Estimating the Cost of a Standard Library for a Mathematical Proof Checker

Freek Wiedijk

freek@cs.kun.nl
University of Nijmegen

Abstract. We estimate the cost of formalizing a proper standard library for proof checking of mathematics in the spirit of the QED project. Apparently it will take approximately *140 man-years*. This estimate does not include the development of the proof checking program, nor does it include work on the metatheory of that program.
This should discourage any individual or small research group to think they can reach anything like the goal of the QED project on their own.

1 Introduction

1.1 Problem

The utopia of the QED manifesto [1] is that in the future all of mathematics will be formalized in a computer. That it will be formalized in such a way that one can be completely sure that it is correct. (One can be completely sure that it is correct, *except* of course for paradoxes in the foundations of the logic and bugs in the kernel of the checking software. But that logic can be very simple and clear and that kernel can be very small.)

This dream has been dreamt many, many times (the Automath project [5], the Mizar project [7], the QED project [1], the MBase project [2], it has been dreamt many times) and until now nothing has happened. A favorite passtime of people who like to dream this dream is to guess how long it will take before the utopia arrives.

Our guess is that when enough mathematics has been formalized, when the *library* of formalized mathematics is powerful enough, people will start routinely formalizing mathematics. So the thing to estimate is: when will this library be sufficiently large?

That's the subject of this note.

1.2 Approach

'Drake's equation' is a way to guess the number of alien civilizations in our galaxy that are able to communicate with us. It guesses various factors and multiplies them together to get a very rough estimate of this number (the point of course is that this estimate is bigger than zero). We will use a similar approach to Drake's equation. We will take various factors and combine them to get an estimate on the number of man-years it will take to get a proper library of formalized mathematics.

The difference between our approach and Drake's equation is that the numbers in Drake's equation are all highly speculative while our numbers are backed by solid experience in research in formal mathematics.

1.3 Related work

As far as we know this estimate has not been published before.

1.4 Contribution

We estimate the amount of work needed to arrive at a library that will be a proper basis for a world in which mathematics can be formalized routinely.

1.5 Outline

In Section 2 we argue what we think should be in a good library for formal mathematics. In Section 3 we study how expensive formalizing mathematics is. In Section 4 we combine these numbers to estimate the cost of a good library for formal mathematics.

2 The union of what all mathematicians know or the intersection of what all mathematicians know?

Projects that dream the QED dream tend to promise the sky. Their goal invariably is to formalize *all* of mathematics. (As an example of this kind of 'grand dream', the people of the MBase project [2] compare their goals to that of the human genome project. Probably they think that 'if you promise more, you get more.' Funding money for their project, that is.)

However this might not be the right approach. Consider the field of *hypertext*. This used to be an utopian field, just like formal mathematics. Only a few researchers were active in this area, dreaming of a future in which all text would be linked together in an interactive net of hypertext. And then, suddenly, there was the World Wide Web. Clearly the thing needed for hypertext to take off was not that it was *perfect*, but that it was *good enough*.

We hope for a similar thing to happen to formal mathematics. So what would *good enough* mean for a library of formal mathematics?

Our guess is that it should contain what working mathematicians take for granted. These people don't want to be bothered with formalizing mathematics that 'everybody knows'. So the library should contain the common knowledge of most mathematicians. A good approximation to this is:

the mathematics in the undergraduate curriculum of a mathematics study

In [8] we investigated the programs of the undergraduate mathematics studies of the six Dutch universities in which you can study mathematics. All six turned out to be almost exactly the same. They all had the following 12 subjects:

a		Algebra
b		Analysis
c		Calculus
d	05	Combinatorics
e	30	Complex Variables
f	34	Differential Equations
g	51	Geometry
h	28	Integration
i	15	Linear Algebra
j	03	Mathematical Logic
k	60	Probability Theory
l	54	Topology

(In the second column is the MSC classification of these subjects, when appropriate.) So this is what we think needs to be in a good library of formal mathematics.

3 How much work is it to formalize something?

In [9] we studied the relation between the size of an informal text of mathematics and its formal counterpart. We called this the *de Bruijn factor*. We measured it for various formalizations in various systems (good ones, like Mizar and HOL) and in all cases it turned out to be close to 4. So this number predicts the size of a formalization.

To predict the amount of work needed to *write* such a formalization we use two data points:

- In the FTA project in Nijmegen [3] a formalization of 1 megabyte of formal text was created with approximately 3 man-years of work.
- In the Mizar library MML the average size of an *article* is 70 kilobytes and writing such an article takes about 1 month. If we include revisions to the article that are made afterwards, we can estimate the total work on an article through time to be 1.5 months. This means it takes is 2 man-years to formalize 1 megabyte of MML text.

Below we use the factor of 2.5 man-years per megabyte.

4 The cost estimate

To get a proper library of mathematics it should include a formalization of textbooks for all 12 subjects from Section 2. This textbook should thoroughly cover the basics of the subject. We guess that such a textbook will have to be about 400 pages long. A page in such a textbook is about 3 kilobytes of text.

