Research Group KOMeT
(funded by the German Federal Ministry of Education and Research eHumanities program)

KOMeT Standoff XML Documentation

Version: 1.0.1
Author: Dr. Amir Zeldes

1. Preamble

For many purposes, TEI XML is the standard of choice for encoding textual information, and particularly in the fields encompassing the Digital Humanities. For the annotation of ancient manuscripts, the EpiDoc subset (see Bodard 2010) of TEI XML has become established as a widespread format to encode primarily physical and organization structure in ancient texts. There is therefore a wide variety of ancient documents available in EpiDoc/TEI XML (see http://papyri.info for a large repository of freely available ancient papyri in EpiDoc).

At the same time, the possibility of linguistically annotating ancient texts in TEI is limited. While it is possible to make some annotations using e.g. the @type attribute of <w> elements for parts of speech, or simple syntax trees using hierarchies of <s> and <phr> elements (see e.g. Höder 2012), more complex annotations (crossing edges, discontinuous constituents, complex coreference and dependency annotation schemes, discourse analyses) are not well-supported in TEI. This means that extending existing TEI projects can only be achieved by abandoning the TEI version and converting to a linguistic annotation format, thereby halting development of the TEI annotation and leading to diverging, potentially incompatible versions of the corpus.

In order to address this situation, KOMeT Standoff offers a format for externally annotating TEI documents using separate PAULA XML files, which do not alter the TEI document. PAULA XML (Dipper 2005) is an established linguistic annotation format, based on arbitrary graph structures, which is powerful and generic enough to express a variety of linguistic information types. Because PAULA files can point to elements in the TEI document, there is no problem with editing both projects at once, as long as these guidelines are followed.

2. Roadmap

Multiple levels of integration for PAULA and TEI are conceivable. These are detailed in the successive roadmap stages below. So far, only Stage 1 has been implemented, which is documented in the following sections.

1. As a first step KOMeT standoff allows PAULA projects to be based on <w> elements with @xml:id's in the TEI project. Instead of using PAULA tokens, defined in the PAULA documentation on the PAULA website (https://www.sfb632.uni-potsdam.de/en/paula.html), all PAULA annotation files ultimately point to an identifiable TEI <w> element. It is possible for <w> elements to contain further annotations, including line breaks within words (<lb/>), but these are not referred to by PAULA. This type of annotation is robust against changes in the TEI document as long as the @xml:id of <w> elements remains constant.
2. As a second step (not yet implemented), it could be possible to use PAULA character range xpointer semantics to annotate individual plain text areas of TEI files. This has the advantage of allowing work with TEI files not containing <w> elements but at the risk of less robustness: any change to the TEI file will require a readjustment of xpointer character ranges.
3. A third and final step left for future planning is to allow any TEI element with @xml:id to be referenced with a PAULA annotation. This will presumably be difficult to implement, but would provide a robust and flexible framework to extend standoff TEI further.

3. The TEI Document

The document being extended via standoff annotation can in principle be any type of valid TEI file, but for convenience we will limit ourselves to EpiDoc files, which have been implemented in the KOMeT demo corpus, Besa.letters.

Any TEI file extended via stage 1 KOMeT standoff must contain <w> elements with @xml:id attributes. Only these elements can be referenced via PAULA annotation structures, though once the reference occurs, any number of complex PAULA structures can refer to the elements. As an example, consider the following fragment from the Letter to Thieving Nuns:

Besa.letters.to_thieving_nuns_KOMeT_TEI.xml

```
<?xml-model href="http://www.stoa.org/epidoc/schema/latest" ?>
<TEI xmlns="http://www.tei-c.org/ns/1.0">
<teiHeader>
…
</teiHeader>
<text>
<body>
```

```
<ab>
<lb/><phr>
<w xml:id="tok_1">[ ]</w>
</phr>
<cb/><lb/><phr>
<w xml:id="tok_2" xml:lang="Greek"><hi rend="large, decorated">ⲱ̃</hi></w>
</phr>
<phr>
<w xml:id="tok_3">ⲧⲉⲧⲛ̄</w>
<w xml:id="tok_4">ⲣⲱⲙ<hi rend="small">ⲉ</hi></w>
</phr>
<lb/><phr>
<w xml:id="tok_6">ⲛ̄ⲧⲱⲧⲛ̄</w>
</phr>
…
```

As the example shows, the text is divided into TEI <phr> elements to signify contiguous bound-groups of Coptic words (cf. Layton 2004: 19-27), while <w> elements designate individual words. The <w> elements all have @xml:id, while some words, borrowed from Greek, receive an additional @xml:lang. Within and between words, we find various EpiDoc annotations, such as line breaks <lb/>, column breaks <cb/>, as well as rendering information in hi@rend (e.g. for the last letter in the word ⲣⲱⲙⲉ).

It should be noted that the values of the @xml:id attributes are not important, as long as they are distinct. The word ⲣⲱⲙⲉ is designated as <w xml:id="tok_4">, while the next word ⲛ̄ⲧⲱⲧⲛ̄ is "tok_6" – this is not problematic.[1] These id's will be referred to in the PAULA annotation project below.

