— SMGloM Blue Note* —
Management of Content in SMGloM

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

April 14, 2015

Abstract

In this note we discuss strategies for management of content in the SMGloM, a semantic, multilingual glossary for mathematical vocabularies. This note discusses organizational foundations of content management in the SMGloM, in particular, the development model, licensing, strategies for lifecycle management and quality control, as well as governance, stewardship, and editorial strategies.

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

Contents

1 Introduction

The SMGloM is a semantic, multilingual glossary for mathematical vocabularies; see [Koh13]. Establishing a communal resource like the SMGloM requires finding a balance between two factors

- *community contributions*: SMGloM can profit from making it easy for individuals and organizations to contribute glossary entries.
- *quality control & reputation*: allowing just anybody to contribute freely to SMGloM could lead to low-quality content[1]. Even the public conception that a stringent quality control might be absent can lead to a loss of reputation of the resource and thus stunt its development.

In this note, we will look at workflows for the curation, change management, and quality control of the glossary content. The main idea we will pursue in the SMGloM is to allow (and encourage) open submission of glossary components (glossary modules or language bindings) to the SMGloM archive on MathHub [MH], but moderate the submitted content after submission (see Section 4). What is more, the SMGloM must support the whole document life cycle of a glossary component from the conception to the deprecation (see Section 6).

2 Content Development Model

SMGloM implements a specific content development model based on the "authority model" of PlanetMath.org [PM] and implements it[2] in the MathHub-based SMGloM system [Ian+14].

2.1 The Authority Model in SMGloM

The SMGloM authority model is characterized by the principles below:

1. Only users registered to the MathHub system can submit glossary entries.
2. An author who starts a new glossary component becomes its owner, that is the only person authorized to edit that glossary component.
3. To make the development more smooth, the owner may also choose to grant editing rights to other individuals or groups.
4. Other users may propose corrections and discuss improvements but the resulting modifications of the article, if any, are always made by the owners.
5. However, if there are long lasting unresolved corrections, the ownership can be removed. More precisely, after 2 weeks the system starts to remind the owner by mail; at 6 weeks any user can "adopt" the article; at 8 weeks the ownership of the component is completely removed (and such an entry is called "orphaned").

The "authority model" has been successful in PlanetMath.org, it strikes a pragmatic balance between the legitimate sense of ownership and responsibility of the author group and agility of processes. In SMGloM the authority model is mitigated by the additional considerations below.

[1] unfortunately, it is often the least qualified to contribute high-quality content who have most time and motivation to do so.

[2] actually, the implementation is still largely to come, but the underlying mechanisms have been implemented in `PlanetMath.org`, which is based on the Planetary System just like MathHub.info. So we anticipate that implementing them mostly a matter of activating the respective drupal modules in `MathHub.info`.

2.2 Distributed Content Management

The SMGloM system is based on MathHub, which uses GitLab [GL] for the management of git repositories, which store the sources of MathHub/SMGloM content. In particular, the MathHub.info accounts coincide with the ones in GitLab, so we can directly use the well-established git functionalities for access rights management and content distribution – the revision control aspect of git is essential for the project, but not in our focus here. To keep the SMGloM content manageable – recall that every glossary entry has $n + 1$ files – we organize it into git repositories, which are registered with the SMGloM system and assembled into a joint repository resource.

In SMGloM we organize repositories by author (or author group) and topic independently, i.e. for every author and topic we assume a new repository – repositories in GitLab are cheap and scalable. An author Authors A can hash out glossary entries in a private repository, he can give authors B and C access to the repository for further development (e.g. translation into other languages) by topic, and eventually, they the glossary entries can be turned over to a glossary collection – a special repository owned by an editorial collective – for curation. This workflow corresponds to the "shared repository mode", which is supported by MathHub via GitLab.

But even without access rights, authors can contribute to a repository via the "fork/pull" model [Git]: the contributor forks an existing repository and push changes to their personal fork without requiring access be granted to the source repository. The changes can then be pulled into the source repository by the project maintainer, which is supported by MathHub trough the underlying GitLab functionality.

We call the this approach to distributed content managment the "**small repositories paradigm**". Multi-repository workflows on MathHub are supported by the lmh tool [LMH] and the sTeX format (see [KGA15, Section 2.10]), so that the reliance on swarms of small repositories become transparent and manageable to authors.

