How to Build a Global Digital Mathematics Library

Stephen M. Watt
David R. Cheriton School of Computer Science
University of Waterloo
Waterloo, Canada N2L 3G1
smwatt@uwaterloo.ca

***Abstract*—As with many other areas of study, mathematical knowledge has been produced for centuries and will continue to be produced for centuries to come. The records have taken many forms, from manuscripts, to printed journals, and now digital media. Unlike many other fields, however, much of mathematical knowledge has a high degree of precision and objectivity that both gives it permanent utility and makes it susceptible to mechanized treatment. We outline a path toward assembling the world's mathematical knowledge. While initially in the form of a comprehensive digital library of page images, we expect evolution toward a knowledge base supporting sophisticated queries and automated reasoning. It is the aim of the nascent International Mathematical Knowledge Trust to provide a framework and to foster a community to make progress in this direction. We can foresee that such a knowledge base will enhance the capacity of individual mathematicians, accelerate discovery and allow new kinds of collaboration.**

I. Introduction

We live in a time of unprecedented access to information and tools to make use of it. Data sets are readily accessible for everything from exoplanets to street views. Several projects provide extensive collections of the written word, including, for example, Project Gutenberg [1], the Internet Archive [2], Google Books [3] and JSTOR [4], each with a different focus.

It is natural to ask what can and should be done in this regard in the area of mathematics. Indeed we will want to go further than thinking about articles and collections and ask how we might curate, manage and use the *knowledge* contained in the articles.

Perhaps more than in any other field, knowledge in mathematics is unique in its precision and enduring utility. The mathematical literature, however, is widely scattered and uses a variety of inconsistent notations and convention. It is difficult for non-specialists to combine results from several papers with confidence in correctness and consistency.

We are interested in the technical questions that arise in the entire process, from digitization and recognition of documents to the mathematical knowledge management. For example:

- How can the existing literature repositories be united?
- What forms of semantic representation are most achievable and useful for mathematical knowledge?
- How can mathematical OCR and natural language processing be used in a semi-supervised machine learning bootstrap process?

We are equally interested in the organizational questions, such as

- How can we build upon existing research projects around the globe?
- How can we most effectively engage relevant commercial enterprises including publishers and software companies?
- How can these efforts be brought to the public in a coherent and sustainable fashion?

II. The Past

The notion of a comprehensive digital mathematics library has occurred to many over time as desirable, but bringing it to fruition has been elusive. Here we outline a few elements from the past that have brought us to the present day. Early proponents of this concept (mid 1990s) included Bernd Wegner of Zentralblatt Math, Keith Dennis of MathSciNet and Paul Ginsparg of arXiv, among others.

Desiderata for such a digital mathematics library were endorsed by the General Assembly of the International Mathematical Union in 2006 [5] and included these points that each item should satisfy:

- Accurate metadata consistent with agreed upon standards.
- A separate list of references (when available) with links to the indexing databases Mathematical Reviews and Zentralblatt Math.
- A high-quality scanned image of each page
- The text derived from optical character recognition (which is normally hidden from the reader, but keyed to the image for searching).

It was identified that the components of the digital mathematics library should be connected, both to each other and to the current literature and it was further proposed to create a registry of digitization projects.

At the end of 2006, Philippe Tondeur, the then past Director of the Division of Mathematical Sciences of the US National Science Foundation, gave a presentation on the current state of work toward this initiative and the size of the problem [6]. He observed that the mathematics inventory then consisted of about 2,300 journals and periodical publications with 2,000,000 articles identified since 1868, growing at about 80,000 additional items per year, for a total of about 50,000,000 pages. He listed some reasons for a world digital mathematics library, including the longevity of relevance of mathematical results and the acceleration of research. He also observed some challenges, including

- decisions on what material to select
- the file formats to be used
- the format of the metadata
- dealing with copyright holders
- the cost of long-term archiving
- avoiding international rivalry
- maintaining compatibility with other disciplines
- the ability to sustain ongoing literature.

Many of these problems had been solved or close to solution at the time, but some still remain today.

The decade of the 2000s saw large digitization projects undertaken, with publishers and other interested parties scanning journal backfiles. One of the notable collections efforts in the 2000s was the French project NUMDAM, NUMérisation de Documents Anciens Mathématiques [7]. NUMDAM had as one of its major objectives to make available to the global scientific community the French heritage of mathematics. It presently shows about 55,000 articles with some 1,000,000 digitized pages. Around 2010 the EuDML initiative [8] was formed with the broader scope of creating an enduring digital collection, developed and maintained by a network of European partner institutions. EuDML currently indexes 253,016 articles from 323 journals as well as about 6000 individual contributions from edited books, monographs, etc.

