XML specifications for Codex Sinaiticus

Initially compiled for website release of XML transcription 11.11.2011, and covering versions 1.0 (6.7.2010) to 1.05 (14.1.2023).
This file adjusted to take account of markup changes to v1.95 (30.8.2024), prepared towards the release of a new website (which will correspond to the v2.0 release).

Header and file structure

With effect from this version (v1.94), the file conforms to the TEI P5 Guidelines, with a few additional attributes which are listed at the end of this document.

A standard TEI header was supplied following the agreement of the Project Board on 28.10.2010. This has been kept up to date with all textual changes recorded in the <revisionDesc> element, along with the version number in which the change was released.

*Layout: a) by *

** ... ** encloses the whole transcript file (i.e. witness).
** ... ** encloses each individual book.
** ... ** encloses each individual chapter.
<ab xml:id="V-...">...</ab> encloses each verse.
<w>...</w> encloses each word of the biblical text. **<pc>...</pc>** encloses punctuation. Occasionally, punctuation elements are found within a word.
All of these are numbered apart from , words outside the normal text block, and most punctuation.
In most books, an initial verse is numbered 0 in order to contain title material.

With effect from v1.94, the following information has been added as attributes to words, in order to assist with searching and translations:

- **corresp**: the dictionary form of the word following;
- **lemma**: the Greek Strong's number (a biblical concordance);
- **morph**: Packard's Morphological Analysis Codes;
- **norm**: a normalised form of the following word, including accentuation as it appears in a critical edition of the biblical text.

This was provided through a restrictive process which mapped the manuscript's text to that of a critical edition. The coverage in v1.94 is about 85% of the total number of words (430,445): there remain 62,791 untagged words which cannot be automatically disambiguated because of Greek homonyms (and overlap within the Strong's number system). Many of the remaining words are spelling errors or erroneous forms in units involving corrections.

Layout: b) by page

<pb/> identifies the beginning of each page.
Pages are identified by quire number, then page number, so Q34-8r is Quire 34, Page 8 recto.
Each <pb> includes the following identifiers:

- **copyist="n"** for the copyist of the page. Two pages feature more than one copyist. Changes to the copyist are indicated by the <**handShift**/> milestone.
- **source="#n"** for the holding library (BL / LUL / SC / NLR)
- **folionum="n"** for the folio number assigned to the page by the holding library

<cb/> identifies the beginning of each column.
Columns are identified by quire number, page number and column number, e.g. Q34-8r-4
<lb/> identifies the beginning of each line.
Lines in the main text are identified with an id using the cumulative system as described, e.g. Q34-8r-4-23; other lines are simply indicated by <lb/>.
Lines may be positioned in the following ways:

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

If a word is broken across two lines, the <lb> has the attribute **break="no".**

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.

Margins

Text or items in margins are enclosed in the following <seg> elements:
<seg type="margin" subtype="page-top-margin">...</seg>
<seg type="margin" subtype ="page-bottom-margin">...</seg>
<seg type="margin" subtype ="page-right-margin">...</seg>
<seg type="margin" subtype ="page-left-margin">...</seg>

<seg type="margin" subtype="column-top-margin">...</seg>
<seg type="margin" subtype="column-bottom-margin">...</seg>

<seg type="margin" subtype="line-left">...</seg>
<seg type="margin" subtype="line-right">...</seg>
<seg type="margin" subtype="line-left-symbol">...</seg>
<seg type="margin" subtype="line-right-symbol">...</seg>

The alignment of items within these elements is indicated by the rend attribute on each item. Thus <note type="running-title" rend="center"> is a center-aligned running title within the page top margin. Margins may include multiple items with different

alignments (left, right or center). If no alignment is specified, then the item is assumed to be left-aligned.

It should be noted that binding marks within <seg type="margin" subtype ="page-right-margin"> and <seg type="margin" subtype ="page-left-margin"> have rend="middle" to indicate their vertical alignment on the page.

The indication "line-left-symbol" and "line-right-symbol" preserve the old <margin type="GL">, although these appear to be identical to "line-left" and "line-right"

Margins may include text and/or graphic elements. Words are not usually 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="orig">...</rdg>**
- Correctors are identified by type="corr", with the name of the corrector:
  Thus **<rdg type="corr" hand="S1">...</rdg>** is a correction by the corrector known as S1.

Words within a correction are enclosed within **<w>...</w>** tags, and it may also include punctuation in **<pc>...</pc>** 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="orig") 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 xml:id in the <rdg> element, e.g. **<rdg type="corr" hand="D" xml:id="AM-B7K11V18-07-1CHR-1" corresp="1Chr. 11:18">**.
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 <seg type="margin" subtype="column-top-margin"> or <seg type="margin" subtype="column-bottom-margin">, it is encoded as <ptr type="appmargin" n="AM-B7K11V18-07-1CHR-1" />, horizontally aligned by rend="left" or rend="right".

