Semantic Markup in TEX/LATEX

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

April 7, 2016

Abstract

We present a collection of TEX macro packages that allow to markup TEX/LATEX documents semantically without leaving the document format, essentially turning TEX/LATEX into a document format for mathematical knowledge management (MKM).

Contents

1 Introduction

The last few years have seen the emergence of various content-oriented XML-based, content-oriented markup languages for mathematics on the web, e.g. OpenMath [Bus+04], content MathML [Aus+03], or our own OMDoc [Koh06]. These representation languages for mathematics, that make the structure of the mathematical knowledge in a document explicit enough that machines can operate on it. Other examples of content-oriented formats for mathematics include the various logic-based languages found in automated reasoning tools (see [RV01] for an overview), program specification languages (see e.g. [Ber89]).

The promise if these content-oriented approaches is that various tasks involved in "doing mathematics" (e.g. search, navigation, cross-referencing, quality control, user-adaptive presentation, proving, simulation) can be machine-supported, and thus the working mathematician is relieved to do what humans can still do infinitely better than machines: The creative part of mathematics — inventing interesting mathematical objects, conjecturing about their properties and coming up with creative ideas for proving these conjectures. However, before these promises can be delivered upon (there is even a conference series [Mkm] studying "Mathematical Knowledge Management (MKM)"), large bodies of mathematical knowledge have to be converted into content form.

Even though MathML is viewed by most as the coming standard for representing mathematics on the web and in scientific publications, it has not not fully taken off in practice. One of the reasons for that may be that the technical communities that need high-quality methods for publishing mathematics already have an established method which yields excellent results: the TEX/LATEX system: and a large part of mathematical knowledge is prepared in the form of TEX/LATEX documents.

TEX [Knu84] is a document presentation format that combines complex page-description primitives with a powerful macro-expansion facility, which is utilized in LATEX (essentially a set of TEX macro packages, see [Lam94]) to achieve more content-oriented markup that can be adapted to particular tastes via specialized document styles. It is safe to say that LATEX largely restricts content markup to the document structure[1], and graphics, leaving the user with the presentational TEX primitives for mathematical formulae. Therefore, even though LATEX goes a great step into the direction of an MKM format, it is not, as it lacks infrastructure for marking up the functional structure of formulae and mathematical statements, and their dependence on and contribution to the mathematical context.

1.1 The XML vs. TEX/LATEX Formats and Workflows

MathML is an XML-based markup format for mathematical formulae, it is standardized by the World Wide Web Consortium in [Aus+03], and is supported by the major browsers. The MathML format comes in two integrated components: presentation MathML presentation MathML and content MathML content

[1] supplying macros e.g. for sections, paragraphs, theorems, definitions, etc.

MathML. The former provides a comprehensive set of layout primitives for presenting the visual appearance of mathematical formulae, and the second one the functional/logical structure of the conveyed mathematical objects. For all practical concerns, presentation MathML is equivalent to the math mode of TEX. The text mode facilitates of TEX (and the multitude of LATEX classes) are relegated to other XML formats, which embed MathML.

The programming language constructs of TEX (i.e. the macro definition facilities[2]) are relegated to the XML programming languages that can be used to develop language extensions. transformation language XSLT [Dea99; Kay00] or proper XML-enabled The XML-based syntax and the separation of the presentational-, functional- and programming/extensibility concerns in MathML has some distinct advantages over the integrated approach in TEX/LATEX on the services side: MathML gives us better

- integration with web-based publishing,
- accessibility to disabled persons, e.g. (well-written) MathML contains enough structural information to supports screen readers.
- reusability, searchabiliby and integration with mathematical software systems (e.g. copy-and-paste to computer algebra systems), and
- validation and plausibility checking.

On the other hand, TEX/LATEX/s adaptable syntax and tightly integrated programming features within has distinct advantages on the authoring side:

- The TEX/LATEX syntax is much more compact than MathML, and if needed, the community develops LATEX packages that supply new functionality in with a succinct and intuitive syntax.
- The user can define ad-hoc abbreviations and bind them to new control sequences to structure the source code.
- The TEX/LATEX community has a vast collection of language extensions and best practice examples for every conceivable publication purpose and an established and very active developer community that supports these.
- There is a host of software systems centered around the TEX/LATEX language that make authoring content easier: many editors have special modes for LATEX, there are spelling/style/grammar checkers, transformers to other markup formats, etc.