A number of research conference series are devoted to the analysis and recognition of documents, with the *International Conference on Document Analysis and Recognition* and *Document Analysis Systems* being of particular interest. In 2008 the series *Towards a Digital Mathematics Library* was initiated, holding events annually up to the present as part of the Conferences on Intelligent Computer Mathematics (CICM). The treatment of mathematical knowledge (as opposed to documents) has been the subject of the *Mathematical Knowledge Management* (MKM) conference series, initiated in 2002 and since 2008 held as part of the CICM. The proceedings of these conferences have been published by Springer Verlag [9]–[16], [16].

The Alfred P. Sloan Foundation has provided considerable support for activities in this area. A grant in 2012 supported the Future World Digital Mathematics Library Symposium to explore more modern aspects of what a digital mathematics library or knowledge base could provide, including such topics as mathematical search [18]. This was followed by support in 2013 to a committee of the US National Research Council to deliberate and write a report on the current state [21].

III. The Present

The next major action was taken by Ingrid Daubechies, then ending her term as President of the International Mathematical Union, at the 2014 International Congress of Mathematicians in Seoul. This was to convene a working group to deliberate the necessary actions to move the prospects forward for the conceived digital library of mathematics. This IMU Working Group resolved to create a not-for-profit organization, the International Mathematical Knowledge Trust (IMKT) [22], to coordinate a network of contributing participants, including a number of existing active research endeavours, to bring the results to a publically useful form.

The creation of the IMKT is supported by the Alfred P. Sloan Foundation, as have been related research activities involving members of the IMU Working Group. The Foundation has supported a workshop on Semantic Capture Language Design [23], held at the Fields Institute and a project to prototype potential library features focussing on continued fractions [19] (now incorporated into Wolfram Alpha). Additionally, a research project in using statistical machine learning models to extract mathematical knowledge from the scientific literature [20] is underway.

There are many fine examples of hand-built mathematical knowledge bases available today. These include

- NIST Digital Library of Mathematical Functions dlmf Online Encyclopedia of Integer Sequences [28]
- Dynamic Dictionary of Mathematical Functions [29]
- Online Integral Calculator [30]
- Inverse Symbolic Calculator [31]
- Atlas of Finite Simple Groups [32]
- L-functions and modular forms database [33]
- Combinatorial Statistic Finder [34]
- A Catalogue of Lattices [35]
- Encyclopedia of Triangle Centers [36]
- the eCF database of continued fraction knowledge [19]
- MathWorld [37]
- Wikipedia Math Project [38]
- the libraries of proof systems such as Mizar, Flyspeck, Coq and Isabelle/HOL
- the algorithm libraries of computer algebra systems such as Maple and Mathematica

Some of these have a long history while others are newly developed.

For the hard problem of automated collection and assembly of mathematical knowledge and expressing the results, there have been some notable results, including the work of Kamareddine [24] and Ganesalingham [25] and going back at least as far as de Bruijn [26]. The problem of parsing mathematical expressions from the printed page has been addressed by various authors and the state of the art continues to advance. Some of the issues there are to resolve ambiguities in grouping and baseline and problems arising from the fact that the same notation may mean many different things and, conversely, one concept may have many different notations.

IV. The Future

Given the success of other large-scale digitization efforts, there remain few technical difficulties to achieving the original goal of a global digital mathematics library encompasing all of the mathematics literature, but there do remain obstacles in the organization of the activity. In particular, aligning the different commercial, social and scientific goals of different stakeholders on issues such as copyright, data licenses, activity ownership, and the like, present challenges.

As we are starting new activities in earnest today, we should ask ourselves what more ambitious goals might we set beyond

achieving a basic set of well-indexed, cross-linked articles. We have already talked about the desire to capture and use the knowledge contained in these pages, but what might this mean?

This was one of the main points addressed in the *Mathematical Information in the Digital Age of Science* workshop held as part of the 2016 Joint Mathematics Meetings held January 2016 [39]. Some of the immediate goals set out there were:

- Determining whether a result is known, where the answer is hoped to be positive so the result can be used. Here, proofs, examples, counter-examples, applications, and so on would be desired.
- Determining whether a result is known, where the answer is hoped to be negative and thus confirm a discovery or avoid unintentional duplication.
- Accelerating the advancement of mathematics.
- Organizing knowledge to accelerate the learning of mathematics.
- Enhancing collaboration.
- Making existing tools more powerful.

Some of the goals that were further afield were:

- Certification (machine validation) of all mathematical knowledge.
- To serve as a mathematical assistant, or as a teacher.
- To identify holes in our mathematical knowledge.
- Conjecture generation.
- Reflection— refactoring or reformulating mathematics for elegance or ease of application.
- Expanding the mathematical capacity of humanity.

