— sTeX Blue Note* —
Rethinking Modules and Semantic Macros in sTeX

Michael Kohlhase
Computer Science, Jacobs University.de

April 3, 2014

Abstract

In this note, we document the state of rethinking the sTeX infrastructure in terms of the SMGloM.

1 Introduction

We have been using sTeX as the encoding for the Semantic Multilingual Glossary of Mathematics (SMGloM; see [Gin+14]). 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.

2 Mixed Presentation/Content Markup

Currently, sTeX produces content markup in the OpenMath encoding. But often sTeX formulae often contain bits of presentational LaTeX, which LaTeXML has to convert into OpenMath heuristically, which often leads to non-optimal results. Therefore we want to rethink the representation of formulae, instead of insisting on homogeneous content markup in OpenMath, we switch to MathML allow mixed presentation/content MathML, which conforms much more closely to user input (preserving presentational bits) and postpones full semantification to later stages of processing. Let us make an example: consider the formula $(a+b)^n$ encoded das `\exp{a+b}n`, where we have a semantic macro `\exp` defined by `\symdef{exp}[2]{#1^{#2}}` in module `arith` Then we should create

```
<math>
  <apply>
    <csymbol cd="arith">exp</csymbol>
    <mrow><ci>a</ci><mo>+</mo><ci>b</ci></mrow>
    <ci>n</ci>
  </apply>
</math>
```

Note that MathML does indeed allow to freely mix content and presentation MathML, here we have an application produced by the semantic macro `\exp` applied to the presentational $a+b$, where a and b are "content identifiers".

A side effect of the switch to MathML is that complex variable names are much nicer in MathML: x_5 is just

```
<ci name="x5"><msub><mi>x</mi><mn>5</mn></msub></ci>
```

*Inspired by the "blue book" in Alan Bundy's group at the University of Edinburgh, sTeX blue notes, are documents used for fixing and discussing ϵ-baked ideas in projects by the sTeX group (see `http://github.com/KWARC/sTeX`). Unless specified otherwise, they are for project-internal discussions only. Please only distribute outside the sTeX group after consultation with the author.

3 sTeX 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[1] 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

Listing 1: Old-Style sTeX

```
\begin{module}[id=foo]
\symdef{bar}{BAR}
\begin{definition}[for=bar]
  A \defiii{big}{array}{raster} ($\bar$) is a\ldots
\end{definition}
\end{module}
```

we now advocate the divided style in Listing 2[1]. 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. The `langbind` 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 to the optional argument of `langbind`, which can specify the primary language binding (the one the others translate from; and which serves as the reference in case of translation conflicts). EdN:1

Listing 2: New-Style sTeX

```
\begin{modsig}{foo}
\symdef{bar}{BAR}
\end{modsig}

\begin{langbind}[creators=miko,primary]{foo}{en}
\begin{definition}[for=bar]
  A \defiii{big}{array}{raster} ($\bar$) is a\ldots
\end{definition}
\end{langbind}
```

We retain the old `module` environment as an intermediate stage (during the)

4 Conclusion

References

[Gin+14] Deyan Ginev et al. "The SMGLoM Project and System". submitted to CICM 2014. 2014. URL: http://kwarc.info/kohlhase/submit/cicm14-smglom-system.pdf.

[Koh14] Michael Kohlhase. "A Data Model and Encoding for a Semantic, Multilingual Glossary of Mathematics". submitted to CICM 2014. 2014. URL: http://kwarc.info/kohlhase/submit/cicm14-smglom-datamdl.pdf.

[1] Thanks to Deyan Ginev for realizing this.

[1] EdNote: MK: the names of the environments are still very much in the air. "modsig" I rather like, but "langbind" is terrible