IGNTP guidelines for XML transcriptions of New Testament manuscripts using the TEI P5

Version 1.3, 28.5.12

This document specifies the XML elements for the encoding of IGNTP manuscript transcriptions. In previous electronic editions, these have been generated automatically by Collate, and vary from project to project. This schema provides a single format for all transcriptions which can be applied to the recent Unicode files, underlie an online transcription tool, and enable consistent publication of transcriptions through SDPublisher and a single TEI-compatible format for deposit in the Institutional Repository. It does not change the way we currently transcribe manuscripts, but only their conversion to XML for publication. It has been compiled based on the Majuscule Edition, the Vetus Latina Iohannes, Codex Sinaiticus and the IGNTP/INTF Transcription Guidelines.
A subset of the TEI P5, called TEI-NTMSS, can be used to parse transcriptions which conform to this schema

A summary of the overall structure, listing the elements included, is provided in an Appendix.

1. TEI Header

This should go at the beginning of every file, replacing the current header in braces. The information in bold should be edited as appropriate: not all may be required or available. Make sure that complete XML elements are deleted if any material is removed. Further information and parameters may be found in TEI P5 part 10 (Manuscript Description) at: http://www.tei-c.org/release/doc/tei-p5-doc/en/html/MS.html
Material in italics is optional, and may be added later or generated from elsewhere.

```
<?xml  version="1.0" encoding="utf-8"?>
<!DOCTYPE TEI [
        <!-- entities used in individual transcriptions will be defined here -->
]>
<TEI xml:id="GA02" xmlns="http://www.tei-c.org/ns/1.0"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.tei-c.org/ns/1.0 TEI-NTMSS.xsd ">
<teiHeader>
<fileDesc>

<titleStmt>
<title>A Transcription of John in Codex Alexandrinus (GA 02)</title>
<title type="document" n="02" key="20002">Codex Alexandrinus, GA 02/title>
<title type="collection" level="s" xml:lang="en">The New Testament</title>
<title type="work" level="m" xml:lang="en" n="4">The Gospel according to
John</title>
<title type="short" level="m" xml:lang="en" n="4">john</title>
<respStmt><resp>Created by</resp><name>the International Greek New Testament
Project</name>
</respStmt>
<funder>The Arts and Humanities Research Council</funder>
<sponsor><name type="org">The Institute for Textual Scholarship and Electronic
Editing</name></sponsor>
</titleStmt>
```

```
<editionStmt>
 <edition n="1.1">Version 1.1, last updated <date>28.1.2011</date></edition>
</editionStmt>

<publicationStmt>
<publisher>
 <name type="org">The Institute for Textual Scholarship and Electronic
Editing</name>
<name type="org">University of Birmingham</name></publisher>
<date>30.1.2010</date>
<availability><p>Available for noncommercial re-use provided attribution is made to
the original creators and the subsequent derivative is licensed as ShareAlike
(Creative Commons by-nc-sa)</p></availability></publicationStmt>

<sourceDesc>
<msDesc>
  <msIdentifier>
    <country>United Kingdom</country>
    <settlement>London</settlement>
    <repository>British Library</repository>
    <idno>Royal 1 D.VIII</idno>
    <msName xml:lang="la">Codex Alexandrinus</msName>
    <altIdentifier type="GA"><idno>02</idno></altIdentifier>
    <altIdentifier type="Liste"><idno>20002</idno></altIdentifier>
    <altIdentifier type="Tischendorf"><idno>A</idno></altIdentifier>
    <altIdentifier type="TM"><idno>62318</idno></altIdentifier>
    <altIdentifier type="LDAB"><idno>3481</idno></altIdentifier>
  </msIdentifier>
```

