Gloss Corpus encoding guidelines

http://www.glossing.org/glosscorpus
Version 1: 18 December 2024

All data is in XML format, following TEI encoding guidelines. See the template files available on the Gloss Corpus website. Attributes are referred to below by the `@` sign.

1. Base texts (primary texts)

1.1 Edition information

Contained within the TEI header. (On segmentation, see next section.)

```
<teiHeader>
  <fileDesc>
    <titleStmt>
      <title>Ars grammatica</title>
      <author>Priscian</author>
    </titleStmt>
    <editionStmt>
      <p>[Statment about origins of this edition, e.g. whether previously published or
newly transcribed.]</p>
    </editionStmt>
    <publicationStmt>
      <publisher>Gloss Corpus</publisher>
        <authority>Pádraic Moran</authority>
      <date>2024</date>
      <idno type="URL">http://www.glossing.org/glosscorpus</idno>
        <availability status="free">
          <licence target="https://creativecommons.org/licenses/by-nc-sa/4.0/">
            <p>Creative Commons BY-NC-SA 4.0</p>
          </licence>
        </availability>
    </publicationStmt>
    <notesStmt>
      <note type="short_title">Priscian</note>
      <note type="segmentation">Segmented by volume, page and line in GL.</note>
    </notesStmt>
    <sourceDesc>
      <p>[Information on the digitisation process.]</p>
    </sourceDesc>
  </fileDesc>
</teiHeader>
```

1.2 Text structure

Text is contained within:

```
<text>
  <body>
```

Each major section has the following structure. The label supplied within `<head>` will appear in the table of contents.

```
<div type="section" n="0">
  <head>Praefatio</head>
```

```
  [content]
</div>
```

The text should be segmented into mangeable units to allow glosses to be displayed beneath in a group of reasonable size. Segments should ideally follow some already established convention, e.g. lines or numbered sections of a printed edition.

Each segment is contained within `<ab>` tags, with a unique `@xml:id`. An `@n` attribute provides a more reader-friendly reference.

```
<ab xml:id="II_1.2" n="GL II 1.2">sapientiae luce praefulgens a Graecorum fontibus
deriuatum Latinos proprio </ab>
```

Where texts are tokenised, each word or punctuation token is encoded with either `<w>` or `<pc>` tags, respectively. Each token will have an `@xml:id`, based on the `@xml:id` of the segment followed by two underscores and a number for the token, e.g.

```
<ab xml:id="II_5.1" n="GL II 5.1">
  <w xml:id="II_5.1__1">philosophi</w>
  <w xml:id="II_5.1__2">definiunt</w>
  <pc xml:id="II_5.1__3">,</pc>
  <w xml:id="II_5.1__4">uocem</w>
</ab>
```

2. Gloss collections

2.1 Edition information

Contained within the TEI header.

```
<teiHeader>
  <fileDesc>
    <titleStmt>
     <title>Glosses on Priscian in St Gall, Stiftsbibliothek, 904</title>
     <editor>Rijcklof Hofman</editor>
    </titleStmt>
    <editionStmt>
      <p>[Statment about origins of this edition, e.g. whether previously published or
newly transcribed, whether part of a larger project, etc.]</p>
    </editionStmt>
    <publicationStmt>
      <publisher>Gloss Corpus</publisher>
      <authority>Pádraic Moran</authority>
      <date when="2024-12-18">18 December 2024</date>
      <idno type="DOI">https://doi.org/10.71555/hofman.2024</idno>
      <idno type="URL">http://glosscorpus/glosscorpus/publications/hofman</idno>
      <availability status="free">
        <licence target="https://creativecommons.org/licenses/by-nc-sa/4.0/">
        <p>Creative Commons BY-NC-SA 4.0</p>
        </licence>
      </availability>
    </publicationStmt>
    <notesStmt>
      <note type="siglum">G</note>
      <note type="hex_colour">e39f29</note>
    </notesStmt>
    <sourceDesc>
      <msDesc>
        <msIdentifier>
        <settlement>St Gall</settlement>
        <repository>Stiftsbibliothek</repository>
        <collection></collection>
```

```
        <idno>904</idno>
        </msIdentifier>
        <physDesc>
          <handDesc><!-- all hand info is optional -->
            <summary><!-- optional general summary here --></summary>
            <handNote xml:id="hand-A">Main glossator, Irish minuscule</handNote>
        </handDesc>
        </physDesc>
      </msDesc>
    </sourceDesc>
  </fileDesc>
</teiHeader>
```

