# 

Michael Kohlhase, Deyan Ginev Jacobs University, Bremen http://kwarc.info/kohlhase

November 5, 2017

#### Abstract

The smultiling package is part of the STEX collection, a version of TEX/LATEX that allows to markup TEX/LATEX documents semantically without leaving the document format, essentially turning TEX/LATEX into a document format for mathematical knowledge management (MKM).

The smultiling package adds multilinguality support for STEX, the idea is that multilingual modules in STEX consist of a module signature together with multiple language bindings that inherit symbols from it, which also account for cross-language coordination.

# Contents

1	Intr	roduction	3		
	1.1	STEX Module Signatures	3		
<b>2</b>	The	User Interface	3		
	2.1	Multilingual Modules	4		
	2.2	Multilingual Definitions and Cross-referencing Terms	4		
	2.3	Multilingual Views	5		
	2.4	Mathematical Keywords	6		
	2.5	GF Metadata	6		
3	Limitations				
	3.1	General babel Integration	7		
	3.2	PDF links on term references are language-dependent	7		
	3.3	Language-Specific Limitations	8		
4	Imp	plementation	9		
	4.1	Class Options	9		
	4.2	Signatures	9		
	4.3	Language Bindings	10		

4.4	Multilingual Statements and Terms	11
4.5	GF Metadata	11
4.6	Miscellaneneous	12

#### 1 Introduction

We have been using STEX as the encoding for the Semantic Multilingual Glossary of Mathematics (SMGloM; see [Gin+16; SMG]). The SMGloM data model has been taxing the representational capabilities of STEX with respect to multilingual support and verbalization definitions; see [Koh14], which we assume as background reading for this note.

### 1.1 ST<sub>E</sub>X Module Signatures

(Monolingual) STEX had the intuition that the symbol definitions (\symdef and \symvariant) are interspersed with the text and we generate STEX module signatures (SMS \*.sms files) from the STEX files. The SMS duplicate "formal" information from the "narrative" STEX files. In the SMGloM, we extend this idea by making the the SMS primary objects that contain the language-independent part of the formal structure conveyed by the STEX documents and there may be multiple narrative "language bindings" that are translations of each other – and as we do not want to duplicate the formal parts, those are inherited from the SMS rather than written down in the language binding itself. So instead of the traditional monolingual markup in Figure 1, we we now advocate the divided style in Figure 2.

```
\begin{module}[id=foo]
\symdef{bar}{BAR}
\begin{definition}[for=bar]
  A \defiii{big}{array}{raster} ($\bar$) is a\ldots, it is much
  bigger than a \defiii[sar]{small}{array}{raster}.
\end{definition}
\end{module}
```

Example 1: A module with definition in monolingual STEX

We retain the old module environment as an intermediate stage. It is still useful for monolingual texts. Note that for files with a module, we still have to extract \*.sms files. It is not completely clear yet, how to adapt the workflows. We clearly need a lmh or editor command that transfers an old-style module into a new-style signature/binding combo to prepare it for multilingual treatment.

#### 2 The User Interface

langfiles

EdN:1

The smultiling package accepts the langfiles option that specifies – for a module  $\langle mod \rangle$  that the module signature file has the name  $\langle mod \rangle$ .tex and the language bindings of language with the ISO 639 language specifier  $\langle lang \rangle$  have the file name  $\langle mod \rangle$ .  $\langle lang \rangle$ .tex.<sup>1</sup>

<sup>&</sup>lt;sup>1</sup>Ednote: implement other schemes, e.g. the onefile scheme.

```
EdN:2
```

```
\usepackage{multiling}
\begin{modsig}{foo}
  \symdef{bar}{BAR}
  \symi[gfc=N]{sar}
\end{modsig}
\begin{modnl}[creators=miko,primary]{foo}{en}
  \begin{definition}
    A \defiii[bar] {big} {array} {raster} ($\bar$) is a \ldots, it is much bigger
    than a \defiii[sar]{small}{array}{raster}.
  \end{definition}
\end{modnl}
\begin{modnl}[creators=miko]{foo}{de}
  \begin{definition}
   Ein \defiii[bar]{gro"ses}{Feld}{Raster} ($\bar$) ist ein\ldots, es
    ist viel gr"o"ser als ein \defiii[sar]{kleines}{Feld}{Raster}.
  \end{definition}
\end{modnl}
```

Example 2: Multilingual STFX for Figure 1.