More optional information can be added here using the <msContents> <physDesc> and <history elements> but these are no longer included in the example. See the example in version 1.1 of these guidelines.

```
</msDesc>
</sourceDesc>
</fileDesc>

<encodingDesc>
<projectDesc><p>This transcription was made by the the International Greek New
Testament Project in its work towards the Editio Critica Maior of John between 2000
and 2010..</p></projectDesc>
<editorialDecl>The initial transcription file was a plain text file for use with
COLLATE, later converted to XML. Further information about the procedures
followed may be found in the electronic edition at
http://www.iohannes.com/majuscule/</p></editorialDecl>
<variantEncoding method="parallel-segmentation"  location="internal"/>
</encodingDesc>

<revisionDesc>
<change n="5" when="2010-09-30">Correction to John 2.4</change>
<change n="4" when="2010-01-15">Version 1.0 published.</change>
<change n="3" when="2006-09-26">Changes incorporated following consultation of
manuscript in London</change>
```

```
<change n="2" when="2006-01-14">Transcription completed.</change>
<change n="1" when="2005-10-02">Transcription begun.</change>
</revisionDesc>

</teiHeader>
<text><body> ...... </body></text></TEI>
```

Observation on texts in Greek and Coptic. *The accepted xml indication of Greek is* xml:lang="el" . *However, this does not distinguish between historical stages of the language. The IANA codes* "grc" *for classical Greek and* "cop" *for Coptic will be adopted here.*[1]

Notes on the TEI header

A version of the header has also been developed for the conversion of INTF transcriptions which simply relies on 6 variables (and so can be automatically generated). This can be made available on request.

The xml:id on the TEI header cannot begin with a number: for Greek the prefix is GA; for Old Latin VL and for Coptic the specific language (sa, fa, mae, cw) plus the usual ID number.

In <title type="document"> the n-attribute is the siglum which the manuscript will be given in collation, and the key-attribute is the Liste number from the Münster database (Latin is assigned the prefix 5- and Coptic the preface 6-).

In the <encodingDesc> section, it is obligatory for TEI compliance to explain that the variant readings are encoded in parallel within the <app> elements:

- <variantEncoding method="parallel-segmentation" location="internal"/>

However, it should be noted that this currently breaks the xsd validation in Roma, so may been to be commented out if validation is performed in this way.

2. Divisions of the work

This occurs within the element <body>*.*

Book (Collate <B 4>): <div type="**book**" n="**B04**" >**.....**</div>
Chapter (Collate <K 1>): <div type="**chapter**" n="**B04K1**">**.....**</div>
Verse (Collate <V 2>): <ab n="**B04K1V2**" >**....**</ab>[2]

Within verses, words are encoded as <w> elements, numbered within each verse: the numbering may be added at a later stage in the process and normally proceeds in units of 2, matching the practice of the *Editio Critica Maior*. Punctuation tokens are encoded as <pc> (but are not necessarily numbered). Certain numbers (usually paratextual) are encoded as <num> (see '4. Paratextual Elements' below).

[1] The BCP47 standard adopted elsewhere in the TEI suggests a private use code, such as xml:lang="el-x-koine". For the wider selection, see http://www.iana.org/assignments/language-subtag-registry. Of course, Coptic needs further subdivisions, not present in IANA...

[2] The use of <ab> rather than <div> for biblical verses is expressly recommended at TEI P5 16.3. As of version 1.2 of these guidelines (March 2012), witness identifiers and xml:id attributes are no longer used in the <div> and <ab> elements, in order to cater for manuscripts with more than one language or lectionaries (and other manuscripts) with repeated verses; however, if unique verse identifiers cannot be generated by CollateX we will have to revisit this.

When a manuscript featuring more than one language is transcribed as a single document, the xml:lang attribute should be added to each <ab> element. The most common values will be "la" (Latin), "grc" (Greek),

When a verse is only partially preserved (or preserved in more than one location), the 'part' attribute may be used, with the values 'I' for the initial portion of the verse, 'M' for a medial portion of text and 'F' for the final portion of text. This may be helpful when collating,

Prefaces, lists of lections, and capitula pertaining to each book are best treated as <div> elements within the element <div type="book">.