In other words, the technical community is is heavily invested in the whole workflow, and technical know-how about the format permeates the community. Since all of this would need to be re-established for a MathML-based workflow, the technical community is slow to take up MathML over TEX/LATEX, even in light of the advantages detailed above.

[2]We count the parser manipulation facilities of TEX, e.g. category code changes into the programming facilities as well, these are of course impossible for MathML, since it is bound to XML syntax.

1.2 A LaTeX-based Workflow for XML-based Mathematical Documents

An elegant way of sidestepping most of the problems inherent in transitioning from a LaTeX-based to an XML-based workflow is to combine both and take advantage of the respective advantages.

The key ingredient in this approach is a system that can transform TeXLaTeX documents to their corresponding XML-based counterparts. That way, XML-documents can be authored and prototyped in the LaTeX workflow, and transformed to XML for publication and added-value services, combining the two workflows.

There are various attempts to solve the TeX/LaTeX to XML transformation problem (see [Sta+09] for an overview); the most mature is probably Bruce Miller's LaTeXML system [LTX]. It consists of two parts: a re-implementation of the TeX analyzer with all of it's intricacies, and a extensible XML emitter (the component that assembles the output of the parser). Since the LaTeX style files are (ultimately) programmed in TeX, the TeX analyzer can handle all TeX extensions, including all of LaTeX. Thus the LaTeXML parser can handle all of TeX/LaTeX, if the emitter is extensible, which is guaranteed by the LaTeXML binding language: To transform a TeX/LaTeX document to a given XML format, all TeX extensions[3] must have "LaTeXML bindings"binding, i.e. a directive to the LaTeXML emitter that specifies the target representation in XML.

1.3 Generating OMDoc from sTeX

The sTeX packages (see Section 2) provide functionalities for marking up the functional structure of mathematical documents, so that the LaTeX sources contain enough information that can be exported to the OMDoc format (Open Mathematical Documents; see [Koh06]). For the actual transformation, we use a LaTeXML plugin [L2O] that provides the LaTeXML bindings for the sTeX packages.

1.4 Conclusion

The sTeX collection provides a set of semantic macros that extends the familiar and time-tried LaTeX workflow in academics until the last step of Internet publication of the material. For instance, an SMGloM module can be authored and maintained in LaTeX using a simple text editor, a process most academics in technical subjects are well familiar with. Only in a last publishing step (which is fully automatic) does it get transformed into the XML world, which is unfamiliar to most academics.

Thus, sTeX can serve as a conceptual interface between the document author and MKM systems: Technically, the semantically preloaded LaTeX documents are transformed into the (usually XML-based) MKM representation formats, but conceptually, the ability to semantically annotate the source document is sufficient.

[3] i.e. all macros, environments, and syntax extensions used int the source document

The sTeX macro packages have been validated together with a case study [Koh04], where we semantically preload the course materials for a two-semester course in Computer Science at Jacobs University Bremen and transform them to the OMDoc MKM format.

1.5 Licensing, Download and Setup

The sTeX packages are licensed under the LaTeX Project Public License [Pro07], which basically means that they can be downloaded, used, copied, and even modified by anyone under a set of simple conditions (e.g. if you modify you have to distribute under a different name).

1.5.1 The sTeX Distribution

The sTeX packages and classes are available from the Comprehensive TeX Archive Network (CTAN [CTAN]) and are part of the primary TeX/LaTeX distributions (e.g. TeXlive [Tex] and MikTeX [Mik]). The development version is on GitHub [sTeX], it can cloned or forked from the repository URL

`https://github.com/KWARC/sTeX.git`

It is usually a good idea to enlarge the internal memory allocation of the TeX/LaTeXexecutables. This can be done by adding the following configurations in `texmf.cnf` (or changing them, if they alreday exist). Note that you will probably need `sudo` to do this.

```
max_in_open = 50        % simultaneous input files and error insertions,
param_size = 20000      % simultaneous macro parameters, also applies to MP
nest_size = 1000        % simultaneous semantic levels (e.g., groups)
stack_size = 10000      % simultaneous input sources
main_memory = 12000000
```

After that, you have to run the

```
sudo fmtutil-sys --all
```

With this installation using sTeX is as painless as using LaTeX, just make sure the sTeX distribution is where `latex` can find it and run `pdflatex` over the main file.

1.5.2 The sTeX Plugin for LaTeXML

For the OMDoc transformation of sTeX documents we use a LaTeXML plugin that provides the LaTeXML bindings for the sTeX packages. For installation and setup follow the instructions at [L2O][1] EdN:1

[1] EdNote: We are working on a CPAN submission that should make installations painless.

