KBSET/Letters Markup (KLM): Parsimonious Descriptive Markup for Scholarly Editions of Correspondences

Jana Kittelmann and Christoph Wernhard

Draft – September 21, 2020

Contents

1 Introduction

KBSET/Letters Markup (KLM) is a collection of descriptive markup elements that is adequate for scholarly editions of correspondences of the 18th and 19th century. Its objectives are:

1. *KLM* source documents should be easy to create and maintain, in particular for users working in the application field.
2. *KLM* source documents should permit direct conversion into high-quality PDFs for printing and on-screen reading.
3. *KLM* source documents should permit direct conversion into high-quality HTML Web presentations.
4. *KLM* source documents should permit conversion into other formats and the extraction of data (e.g., meta information) represented in other formats.

KLM meets these objectives by using LaTeX as representation language of the markup and by identifying a small tightly constrained set of markup elements of relevance for the application field. Users then work essentially with these elements, focusing on concerns of the application. They can create documents with a text editor that supports LaTeX. Requirement 2 can be realized by implementing the markup elements as LaTeX commands. The elements are sufficiently constrained to permit the generation of high-quality PDFs. Meeting requirements 3 and 4 involves parsing the *KLM* source documents, where, however, the parser needs not to support LaTeX in full. Again, the tightly constrained markup facilitates the generation of high-quality outputs.

This document is at the present stage a draft. Nevertheless, the described processing of *KLM* documents as specified here has been implemented as free software and is successfully applied in a comprehensive scholarly edition project.[1] Hence, this specification comes already accompanied by a practical environment that verifies reachability of the stated objectives. This environment, *KBSET (**K**nowledge-**B**ased **S**upport for Scholarly **E**diting and **T**ext Processing)*, is available from

`http://cs.christophwernhard.com/kbset`.

The distribution and the system Web page include examples that illustrate the use of this specification in source documents and, based on it, the generation of high-quality PDFs and Web pages, as well as fact bases for future semantics-oriented applications.

TODO: Support for other languages than German.

TODO: Permitted and not-permitted nesting of markup should be made more precise.

[1] The correspondence of Johann Georg Sulzer (1720–1779) and Johann Jakob Bodmer (1698–1783), transcribed and annotated by J. Kittelmann with assistance of B. Baumann, which is going be published in 2020 in print as volume 10 of Sulzer's *Gesammelte Schriften*, edited by H. Adler and E. Décultot.

2 Preliminaries

2.1 Encoding and Special Characters

Documents are encoded in UTF-8. The following LaTeX commands for special characters are permitted: `\ldots`, `\vdots`, `--`, `---`, `\textbackslash`, `\{`, `\}`, `\$`, `\&`, `\#`, `\textdegree`, `\^{}`, `\~{}`, `\_`, `\%`, `~` (non-breaking space), `\\` (new line)

TODO: Further such commands? `"‘`, `"’`?

2.2 Uses of Symbolic Identifiers

Symbolic identifiers (briefly *identifiers*) used as unique identifiers of entities like persons, location, bibliographic references, letters and text constituents such as occurrences of pieces of text.

Identifiers are not global (like *GND* identifiers or URIs of articles in the German Wikipedia) but specific to the project. They can have mnemonic names. Global identifiers such as URIs can be derived from them.

An *identifier* is introduced and used with respect to a *type*, for example, *letter*, *person*, *geo*, *text-place*. The same identifier can be used (with obviously different meanings) for objects of different types. A derived global identifier needs to incorporate a type designator. For example, for `sulzer` as an identifier of type *person*, the URI `http://www.sulzer-digital.de/person-sulzer.html` might be a derived global identifier. In addition, `sulzer` could be used, for example, as identifier of a position in a text document, for which a different global identifier would be derived.

The scope of identifiers includes all documents of the project, except for identifiers of type *text-place* (i.e., those introduced by `\xl` and `\xll`), which are relative to a `letter` environment.

Identifiers are *introduced* with `\def...` statements (Sect. 9.1), with the `letter` environment (Sect. 6.1), and with `\xl` and `\xll` statements (Sect. 6.2).

2.3 Syntax of Symbolic Identifiers

Arguments ending with *Id* in element specifications denote an identifier.

Identifiers are only allowed to contain alphanumeric ASCII characters, “`:`” and “`-`”. Hence, umlauts and spaces are, for example, not permitted.

In XML/HTML representations these identifiers are translated to components of URIs and values of the `id` attribute, where “`:`” is usually not permitted. It is then translated into “`-`”.