- e.g. <div type="**preface**"><p>**....**</p></div>
  or <div type="**capitula**" n="**1**"><ab>**....**</ab></div>

Alternatively, if it is desired to treat these as different works (e.g. the Letter to Carpianus), the <group> element may be used, a division which precedes <text> and <body> (see TEI P5 4.3).

The incipit and explicit of the work (previously treated as chapter 0 and chapter 22) are instead to be named <div> elements:

- e.g. <div type="**incipit**"><ab><w n="**1**">**incipit**</w><w n="**2**">**euangelium**</w></ab></div>
- <div type="**explicit**"><ab>**....**</ab></div>

When a manuscript is fragmentary, and begins in the middle of a verse, the <div> or <ab> elements should be placed after the opening page layout elements or a ghost page added (see 6. Lacunae below).

3. Page layout

Quire (optional, not marked in Collate): <gb n="**3**" />
Page (Collate |P 121|): <pb n="**121**" type="**page**" xml:id="**P121-*wit***" />
Folio (Collate |F 3v|): <pb n="**3v**" type="**folio**" xml:id="**P3v-*wit***" />
Column (Collate |C 2|): <cb n="**2**" xml:id="**P3vC2-*wit***" />
Line (Collate |L 37|): <lb n="**37**" xml:id="**P3vC2L37-*wit***" />

Note that these elements simply mark the breaks: they are not allowed to contain any text. For information on marginal material, see '10. Marginal material' below.

The attribute "facs" may be used on a <pb> element to give a fully-qualified URL to a digital image of that page.

When a word is broken over a line/column/page, the <w> element should be left open, and the following layout element should include this attribute: break="no" (this is an innovation in TEI P5 in 2011 and replaces the &hyphen; entity). The break attribute should be added to the first layout element, so if a word is broken over a page it should be added to the <pb> not the <cb> or <lb>.

Indentations and line justification should be expressed with the 'rend' attribute of the <lb> element:
Centre-justified (Collate [cj]): <lb ... rend="centerJust" />
Right-justified (Collate [rj]): <lb ... rend="rightJust" />
Indented (Collate &indent;): <lb ... rend="indent" />
Hanging line (Collate &hang;): <lb ... rend="hang" />

4. Paratextual elements

This covers elements of the presentation and *mis en page* supplementary to the base text, such as running titles, quire signatures, chapter numbers and titles. Many of these will appear in margins (for formatting, see below '10. Marginal material').

The following numerals are to be represented by <num> elements (with the numerical value given in the 'n' attribute):
Chapter numbers (Collate [cn]): <num type="**chapNum**" n="**10**">**x**</num>
Ammonian sections (Collate [as]): <num type="**AmmSec**" n="**32**">**xxxii**</num>
Eusebian canons (Collate [ec]): <num type="**EusCan**" n="**8**">**viii**</num>

Other elements of the *mis en page* are expressed by <fw> elements, with the following types:
Page numbers: <fw type="**pageNum**" n="**29**">**xxix**</fw>
Quire signatures (Collate {-qs- ...}): <fw type="**quireSig**" n="**14**">**Q.XIIII**</fw>
Running titles (Collate {-rt- ...}):<fw type="**runTitle**">**....**</fw>
Chapter titles (Collate {-ct- ...}):<fw type="**chapTitle**">**....**</fw>
Lectionary headings (Collate {-lh- ...}):<fw type="**lectTitle**">**....**</fw>

When paratextual information is not present on the page but is added by the transcriber/editor, this should be a <note> element. For example, in tables of canon headings, the verse reference is sometimes provided:
Canon reference (Collate {+cref+ ...}): <note type="canonRef">....</note>

If there is a change of hand in the manuscript, this should be signalled informally by an editorial note (see below), accompanied by the <handShift/> element. The <handShift/> element should have the attribute n identifying the hand, with values such as "original" or "supplement", or "A" and "D" etc. where the hands are known.

For punctuation, see '8. Unusual characters'.