4. The PAULA files

PAULA XML works by separating each layer of information into a separate file. Some files define markable areas for annotation, while others give annotation values to such areas. For example, the following fragment delimits areas for normalization:

komet.to_thieving_nuns.normSeg.xml

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>

<!DOCTYPE paula SYSTEM "paula_mark.dtd">
<paula version="1.0">
<header paula_id="komet.to_thieving_nuns_normSeg"/>
```

[1] The actual reason for gaps in numbering in this example corpus, is that the data has been generated via annotation software which contained a placeholder for the annotation border corresponding to the intervening hi@rend annotation. In the TEI output, this is not a <w> element, so no corresponding "tok" element is generated. Naturally, sequential numbering is also allowed when intervening elements are present, and the @xml:id values are completely arbitrary.

```
<markList xmlns:xlink="http://www.w3.org/1999/xlink" type="normSeg"
xml:base="Besa.letters.to_thieving_nuns_KOMeT_TEI.xml">
…
        <mark id="norm_2" xlink:href="#tok_2"/><!-- ῶ -->
…
```

The entire file takes as its xml:base the name of the TEI file from the previous section: "Besa.letters.to_thieving_nuns_KOMeT_TEI.xml". The markable element with the id "norm_2" points to "#tok_2" in that file, meaning this area will be defined as a "normSeg", as determined by the @type of the <markList> above. This area, now identified as "norm_2", can be given an annotation in a feature file specifying the values for each annotation, as shown below:

komet.to_thieving_nuns.normSeg_norm.xml

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>

<!DOCTYPE paula SYSTEM "paula_feat.dtd">
<paula version="1.0">

<header paula_id="komet.to_thieving_nuns_norm"/>

<featList xmlns:xlink="http://www.w3.org/1999/xlink" type="norm"
xml:base="komet.to_thieving_nuns.normSeg.xml">
…
        <feat xlink:href="#norm_2" value="ω"/><!-- ῶ -->
…
```

This file points to "komet.to_thieving_nuns.normSeg.xml" as its @xml:base, and contains a <feat> pointing to "#norm2" with the value ω. Thus the file defines a normalization for the <w> element 'ῶ', by assigning a norm annotation with the value 'ω' (without the circumflex accent). Note that the TEI document has not been modified in any way to allow for this annotation.

As a more complex example, consider the addition of translation annotations, spanning multiple <w> elements. The first PAULA file again defines areas for annotation:

komet.to_thieving_nuns.translationSeg.xml

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<!DOCTYPE paula SYSTEM "paula_mark.dtd">
<paula version="1.0">
<header paula_id="komet.to_thieving_nuns_translationSeg"/>
```

```
<markList xmlns:xlink="http://www.w3.org/1999/xlink" type="translationSeg"
xml:base="Besa.letters.to_thieving_nuns_KOMeT_TEI.xml">
<mark id="translation_2" xlink:href="#xpointer(id('tok_2')/range-to(id('tok_19')))"/>
…
```

This area spans the text from the <w> element with @xml:id="tok_2" till "tok_19". The translation features value for that sequence of words is given in the next file:

komet.to_thieving_nuns.translationSeg_translation.xml

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>

<!DOCTYPE paula SYSTEM "paula_feat.dtd">
<paula version="1.0">

<header paula_id="komet.to_thieving_nuns_translation"/>

<featList xmlns:xlink="http://www.w3.org/1999/xlink" type="translation"
xml:base="komet.to_thieving_nuns.translationSeg.xml">
<feat xlink:href="#translation_2" value="Oh you guilty on your part who allow Satan to come in,"/>
```

Again, the TEI document is left untouched, and external annotations can be added via PAULA XML. The possibilities of annotation via PAULA exceed simple span annotation substantially, and any number or repeating, conflicting, hierarchical or pointing relations can be added. Limitations on nesting or naming conventions which constrain TEI annotations are not problematic in PAULA XML. For further information on complex annotations in PAULA, the reader is referred to the PAULA documentation. From the point of view of KOMeT standoff, the identifiability and addressability of <w> elements is the crucial factor.

The entire corpus used in these examples can be downloaded from the KOMeT website, currently at: http://korpling.german.hu-berlin.de/komet/

References

Bodard, G. 2010. EpiDoc: Epigraphic documents in XML for publication and interchange. In Feraudi-Gruenais, F. (ed.) *Latin on Stone: Epigraphic Research and Electronic Archives*. Lanham, MD: Lexington Books, 101–118.

Dipper, S. 2005. XML-based stand-off representation and exploitation of multi-level linguistic annotation. In *Proceedings of Berliner XML Tage 2005 (BXML 2005)*. Berlin, Germany, 39–50.

Höder, S. 2012. Annotating ambiguity: Insights from a corpus-based study on syntactic change in old swedish. In Schmidt, T./Wörner, K. (eds.) *Multilingual Corpora and Multilingual Corpus Analysis*. Amsterdam/Philadelphia: Benjamins, 245–271.

Layton, B. 2004. *A Coptic Grammar*. Second Edition, Revised and Expanded. Wiesbaden: Harrassowitz.