#### 2.1 Multilingual Modules

modsig

There the modsig environment works exactly like the old module environment, only that the id attribute has moved into the required argument – anonymous module signatures do not make sense.

modnl

The modnl environment takes two arguments the first is the name of the module signature it provides language bindings for and the second the ISO 639 language specifier of the content language. We add the primary key modnl, which can specify the primary language binding (the one the others translate from; and which serves as the reference in case of translation conflicts).<sup>2</sup>

\symi

\symi\*

\symii \symiii \symiv There is another difference in the multilingual encoding: All symbols are introduced in the module signature, either by a \symdef or the new \symi macro. \symi[\langle keys\] {\langle name}\ takes a symbol name \langle name\ as an argument and reserves that name. The variant \symi\*[\langle keys\] {\langle name}\ declares \langle name\ to be a primary symbol; see [Koh14] for a discussion. STEX provides variants \symii, \symii, \symiii, and \symiv - and their starred versions - for multi-part names. The key-value interface \langle keys\ does not have any effect on the LATEX rendering, it can be used to embed metadata. See for instance Subsection 2.5.

#### 2.2 Multilingual Definitions and Cross-referencing Terms

We do not need a new infrastructure for defining mathematical concepts, only the realization that symbols are language-independent. So we can use symbols for the coordination of corresponding verbalizations. As the example in Figure 2 already

 $<sup>^2\</sup>mathrm{EdNote}$ : @Hang: This needs to be implemented in LaTeXML

shows, we can just specify the symbol name in the optional argument of the \defi macro to establish that the language bindings provide different verbalizations of the same symbol.

For multilingual term references the situtation is more complex: For single-word verbalizations we could use \atrefi for language bindings. Say we have introduced a symbol foo in English by \defi{foo} and in German by \defi[foo]{foo}. Then we can indeed reference it via \trefi{foo} and \atrefi{foo} and \atrefi{foo}. But one the one hand this blurs the distinction between translation and "linguistic variants" and on the other hand does not scale to multi-word compounds as bar in Figure 2, which we would have to reference as \atrefiii{gro"ses Feld Raster}{bar}. To avoid this, the smultiling package provides the new macros \mtrefi, \mtrefii, and \mtrefiii for multilingual references. Using this, we can reference bar as \mtrefiii[?bar]{gro"ses}{Feld}{Raster}, where we use the (up to three) mandatory arguments to segment the lexical constituents.

\mtref

The first argument it syntactically optional to keep the parallelism to  $\ensuremath{\mbox{\tt \*tref*}}$  it specifies the symbol via its name  $\langle name \rangle$  and module name  $\langle mod \rangle$  in a MMT URI  $\langle mod \rangle$ ?  $\langle name \rangle$ . Note that MMT URIs can be relative:

- 1. foo?bar denotes the symbol bar from module foo
- foo the module foo (the symbol name is induced from the remaining arguments of \mtref\*)
- 3. ?bar specifies symbol bar from the current module

Note that the number suffix i/ii/iii/iv indicates the number of words in the actual language binding, not in the symbol name as in \atref\*.

Finally note that hyperlinks on term references only have information on the underlying symbol and module names – i.e. signature information – and we need to cross-reference into the language bindings. To do this, we need to know the base language of the document. To ensure basic functionality we set this to en and provide the \sTeXlanguage macro to set it.

