White Paper of the Semantic Representation of Mathematical Knowledge Workshop

Overview

The Semantic Representation of Mathematical Knowledge Workshop was held February 3–5, 2016 at the Fields Institute located on the campus of the University of Toronto, Canada (https://www.fields.utoronto.ca/programs/scientific/15-16/semantic). The Workshop was made possible through a grant from the Alfred P. Sloan Foundation, by staff and resources provided by the Wolfram Foundation, and by staff and resources for local arrangements provided by the Fields Institute (http://www.wolframfoundation.org/programs/computable-archive-of-mathematics.html). It was inspired by the "Developing a 21st Century Global Library for Mathematics Research" report by the Committee on Planning a Global Library of the Mathematical Sciences (http://arxiv.org/abs/1404.1905). The Workshop was organized by the Wolfram Foundation, represented by Michael Trott and Eric Weisstein; the Fields Institute and its Director Ian Hambleton; and by the Global Digital Mathematics Library Working Group of the International Mathematical Union's Committee on Electronic Information and Communication, represented by Patrick Ion working for Ingrid Daubechies.

Goal and expected outcomes/products

The goal of the Workshop was, through a set of talks and discussion sessions among mathematicians who are experts in their fields and experts in semantic language design and markup for mathematics, to reach consensus that would, in turn, enable the creation and application of a semantic language for all of mathematics. The Workshop was intended to produce: (1) a white paper outlining the structure of the proposed semantic language; (2) a plan for creating an explicit semantic language that would be used to mark up results in a specific area of mathematics; and (3) internet publication of the white paper together with examples of encoded material.

In addition, provisions were made to produce a video documentary to publicize not only the Workshop itself, but also the importance of the issues it considered. A feature-length documentary about the Workshop (including interview excerpts from many participants) is nearing completion and is expected to be available very soon on the Wolfram Foundation website as well as on YouTube.

The expected outcome of the Workshop was an agreed-upon way forward to finalizing a design and initiating the processing of significant literature samples into the new semantic markup. The huge amount of research, development, and labor required to carry out that plan was then envisioned to be carried out as separate projects to be proposed to various funding agencies.

This White Paper presents a summary of what was discussed at this Workshop, outlines what was (and was not) achieved, recommends a general direction toward a Global Digital Mathematics Library, and suggests some small but concrete steps that could be taken to further its goals.

The Workshop

A total of 36 invited participants attended the Workshop. A full list of attendees is available on the Fields Institute website for the event at http://www.fields.utoronto.ca/programs/scientific/15-16/semantic. Participants included eminent individuals from relevant fields, often including some of their foremost practitioners. Most attendees gave 30- or 45-minute presentations (including discussion), which were recorded and are now archived at http://www.fields.utoronto.ca/video-archive/event/2053.

The large number of talks had the advantage of allowing all participants to present interesting work and discuss the issues, questions, and other implications of their work as it relates to: (1) the semantic representation of mathematics; and (2) the context of the larger goal of achieving a Global Digital Mathematics Library. It had the disadvantage that while many presentations made clear connections to the goal of designing and building a semantic language for mathematics, the breadth of material covered meant that no common focus developed on what should be done to achieve the Workshop's detailed goals.

A great deal of intensive knowledge transfer took place during talks, post-talk discussions, and discussions during breaks. A number of these seem likely to produce, or have already resulted in, collaborations among small groups of participants. In particular, many fruitful and opinionated discussions developed around the questions of set theory vs. type theory, language vs. library, scoping and evaluation, efficient notation, partiality, and work flow concurrency, as well as in connection with other topics. These discussions offered some insight into the degree of varying opinions among participants and served to highlight specific challenges that will be important to address in any future work. However, these discussions did not coalesce into a concise, directly actionable, clearly defined, and agreed-upon set of issues, priorities, or design principles for a semantic language of mathematics.

