— KWARC Blue Note* —
Editing Workflows in SMGloM

Michael Kohlhase & Constantin Jucovschi
Computer Science, Jacobs University Bremen
http://kwarc.info

July 22, 2014

Abstract

In this note we discuss editing workflows for SMGloM.

Contents

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

1 Introduction

The SMGloM is a semantic, multilingual glossary for mathematical vocabularies; see [Koh13]. In this note, we will look at workflows for the curation, change management, and quality control of the glossary content.

2 Manual Curation vs. Text Mining

Most of the glossaries mentioned in Section **??** are manually curated, but that does not preclude automatic harvesting component for the glossary. But with the SMGloM we are after a different resource than we can create with automated methods: Here is the kind of data seem hard to create automatically in sufficient quality:

synchronized symbols/verbalizations/notations The SMGloM data model foresees that we have symbols for concepts that may have multiple verbalization definitions and notation definitions (all synchronized). Some work into this direction (extraction of notations and verbalizations - but not synchronization) has been done in [Aut+07].

synchronized multi-lingual markup The SMGloM data model foresees that we have synchronized markup in multiple languages. This is very hard to do automatically, even though there is some (non-math) work on such things in the YAGO system [Hof+b; Hof+a].

content markup in the formulae in the definitions.

Theory structures and conceptual relations but this is maybe the least important of the four.

Therefore, we go for a community effort project with SMGloM instead, where the glossary entries are hand-curated and we supply editor support for these four kinds of data (cf. section 3.2.2). In particular, we counter the argument "human glossary curation is work" with "human curation creates value". After all, resources like WordNet (and MSC/ZBMath) were created by human curation over decades, and are the valuable resource they are today exactly because that. We need resources like that for Math understanding.

There no objection against combining SMGloM with automatic glossary extraction. Indeed that has been the plan from the start. In particular, the SMGloM is the ideal platform to try them out. Automatic extraction methods would be considered "semantic services" in SMGloM. For instance

- Automatically generated glossary entries can be considered system-generated drafts, which can then be "completed" by human curators.
- glossary extraction could serve as a "entry-suggestion service" to guide human curators
- automated extraction methods could serve as coverage test services, the results of this can be integrated into the human-curated part as additional notation definitions, and
- the human curators are the ideal evaluators for the automated services, since they can judge the quality of the extraction.

So the investigation of automated glossary entry harvesting via text mining is not a contradiction, but an extension of the manual curation approach inscribed into the SMGloM project.

3 Managing the SMGloM

[1] EdN:1

3.1 Direct access to the SMGloM via MathHub.org

The SMGloM content is hosted on `MathHub.info`[1], a `github.org`-like web-based hosting service for mathematical document collections using the Git revision control system. The content is publicly available for checkout at `http://gl.mathhub.info/smglom/smglom.git`. Parties interested in contributing can clone the repository add new items, make a Git patch, and send it to `m.kohlhase@jacobs-university.de` who will apply them to the master branch. Regular committers can apply for Git push rights on `MathHub.info`.

For documentation of MathHubauthoring see [MHA].

Direct access to the repository is optimal for large-scale edits across multiple files. For more casual access we recommend the Planetary-based structure editor under development at Jacobs University.

3.2 A Planetary Instance for SMGloM

The Planetary system [Koh12; Koh+11; PDFm] is a framework for building mathematical knowledge portals. It presents its content as **active documents** – documents that are instrumented with semantic services to become interactive and react to the reader's context. The portal supports editing LaTeX and STeX documents that are then converted to OMDoc and activated on the fly.

3.2.1 Presenting the Glossary

For the reader, the SMGloM system presents the glossary by giving access to the glossary terms e.g. alphabetically (possibly in a faceted search via other classification criteria). For cases, where the glossary term has homonyms, the system generates a wikipedia-style disambiguation page.[2] [3] [4] EdN:2 EdN:3 EdN:4

3.2.2 Editing Services

The SMGloM portal extends Planetary by a structural editor for module definitions (see Figure 1) that combine module definition and language bindings (in the language tabs on the

[1] EdNote: should say something about related work: in the non-math field we have e.g. Anchovy [AN], there must be others, research them.

[1] `MathHub.info` will offer public repositories for open document collections (e.g. SMGloM) and escrow repositories (the repository is is private for three months and is published afterwards unless the private period is re-set). SMGloM is a test project

[2] EdNote: MK: from what information? Do we have a "short definition"? Probably a classification by area would also be very helpful. The disambiguation page could also be hierarchical. We should think about this carefully

[3] EdNote: MK@CJ: make a proposal, add a figure/screenshot.