3 Uniform Legal Framework

One of the most important considerations in the creation of a global community resource is the establishment of a uniform legal framework for the intellectual property involved. As glossary components consist of mathematical terminology and notations in common use, definitions that have already been published in the mathematical literature, structurally annotated, and equipped with explicit semantic relations, their expression in SMGloM constitutes copyrightable work by the author or author group.

The SMGloM project provides a uniform legal framework with the SMGloM Public License (SPL [SPL]), which is a prerequisite for the submission of glossary components to the SMGloM. We will briefly discuss copyright and licensing approaches before presenting the SPL in subsection 3.3 below.

3.1 Copyright

There are three models for copyright management for public resources like the SMGloM: The glossary entry author who is the original copyright holder can *i*) release it into the public domain *ii*) assign it to some kind of organizing entity, or *iii*) keep it. The consequences of these alternatives are mediated by the license chosen for the content. Given an open license

copyright attribution is less of an issue, since all parties involved: content authors, aggregators, and users have all the rights they need. However, projects like the SMGloM often need some stewardship by an organization that can provide server space, bandwidth, organizational leadership, and legal representation. For this to be effective, assigning copyright to a steward organization that assumes organizational role for a sub-glossary (see Section 5.3) and promises to keep the content open; often simplifies the growth and health of a data collection.

Note that all the organizational stipulations in this note are independent of copyright ownership, so we will assume that authors keep their copyright in the following as the default case.

3.2 Licensing

An open resource is created by licensing the content with an open license, which explicitly allows derived works and often contains a share-alike clause. As the SMGloM is almost certainly going to be a community effort, an open license seems a critical requirement for the success of the initiative.

There are many open licenses, some – e.g. the General Public Licenses [GNUL] (GPL) – apply to software and its documentation, whereas others – e.g. the Creative Commons (CC) Licenses [CCL] – are designed for literary and artistic works. Finally, there is the Open Database License [ODbL] (ODL) that is designed for data collections.

Glossary entries are somewhere in-between programs, documents, and data. Additionally, the license must warrant the semantic stability of the glossary entries over time. As glossary entries are referenced by module name (and URI), the meaning (i.e. the structure of the definition) may not change over time, and must be independent of the language used to assess them. The stability issue has been addressed by the LaTeX Project Public license [Pro07] (LPPL) and the licenses of OpenMath Content Dictionaries (CDs), which make sure that package names and CD names are changed for derived works.

All of these licenses are "viral" in the sense that the mandate that any derived work must be licenses under (a variant of) the original license. This clause originated in the GPL and is called the "copyleft" or "share-alike" clause.

3.3 The SMGloM Public License (SPL)

The SPL [SPL] combines aspects of the all the licenses discussed in the last section: Clause D1 ensures attribution in the sense of the CC licenses, clause D2 ensures name stability as the LPPL does, and finally the copyleft clause (D3) makes sure that SMGloM glossary entries and any improvements stay in the digital mathematical commons. The last two paragraphs (Maintenance/Translation) reflect the particular requirements of glossary entries introduced in this note.

4 Community-Based Quality Control

For quality control SMGloM adopts the "lenses model" [CNXa] of Connexions Project [CNXb], a global repository of educational content provided by volunteers. Lenses are special annotations that enable both organizations and individuals to give their stamps of approval to glossary components. They have been introduced [Fle07] to organize a post-publication re-

view process for the course modules and collections in Connexions. Lenses serve multiple purposes in SMGloM:

endorsement Individuals and organizations, such as professional societies, can create endorsement lenses containing content they have carefully reviewed and deem to be of high quality. The individuals and organizations use their own criteria for endorsement and describe their selection process on the lens home page.

reputation Author groups can apply, and readers can nominate glossary entries for endorsement to boost visibility of their glossary entries. In this way, endorsements can bring reputation to SMGloM authors.

filtering lenses can be used for filtering SMGloM content.

Personal lenses as used in Connexions for bookmarking and commenting seem of less importance in SMGloM. But many content management operations may be parameterized by a lens or organization and thus be given legitimacy.

5 Distributed Governance

As a consequence of the open content development model paired with the community-driven quality control, the SMGloM can do with a minimal governance structure. The only structure needed is to maintain the set of organizational rules set out in this document and to ensure a technical infrastructure for supporting the content development model itself. We will discuss the various aspects of distributed governance by their function.