2.2 Glosses (structure)

All glosses are contained within:

```
<text>
  <body>
    <div>
```

Example gloss entries:

```
<!-- Example 1 -->
<note n="1a2 a" target="#II_1.1">
  <term>omnis</term>
  <gloss>.i. adit conatus sum usque grammaticorum</gloss>
</note>

<!-- Example 2 -->
<note n="1a10 e" target="#II_1.3__9" facs="&facsBase;1" ana="#211 #212"  hand="#hand-A"
place="m.l.">
  <term>libralibus</term>
  <gloss>.i. libardaib</gloss>
  <note type="translation">i.e. bookish (arts)</note>
  <note type="editorial">Misunderstanding based on corruption of the lemma.</note>
  <ref target="&ref2Base;049">Thes. 1a1</ref>
</note>
```

Example 1 represents a minimal entry:

- `@n`: A human-readable reference for this gloss.
- `@target`: Refers to the `@xml:id` of a line or word in the basetext file (prefixed with #).
- `<term>`: The lemma in the main text. (Optional, but recommended, especially for references to lines, not words.) Variants from the standard edition of the base text can be recorded here.
- `<gloss>`: Content of the gloss.

Example 2 illustrates additional options:

- `@facs`: URL (web address) for a manuscript image, either whole or abbreviated (see below).
- `@ana`: Number references for categories. (Still under implementation.)
- `@hand`: Refers to the `@xml:id` in `<handNote>` above (prefixed with #).
- `@place`: Location, if not interlinear. (Suggest: m.s., m.i., m.l., m.d. for upper, lower, left and right margins, respectively.)
- `<note type="translation">`: Translation.

- `<note type="editorial">`: Editorial notes.
- `<ref target="">`: URL for other resource for this gloss (e.g. a previous edition, online), either whole or abbreviated (see below).

Since URLs will probably be often repeated and can be long and cumbersome, you can optionally abbreviate them, by designating a common base with `&facsBase;` or `&ref2Base;` and specifying what these abbreviations stand for in entity declarations at the top of the XML file, e.g.

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE TEI [
  <!ENTITY facsBase "https://www.e-codices.unifr.ch/en/csg/0904/">
  <!ENTITY ref2Base "https://archive.org/details/thesauruspalaeo02stok/page/">
]>
```

With the example above, `&facsBase;37` would be interpreted as:
`https://www.e-codices.unifr.ch/en/csg/0904/37`

2.3 Manuscript transcription and editorial mark-up

Tags for manuscript transcription:

- `<add>…</add>`: Secondary addition in manuscript
- `<del>…</del>`: Text cancelled in manuscript
- `<g>…</g>`: Graphical symbol in manuscript (represented in transcription with a similar keyboard/Unicode character)
- `<lb/>`: Line break in manuscript
- `<space/>`: Blank space in manuscript. (You may optionally indicate an extent: e.g. `<space extent="1 line"/>`.)

Tags that document editorial interventions:

- `<corr>…</corr>`: Editorial correction
- `<ex>…</ex>`: Editorial expansion
- `<gap/>`: Gap in editorial transcription due to illegibility, damage, etc. (You may optionally indicate an extent or reason: e.g. `<gap extent="3 letters" reason="illegible"/>`.)
- `<sic>…</sic>`: Editorial marking of apparent errors
- `<supplied>…</supplied>`: Editorially supplied text
- `<surplus>…</surplus>`: Editorial marking of redundant text
- `<unclear>…</unclear>`: Editorial uncertainty in transcription

3. Publications

The publication XML file provides a way to manage publication information that might apply to several gloss collections published as a single publication (with a single DOI and persistent link).

Where the publication comprises a single collection, information can be copied from the gloss collection XML file.

```
<publication>
  <citation>Nike Stam, 'Glosses on Óengus mac Óengobann, <i>Félire Óenguso</i>, in
Oxford, Bodleian Library, Rawlinson B505'</citation>
  <doi>10.71555/stam2024</doi>
  <publish_date>2024-09-05</publish_date>
  <about>
    <p>This transcription was produced under the auspices of the NWO Vrije Competitie
Project
‘<a href="https://medievalirishbilingualism.sites.uu.nl/publications/">Medieval Irish
Bilingualism</a>’, which
ran at Utrecht University from 2012–2016.</p>
  </about>
</publication>
```

The `<citation>` tags contain the editor name and publication title only. Atribution to Gloss Corpus, the date, and DOI will be generated automatically.

Since this is not a TEI document, it may contain HTML tags, e.g. `<i>…</i>` for italics or `<a href="http://url.com">…</a>` for web links.