5. Text formatting

Formatting which is purely decorative should be expressed by <hi> elements, as follows:

- rubrication (Collate [rub]): <hi rend="rubric">...</hi>
- other coloured ink: add appropriate colour to rend attribute.
- Outsize capitals (Collate [cap4] etc.): <hi rend="cap" height="**4**">...</hi>
  (the height attribute refers to the number of lines covered by each capital letter)
- Overlines (Collate combining overline - when this does not indicate an abbreviation): <hi rend="ol">...</hi>

6. Lacunae, spaces, supplied and damaged text

For lacunae, when the writing material is not present, (Collate entity &lac;), a <gap> element should be used.[3] Where possible, in an XML transcription, more details can be added about the extent:

[3] The elements <lacunaStart/> and <lacunaEnd/> are for use in a critical apparatus rather than a transcription, and can only appear within the <rdg> and <lem> elements.

- e.g. <gap reason="lacuna" unit="line" extent="5" />

When the Collate tag [º] contains numbers of missing characters rather than supplied letters, it should be rendered in the same way.

- e.g. <gap reason="**lacuna**" unit="**char**" extent="**2**" />

When a manuscript is fragmentary and begins in the middle of a verse, a "ghost page" may be added in the XML to ensure that the lacuna marking does not erroneously appear at the beginning of a complete page. This should be indicated by the following element (after the page layout and verse block markers): <gap reason="absent" unit="page" extent="1">.

- e.g. <pb .../><cb .../><lb .../> <div type="book"...><div type="chapter"...><ab ...><gap reason="absent" unit="page" extent="1" />.

Note that <gap> should not be used where a blank space has been left by the copyist: these should be expressed by <space> elements:

- Blank spaces (Collate &spa1; etc.): <space unit="**char**" extent="**1**" />

Text may be supplied by the transcriber in places where the manuscript is not extant or is no longer legible. This is expressed by <supplied> elements, where the reason and the source for the supplied text should be specified as attributes:

- Text supplied by the transcriber for a lacuna (Collate [º]):
  <supplied reason="lacuna" source="**transcriber**"> ... </supplied>
- Unreadable text where the writing material is still extant (Collate [ill], or in earlier transcriptions, [unr] or [re]):
  <supplied reason="illegible" source="**transcriber**"> ... </supplied>

When text is supplied from another source (e.g. an earlier edition or a published text) this should be specified (e.g. <supplied source="**Tischendorf**"> or <supplied source="**NA27**">).

Doubtful readings (Collate [dub] or combining dot below): <unclear>....</unclear>.
The reason may also be specified as an attribute, e.g.

- <unclear reason="**damage**">....</unclear>
  or <unclear reason="**poor photo**">....</unclear>

It is also possible to use the <damage> element to enclose sections of text, but the use of <supplied> and <unclear> is preferred in these guidelines. For further guidance given in the TEI on when to use <gap>, <supplied>, <unclear> and <damage> and how these may be combined see TEI P5 11.5.2.

7. Corrections and alternative readings

All corrections must be enclosed within the <app> element, with <rdg> in chronological sequence.

Collate [app]: <app> .... </app>
Collate [*]: <rdg type="orig" hand="firsthand"> .... </rdg>
Collate [C]: <rdg type="corr" hand="corrector"> .... </rdg>
Collate [C*]: <rdg type="corr" hand="firsthand"> .... </rdg>
Collate [C2]: <rdg type="corr" hand="corrector2"> .... </rdg>
Collate [K]: <rdg type="comm" hand="firsthand"> .... </rdg>
Collate [A]: <rdg type="alt" hand="firsthand"> .... </rdg>
and so on for all the correctors identified in each manuscript.

Where a correction is written in the margin, the text within that <rdg> element may be tagged with <seg type="margin"> (see '10. Marginal material' below).