TODO: If both “`:`” and “`-`” are mapped in the XML representation to “`-`” we have no purely syntactic reverse translatability.

2.4 Comparison to LaTeX Syntax

In *KLM* arguments must always be given explicitly in curly braces `{}`.

3 Inline Properties

3.1 Syntactic Properties

`\xlift{`*Text*`}` Legacy version: `\xhoch{`*Text*`}`
`\xls{`*Text*`}`
`\xsout{`*Text*`}`
`\xul{`*Text*`}`
`\xulul{`*Text*`}`

Text appears as superscript, with extended letter spacing (*gesperrt*), lined through, underlined, and underlined twice, respectively.

`\xfrac{`*Numerator*`}{`*Denominator*`}`

Represents the fraction $\frac{Numerator}{Denominator}$.

3.2 Quotations

`\xqwork{`*Text*`}` Legacy version: `\xqwerk{`*Text*`}`
`\xquote{`*Text*`}` Legacy version: `\xqzitat{`*Text*`}`

`\xqwork{`*Text*`}` inidcates that *Text* is a title, e.g., of a work. `\xquote{`*Text*`}` indicates that *Text* is a quotation. `\xqoute` can be nested once.

3.3 Languages

`\xenglish{`*Text*`}` Legacy version: `\xeng{`*Text*`}`
`\xfrench{`*Text*`}` Legacy version: `\xfranz{`*Text*`}`
`\xgreek{`*Text*`}` Legacy version: `\xgriech{`*Text*`}`
`\xlatin{`*Text*`}` Legacy version: `\xlatein{`*Text*`}`

For *Text* in English, French, Greek and Latin, respectively. Note that language specific special characters as well as Greek letters can be written directly as UTF-8 symbols.

`\xlatin` is also used for text that is written in Latin (Roman) letters. TODO: What does this exactly imply? Maybe there should be another element for this case.

3.4 Text Status

`\xus{`*Text*`}`

Transcription as *Text* is unsure, just conjectured.

`\xxx`

Represents text that is not readable or lost.

3.5 Explicit Hyperlinks

`\xhref{`*URL*`}{`*Text*`}`

Text is displayed, associated with a hyperlink to the URL *URL*.

`\url{`*URL*`}`

The URL *URL* is displayed, associated with a hyperlink.

4 Block Properties

4.1 Paragraph Formatting

`\xpar`

New paragraph.

`\noindent`

At the start of a paragraph: expresses that it should not be indented. Taken directly from LaTeX.

`\begin{xverse}` `\end{xverse}` Legacy version: `\begin{xvers}` `\end{xvers}`

Verse. Line-breaks can be marked with `\\`

`\begin{xsequence}` `\end{xsequence}`

Sequence of lines, similar to verse. Line-breaks can be marked with `\\`. Typically used for lists that should have no bullets.

`\xcenter{`*Text*`}`
`\begin{center}` `\end{center}`

Both versions, the command and the environment indicate the same: *Text* appears centered and with space at the top and bottom.. The `center` environment is taken directly from LaTeX.

4.2 Tables

`\begin{tabular}{`*TabSpec*`} \end{tabular}`

Directly from LaTeX, except that only certain column definitions are allowed: `c`,`l`,`r`, `C{`*Length*`}`, `L{`*Length*`}`, `R{`*Length*`}`, and `|`, where *Length* must be explicitly specified in `em`. `C{`*Length*`}`, `L{`*Length*`}`, `R{`*Length*`}` represent columns of the specified width with paragraph formatting and aligned in the center, left justified, or right justified, respectively.

5 Special Symbols

See also Sect. 2.1.

5.1 Historic Currency Symbols

```
\xGulden
\xReichstaler
\xPfennige
\xPfund
\xGuldenPeriod
\xReichstalerPeriod
\xPfennigePeriod
\xPfundPeriod
```

The `Period` variants include a trailing period. They are intended for use at the end of a sentence, to prevent double periods in case the symbols are not represented by abbreviating words. TODO: Are the `Period` variants really needed?

