- **scribe="..."**
- **comment="..."** (usually to record corrections)

8. **<note type="booktitle">...</note>** for book titles (at the beginning of each book).
There are two optional attributes :

- **scribe="..."**
- **comment="..."** (usually to record corrections)

9. **<note type="section" n="n">...</note>** for section numbers.
The attribute comment="..." is optional, and is used to record corrections.
10. **<note type="lectionary">...</note>** . This is a subset of glosses, which have been added to show the beginning and end of passages read in the liturgy. There are two optional attributes :

- **scribe="..."**
- **reading="..."**

Character rendering

The character set of the transcripts is Unicode UTF-8.
Some letters are made up of two characters (e.g. combining underdots).
Particular attention should be paid to the following characters:
c : Greek lunate sigma, Unicode 03F2
◌̣ : Combining dot below, Unicode 0323
◌̈ : Combining diaeresis, Unicode 0308
◌̇ : Combining dot above, Unicode 0307
◌̃ : Combining tilde, Unicode 0303 (used with some Greek characters)

Characters may be modified as follows within a **<hi>...</hi>** element:

- **<hi rend="red">...</hi>** rubricated text

- **<hi rend="ul">...</hi>** underlined text
- **<hi rend="ol2">...</hi>** overlined text
- **<hi rend="ol">...</hi>** a "joined-up diaeresis" (rendered on the website by a normal combining diaeresis above the letter, Unicode 0308)
- **<hi rend="ul-ol2">...</hi>** text with lines above and below
- **<hi rend="kwhyphen" />** This indicates a word split over one or more lines, and is not rendered in the display.

The following characters, within <w> tags, render graphic elements in the text:

- <w n="bindingmark">⸗</w> (Unicode FE34)
- <w n="blackcross">+</w>
- <w n="blackstaurogram">P̶</w> (Capital Greek Rho and Unicode 0336)
- <w n="coronis"> ̓</w>
- <w n="crosswithdots">※</w> (Unicode 203B)
- <w n="fourdots">⁖</w>
- <w n="paragraphus">⁋</w> (Unicode 203E, 203E and 0337)
- <w n="paragraphusred"><hi rend="red">⁋</hi></w>
- <w n="pgline">‾</w> (Unicode 203E and 203E)
- <w n="pgtilde">˜</w> (Unicode 02DC)
- <w n="sandline">s̶</w> (s and Unicode 0335)
- <w n="squiggle">≀</w> (Unicode 2240)
- <w n="threedots">⁖</w> (Contents may vary depending on arrangement)
- <w n="threedotsandline">⁖–</w> (Contents may vary depending on arrangement)
- <w n="wrsymbol">ω̲</w> (Greek omega and Unicode 1D68)
- <w n="wrsymbolwithcrosspiece">ω̲</w>
- <w n="diple">></w> ONLY for use in margins

Text tagged as **<supplied>...</supplied>** has been reconstructed by editors.

Missing text and fragments

Gaps are treated as separate elements as follows:
<gap extent="1" units="chars" /> displays a space of 1 character within a line
<gap extent="10" units="lines" /> displays a space of 10 lines
Note also the following type of gap:
<gap reason="unreadable" units="chars" extent="5" /> which indicates unreadable characters.