5.1 Content Selection and Editorial Strategies

An important aspect in the establishment of a large communal resource as such as the SMGloM is the question about the SMGloM content, i.e. what terms/symbols constitute legitimate glossary entries. Other resources have an editorial board that defines a content strategy. For a resource with a universal scope as the SMGloM, this is difficult. Therefore we invite mathematical societies and interest groups to assume responsibility for various parts of the glossary. They can develop their own content strategies, motivate members to contribute, and set up quality control strategies and policies with the help of MathHub lenses. SMGloM will only provide a uniform data model, infrastructure, and added-value services. The uniform legal framework of SMGloM makes sure that the resource stays communal, forming the base of a flexiformal mathematical content commons.

5.2 System Infrastructure

The KWARC [KWARC] group at Jacobs University Bremen is developing a reference infrastructure for the SMGloM based on their MathHub system [MH; Ian+14]. But the open licensing framework of the SMGloM allows any organization to mirror the content and provide mathematical services based on it. In particular, in the a situation where the KWARC group be unable or unwilling to provide sufficient infrastructure support for the SMGloM, other organizations can assume the task and carry on, even with the same software as it is open source as well. Even a distributed infrastructure like the CTAN [CTAN] is easily conceivable.

5.3 Stewardship and Project Board

The lightweight nature of governance notwithstanding, the SMGloM project will work towards the establishment of a scientific advisory board and an executive board that will spread the responsibility for the SMGloM commensurate with the large scope of the project. To facilitate that, the project will seek a stewardship organization which can offer administrative support, legal advice if necessary, and some sustainability. Possible stewardship organizations that come to mind could be an existing organization like PlanetMath.org, the Open Math Society, or one of the mathematical societies (e.g. SIAM, AMS, EMS, IMU, ...), a university, a charitable foundation who funds the effort, or even a custom made foundation.

6 Supporting the Glossary Entry Life Cycle

It is important to realize that even small documents like glossary entries (see [Koh14] for a discussion of their anatomy) are subject to a life-cycle that must be managed. At the system level, we make use versioning and editing workflows of the MathHub portal [MHA].

For the following recall from [Koh14] that a glossary entry consists of

1. a **signature** – a file containing LaTeX markup for the symbols and notations introduced in the glossary module. Its file name coincides with the glossary module name.
2. multiple **language bindings** – LaTeX files indexed by language that contain the (natural language) definitions and verbalization definitions for the symbols from the glossary entry signature. For consistency management purposes, one language bindings – usually the English one as it is most accessible – is characterized as primary.

The special characteristics of glossary entries, which have to balance stability with agility are supported with a system of states that indicate the place in the glossary life-cycle. Note that states apply to glossary components (modules and their language bindings) separately.

6.1 Authoring States

The first set of states are given by the authors of the components, they are inspired by the states of OpenMath content dictionaries [Bus+04].

experimental this is the initial state of any glossary component, it indicates that the component is still under development and may change at any time.

complete if the authors mark a glossary component as "completed", then they indicate that they have finished the initial draft and invite comments, endorsements, and translations. A glossary module can only be marked as "complete", if its primary language binding is marked as "complete". At this time, the glossary module name is pre-reserved by the SMGloM system – as the symbol names are referenced via the module name, they are implicitly reserved as well. Note that a complete glossary entry can still be deleted, freeing up the pre-reserved name.

stable this state indicates that the authors have committed to the state of the glossary component and have declared it as semantically stable. This means that the meaning of symbols declared in the glossary module will never change; only conservative extensions – e.g. addition of verbalizations, notations, new symbols and their definitions, and even language bindings are admissible.

Note that the SMGloM system permanently reserves the module (and symbol) names at this stage. Stable glossary entries can never be removed from the SMGloM system

any more, they can only be deprecated (see below).

As a consequence "stable" glossary modules can be safely imported into other modules, "stable" language bindings can safely be referenced[3] in other language bindings. In particular "stable" glossary modules and primary language bindings can never be deleted (they might still be imported somewhere). Secondary language bindings can change more freely, but only in ways that are semantically consistent with the primary one.

A glossary module may not be labeled "stable", unless it has at least three language bindings (to ensure multilinguality and ease of translation).

deprecated in some situations, "stable" glossary entries can become obsolete, e.g. because they have been found to be gravely erroneous or duplicates of other entries. As they cannot be deleted, the state of the module and language bindings is changed to "deprecated", and they are not shown to the user any more (unless directly referenced).

