Mathematical Documents want to Active Digital Math Libraries want to be Semantic — Position paper for WDML 2012

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

April 29, 2012

Introduction

Mathematics plays a fundamental role in science, technology, and engineering (STEM). Mathematical knowledge is rich in content, sophisticated in structure, and technical in presentation, its conservation, dissemination, and utilization constitutes a challenge for the community and an attractive line of inquiry. In this position paper, we will take the position of the MKM (Mathematical Knowledge Management) community, an emerging interdisciplinary field at the intersection of mathematics, computer science, Semantic Web, library science, and scientific publishing that develops representation formats, methods, and tools to facilitate the creation of a "universal digital library of mathematics" and empower its users with added value services (see [Far05] for an introduction). It is a driving intuition for the MKM community that mathematical knowledge constitutes an attractive "test tube" for structure research and tool development and that results and tools can be generalized to all of STEM.

Most of mathematical knowledge is currently recorded in the form of informal (see below) documents – ranging from journal papers over preprints to sketches on blackboards. Efforts are currently under way to digitize large parts of the former and turn them into a generally accessible "World Digital Mathematical Library" (WDML). We claim that *the digitization effort should be complemented by a flexiformalization effort that makes the WDMS semantically accessible and turns mathematical documents into active documents*.

Formality?, Informality?, Flexiformality!

Since the foundation debate in mathematics almost a century ago, "formal mathematics" has been defined as reifications mathematical knowledge expressed in a "formal system", i.e. in a well-defined logical language with a syntactic proof system, where grammaticality of expressions and the verification of proofs is decidable. Moreover, formal systems are usually expected to have a well-defined model theory, into which expressions interpreted compositionally. Formal developments of mathematics fix a foundation: a logical system $\mathcal{L}$ and a foundational theory $\mathcal{F}$, e.g. first-order logic with descriptions and ZFC set theory. Based on this foundation, mathematical objects are specified via axioms and/or definitions (special $\mathcal{L}$-expressions), and their properties stated in form of "assertions"

($\mathcal{L}$-expressions again) which are justified by proofs (again $\mathcal{L}$ expressions; we assume $\mathcal{L}$ to contain a proof system).

In this sense, almost all mathematical documents are informal in at least three ways:

I1. *the foundation is unspecified*: mathematical documents usually leave the foundation open,

I2. *the language is informal*: mathematical vernacular (MV) is a mixture of natural language (NL) with formulae and discourse-level cues on the epistemic status of text fragments. This is informal, as we do not have decision procedures for grammaticality or interpretation,

I3. even *formulae are informal*: as they are in presentation markup that specifies the layout, and not the logic/functional structure of the mathematical object or property represented, and finally

I4. *context references are underspecified*: mathematical objects and concepts are often identified by name without making the references to context explicit. This applies both to the natural language part, formulae, and at the statement level (citations of definitions, theorems, and proofs).

In a world, where mathematical documents are exclusively addressed at human readers, all of these informalities are features, not bugs, since they avoid spurious over-specifications (e.g. most foundations are essentially equivalent, and most arguments can be formalized into most foundations). The mathematical community has developed the standard of "rigorous developments" for the subset of documents that could be formalized in some foundation given enough resources. In [KK11] we have introduced the concept of *flexiforms* for representations of mathematical knowledge of flexible formality, and the concept of *flexiformalization* for any act of disambiguation by explicit markup.

As all machine support is based on syntactic manipulations (until we achieve artificial intelligence) we need some formalization if we want to enlist computers in mathematics. Machine support in mathematics and STEM is advantageous, since humans and machines have very different strengths and weaknesses. Humans have unmatched abilities in exploring mathematical theories while developing deep insights into the key properties and inherent invariants, which allow to conjecture key statements and drive proofs via accurate intuitions. Machines excel at systematic analysis of large structures – e.g. for verifying large and convoluted proofs or indexing large datasets for search. We claim that mathematics research and application will be strongest, if we employ a combination of human and machine strengths.

Active Documents, Semantic Libraries

To enable an optimal collaboration between man and machine, we need at the same time to keep close to established workflows of mathematicians and give algorithms the explicit representations needed for computation. For the first goal we want to keep traditional documents as "user interface" and augment them with embedded services that activate the content for interaction and adaption (we call such enhanced documents **active documents**). Whereas the documents themselves are essentially tree-structured, the knowledge reified in them is best structured as a hypergraph, where the nodes are mathematical objects, statements, and theories, whereas the edges are given by the content-structural relations among them – e.g. the "inheritance" and "views"

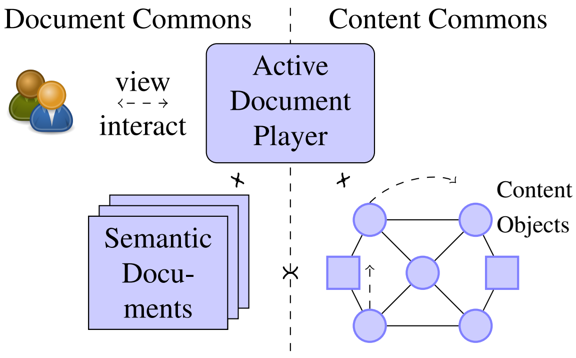

Figure 1: Active Documents Paradigm

relations between theories, the "justification" of theorems by proofs, and the "dependency" of new concepts on the concepts in the definienda.