Now we can put our numbers together. 'Drake's equation' to estimate the cost of a good basic library for formal mathematics is:

$$\textit{number of subjects} \cdot \textit{textbook pages per subject} \cdot \textit{kilobytes text per page} \cdot \\ \textit{de Bruijn factor} \cdot \textit{man-years per kilobyte of formalization}$$

If we plug in the numbers we get:

$$12 \cdot 400 \cdot 3 \cdot 4 \cdot \frac{2\frac{1}{2}}{1024} \approx 140$$

So it will take 140 man-years to create a good basic library for formal mathematics. Here are some numbers *per subject* that are related to this estimate, as a sanity check:

Amount of text in a solid textbook	1.2 megabytes
Size of a formalization of such a textbook	5 megabytes
Number of MML article equivalents	72 articles
Cost to formalize such a textbook	12 man-years
Number of Ph.D. equivalents to do this	6 Ph.D.s

(We assume here that while doing a Ph.D., about 2 years work is spent on formalizing, and the rest of the time is spent on other things like writing the thesis. So each Ph.D. equivalent would write 12 MML article equivalents in that time.)

5 Conclusion

5.1 Discussion

Utopian people tend to think they can change the world alone. People who dream the QED dream like to think that they can make the *perfect proof checker* by themselves and when it's there the QED utopia will arrive.

We claim that the library is much more important that the system. A good system without a library is useless. A good library for a bad system still is very interesting (the system might be improved or the library might be ported to a different, better, system). So the library is what counts.

Therefore the cost estimate from this note should discourage people or even small research groups from thinking that *they can change the world alone*. They can't. It just is too much work.

It is interesting to compare the numbers of this note to the size of the Mizar library MML. It consists of 700 articles which took about 90 man-years to create. It might seem that the 'standard library' as we describe it in this note is already there. However most articles in the Mizar library are about topics that would not be in the library that is proposed here. In the MML many articles are about more advanced mathematics than are treated in an undergraduate curriculum and many MML articles are about applications from computer science. Also the Mizar library has not been created by translating textbooks. Therefore it is not as complete and well organized as a systematic translation of a good textbook would be.

One might try to modify the numbers in our equation based on optimistic thoughts about what the future might bring. Maybe the proof checkers of the future will be so much better that the de Bruijn factor drops significantly (maybe even below 1!) Or they become so much easier to use that a person can produce much more formalized text in the same time.

One should realize that there are reasons to think the numbers might have to be adjusted in the other direction. In a textbook basic results often are just taken for granted. For instance in a book about complex variables (subject 'e' in the list on page 3) the *Jordan curve theorem* probably is passed over in a few lines. In the Mizar project people have been working on a formalization of this theorem since 1992 [4] and still they're not finished. Currently there are about 30 articles in this formalization, so until now it has taken around 4 man-years of work. (Admittedly, if they had chosen a more abstract proof, like the one in [6], they would probably have been finished by now.)

Apparently formalizing only a few lines in a textbook can sometimes take a very long time.

5.2 Future work

Create a library as described in this note.

References

1. R. Boyer. The QED Manifesto. In A. Bundy, editor, *Automated Deduction - CADE 12*, volume 814 of *LNAI*, pages 238–251. Springer-Verlag, 1994.
   URL: `<http://www.cs.kun.nl/~freek/qed/qed.html>`.
2. Andreas Franke and Michael Kohlhase. System Description: MBase, an Open Mathematical Knowledge Base. In *CADE-17, Pittsburgh*, 2000.
3. H. Geuvers, F. Wiedijk, and J. Zwanenburg. Equational Reasoning via Partial Reflection. In *Theorem Proving in Higher Order Logics, 13th International Conference, TPHOLs 2000*, volume 1869 of *LNCS*, pages 162–178, Berlin, Heidelberg, New York, 2000. Springer Verlag.
4. Yatsuka Nakamura and Jarosław Kotowicz. The Jordan's Property for Certain Subsets of the Plane. *Journal of Formalized Mathematics*, 4, 1992. MML Identifier: `JORDAN1`.
5. R.P. Nederpelt, J.H. Geuvers, and R.C. de Vrijer. *Selected Papers on Automath*, volume 133 of *Studies in Logic and the Foundations of Mathematics.* Elsevier Science, Amsterdam, 1994.
6. M.H.A. Newman. *Elements of the Topology of Plane Sets of Points.* Dover Publications, 1992.
7. F. Wiedijk. Mizar: An Impression.
   URL: `<http://www.cs.kun.nl/~freek/mizar/mizarintro.ps.gz>`, 1999.

8. F. Wiedijk. Selecting the Domain of a Standard Library for a Mathematical Proof Checker.
URL: <http://www.cs.kun.nl/~freek/notes/mathstdlib.ps.gz>, 1999.
9. F. Wiedijk. The De Bruijn Factor.
URL: <http://www.cs.kun.nl/~freek/notes/factor.ps.gz>, 2000.