The Workshop was held on the assumption that, during the Workshop, a plan would be made for the creation of a "semantic capture language," which would then be written up in this Paper. However, as circumstances transpired, this did not happen. While the final session of the Workshop was dedicated to precisely this purpose, no consensus crystallized during this session on what the focus of a concerted effort should be. This Paper, therefore, attempts to synthesize and expand on the thoughts and opinions expressed at the Workshop, its white paper session, and additional discussions that continued following the close of the Workshop, primarily involving Workshop organizers; a small group assigned to the task of compiling a white paper; and members of the GDML working group, in particular James Davenport, Patrick Ion, Michael Trott, Stephen Watt, Eric Weisstein, and Freek Wiedijk (but also including a number of other contributors). While this Paper attempts to represent the comments, thoughts, conclusions, and priorities of all attendees concisely and with as much consensus as possible, it is anticipated that it may be appropriate to collect and publicize other material worthy of consideration through a more detailed Appendix document.

Semantic language

One of the most fundamental outcomes of the Workshop was the possibility that a "semantic capture language" may not, in fact, be the best framework or direction in which to pursue progress toward a Global Digital Mathematics Library. Some underlying reasons that emerged for this are:

- It turned out not to be clear what should be understood by the term "semantic capture." Even people who considered this term unproblematic discovered that others understood the term in a very different way.

- It is not clear that defining a language is the best first step towards a semantic version of a Global Digital Mathematics Library. To be good, a language should be developed together with a system to process it. And for a system to be good, it should be developed together with a library of content to exercise it and discover its strong and weak points. This means that it might be more beneficial to focus on a "library" instead of a "language."
- There are already several languages in existence that could be called "semantic capture languages." These include Michael Kohlhase's OMDoc language and Fairouz Kamareddine's MathLang language. It did not become specifically clear during Workshop discussions what benefits any newly designed language would have over these or any of the several other similar languages.
- "Semantic capture" supposes a workflow in which existing informal documents are annotated and enhanced towards a more "semantic" version. However, it is not clear whether such processing is the best place to begin work toward a semantic Global Digital Mathematics Library. Eliminating other approaches at the start by using this term might be overly restrictive.
- "Semantic capture" suggests using semantic web technologies to organize a Global Digital Mathematics Library. These technologies might not be the most appropriate choice because they are less expressive than other existing technologies, which are based on stronger logical foundations.

For these reasons, it is perhaps best to avoid even using of the term "semantic capture." Also, for these and other reasons, not only did the Workshop fail in its stated goal to create a semantic language for mathematics, but it also did not provide a specific direction for this supposed first step or for the ultimate goal of achieving a Global Digital Mathematics Library.

The organizers believe that the failure of the Workshop to produce a concrete plan for creating an explicit semantic language (or, in fact, to reach consensus on most detailed aspects) can be attributed primarily to two factors. The first factor was the Workshop's size, which, at 36, was sufficiently large that it became difficult to reach closure in discussions. The second factor was related to inertia, in the sense that currently extant approaches and philosophies for encoding mathematical knowledge are sufficiently specialized and entrenched among different fields and interest groups.

This latter factor is perhaps the more significant of the two, since it suggests that even had the Workshop consisted of a smaller number of more like-minded participants who were able to reach consensus over the three days of meetings, any subsequent attempt to promote and implement the "agreed upon" semantic capture language would almost certainly have been neither widely endorsed nor adopted by relevant practitioners in the field. It should also be noted that even within the "single" community of theorem provers, the variety of approaches to representing mathematical theorems is so diverse in design and detail that automated translations from one proof system to another remain an active research project (the OpenTheory project)—this despite the fact that each system already encodes mathematical meaning in an unambiguous and computer-readable syntax.

A number of systems and languages were represented at the Workshop, including systems for symbolic computation:

- Mathematica/The Wolfram Language
- Maple
- Magma

Interactive theorem proving systems (and languages):

- Lean
- HOL Light
- Coq

- Isabelle
- Mizar
- Theorema
- MathScheme

And languages for mathematical knowledge management:

- OMDoc/OpenMath
- MathLang
- OntoMath

While none of these systems or languages emerged as an obvious or viable choice for semantic representation of general mathematics, what did become clear in the course of talks and discussions is that increased interoperability between existing systems and languages is potentially the single simplest and most efficient means of making mathematics more computable and its representation more robust and standardized. As a result, proponents of each of the above (as well as other) systems should strive to maximize the capabilities of their systems with an eye toward interoperability, functionality, flexibility, and relevance to any potential future Global Digital Mathematics Library. The widespread use and extremely flexible nature of the Wolfram Language make it an attractive candidate for such work, although a subset of Workshop organizers, participants, and White Paper contributors must admit an obvious bias in that direction.