2 The Packages of the sTeX Collection

In the following, we will shortly preview the packages and classes in the sTeX collection. They all provide part of the solution of representing semantic structure in the TeX/LaTeX workflow. We will group them by the conceptual level they address. Figure 1 gives an overview.

2.1 The sTeX Distribution

The `stex` package provides `stex.sty` that just loads all packages below and passes around the package options accordingly and `stex-logo.sty` that provides the
\sTeX macros `\sTeX` and `\stex` that typeset the sTeX logo.
\stex

2.2 Content Markup of Mathematical Formulae in TeX/LaTeX

2.2.1 `cmath`: Building Content Math Representations

The `cmath` packge (see [Koh16a]) supplies an interface for building content math representations. It gives special macros for marking up variables, applications and bindings.

2.2.2 `presentation`: Flexible Presentation for Semantic Macros

The `presentation` package (see [KG16]) supplies an infrastructure that allows to specify the presentation of semantic macros, including preference-based bracket elision. This allows to markup the functional structure of mathematical formulae without having to lose high-quality human-oriented presentation in LaTeX. Moreover, the notation definitions can be used by MKM systems for added-value services, either directly from the sTeX sources, or after translation.

2.3 Mathematical Statements

2.3.1 `statements`: Extending Content Macros for Mathematical Notation

The `statements` package (see[Koh16e]) provides semantic markup facilities for mathematical statements like Theorems, Lemmata, Axioms, Definitions, etc. in sTeX files. This structure can be used by MKM systems for added-value services, either directly from the sTeX sources, or after translation.

2.3.2 `sproof`: Extending Content Macros for Mathematical Notation

The `sproof` package (see [Koh16d]) supplies macros and environment that allow to annotate the structure of mathematical proofs in sTeX files. This structure can be used by MKM systems for added-value services, either directly from the sTeX sources, or after translation.

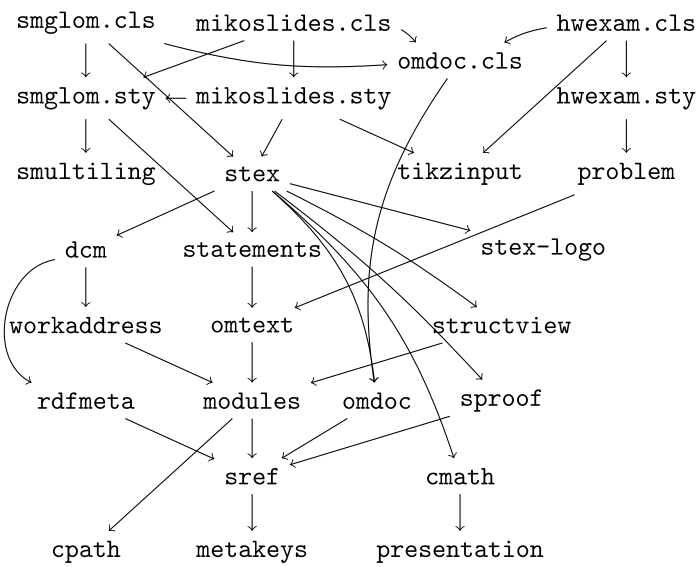

Figure 1: The sTeX packages and their dependencies.

2.3.3 omtext: Mathematical Text

EdN:2 [2]

2.4 Context Markup for Mathematics

2.4.1 modules: Extending Content Macros for Mathematical Notation

The `modules` package (see [KGA16]) supplies a definition mechanism for semantic macros and a non-standard scoping construct for them, which is oriented at the semantic dependency relation rather than the document structure. This structure can be used by MKM systems for added-value services, either directly from the sTeX sources, or after translation. A side effect of this is that

2.4.2 smultiling: Multilingual Mathematical Modules

EdN:3 [3]

2.4.3 structview: Structures and Views

EdN:4 [4]

[2] EdNote: say something
[3] EdNote: Say something
[4] EdNote: Say something

2.5 Mathematical Document Classes

2.5.1 OMDoc Documents

EdN:5 The omdoc package provides an infrastructure that allows to markup OMDoc documents in LaTeX. It provides omdoc.cls, a class with the and omdocdoc.sty[5]

2.5.2 hwexam: Homeworks and Exams

The hwexam package [Koh16b] provides hwexam.cls and hwexam.sty for marking up homework assignments, and exams. The content markup strategy employed in sTeX allows to speficy – and profit from – administrative metadata such as time and point counts. This package relies on the problem package [Koh16c] which provides markup for problems, hints, and solutions.