8. Abbreviations and non-standard characters

Nomina sacra are not to be expanded but treated as words marked with a supplementary <abbr> element. Instead of representing the overline by a separate character, the <hi> element should apply to the whole word:

- e.g. <w><abbr type="nomSac"><hi rend="ol">dms</hi></abbr></w>

If the overline is missing, then the <hi> element should be omitted.

Numerals in the biblical text should be recorded as they are written, identified with the <abbr type="num"> element. If they are marked with a superline, this should be included separately.

- e.g. <w><abbr type="num"><hi rend="ol">ιβ</hi></abbr></w>

Other abbreviations should be expanded. This may be silent (and recorded in the <encodingDesc> part of the TEI header), or with <ex> elements corresponding to the use of brackets in Collate:[4]

- e.g. s(unt) n(ost)r(u)m becomes s<ex>unt</ex> n<ex>ost</ex>r<ex>u</ex>m

An exception may be made for superlines replacing 'n' or 'm', since these are currently encoded as separate characters after the letter to which they apply. It seems best to render these as an entity &nbar;. Note that some Latin manuscripts distinguish between n- and m- bars (the latter have a dot above or below): the latter will be rendered as &mbar;.

- e.g. co&nbar;te&nbar;du&nbar;t uerbu&mbar;

Similarly, e-caudata should be transcribed as &ecaud;
Otherwise, where possible, the appropriate Unicode glyph should be used for combining letters.

- e.g. œ æ ç é

Other combined characters should be treated as entities.

- e.g. &nt; &ns; &unt; &or;

Note that the use of &om; in Collate for an omission means that this entity should be avoided in plain text transcriptions.

Non-standard abbreviating symbols should be treated as entities. The following are common in Latin:

- &est; (for ÷) (where the 'e' is visible, this should be e(st)
- &autem; (for the symbol shaped like 'hr')
- &enim; (for the symbol shaped like '++')
- &et; (for an ampersand '&'. Where this is equivalent to a capital letter, it may be transcribed as &Et;)
- &et7; (for the symbol shaped like '7')

Others may be added; note that entities may not begin with a numeral.

In the Latin manuscripts, we have tagged Greek words as [gk]. These may be expressed as <foreign xml:lang="grc"><w n="1">κατα</w> <w n="2">Ιωαννην</w></foreign>.

[4] Note that <ex> is intended for editorial expansions which add a sequence of letters, whereas <expan> is intended to comprise whole words and is not suitable here.

Word-break hyphens should be replaced by the attribute break="no" on the <lb> <cb> or <pb> element which divides the word. As noted above, if more than one of these divides the word then break should go only on the highest break element so where a <pb>, <cb> and <lb> divide the break attribute goes on the pb only.

Punctuation should be recorded using standard characters, so far as is possible, although it may be necessary to use entities. The following entities are reasonably common:

&paraph;	paragraphus
&obelos;	obelus
&diple;	diple (>)
&semicolon;	(semicolons conflict with the entity expression!)

Where practicable, these should be treated as <pc> elements.

Declaring entities
For a single transcription to be valid TEI XML, all the entities must be declared in the <!DOCTYPE> part of the TEI header (see above). However, it is possible to point several transcriptions which share entities towards a file which lists these only once, e.g.:

```
<!DOCTYPE TEI [
        <!ENTITY % vetuslatina SYSTEM "vetuslatina.ents">
        %vetuslatina;
]>
```

9. Editorial notes

The system of putting notes into a separate, non-Collate file is very unsatisfactory. All editorial notes should be included at the appropriate point in the transcription.

Notes should be expressed with the <note> element, linked to the relevant feature with an id attribute, and categorised by the type attribute.

- e.g. <note type="editorial" xml:id="B4K3V5-*wit*-1">This word has been reinked by a later hand</note>

The ID may, equally, refer to the page layout:

- <note type="editorial" xml:id="P3vC2-*wit*-1">There is no text in this column</note>

The -1 suffix is to allow more than one note to refer to the same location; as xml:id values must be unique, this should be increased to -2, -3, -4, -5 for multiple references to the same location.