\sTeXlanguage

#### 2.3 Multilingual Views

Views receive a similar treatment as modules in the smultiling package. A multilingual view consists of

viewsig

 a view signature marked up with the viewsig environment. This takes three required arguments: a view name, the source module, and the target module. The optional first argument is for metadata (display, title, creators, and contributors) and load information (loadfrom and loadto) and

viewnl

2. multiple language bindings marked up by the viewnl environment, which takes two required arguments: the view name and the language specifier. The optional first key/value argument takes the same keys as viewsig except the last two.

```
EdN:3
```

```
\begin{viewsig}[creators=miko] {norm-metric}{metric-space}{norm}
   \vassign{base-set}{base-set}
   \fassign{x,y}{\metric{x,y}}{\norm{x-y}}
\end{viewsig}
```

Views have language bindings just as modules do, in our case, we have

# 2.4 Mathematical Keywords

For translations of the mathematical keywords, the statements and sproofs packages in STEX define special language definition files, e.g. statements-ngerman.ldf.<sup>34</sup> There is currently only very limited support for this.

#### 2.5 GF Metadata

Several STEX macros and environments allow keys for syntactical information about the objects declared.

gfc

The symbol-declaring macros \symi and friends as well as \symdef allow gfc key allows to specify the grammatical category in terms of the Resource Grammar of the Grammatical Framework [GFR].

The verbalization-defining macros \defi and friends allow the gfa (GF apply) and gfl (GF linearization) keys.

A definiendum of the form \defii[gfa=mkN]{empty}{set} generates the GF linearization empty\_set = mkN "empty set". Some what less conveniently, \defii[name=datum,gfl={mkN "Datum", "Daten"}{Datum} can be used if the GF linearization is more complex than simply applying a "make command" to the verbalization.

#### 3 Limitations

We list the limitations of the smultiling package.

<sup>&</sup>lt;sup>3</sup>EdNote: say more about this

 $<sup>^4\</sup>mathrm{EdNote}\colon$  There is the translator package which belongs to beamer, maybe we should switch to that.

#### 3.1 General babel Integration

There is currently no integration with the babel package that handles language-specific aspects in LATEX. In particular, selecting the right language must be done manually. In particular, the example from Figure ?? would really have the form given in Figure 3 – see the \usepackage[usenglish,ngerman]{babel} in line 2, and the \selectlanguage statements in lines 6 and 13.

```
\usepackage{multiling}
\usepackage[usenglish,ngerman]{babel}% babel support
\begin{modsig}{foo}
  \symdef{bar}{BAR}
  \symi{sar}
\end{modsig}
\selectlanguage{english}% english version follows
\begin{modnl}[creators=miko,primary]{foo}{en}
  \begin{definition}
    A \defiii[bar]{big}{array}{raster} ($\bar$) is a\ldots, it is much bigger
    than a \defiii[sar]{small}{array}{raster}.
  \end{definition}
\end{modnl}
\selectlanguage{german}% german umlauts please
\begin{modnl}[creators=miko]{foo}{de}
  \begin{definition}
    Ein \defiii[bar]{gro"ses}{Feld}{Raster} ($\bar$) ist ein\ldots, es
    ist viel gr"o"ser als ein \defiii[sar]{kleines}{Feld}{Raster}.
  \end{definition}
\end{modnl}
```

Example 3: Multilingual STFX with babel

For the langfiles setup, which assumes that module signatures and language bindings are in separate files, babel integration can be simplified by providing a language-specific preamble file with \usepackage{\language}}{\text{babel}} which is pre-pended to all language binding files when formatted. This preamble can also contain the other language-specific packages (e.g. for font encodings, etc.).

#### 3.2 PDF links on term references are language-dependent

Given the langfiles mode, we need the intended language to generate PDF links on term references. But we cannot infer this for top-level "papers" (we do in the language bindings). So it has to be specified via \sTexlanguage, and we do not really had a way to check that it is. Unfortunately, the only place it would be natural to do so is in \mod@component, but the \PackageError there had to be commented out, since it leads to serious errors. Thus we set the language to en by default, which is sub-optimal. Maybe there is a way to infer the document language from the babel settings.