2.5.3 mikoslides: Slides and Course Notes

The mikoslides package provides a document class from which we can generate both course slides and course notes in a transparent way.

2.6 Metadata

2.6.1 rdfmeta: RDFa Metadata for sTeX

EdN:6 [6]

2.6.2 dcm: Dublin Core Metadata

EdN:7 [7]

2.6.3 workaddress: Markup for FOAF Metadata

EdN:8 [8]

2.7 Support for MathHub.info

The mathhub package provides the supplementary packages mikoslides-mh, modules-mh.sty, omtext-mh.sty, problem-mh.sty, smultiling-mh.sty, structview-mh.sty, and tikzinput-mh.sty with variants of the user-visible macros that are adapted to the MathHub system – see Section 3.1 for details.

[5]EdNote: continue
[6]EdNote: Say something
[7]EdNote: Say something
[8]EdNote: Say something

2.8 Auxiliary Packages

2.8.1 metakeys: An extended key/value Interface

EdN:9 [9]

2.8.2 pathsuris: Managing Relative/Absolute File Paths

EdN:10 [10]

2.8.3 tikzinput: External TIKZ Pictures as Standalone Images

EdN:11 [11]

3 Workflows and Best Practicies

3.1 Ulities

`sms` computes the sTeX module signatures for a give sTeX file

3.2

[9] EdNote: Say something
[10] EdNote: Say something
[11] EdNote: Say something

4 The Implementation

4.1 Package Options

The first step is to declare (a few) package options that handle whether certain information is printed or not. They all come with their own conditionals that are set by the options.

```
⟨*package⟩
\DeclareOption*{\PassOptionsToPackage{\CurrentOption}{statements}
  \PassOptionsToPackage{\CurrentOption}{structview}
  \PassOptionsToPackage{\CurrentOption}{sproofs}
  \PassOptionsToPackage{\CurrentOption}{omdoc}
  \PassOptionsToPackage{\CurrentOption}{cmath}
  \PassOptionsToPackage{\CurrentOption}{dcm}}
\ProcessOptions
```

Then we make sure that the necessary packages are loaded (in the right versions).

```
\RequirePackage{stex-logo}
\RequirePackage{statements}
\RequirePackage{structview}
\RequirePackage{sproof}
\RequirePackage{omdoc}
\RequirePackage{cmath}
\RequirePackage{dcm}
⟨/package⟩
```

4.2 The sTeX Logo

To provide default identifiers, we tag all elements that allow `xml:id` attributes by executing the `numberIt` procedure from `omdoc.sty.ltxml`.

```
⟨*logo⟩
\RequirePackage{xspace}
\def\stex{%
  \@ifundefined{texorpdfstring}%
  {\let\texorpdfstring\@firstoftwo}%
  {}%
  \texorpdfstring{\raisebox{-.5ex}S\kern-.5ex\TeX}{sTeX}\xspace%
}
\def\sTeX{\stex}
⟨/logo⟩
```

Index

Numbers written in italic refer to the page where the corresponding entry is described; numbers underlined refer to the code line of the definition; numbers in roman refer to the code lines where the entry is used.

References

[Aus+03] Ron Ausbrooks et al. *Mathematical Markup Language (MathML) Version 2.0 (second edition)*. W3C Recommendation. World Wide Web Consortium (W3C), 2003. URL: `http://www.w3.org/TR/MathML2`.

[Ber89] J. A. Bergstra. *Algebraic specification*. ACM Press, 1989.

[Bus+04] Stephen Buswell et al. *The Open Math Standard, Version 2.0*. Tech. rep. The OpenMath Society, 2004. URL: `http://www.openmath.org/standard/om20`.

[CTAN] *CTAN the Comprehensive TEX Archive Network*. URL: `http://ctan.org` (visited on 12/11/2012).

[Dea99] Stephen Deach. *Extensible Stylesheet Language (XSL) Specification*. W3C Working Draft. World Wide Web Consortium (W3C), 1999. URL: `http://www.w3.org/TR/WD-xsl`.

[Kay00] Michael Kay. *XSLT Programmers Reference*. Wrox, 2000.