Comments in braces (ignored by Collate) are not supported in XML. Instead these should be expressed as a type <note> (id attributes are not strictly necessary as these notes will never be displayed in the edition, but it is good practice to include them and may help the editor work out the location of the problem!):

- <note type="transcriberquery" xml:id="B4K3V6-*wit*-1">I can't decide what the first hand wrote</note>

The {change of hand} note should be recorded in an editorial note, followed by the <handShift> element (see Section 4 above).

10. Marginal material

Text and numbers appearing within the margins will be encoded at the point to which they refer in the text: corrections will appear within the <app> element, chapter numbers and Eusebian apparatus will be included at the appropriate place in the verse; *diplai* will appear at the beginning of each line; page numbers and running titles will follow the <pb> element, while quire signatures in the bottom margin will be placed after the final line of the page.

In order to permit marginal material to be collated automatically and display correctly, marginal text will be enclosed by the element <seg type="margin">. The following attributes should be specified for subtypes:

- pagetop pagebottom pageleft pageright
- coltop colbottom colleft colright
- lineleft lineright

In addition, the 'n' attribute will enable the correct positioning of the material by including the xml:id of the relevant <pb> <cb> or <lb>.
Generally speaking <seg type="margin">...</seg> should enclose all the elements to which it refers; for corrections, however, it needs to come within the <rdg> element.

Here is an example of a correction added in the left margin of folio 7r, col. 1, line 45 of manuscript 33:

- <rdg type="corr" hand="B"><seg type="margin" subtype="lineleft" n="@F7rC1L45-33"><w n="1">et</w><w n="2">dicit</w></seg></rdg>

Here is an example of a running title in the top margin of page 321 in manuscript 3:

- <pb n="321" type="page" xml:id="P321-3" /><seg type="margin" subtype="pagetop" n="@P321-3"><fw type="runTitle">secundum</fw></seg>

Appendix: Summary

This is only a graphic overview: a commented xsd schema will provide a machine-readable declaration of these guidelines for use in parsing.

```
<teiHeader>
  <fileDesc>
    <titleStmt />
    <editionStmt />
    <publicationStmt />
    <sourceDesc>
      <msDesc />
    </sourceDesc>
  </fileDesc>
  <encodingDesc />
  <revisionDesc />
</teiHeader>

<text>
   <body>
      <div type="book" n="" >
             Optional <div type="preface"> <div type="capitula/kephalaia">
                           <div type="incipit">  <div type="explicit">
      <div type="chapter" n="" >
      <ab n="" >
                Contents of <ab> may include:
                       <w />  <pc />  <num />  <note />   <gap />
                The elements <w> <pc> <num> may contain or be contained within:
                       <hi>  (rend="colour", rend="cap", height="")
                       <unclear> (reason="")
                       <supplied> (source="" reason="")
                       <foreign xml:lang="">
                       <seg type="margin" subtype="" n="@...">
                The elements <w> <pc> <num> may be contained within:
                       <app><rdg type="" n="" id="">....</rdg></app>
                The element <w> may contain:
                       <abbr type="nomSac">  <abbr type="num">
                The element <note> must contain type="" and may contain id=""
                       note-types: editorial transcriber

       <gb n="">  (optional quire element)
       <pb type="page" n="" xml:id="">  (alternative type="folio")
       <cb n="" xml:id="">
       <lb n="" xml:id="">
                  All the above location elements <pb> <cb> <lb> may contain:
                            <seg type="margin" subtype="" n="@...">
                                   subtypes: pagetop pagebottom pageleft pageright
                                   coltop colbottom colleft colright lineleft lineright
                            <fw type="" />
                                   types: pageNum quireSig runTitle chapTitle lectTitle
                            <num type="" n="" />
                                   types: chapNum AmmSec EusCan

                            <note type="" n="" />
                            types: canonRef editorial transcriber
                  <lb> may include the following rend attributes:
                            centerJust rightJust indent hang

      </ab></div></div>
   </body>
</text>
</TEI>
```