Most immediately, the International Mathematical Knowledge Trust is launching a set of seed projects that should each be useful in their own right while also providing insights in setting an overall roadmap. These seed projects are:

Concordance of the vocabulary of special functions.

While there is general agreement on the terminology and conventions for the classical special functions, there remain differences among written works and software systems on such matters as notation, argument order, sign conventions, normalization factors and branch cuts for complex inverse functions. But as there is already a good degree of agreement in the subject, this is a good place to start to study the issues arising in designing ontologies and the operations upon ontologies. A concordance among the software systems Mathematica, Maple, Matlab, Sage and the NAG libraries and reference works such as the NIST Digital Library of Mathematical Functions would be a practical goal. It is recognized before we begin that there are some areas where these systems and written works are under-specified and that it will be undecidable to determine zero equivalence of symbolic expressions using mappings to standard ontology definitions.

Formalization: FAbstracts and systems harmony.

Some set as an ultimate goal the full formalization of usable mathematics appearing in the literature. With some thought, a number of issues come to the fore: What should be formalized? Each paper, as written? Papers after the obvious errors are corrected? How far should this go? Secondly, the degree of desired formalization varies by user and application. For example, an easy position would be to say that all statements must be validated with a proof assistant, such as Coq, Flyspeck, etc. Another view would be that some level of imprecision and ambiguity would allow much more of the literature to become available in a shorter time frame and would nevertheless be useful. A useful approach would be to allow the level of formalism to vary, with only some things being validated rigorously. This is the basis of the "flexiformalist" approach [40].

A second useful approach would be to assemble a complete set of formal definitions and only the statements of the theorems in a corpus of articles, leaving the proofs of the theorems aside. This is the notion of formal abstracts (FAbstracts) that has been discussed in the formal mathematics community for some time. The second seed project is to develop a set of formal abstracts for a well identified subset of the literature. This would be done entirely by hand initially, but it is anticipated that far down the line this could become a semi-supervised activity.

Another project that falls under formalization is to align the pieces of mathematics in overlaps between the libraries around different proof systems. This is not just a matter of the way the proofs proceed or the choice of conventions but may involve subtleties in the semantics resulting from different foundational bases.

Mathematical n-gram analysis.

As with the formalization, the language processing of articles can proceed in stages. There are several sorts of document analysis used in other settings that could be applied to mathematical documents, and a promising one of these is n-gram analysis. In previous work, we have examined symbol and n-gram frequencies in the mathematical expressions appearing in two corpora: the mathematics articles submitted to the arXiv and second year engineering mathematics textbooks [41], [42]. This seed research activity would use n-gram analysis techniques to the zbMath corpus. An interesting extension would be to allow wild cards, to recognize common shapes of expressions via anti-unification. As a basis for machine learning this is a first step.

V. Conclusions

The development of an *international digital mathematics library* is underway. The new International Mathematical Knowledge Trust endeavours to assist the multiple interested groups in coordination and has seed projects of its own. There are many immediate and compelling uses for a comprehensive digital mathematics library, ranging from mathematics research to sound application of well known math. The more open-ended goals of an *international mathematics knowledge base* are tantalizing in their potential. The long useful lifetime

of mathematical knowledge and its susceptibility to mechanized treatment make mathematics an ideal target. We hope to see much more activity in this area as time goes on.

References

[1] *Project Gutenberg* http://www.gutenberg.org/.

[2] *Internet Archive* http://archive.org/.

[3] *Google Books* http://books.google.com/.

[4] *JSTOR* http://www.jstor.org/.

[5] *Digital Mathematics Library: A Vision for the Future*, endorsed on August 20, 2006 by the General Assembly of the International Mathematical Union http://www.mathunion.org/fileadmin/IMU/Report/dml_vision.pdf

[6] Ph. Tondeur, *WDML: The World Digital Mathematics Library*, slides presented at IMA Workshop "The Evolution of Mathematical Communication in the Age of Digital Libraries", http://www.math.uiuc.edu/ tondeur/WDML_IMA_DEC2006.pdf 2006.

[7] *NUMDAM*, http://www.numdam.org/?lang=en.

[8] *EuDML*, http://eudml.org.

[9] *Intelligent Computer Mathematics*, Springer Verlag LNAI 5144, 2008.

[10] *Intelligent Computer Mathematics*, Springer Verlag LNAI 5625, 2009.

[11] *Intelligent Computer Mathematics*, Springer Verlag LNAI 6167, 2010.