[KG16] Michael Kohlhase and Deyan Ginev. *`presentation.sty`: An Infrastructure for Presenting Semantic Macros in sTeX*. Self-documenting LATEX package. Comprehensive TEX Archive Network (CTAN), 2016. URL: `http://mirror.ctan.org/macros/latex/contrib/stex/sty/presentation/presentation.pdf`.

[KGA16] Michael Kohlhase, Deyan Ginev, and Rares Ambrus. *`modules.sty`: Semantic Macros and Module Scoping in sTeX*. Self-documenting LATEX package. Comprehensive TEX Archive Network (CTAN), 2016. URL: `http://www.ctan.org/get/macros/latex/contrib/stex/modules/modules.pdf`.

[Knu84] Donald E. Knuth. *The TEXbook*. Addison Wesley, 1984.

[Koh04] Michael Kohlhase. "Semantic Markup for TEX/LATEX". In: *Mathematical User Interfaces*. Ed. by Paul Libbrecht. 2004. URL: `http://www.activemath.org/~paul/MathUI04`.

[Koh06] Michael Kohlhase. OMDoc – *An open markup format for mathematical documents [Version 1.2]*. LNAI 4180. Springer Verlag, Aug. 2006. URL: `http://omdoc.org/pubs/omdoc1.2.pdf`.

[Koh16a] Michael Kohlhase. *`cmath.sty`: An Infrastructure for building Inline Content Math in sTeX*. Self-documenting LATEX package. Comprehensive TEX Archive Network (CTAN), 2016. URL: `http://mirror.ctan.org/macros/latex/contrib/stex/sty/cmath/cmath.pdf`.

[Koh16b] Michael Kohlhase. *`hwexam.sty/cls`: An Infrastructure for formatting Assignments and Exams*. Self-documenting LATEX package. Comprehensive TEX Archive Network (CTAN), 2016. URL: `http://mirror.ctan.org/macros/latex/contrib/stex/sty/hwexam/hwexam.pdf`.

[Koh16c] Michael Kohlhase. *problem.sty: An Infrastructure for formatting Problems*. Self-documenting LaTeX package. Comprehensive TeX Archive Network (CTAN), 2016. URL: http://mirror.ctan.org/macros/latex/contrib/stex/sty/problem/problem.pdf.

[Koh16d] Michael Kohlhase. *sproof.sty: Structural Markup for Proofs*. Self-documenting LaTeX package. Comprehensive TeX Archive Network (CTAN), 2016. URL: http://mirror.ctan.org/macros/latex/contrib/stex/sty/sproof/sproof.pdf.

[Koh16e] Michael Kohlhase. *statements.sty: Structural Markup for Mathematical Statements*. Self-documenting LaTeX package. Comprehensive TeX Archive Network (CTAN), 2016. URL: http://mirror.ctan.org/macros/latex/contrib/stex/sty/statements/statements.pdf.

[L2O] GitHub repository. URL: https://github.com/KWARC/LaTeXML-Plugin-sTeX.

[Lam94] Leslie Lamport. *LaTeX: A Document Preparation System, 2/e*. Addison Wesley, 1994.

[LTX] Bruce Miller. *LaTeXML: A LaTeX to XML Converter*. URL: http://dlmf.nist.gov/LaTeXML/ (visited on 03/12/2013).

[Mik] *MiKTeX... typesetting beautiful documents...* URL: http://miktex.org (visited on 12/11/2012).

[Mkm] *Meetings and Conferences on Mathematical Knowledge Management*. Project homepage. URL: http://www.mkm-ig.org/meetings/ (visited on 03/12/2007).

[Pro07] LaTeX Project. *The LaTeX project public license*. Software License. 2007. URL: http://www.latex-project.org/lppl/.

[RV01] Alan Robinson and Andrei Voronkov, eds. *Handbook of Automated Reasoning*. Vol. I and II. Elsevier Science and MIT Press, 2001.

[Sta+09] Heinrich Stamerjohanns et al. "MathML-aware article conversion from LaTeX, A comparison study". In: *Towards Digital Mathematics Library, DML 2009 workshop*. Ed. by Petr Sojka. Masaryk University, Brno, 2009, pp. 109–120. URL: http://kwarc.info/kohlhase/papers/dml09-conversion.pdf.

[sTeX] *KWARC/sTeX*. URL: https://github.com/KWARC/sTeX (visited on 05/15/2015).

[Tex] *TeX Live*. URL: http://www.tug.org/texlive/ (visited on 12/11/2012).