In the **Active Documents Paradigm** (ADP see Figure 1 for an overview and e.g. [Koh+11] for details) both structures are read by a **document player**: a software application that generates document presentations that are instrumented with controls for user interactions (the active documents). It is crucial for the ADP that the documents are semantically annotated (we call them **semantic documents**), usually by classifications of text fragments and references into the content commons that serves as an explicit semantic context. For instance, symbols in formulae or technical terms in a text could be linked to rigorous definitions; the link is an example of a formalization of the informal context references discussed in clause **I4** above. Having such formal links directly translates into an active-document service: "*definition lookup*", which displays the definition induced by a click (or hover) on the symbol or technical term and invites the reader to explore the context from the definition.

We call a collection of semantic documents together with a content commons a **semantic library** and remark that active documents only make sense as members of semantic libraries. Semantic libraries can arise in multiple ways and depths (of formalization). One way is to analyze digital mathematical documents and recover semantic structures, e.g. by transforming LaTeX documents into HTML5, transforming the relations given in the functional LaTeX markup into RDFa annotations which can then be harvested into a content commons realized as a triple store (RDF database); see [Sta+10] for details and [Gin+09] for linguistically-based analysis methods. On the other side of the flexiformality spectrum are presentation workflows that start from completely formal representations and generate semantic documents from that: for instance, the Mizar Mathematical Library [Miz] which contains over 1000 "articles" with over 50000 theorems and over 10000 definitions is published in the *Journal of formalized Mathematics* (JFM [Jfm]), whose articles are generated from "Mizar articles" in an automated presentation process (the JFM still misses out on the chance to make them active, but the Mizar Wiki [Urb+10] does not). Our group has developed the Planetary system [Pla], a generic active document player that can be instantiated to all levels of flexiformality.

Conclusion: An Active, Semantic Layer for the WDML

Mathematical documents are at the same time precision tools optimized for the efficient communication of mathematical knowledge among specialists who share a common knowledge context and at the same time formidable obstacles to be overcome to build up just this shared shared context which is a prerequisite for understanding. This is aggravated by the fact that mathematical knowledge has been growing ever more diverse and intricate over time. The WDML digitization efforts go a first step towards wider adoption of mathematical knowledge by providing *universal access to the mathematical literature*. We claim that with the emerging technologies of flexiformal, semantic libraries and active documents, we have a way to make the mathematical literature more *accessible to non-specialists*[1], by giving access to crucial aspects of the context at the "points of pain" (i.e. in the documents) at the cost of partial flexiformalization of documents and the establishment of a content commons.

This already reveals the main non-technical problem involved in semantic mathematical libraries: unless there is an initial investment into a core content commons to link into, the cost of semantic annotation of documents outweighs the benefit from active documents. We claim that by an act of technology adoption by a major player (the WDML project), we can achieve method standardization and a critical mass of content that kickstarts active mathematical documents and semantic libraries. There is precedent in this: a bold move of the AMS of requiring TeX/LaTeX in its journals

[1] and we are all non-specialists for most of mathematics

brought about the improvement in mathematical typesetting, we still profit from. We conjecture that the induced network effect will lead to widespread flexiformalization, and that we will see additional synergy effects, such as the following one: As soon as a larger body of mathematical theories (by marking up concepts and axioms) we can automatically search for "views" (aka. representation theorems) that allow to import all the theorems of the source theory of the view into the target theory (after translation with the view's signature morphism). We conjecture that systematic automated search will reveal many long-distance views that could not have been found otherwise, as the chance that humans know source and target theories well enough to notice the structural similarities (one-brain constraint) is stlim in today's highly specialized sciences. Methods for this search exist [NK07], we only lack the semantic libraries.

References

[Far05] William M. Farmer. "Mathematical Knowledge Management". In: *Encyclopedia of Knowledge Management*. Ed. by David G. Schwartz. Idea Group Reference, 2005, pp. 599–604.

[Gin+09] Deyan Ginev et al. "An Architecture for Linguistic and Semantic Analysis on the arXMLiv Corpus". In: *Applications of Semantic Technologies (AST) Workshop at Informatik 2009*. 2009. URL: http://www.kwarc.info/projects/lamapun/pubs/AST09_LaMaPUn+appendix.pdf.

[Jfm] *Journal of Formalized Mathematics*. URL: http://www.mizar.org/JFM.

[KK11] Andrea Kohlhase and Michael Kohlhase. "Towards a Flexible Notion of Document Context". In: *Proceedings of the 29th annual ACM international conference on Design of communication (SIGDOC)*. 2011. URL: http://kwarc.info/kohlhase/papers/sigdoc2011-flexiforms.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 Papers Challenge, pp. 598–607. DOI: 10.1016/j.procs.2011.04.063.

[Miz] *Mizar Mathematical Library*. URL: http://www.mizar.org/library (visited on 12/02/2009).

[NK07] Immanuel Normann and Michael Kohlhase. "Extended Formula Normalization for ϵ-Retrieval and Sharing of Mathematical Knowledge". In: *Towards Mechanized Mathematical Assistants. MKM/Calculemus*. Ed. by Manuel Kauers et al. LNAI 4573. Springer Verlag, 2007, pp. 266–279.

[Pla] *Planetary Developer Forum*. URL: http://planetary.mathweb.org/ (visited on 09/08/2011).

[Sta+10] Heinrich Stamerjohanns et al. "Transforming large collections of scientific publications to XML". In: *Mathematics in Computer Science* 3.3 (2010): *Special Issue on Authoring, Digitalization and Management of Mathematical Knowledge*. Ed. by Serge Autexier, Petr Sojka, and Masakazu Suzuki, pp. 299–307. URL: http://kwarc.info/kohlhase/papers/mcs10.pdf.

[Urb+10] Josef Urban et al. "A wiki for Mizar: Motivation, considerations, and initial prototype". In: *Intelligent Computer Mathematics*. Ed. by Serge Autexier et al. LNAI 6167. Springer Verlag, 2010, pp. 455–469. arXiv:1005.4552v1 [cs.DL].