6 Letters and Annotations

6.1 The `letter` and `annotation` Environments

`\begin{letter}{`*LetterId*`}{`*FromId*`}{`*ToId*`}{`*LocationId*`}{`*Date*`} \end{letter}`
Legacy version: `brief`

Environment for the content of a letter. *LetterId* is the symbolic identifier of the letter. The remaining arguments specify basic meta information about the letter: author, addressee, place and date where and when, respectively, it was written (or sent). *FromId* and *ToId* are symbolic person identifiers. *LocationId* is either a symbolic location identifier, text, or `xwithoutLocation`. *Date* is parsable natural language text of the following form:

Date := *ProperDate*
| ‘nach dem ’ *ProperDate*
| ‘nach ’ *ProperDate*
| ‘vor dem ’ *ProperDate*
| ‘vor ’ *ProperDate*
| ‘zwischen dem ’ *ProperDateOrDay* ‘ und ’ *ProperDate*
| ‘zwischen ’ *ProperDateOrDay* ‘ und ’ *ProperDate*
| *ProperDateOrDay* ‘ - - ’ *ProperDate*
| `xwithoutDate`

ProperDate matches various common forms to specify dates in natural language (see KBSET base module *dates* for details).

TODO: *Date* should consider other languages than German. How would cases be handled where, e.g., the receiver is not known or there are multiple senders or receivers?

TODO: Some semantic aspects seem actually complicated. For example: Date given by the author, researched real date, date when written, or when sent. Is it of relevance to consider such aspects more in detail?

`xwithoutLocation` Legacy version: `xohneort`
`xwithoutDate` Legacy version: `xohnedatum`

Special symbolic identifiers for use as arguments in the *begin* phrase of letter environments. They represent that the place or date, respectively, is not determinable or not specified.

`\begin{annotation}{`*LetterId*`} \end{annotation}`
Legacy version: `kommentar`

Environment for the annotation block to the letter identified by *LetterId*.

TODO: Elements for specifying additional meta information in `letter` or `annotation` environments might be useful (e.g., to specify things like *Box X, Folder Y*).

6.2 Letter-Relative References in Annotations

`\xl{`*PosId*`}{`*Text*`}`
`\xll{`*PosId*`}{`*Text*`}`
`\begin{klist} \end{klist}`
`\kitem{`*PosId*`}`

`\xl` and `\xll` are used within the `\letter` environment. They associate *PosId* with the respective text position. *PosId* is a symbolic identifier which is relative to the respective `\letter` environment. Hence, the same identifier can be used as *PosId* in different letters.

`\xl{`*PosId*`}{`*Text*`}` effects that *Text* is printed in the letter and is additionally stored for use with `\kitem`. `\xl{`*PosId*`}{`*Text*`}` effects just that *Text* is stored for use with `\kitem`, but not that it is printed in the letter.

`\xl` and `\xll` are not permitted in semantic entity markup (Sect. 8). However, semantic entity markup (and inline properties – Sect. 3) is permitted in the *Text* argument of `\xl` and `\xll`.

`\klist` ist a list environment for the annotations that refer to specific positions in the text of letters (*Stellenkommentare*). It is for use in the `annotation` environment. The particular annotations for text positions are expressed wit `\kitem{`*PosId*`}` as list elements. *PosId* refers to a position that had been introduced with `\xl{`*PosId*`}{`*Text*`}` or `\xll{`*PosId*`}{`*Text*`}` in the letter associated with the `annotation` environment. In a print presentation, the list element is then typically introduced with a reference to the position (page, line) of the stored *Text*. A Web presentation would represent the reference as a hyperlink.

6.3 Structuring of the Annotation Environments

`\ksection{`*Text*`}`

Section header for use in annotation environments. Typical such sections (in a German edition) are: *Überlieferung*, *Entstehungsvarianten*, *Anschrift*, *Handschrift*, *Stellenkommentar*.

6.4 Global References to Letters

`\xletterRef{`*LetterId*`}` Legacy version: `\briefref{`*LetterId*`}`
`\xletterTitleRef{`*LetterId*`}` Legacy version: `\brieftitelref{`*LetterId*`}`

LetterId ist a symbolic identifier of a letter (see Sect. 6.1). In a print representation `\xletterRef` outputs the number of the letter (in a continuous numbering of letter environments in the document). `\letterTitleRef` outputs the short title of the letter, that is: *From* an *To*, *Date*.

6.5 Enclosures, Inclosures

`\begin{enclosure}{`*EnclosureId*`}{`*Title*`}` `\end{enclosure}`
Legacy version: `\begin{beilage}{`*Title*`}{`*EnclosureId*`}` `\end{beilage}`

For use in the letter environment to represent an enclosure (*Beilage*).

`\begin{inclosure}` `\end{inclosure}`
Legacy version: `\begin{einschluss}` `\end{einschluss}`

For use in the annotation environment to represent an inclosure (*Einschluss*).