# 3.3 Language-Specific Limitations

Some languages have more problems than others

Turkish makes = an active character (to give better spacing); this interacts unfavourably with the keyval package which needs = as key/value separator (and gives it a different category code). Therefore we need to prohibit this by restricting the shorthands option: use \usepackage[turkish,shorthands=:!]{babel}.

Chinese needs special fonts and xelatex<sup>5</sup>.

EdN:5

 $<sup>^5\</sup>mathrm{EdNote}$  get Jinbo to document this

# **Implementation**

#### Class Options 4.1

```
1 (*sty)
2 \newif\if@smultiling@mh@\@smultiling@mh@false
3 \DeclareOption{mh}{\@smultiling@mh@true}
4 \newif\if@langfiles\@langfilesfalse
5 \DeclareOption{langfiles}{\@langfilestrue}
6 \DeclareOption*{\PassOptionsToPackage{\CurrentOption}{modules}}
7 \ProcessOptions
   We load the packages referenced here.
8 \if@smultiling@mh@\RequirePackage{smultiling-mh}\fi
9 \RequirePackage{etoolbox}
10 \RequirePackage{structview}
```

#### Signatures

modsig

The modsig environment is just a layer over the module environment. We also redefine macros that may occur in module signatures so that they do not create markup. Finally, we set the flag  $\mbox{mod}\mbox{@multiling}$  to true.

```
11 \newenvironment{modsig}[2][]{\def\@test{#1}%
```

- 12 \ifx\@test\@empty\begin{module}[id=#2]\else\begin{module}[id=#2,#1]\fi%
- 13 \expandafter\gdef\csname mod@#2@multiling\endcsname{true}
- 14 \ignorespacesandpars}
- 15 {\end{module}\ignorespacesandparsafterend}

\mod@component

We redefine the macro from the modules package that computes the module component identifier for external links on term references. If  $\mbox{\em mod }\mbox{\em cmultiling}$ is true, then we make the component identifier  $\langle lanq \rangle$ , which can be customized by the next macro below.

```
16 \renewcommand\mod@component[1]{%
```

- 17 \expandafter\ifx\csname mod@#1@multiling\endcsname\@true%
- 18 \@ifundefined{smultiling@language}{}
- 19 % for some reason this error message bombs big time; so we leave it out.
- 20 % {\PackageError{smultiling}%
- {No document language specified for term reference links}
- 22 % {use \protect\sTeXlanguage to specify it!}}
- 23 {.\smultiling@language}%
- 24 \fi}

\sTeXlanguage

This macro sets the internal flag \smultiling@language, we set the default to en, since otherwise hyper-references on term references do not work.

- 25 \newcommand\sTeXlanguage[1]{\def\smultiling@language{#1}}
- 26 \sTeXlanguage{en}

viewsig The viewsig environment is just a layer over the view environment with the keys suitably adapted.