As a result of the above, in the rest of this Paper, we abandon focus on the possible development of a "language" and instead discuss the more general problems of how, in light of the discussions held and expertise represented at the Workshop, to best work toward a Global Digital Mathematics Library.

Global Digital Mathematics Library

As was already appreciated, but became again abundantly clear during the Workshop, any vision of a Global Digital Mathematics Library will not be trivial to realize. Two especially significant obstacles are:

- Foremost, collecting all published mathematical papers and books in an all-encompassing library will be impossibe due to copyright issues. Publishers consider the documents containing the mathematics to be their property and might not be willing to cooperate with such an initiative. This is an issue that has to be addressed in a collaborative way from the start.
- Transforming mathematical content to a form that is digestible by many systems of computer mathematics will be very labor intensive. Any such task must be performed carefully to avoid pitfalls, such as the dilution of precise and technical results into approximations that can be more easily represented but are mathematically not always correct.

Fortunately, there are also some possible ways to address these obstacles:

- Platforms like Wikipedia and arXiv are successful because they are populated with content using crowdsourcing. This is the most attractive way to get a Global Digital Mathematics Library into nontrivial shape. However, for this to succeed, it must be attractive for a “crowd” to contribute to the project.
- A combination of strong automation and machine learning technologies should reduce the amount of work significantly. However, one does not want to risk "polluting" the library with mathematically incorrect semi-results. Thorough application of theorem proving might be a protection against this.

- To reduce the problems with intellectual property, a separation between different parts of mathematical content may be needed. Specifically, there is the separation between, for instance, a PDF file that has to be bought from a publisher, and related data that is publicly present, but implicitly so. Since it might not be possible to obtain a custom version of a file from the publisher, a way of "anchoring" data to the unmodified published file must be devised.

Realizing a global library that contains the full extent of current mathematical knowledge, all in a fully digital form, is a very attractive but utopian-sounding goal. Fortunately, the existence of successful projects like Wikipedia and arXiv show that a Global Digital Mathematics Library should be achievable in practice. Ideally, such a library should have properties like the ones listed below. These properties should not be taken as absolute requirements but should be considered a "point at infinity" to be aimed for. It should also be emphasized that the Workshop resulted in consensus that the various coordinates of the "point at infinity" are useful and desirable, even though differences inevitably arose as to what path should be taken to reach it.

- The library should contain the full mathematical literature (peer-reviewed research mathematics), including all published versions of each work. It has been estimated that more than 200,000 theorems are proved each year, which is a very large amount of mathematics, though small in relation to other corpuses treated today, such as medical records.
- The library should be able to cope with all major systems for computer mathematics, not only symbolic computation systems and theorem provers, but also systems like LaTeX and MathML processors needed to process the documents.
- All formulas (terms, statements) in the mathematics should be usable in various computer systems for doing mathematics.
- All algorithms in the mathematics should be available in an executable form.
- All structures in the mathematics should be made explicit. For example, links between articles, references within the articles, and the substructure of proofs (like scopes of variables and assumptions) are all to be explicit.
- All mathematics should be available in a fully verified form, with verification to the highest possible standards.
- The relation between the human- and machine-readable forms (including the fully verified form) of the mathematics should be clear and explicit.
- The library should be fully accessible to the whole world, making the intellectual property rights of everything fully explicit and aiming for maximal availability.

As a cautionary note, it should also be emphasized that creating a library like this is not a completely new idea. Leibniz, Peano, and, as a more recent example, the Bourbaki project already had the dream of organizing all known mathematics into a comprehensive library. The QED Manifesto also describes a closely related vision. Sadly, both the Bourbaki and QED Manifesto projects can be said to have not yet realized their ambitious visions. Those passionate about a Global Digital Mathematics Library should take note of these facts and learn from them, while at the same time pressing forward and attempting to succeed where other similarly ambitious projects have heretofore not done so.