6.6 Special Formatting for Use in Letters

`\xsalute{`*Text*`}` Legacy version: `\xanrede{`*Text*`}`
`\xdate{`*Text*`}` Legacy version: `\xdatum{`*Text*`}`
`\xdateEnd{`*Text*`}` Legacy version: `\xdatumend{`*Text*`}`
`\xcloser{`*Text*`}` Legacy version: `\xschluss{`*Text*`}`

Salutation (*Anrede*), date specifications (*Datumsangaben*) and closer (*Schluss*) in a letter. `\xdateEnd` is for date specifiations at the end of the letter, where they should be displayed with left alignment.

6.7 Symbolic Marks for Use in Annotations

`\xins{`*Text*`}`

Text has been inserted afterwards above or below the line.

`\xinl{`*Text*`}`

Text has been inserted afterwards into the text (in the same line).

`\xmar{`*Text*`}`

Text has been inserted afterwards to the margin (at a side, the top, or the bottom).

`/`
`//`

End of a line or stanza, respectively, in verses.

7 Citations

It is suggested to use *BibLaTeX* as bibliography format with *KLM*. The *Biber* processor can emit XML that can be processed further to generate bibliographies in other formats than LaTeX.

`\xciteAY[`*Text*`]{`*BibId*`}`
`\xciteST[`*Text*`]{`*BibId*`}`
`\xnocite{`*BibId*`}`

`\xciteAY` formats the citation as *Author Year*. `\xciteST` formats the citation as *Author, ShortTitle, Year*. *Author* can there also be the editor (see `Labelname` mechanism of *BibLaTeX*). If `shorttitle` is not specified as attribute for the cited entry, then its `title` attribute is used as *ShortTitle*. `\xnocite` serves like the LaTeXcommand `\nocite` to include an entry that is not explicitly cited into a bibliography.

8 Entity References

\xcorp{*Id*}{*Text*}
\xcorpmain{*Id*}{*Text*}
\xdateEntity{*Id*}{*Text*} TODO: useful?
\xevent{*Id*}{*Text*}
\xeventmain{*Id*}{*Text*}
\xgeo{*Id*}{*Text*}
\xgeomain{*Id*}{*Text*}
\xjournal{*Id*}{*Text*}
\xjournalmain{*Id*}{*Text*}
\xjournalq{*Id*}{*Text*}
\xperson{*Id*}{*Text*}
\xpersonmain{*Id*}{*Text*}
\xsubject{*Id*}{*Text*}
\xsubjectmain{*Id*}{*Text*}
\xwork{*Id*}{*Text*} Legacy version: \xwerk{*Id*}{*Text*}
\xworkmain{*Id*}{*Text*} Legacy version: \xwerkmain{*Id*}{*Text*}
\xworkq{*Id*}{*Text*} Legacy version: \xwerkq{*Id*}{*Text*}

Markup of text phrases that denote a corporation (*Körperschaft oder Institution*), date, event, geographical location, journal, person, subject (*Sachbegriff*) and work, respectively. *Id* is a symbolic identifier of the respective type that is introduced with an assertion specified in Sect. 9.1.

The `main` variants give special importance to the marked-up occurrence, for example as a main reference in the annotations where a person is comprehensively introduced. The main references are typically emphasized (e.g., shown in bold font weight) in indexes. The variants `\xworkq` und `\xjournalq` indicate that *Text* should be quoted as by `\xqwork`.

It is not permitted to nest semantic entity markup. *Text*, can however contain markup for inline properties (Sect. 3).

Note: Occurrences of geographical names as parts of person names (e.g., *Hermann von Pückler-Muskau*, *Hess von Neftenbach*) are not marked up with `xgeo`.

TODO: A single reference to several entities (e.g., *the two brothers*).

9 Fact Bases

Certain kinds of meta information can be represented in *KLM* as fact bases, that is, sets of assertions with a predicate and arguments that can be symbolic identifiers (denoting entities) and text values. We distinguish between two kinds of such assertions, assertions that introduce an identifier and assertions that only express information about (the entities represented by) previously introduced identifiers.