27 \newenvironment{viewsig}[4][]{\def\@test{#1}\ifx\@test\@empty%

```
\begin{view}[id=#2,ext=tex]{#3}{#4}\else\begin{view}[id=#2,#1,ext=tex]{#3}{#4}\fi%
             \ignorespacesandpars}
         29
             {\end{view}\ignorespacesandparsafterend}
         has a starred form for primary symbols. The key/value interface has no effect on
 \@sym*
         the LATEX side. We read the to check whether only allowed ones are used. We also
         add the gfc key to \symdef, since that makes a
         31 \newcommand\symi{\@ifstar\@symi@star\@symi}
         32 \newcommand\@symi[2][]{\metasetkeys{symi}{#1}%
             \if@importing\else\par\noindent Symbol: \textsf{#2}\fi\ignorespacesandpars}
         34 \newcommand\@symi@star[2][]{\metasetkeys{symi}{#1}%
             \if@importing\else\par\noindent Primary Symbol: \textsf{#2}\fi\ignorespacesandpars}
         36 \newcommand\symii{\@ifstar\@symii@star\@symii}
         37 \newcommand\@symii[3][]{\metasetkeys{symi}{#1}%
             \if@importing\else\par\noindent Symbol: \textsf{#2-#3}\fi\ignorespacesandpars}
         39 \newcommand\@symii@star[3][]{\metasetkeys{symi}{#1}%
             \if@importing\else\par\noindent Primary Symbol: \textsf{#2-#3}\fi\ignorespacesandpars}
         41 \newcommand\symiii{\@ifstar\@symiii@star\@symiii}
         42 \newcommand\@symiii[4][]{\metasetkeys{symi}{#1}%
             \if@importing\else\par\noindent Symbol: \textsf{#2-#3-#4}\fi\ignorespacesandpars}
         44 \newcommand\@symiii@star[4][]{\metasetkeys{symi}{#1}%
             \if@importing\else\par\noindent Primary Symbol: \textsf{#2-#3-#4}\fi\ignorespacesandpars}
         46 \newcommand\symiv{\@ifstar\@symiv@star\@symiv}
         47 \newcommand\@symiv[5][]{\metasetkeys{symi}{#1}%
             \if@importing\else\par\noindent Symbol: \textsf{#2-#3-#4-#5}\fi\ignorespacesandpars}
         49 \newcommand\@symiv@star[5][]{\metasetkeys{symi}{#1}%
             \if@importing\else\par\noindent Primary Symbol: \textsf{#2-#3-#4-#5}\fi\ignorespacesandpars}
               Language Bindings
         4.3
modnl:*
         51 \addmetakey{modnl}{load}
         52 \addmetakey*{modnl}{title}
         53 \addmetakey*{modnl}{creators}
```

```
54 \addmetakey*{modnl}{contributors}
55 \addmetakey{modnl}{srccite}
56 \addmetakey{modnl}{primary}[yes]
```

The module environment is just a layer over the module environment and the \importmodule macro with the keys and language suitably adapted.

```
57 \newenvironment{modnl}[3][]{\metasetkeys{modnl}{#1}%
```

- \def\@test\#1\\ifx\@test\@empty\begin\module\[id=#2.#3]\else\begin\module\[id=#2.#3,#1]\fi% 58
- \def\smultiling@language{#3}% 59
- \if@langfiles\importmodule[load=#2,ext=tex]{#2}\else 60
- \ifx\modnl@load\@empty\importmodule{#2}\else\importmodule[ext=tex,load=\modnl@load]{#2}\fi% 61
- \fi% 62
- \ignorespacesandpars}
- 64 {\end{module}\ignorespacesandparsafterend}

```
EdN:6
```

viewnl The viewnl environment is just a layer over the view environment with the keys and language suitably adapted.<sup>6</sup>

```
65 \newenvironment{viewnl}[5][]{\def\@test{#1}\ifx\@test\@empty%
```

- 66 \begin{view}[id=#2.#3,ext=tex]{#4}{#5}\else%
- 67 \begin{view}[id=#2.#3,#1,ext=tex]{#4}{#5}\fi%
- 68 \ignorespacesandpars}
- 69 {\end{view}\ignorespacesandparsafterend}

#### 4.4 Multilingual Statements and Terms

```
we first first define an auxiliary conditional \@instring that checks of ? is in the
first argument. \mtrefi uses it, if there is one, it just calls \termref, otherwise
it calls \@mtrefi, which assembles the \termref after splitting at the ?.

70 \def\@instring#1#2{TT\fi\begingroup\edef\x{\endgroup\noexpand\in@{#1}{#2}}\x\ifin@}
71 \def\@mtref#1?#2\relax{\@mtref{#1}{#2}}

72 \newcommand\@mtref[3]{\def\@@cd{#1}\def\@@name{#2}%
73 \ifx\@@cd\@empty%
```

74 \ifx\@@name\@empty\termref[]{#3}\else\termref[name=\@@name]{#3}\fi% 75 \else%

76 \ifx\@Qname\Qempty\termref[cd=\QQcd]{#3}\else\termref[cd=\QQcd,name=\QQname]{#3}\fi% 77 \fi}