- When the <ptr> appears on the far left or far right margin of any opening (<seg type="margin" type="page-left-margin"> or <seg type="margin" subtype="page-right-margin">, again horizontally aligned by rend="left" or rend="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 'corresp' attribute. Thus <ptr type="appmargin" n="AM-B15K7V23-15-JER-1" corresp="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:
<graphic type="leipzig-libstamp"/> *for Leipzig library stamp on many Leipzig leaves*
<graphic type="nlr-libstamp"/> *for St Petersburg library sticker on 3-4r*
They may include 'rend' attributes to indicate their alignment and have also been supplied with a url in order to provide direct access to the relevant images.

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:
   hand="x": Scribe: x (optional)
   rend="…": for alignment
   reading="x": Reading: x (optional - usually for Arabic glosses)
   translation="x": Translation: x (optional - usually for Arabic glosses)
   comment="x": Comment: x (optional)
3. **<note type="colophon" hand="n">...</note>** for colophons to biblical books.
4. **<note type="Eusebian">...</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 by the following attributes:
   - **section="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="running-title">...</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 three optional attributes :

- **hand="..."**
- **rend="center"** (or other alignment)
- **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.
11. **<note type="hyphen"/>** This has been included in the text where words are broken across lines simply in case it is required for display purposes (the **break="no"**) attribute on the <lb> element and unclosed <w> element are sufficient for encoding.
In addition, notes may have 'rend' attributes to indicate their horizontal alignment.

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 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="underline">...</hi>** underlined text
- **<hi rend="overline">...</hi>** overlined text
- **<hi rend="joined-diairesis">...</hi>** a "joined-up diaeresis" (rendered on the website by a normal combining diaeresis above the letter, Unicode 0308)
- **<hi rend="overline underline">...</hi>** text with lines above and below

Note that <w> elements are on the outside of <hi> elements, but <pc> elements occur within <hi> elements.

The following characters, within <pc> or <fw> tags, render specific details in the manuscript or text:

- <fw n="bindingmark">⸗ </fw> (Unicode FE34)
- <fw n="coronis">⸗ ‿‿</fw>
- <pc n="blackcross">+</pc>
- <pc n="staurogram">𐙐</pc> (Unicode 2CE7)
- <pc n="crosswithdots">※</pc> (Unicode 203B)
- <pc n="fourdots">⁘</pc> (Unicode 2058)
- <pc n="paragraphus">⁀ ⁄ </pc> (Unicode 203E, 203E and 0337)
- <pc n="pgline">‾</pc> (Unicode 203E and 203E)
- <pc n="pgtilde">˜</pc> (Unicode 02DC)
- <pc n="sandline">s̲</pc> (s and Unicode 0335)

- <pc n="squiggle">⁝</pc> (Unicode 2240)
- <pc n="threedots">⁖</pc> (Unicode 2056)
- <pc n="threedotsandline">⁖–</pc> (Unicode 2056 plus line)
- <pc n="wrsymbol">ω᷒</pc> (Greek omega and Unicode 1D68)
- <pc n="wrsymbolwithcrosspiece">ωᷓ</pc>
- <pc n="diple">›</pc>

The following characters are used for punctuation in the line of the text (some are also numbered in the word sequence):

- High dot: ˙ (dot above: Unicode 02D9)
- Low dot: . (full stop: Unicode 002E)
- Middle dot: · (Ano teleia: Unicode 0387)
- Colon: :
- Comma: ,
- Apostrophe: ’ (Unicode 02BC)
- Slash: /
- Numeral signs: ʹ and ͵ (Unicode 0374 and 0375)

Abbreviations

Nomina sacra and numerals are indicated by the <abbr> element along with overlines as they appear in the text. These are usually complete words, although in some cases may be part words.

- <w><**abbr type="nomSac"**>θ<hi rend="overline">υ</hi></abbr></w>
- <w><**abbr type="num"**><hi rend="overline">ιβ</hi></abbr></w>

Abbreviations indicated by special characters are indicated with <ex> tags, with a rend attribute which specifies the character used. For display purposes, it is preferable to replace the enclosed characters with the rend value, but for searching purposes the full text is required. There are two in this category:

- **<ex rend="ϗ">και</ex>** *kai-compendium*
- **<ex rend="̄">ν</ex>** superline *nu* (indicated by a combining overline on the previous letter).

Other abbreviations are indicated by the presence of **<ex>…</ex>** enclosing the missing letters, and should be rendered by parentheses, e.g. μ(ου)

Missing and hard to read text

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

Text tagged as **<supplied>...</supplied>** has been reconstructed by editors. The custom is to display this within square brackets, e.g. t[hus]
Text tagged as **<unclear>...</unclear>** is difficult to read. Each character within such tags is normally indicated by a dot below the letter (Unicode 0323), e.g. tḥuṣ

Adjustments to the TEI P5 schema

The transcription should validate against the complete TEI-P5 with the following additions:

- <lb> add the attribute: vnumber (CDATA #IMPLIED)
- add the attribute: title (CDATA #IMPLIED)
- <pb> add the attributes: hand folionum (CDATA #IMPLIED)
- <note> add the attributes: translation reading comment section canon (CDATA #IMPLIED)
- <w> add the attributes: norm morph lemma (CDATA #IMPLIED)