9.1 Assertions That Introduce Identifiers

`\defcorp{`*CorpId*`}{`*Text*`}`
`\defevent{`*EventId*`}{`*Text*`}`
`\defeventSee{`*EventId*`}{`*Text*`}`
`\defgeo{`*GeoId*`}{`*Text*`}`
`\defjournal{`*JournalId*`}{`*Text*`}`
`\defperson{`*PersonId*`}{`*FullPersonName*`}`
`\defsubevent{`*EventId*`}{`*EventId*`}{`*Text*`}`
`\defsubgeo{`*GeoId*`}{`*GeoId*`}{`*Text*`}`
`\defsubject{`*SubjectId*`}{`*Text*`}`
`\defworkThree{`*WorkId*`}{`*PersonId*`}{`*PersonId*`}{`*PersonId*`}{`*Text*`}`
`\defworkTwo{`*WorkId*`}{`*PersonId*`}{`*PersonId*`}{`*Text*`}`
`\defwork{`*WorkId*`}{`*PersonId*`}{`*Text*`}`

These statements *introduce* the identifier supplied as first argument, that is, declare it as a valid reference of the respective type and associate some meaningful information with it. In print presentations, *Text* is typically used in indexes that register places with entity annotations (Sect. 8).

The `Sub` variants specify an entity (the first argument) as a sub-entity of another (the second argument), which may be reflected in two-leveled indexes.

The *FullPersonName* argument of `\defperson` is formed according to special rules, where the first part, *PersonName* matches person name specifications as used by the *GND*:

$$
\begin{array}{rcl}
\mathit{FullPersonName} & := & \mathit{PersonName} \\
 & | & \mathit{PersonName}\ \text{`\ '}\ \mathit{LivingDates} \\
\mathit{PersonName} & := & \mathit{LastNames}\ \text{`, '}\ \mathit{FirstNames} \\
 & | & \mathit{Names}\ \text{`, '}\ \mathit{Territorium}\ \text{`, '}\ \mathit{Titel} \\
\mathit{LivingDates} & := & \text{`('}\ \mathit{Year}\ \text{`- -'}\ \mathit{Year}\ \text{`)'}
\end{array}
$$

Year is a number (without leading zeros) or the symbol `?`. If year of birth as well as year of death are unknown, the form without *LivingDates* is used. Form *Names* ‘, ’ *Territorium* ‘, ’ *Titel* matches, e.g., *Friedrich II., Preußen, König.*

This notation of person names facilitates to associate the *GND* identifier in an automated way, just based on `defperson` statements, and thus to enrich the information on persons from other sources.

Suggestion for practice: In German documents for writers from classical antiquity the common German form may be used, differently from the *preferred* name in the *GND*, instead of the Latinized form (e.g., *Homer* instead of *Homerus*, *Cäsar, Gaius Julius* instead of *Caesar, Gaius Iulius.*

TODO: How to express years B.C.?

TODO: Make correspondence to *GND* more precise.

TODO: Explain the individual entity declarations.

9.2 Assertions That Do Not Introduce Identifiers

`\defcorpSee{`*CorpId*`}{`*Text*`}`
`\defgeoSee{`*GeoId*`}{`*Text*`}`
`\defjournalSee{`*JournalId*`}{`*Text*`}`
`\defpersonSee{`*PersonId*`}{`*Text*`}`
`\defsubjectSee{`*SubjectId*`}{`*Text*`}`

These statements effect index entries with *see* redirections in print representations and may appear as *other Names* of *other Designators* in Web pages representing the associated entities.

TODO: Should `See` interact with `Sub` entries?

`\defheadername{`*Id*`}{`*Text*`}`

`\defheadername` associates with *Id* a string *Text* intended for example to represent the entity in the headers of letters. That is, while `\defperson` specifies the person by a full name, `\defheadername` provides a short name for use in certain contexts such as letter headers.

TODO: Perhaps use a different prefix than *def* for the *assertion* statement in this subsection.

10 Meta-Level Markup Concerning Contributors to the Edition and the Status of the Edition

`\xmeta{`*Key*`}{`*Text*`}`

Permitted only in `letter` and `annotation` environments, where the effect in both environments is the same: Associates *Text* with the letter of the environment and *Key*. Multiple statements for the same letter and *Key* are not permitted.

Supported values for *Key*:
`transliteration`: Creator(s) of the transliteration.
`annotationAuthor`: Author(s) of the annotations.
`status`: Current status of the edition. For example, to indicate that the edition of the letter is not completed but still in progress.

11 Meta-Level Markup Concerning the Editing Process

`\xmissing` Legacy version: `\xfehlt`

Needs revision because some part of the text is still missing.

`\xnotice{`*Text*`}` Legacy version: `\xhinweis{`*Text*`}`

Needs revision, with notice *Text*.

Acknowledgment

The authors thank Baptiste Baumann for helpful suggestions.