78 \newcommand\mtref[2][]{\if\@instring{?}{#1}\@mtref #1\relax{#2}\else\termref[cd=#1]{#2}\fi}

#### \mtrefi\*

```
79 \newcommand\mtrefi[2][]{\if\@instring{?}{#1}\@mtref #1\relax{#2}%
80 \leq f(d=#1){#2}\fi
81 \newcommand\mtrefis[2][]{\mtrefi[#1]{#2s}}
82 \newcommand\Mtrefi[2][]{\if\@instring{?}{#1}\@mtref #1\relax{\capitalize{#2}}%
83 \else\termref[cd=#1]{\capitalize{#2}}\fi}
84 \newcommand\Mtrefis[2][]{\Mtrefi[#1]{#2s}}
85 \newcommand\mtrefii[3][]{\mtrefi[#1]{#2 #3}}
86 \newcommand\mtrefiis[3][]{\mtrefi[#1]{#2 #3s}}
87 \newcommand\Mtrefii[3][]{\Mtrefi[#1]{#2 #3a}}
88 \newcommand\Mtrefiis[3][]{\Mtrefi[#1]{#2 #3s}}
89 \newcommand\mtrefiii[4][]{\mtrefi[#1]{#2 #3 #4}}
90 \newcommand\Mtrefiiis[4][]{\Mtrefi[#1]{#2 #3 #4s}}
91 \newcommand\Mtrefiii[4][]{\Mtrefi[#1]{#2 #3 #4}}
92 \newcommand\mtrefiiis[4][]{\mtrefi[#1]{#2 #3 #4s}}
93 \newcommand\mtrefiv[5][]{\mtrefi[#1]{#2 #3 #4 #5}}
94 \newcommand\mtrefivs[5][]{\mtrefi[#1]{#2 #3 #4 #5s}}
95 \newcommand\Mtrefiv[5][]{\Mtrefi[#1]{#2 #3 #4 #5}}
```

#### 4.5 GF Metadata

gfc We add the gfc key to various symbol declaration macros.

96 \newcommand\Mtrefivs[5][]{\Mtrefi[#1]{#2 #3 #4 #5s}}

 $<sup>^6\</sup>mathrm{EdNote}$ : MK: we have to do something about the if@langfiles situation here. But this is non-trivial, since we do not know the current path, to which we could append . $\langle lang \rangle$ !

```
97 \addmetakey{symi}{gfc}

98 \addmetakey{symdef}{gfc}%

gfa/1

99 \addmetakey{definiendum}{gfa}

100 \addmetakey{definiendum}{gf1}
```

#### 4.6 Miscellaneneous

the  $\t$  macro (to-translate) is used to mark untranslated stuff. We need a better LATEXML treatment of this eventually that is integrated with Math Hub.info.

# References

- [GFR] B. Bringert, T. Hallgren, and A. Ranta. *GF Resource Grammar Library: Synopsis*. URL: http://www.grammaticalframework.org/lib/doc/synopsis.html (visited on 09/27/2017).
- [Gin+16] Deyan Ginev et al. "The SMGloM Project and System. Towards a Terminology and Ontology for Mathematics". In: Mathematical Software ICMS 2016 5th International Congress. Ed. by Gert-Martin Greuel et al. Vol. 9725. LNCS. Springer, 2016. DOI: 10.1007/978-3-319-42432-3. URL: http://kwarc.info/kohlhase/papers/icms16-smglom.pdf.
- [Koh14] Michael Kohlhase. "A Data Model and Encoding for a Semantic, Multilingual Terminology of Mathematics". In: Intelligent Computer Mathematics. Conferences on Intelligent Computer Mathematics. (Coimbra, Portugal, July 7–11, 2014). Ed. by Stephan Watt et al. LNCS 8543. Springer, 2014, pp. 169–183. ISBN: 978-3-319-08433-6. URL: http://kwarc.info/kohlhase/papers/cicm14-smglom.pdf.
- [SMG] SMGloM Glossary. URL: http://mathhub.info/mh/glossary (visited on 04/21/2014).