Note that the authoring states of glossary modules and primary language bindings are monotonous, i.e. the state can only be changed to a successor state in the ordering of the list above. Non-primary language bindings can be changed non-monotonically e.g. in the case of semantic inconsistency with the primary bindings.

Note furthermore, that there is no technical means of checking "semantic stability" – even defining that rigorously is difficult – so ensuring this must be a manual effort. For the moment, we will leave this to the authoring community and only regulate things, when this causes problems.

Finally note that there is not state "official" or "published" as e.g. for OpenMath content dictionaries. We delegate endorsement to the lenses system discussed in Section 4.

6.2 Computed States

Apart from the author-supplied states, the SMGloM system computes a set of states from the internal relations to inform the user and allow tailored user access. The eventual set of system-computed states will need more experience, and will directly be supported by views in the MathHub.info system, the list below is just given for fortifying our intuition about life-cycle support functionality.

needs-review to indicate that a complete module invites reviews (state=complete, #(reviews)< 3)

needs-translation to indicate that a complete module invites translations to reach stability (state=complete #(bindings)< 3)

buggy? there is a correction request that has not been resolved.

orphaned a glossary component that has been put up for adoption by its current maintainer group or whose maintainer group has been disowned.

proposed someone has proposed the glossary entry to a lens. (state=stable)

6.3 Encoding of States

The authoring states described in Section 6.1 can be encoded in the LaTeX sources of the module components as values of the state key of the optional argument of the modsig (for module signatures) and modnl (for language bindigns) environments. The computed states are handled by the SMGloM system based on the authoring states.

[3]Note that technically only the symbols of a glossary module are referenced, so the number (and quality) of non-primary language bindings is not an issue for this.

7 Conclusion

We have discussed strategies for management of content in the SMGloM. This note discusses organizational foundations of content management in the SMGloM, in particular, the development model, licensing, strategies for lifecycle management and quality control, as well as governance, stewardship, and editorial strategies. The next development steps are the implementation of the the necessary system support in the SMGloM system.

References

[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.

[CCL] *About the Licenses – Creative Commons*. URL: http://creativecommons.org/licenses (visited on 05/31/2013).

[CNXa] *Connexions – Lenses*. URL: http://cnx.org/help/viewing/lenses (visited on 02/01/2014).

[CNXb] *Connexions – Sharing Knowledge and Building Communities*. URL: http://cnx.org (visited on 02/01/2014).

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

[Fle07] Katherine Fletcher. *Lenses: Proposed functional description and high level design*. Tech. rep. June 2007. URL: http://internal.rhaptos.org/docs/architecture/design/lenses/.

[Git] GitHub. *Using Pull Requests*. URL: https://help.github.com/articles/using-pull-requests (visited on 04/27/2014).

[GL] *GitLab*. URL: http://gitlab.org (visited on 02/24/2014).

[GNUL] *Licenses - GNU Project - Free Software Foundation (FSF)*. URL: https://www.gnu.org/licenses/licenses.html (visited on 05/31/2013).

[Ian+14] Mihnea Iancu et al. "System Description: MathHub.info". 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. 431–434. URL: http://kwarc.info/kohlhase/submit/cicm14-mathhub.pdf.

[KGA15] Michael Kohlhase, Deyan Ginev, and Rares Ambrus. *modules.sty: Semantic Macros and Module Scoping in sTeX*. Tech. rep. 2015. URL: https://github.com/KWARC/sTeX/raw/master/sty/modules/modules.pdf.

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

[Koh14] Michael Kohlhase. "A Data Model and Encoding for SMGloM". SMGloM Blue Note. 2014. URL: http://gl.kwarc.info/smglom/blue/raw/master/datamdl/note.pdf.

[KWARC] *Knowledge Adaptation and Reasoning for Content*. URL: http://kwarc.info (visited on 05/12/2011).

[LMH] *Local Editing Workflows*. URL: http://mathhub.info/help/lmh-workflows (visited on 04/14/2014).

[MH] *MathHub.info: Active Mathematics*. URL: http://mathhub.info (visited on 01/28/2014).

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

[ODbL] *Open Data Commons Open Database License (ODbL)*. URL: http://opendatacommons.org/licenses/odbl/ (visited on 05/31/2013).

[PM] *PlanetMath.org – Math for the people, by the people*. URL: http://planetmath.org (visited on 11/11/2012).

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

[SPL] Michael Kohlhase. *The SMGloM Public License (SPL) Version 0.1*. URL: http://mathhub.info/help/smglom/spl0.1.