[4] EdNote: MK: we need to take the "current language" into account; after all, we are multilingual

top). [5] EdN:5

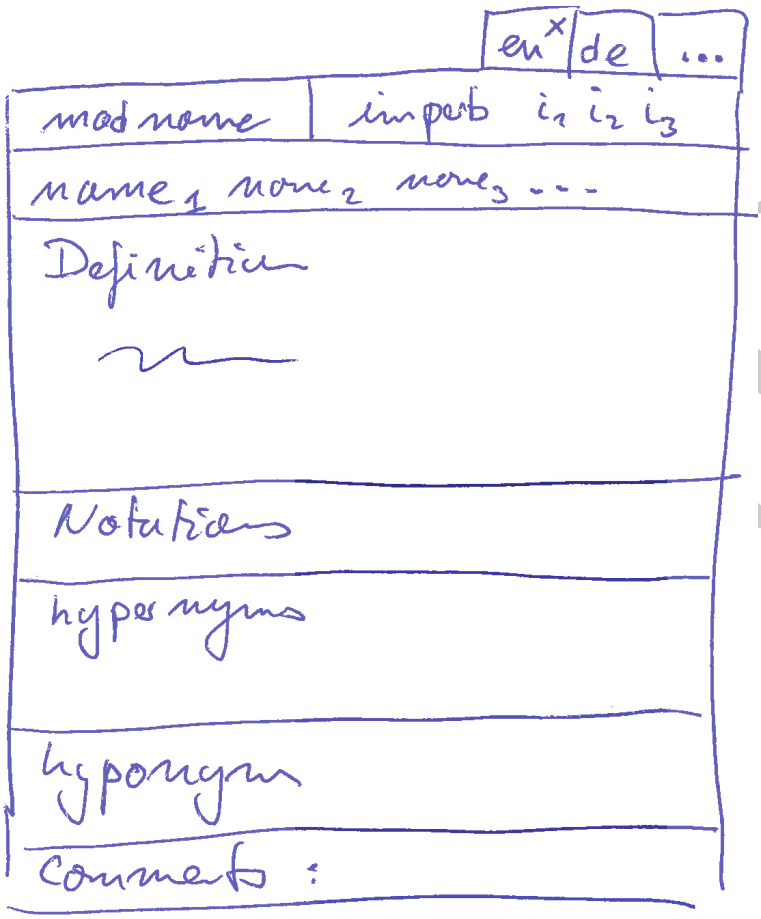

Figure 1: A structure editor for SMGloM

Glossary Auto-linking and Import Declaration Generation The SMGloM editor integrates NNexus autolinker [**GinCor:nnexusCICM13**]: whenever the glossary author types a glossary term, the SMGloM system prompts the user for disambiguation (to determine the symbol it denotes). Once this is determined, the editor adds a (glossary) term reference to the phrase and adds the imports declaration to the module definition if necessary – possibly eliminating transitively redundant declarations from the entry and the ones that import from it. The autolinker is a major source of imports declarations in practice.

Wiki-Style Dangling Links Often the glossary author can identify additional glossary entries terms when writing a definition. Such a term can be marked up via \trefi[foo]{bar}, even though there may not be a symbol bar in module foo. The SMGloM editor issues a warning and interprets this as a wiki-style dangling link. Such links are presented in the system as links to a glossary term creation form – Essentially a call to the structure editor. Note that the particular state of the editor depends on whether the module foo already exists (then the

[5] EdNote: MK@CJ: please continue and add preliminary screenshots..

module is just extended by the symbol in question) or if the symbol already exists, but has no language binding for the current language.

The SMGloM system supports the glossary completion process by recording such links in help-wanted listings, which authors can subscribe to.

4 Conclusion

References

[AN] *Anchovy*. URL: http://www.maxprograms.com/products/anchovy.html (visited on 02/01/2014).

[Aut+07] Serge Autexier et al. "Supporting User-Defined Notations When Integrating Scientific Text-Editors with Proof Assistance Systems". In: *MKM/Calculemus*. Ed. by Manuel Kauers et al. LNAI 4573. Springer Verlag, 2007, pp. 176–190.

[Hof+a] J. Hoffart et al. "YAGO2: A Spatially and Temporally Enhanced Knowledge Base from Wikipedia." In: *AI journal* (). URL: http://www.mpi-inf.mpg.de/yago-naga/yago/publications/aij.pdf.

[Hof+b] J. Hoffart et al. "YAGO2: Exploring and Querying World Knowledge in Time, Space, Context, and Many Languages". In: *WWW 2011*. URL: http://www.mpi-inf.mpg.de/~gdemelo/papers/hoffart-yago2-www2011.pdf.

[Koh+11] Michael Kohlhase et al. "The Planetary System: Web 3.0 & Active Documents for STEM". In: *Procedia Computer Science* 4 (2011): *Special issue: Proceedings of the International Conference on Computational Science (ICCS)*. Ed. by Mitsuhisa Sato et al. Finalist at the Executable Paper Grand Challenge, pp. 598–607. DOI: 10.1016/j.procs.2011.04.063. URL: http://kwarc.info/kohlhase/papers/epc11.pdf.

[Koh12] Michael Kohlhase. "The Planetary Project: Towards eMath3.0". In: *Intelligent Computer Mathematics*. Conferences on Intelligent Computer Mathematics (CICM). (Bremen, Germany, July 9–14, 2012). Ed. by Johan Jeuring et al. LNAI 7362. Berlin and Heidelberg: Springer Verlag, 2012, pp. 448–452. arXiv: 1206.5048 [cs.DL].

[Koh13] Michael Kohlhase. "SMGloM: a Semantic Multilingual Glossary System for Mathematics". SMGloM Blue Note. 2013. URL: http://gl.mathhub.info/smglom/smglom-doc/raw/master/source/blue/smglom/note.pdf.

[MHA] *MathHub: Authoring Resources & Workflows*. URL: http://mathhub.info/help/authoring (visited on 02/01/2014).

[PDFm] *Planetary Developer Forum*. URL: http://planetary.mathweb.org/ (visited on 09/15/2012).