Questions

In order to decide in what direction to proceed, three pointed questions must be answered:

- Who will do the work?
- Who will benefit from it in the short term?

- What organizational structure will allow organizing the long-term development of the needed projects over the next decades?

Putting a framework in place and hoping that it will be taken up by others will only be a successful strategy once there is a reason for others to invest time and energy into using that framework. For this to be attractive, there will have to be a payoff for them. This was the most commonly recurring theme in the internal discussions at the Workshop: it is important to start investigating use cases before continuing with actual work.

There also needs to be a measure of agreement on the guiding principles for design, which may only be possible after more pilot experiments. Depending on the clients intended to be served (directly human, computational systems, etc.), one may need different levels of modularity and universality. In basic practical terms, one can formulate a list of simple initial questions concerning how progress will take place:

- How will the math literature be partitioned for consideration by parallel projects?
- How will expert groups be assembled?
- How will different groups exchange requirements?
- How will straw-man designs be evaluated?
- Who will be the deciding authority for different areas?
- Who will pay for different areas?
- How will the security and integrity of the data be ensured?
- How will the openness, maintenance, and sustainability of semantically encoded mathematics be ensured?

It should be noted that while definite answers to these questions are not mandatory in order for further work toward a digital library of mathematics to be initiated, they must be resolved before such work can be unified, streamlined, and made into something significant, useful, and lasting for the good of the mathematical community.

Possible next steps

After having raised a number of fundamental and difficult issues identified at the Workshop, this Paper concludes with three specific suggestions for small, concrete steps that could be taken to keep the discussion going and possibly even help advance the goal of reaching a Global Digital Mathematics Library in some small way. These were suggested at Workshop sessions, in discussions at the Workshop, and in the days after the Workshop. They were not universally endorsed by every participant but do reflect the opinions of various groups of participants.

Some participants suggested holding a recurring Workshop bringing together the computational mathematics and theorem proving communities at intervals of every two years. Since any semantic language for mathematics and its software implementation must be used by working mathematicians, development must certainly be carried out in close collaboration involving these and other groups, in particular guided by the experience of the mathematical knowledge management community. Various Workshop participants were convinced that modern machine learning techniques will substantially accelerate the development of the theorem prover technology over the coming years. While it is unclear at the present time what organization or organizations could take the lead in organizing such regular conferences, a follow-up conference every few years might prove fruitful.

Another shorter-term, concrete, and straightforward step toward raising awareness and interest and fostering discussions about the semantic representation of mathematics would be to start a Google Group on this topic. For example, it is worth noting that even the quite specialized group on homotopy type theory (https://groups.google.com/forum/#!forum/homotopytypetheory) has a steady flow of questions, answers, and ideas.

A final proposal advanced by a number of Workshop participants was to investigate which existing systems would best serve as a starting point for developing a Global Digital Mathematics Library by means of a competition. One outline according to which such a competition could be organized is as follows:

- A panel of eminent research mathematicians (*not* computer scientists, logicians, theorem provers, etc.) would be convened on a volunteer basis to select the content for the competition and judge submissions.
- The panel would select a small number of mathematical papers (possibly 3-4) as the focus of the contest. To be fair, the papers chosen should contain mathematics that is both computational and theoretical, concrete and abstract, and analytical and algebraic.
- Contestants would use a collection of particular systems, languages, and libraries to faithfully encode, as best as possible, the contents of these papers (or some subset of them) and submit the corresponding computer files together with an explanatory statement.
- The panel (possibly with technical assistance from experts in the use of certain systems) would then judge the entries, and a prize would be awarded to the entry that is judged the best. Some sort of funding would be required to cover the cost of a monetary or another type of prize. All entries would subsequently be published on the web for discussion.

While not all Workshop participants endorsed or agreed with this proposal, it seems likely that a competition of this nature could garner not insignificant attention at least in math blogs and newsgroups.

May 2016

Patrick Ion
Michael Trott
Eric Weisstein
Freek Wiedijk