[12] *Intelligent Computer Mathematics*, Springer Verlag LNAI 6824, 2011.

[13] *Intelligent Computer Mathematics*, Springer Verlag LNAI 7362, 2012.

[14] *Intelligent Computer Mathematics*, Springer Verlag LNAI 7961, 2013.

[15] *Intelligent Computer Mathematics*, Springer Verlag LNAI 8543, 2014.

[16] *Intelligent Computer Mathematics*, Springer Verlag LNAI 9150, 2015.

[17] *Intelligent Computer Mathematics*, Springer Verlag LNAI 9791, 2016.

[18] *The Future World Heritage Digital Mathematics Library Symposium*, report http://www.mathunion.org/fileadmin/IMU/News/WDML_Symposium_Final_Report.pdf, 2012.

[19] Wolfram Foundation, *The eCF-Project—Final Report*, http://www.wolframfoundation.org/programs/FinalReport.pdf 2013.

[20] Alfred P. Sloan Foundation, *To accelerate scientific discovery by using statistical machine learning to enable advanced search of mathematical literature*, http://sloan.org/grant-detail/6703

[21] National Research Council, *Developing a 21st Century Global Library for Mathematics Research*, The National Academies Press, Washington DC, USA, http://arxiv.org/ftp/arxiv/papers/1404/1404.1905.pdf, 2014.

[22] *IMKT* http://www.imkt.org

[23] The Fields Institute, *Semantic Representation of Mathematical Knowledge Workshop*, http://www.fields.utoronto.ca/programs/scientific/15-16/semantic/, 2016.

[24] F. Kamareddine and J.B. Wells, *Computerizing Mathematical Text with MathLang*, Proc Second Workshop on Logical and Semantic Frameworks with Applications, Electronic Notes in Theoretical Computer Science, 205 (2008) 5-30.

[25] M. Ganesalingam, *The Language of Mathematics: A linguistic and Philosophical Investigation*, Springer Verlag LNCS 7805, 2013.

[26] N.G. de Bruijn, *Automath: a language for mathematics*, Montreal: Presses de l'Université de Montréal, 1973 .

[27] National Institute of Standards and Technology, *NIST Digital Library of Mathematical Functions* http://dlmf.nist.gov, 2009.

[28] N.J. Sloane, *Online Encyclopedia of Integer Sequences*, http://oeis.org (founded 1964, retrieved Dec 2016).

[29] INRIA, *Dynamic Dictionary of Mathematical Functions*, http://ddmf.msr-inria.inria.fr.

[30] Wolfram Alpha LLC, *Online Integral Calculator*, http://www.wolframalpha.com/calculators/integral-calculator.

[31] The CARMA Group, *Inverse Symbolic Calculator*, http://isc.carma.newcastle.edu.au.

[32] R. Wilson, P. Walsh, J. Tripp, I. Suleiman, R. Parker, S. Norton, S. Nickerson, S. Linton, J. Bray and R. Abbott, *Atlas of Finite Simple Groups, Version 3*, http://brauer.maths.qmul.ac.uk/Atlas/v3.

[33] *L-Functions and Modular Forms Database*, http://www.lmfdb.org.

[34] *Combinatorial Statistic Finder*, http://www.findstat.org.

[35] G. Nebe, *A Catalogue of Lattices* http://www.math.rwth-aachen.de/Gabriele.Nebe/LATTICES.

[36] C. Kimberling, *Encyclopedia of Triangle Centers*, http://faculty.evansville.edu/ck6/encyclopedia/ETC.html

[37] Wolfram Research, *Wolfram MathWorld: the web's most extensive mathematical resource*, http://mathworld.wolfram.com.

[38] Wikipedia, *WikiProject Mathematics*, http://en.wikipedia.org/wiki/Wikipedia:WikiProject_Mathematics.

[39] P.D.F. Ion, O. Teschke, S.M. Watt, *MIDAS: Mathematical Information in the Digital Age of Science*, http://jointmathematicsmeetings.org/amsmtgs/2181_abstracts/1116-00-2114.pdf

[40] M. Kohlhase, *The Flexiformalist Manifesto*, Proc Symbolic and Numeric Algorithms for Scientific Computing (SYNASC 2012), IEEE, 2012.

[41] Clare So and S.M. Watt, *Determining Empirical Properties of Mathematical Expression Use* , pp. 361-375, Proc. Fourth International Conference on Mathematical Knowledge Management, Springer Verlag LNCS 3863, 2005.

[42] S.M.Watt, *Mathematical Document Classification via Symbol Frequency Analysis*, pp. 29-40, Proc. Towards Digital Mathematics Library, (DML 08), IEEE Computer Society, 2008.