A digital library based on Mizar

Jeremy Gow[⋆]
[⋆]UCL Interaction Centre
London
j.gow@ucl.ac.uk

Paul Cairns[†]
[†]UCL Interaction Centre
London
p.cairns@ucl.ac.uk

Abstract

Many mechanized reasoning systems have facilities for presenting formal content in a more human-oriented way, from outputting web pages to generating verbalisations of proofs. However, these are still some way from producing materials suitable for a mainstream mathematical audience. As an exploratory investigation into this mismatch, we have built a prototype digital library based on the Mizar Mathematical Library. Understanding how it fails helps us to define more specific requirements for formal libraries and presentation technologies.

1 Introduction

The idea that digital resources for mathematics could be founded on a formal basis has received a lot of attention in recent years (Borwein and Farmer, 2006). Formal approaches have great potential, as mathematical proof can be guaranteed valid and content is far more amenable to automatic processing. However, a key problem is the *representational gap* between formal and informal mathematics: the vast majority of mathematics is carried out not in logic but in specialist natural language, formulae and diagrams. Most users of mathematics are not versed in formal mathematics and, even if they were, it is not yet clear that it could support their activities adequately. In principle translation across the representational gap is generally possible. In practice it requires skilled effort for either a person or a machine. The state-of-the-art in machine understanding and presentation is some way short of a mainstream audience, e.g. Zinn (2004).

Work on the presentation of formal mathematics has mainly focused on the hard problem of generating natural language proofs (Horacek, 1999; Holland-Minkley et al., 1999) and has avoided issues such as the way material is commented, structured and organised. These may also be important to user understanding and acceptance, and perhaps simpler to address. There has also been little attention given to identifying actual problems mainstream users encounter with formal material, partly because the mismatch between systems and mathematicians is often still too large.

As an exploratory investigation into these issues, we have built a simple digital library based on the Mizar Mathematical Library (MML). The library is aimed (in theory) at mainstream mathematicians and its content is automatically generated from MML articles. Although Mizar and its associated tools supports its target users extremely well, our library — as we expected — does not suit the needs of its intended audience. Our aim is to observe how the resource fails and feed this back into requirements for presentation and formalisation. We also hope it will provide a realistic context of use to test out different presentations and conduct user studies.

2 The Library

The MML offers one of the largest resources of formally proven mathematics[1]. It covers a wide range of topics in from computation through abstract algebra to topology and set theory. The representational gap can be considered slightly narrower than with other systems, as i) Mizar uses a rich language that can strongly resemble the natural usage of mathematicians and ii) some MML articles are already automatically presented (minus proofs) for the Journal of Formalised Mathematics (Bancerek, 2006).

We used the Greenstone system (Witten et al., 2000) to build a digital library from Mizar's XML output, based on Josef Urban's semantic presentation stylesheet. Digital libraries are a standard means of delivering collections of (informal) content electronically: their key feature is that content is *classified* and *indexed*. They are used widely in publishing diverse collections and are a proven technology that provides valuable resources to knowledge workers in many disciplines.

[1]Currently fourth in Formalizing 100 Theorems (www.cs.ru.nl/~freek/100/).

3 Preliminary Results

The prototype version of our library includes the whole of the MML. As expected, it suffers from some severe shortcomings. Some have been addressed by previous work on proof presentation, e.g. readability and granularity, and the library would benefit from using such an approach. However, we have found a number of other general issues with navigation and the lack of structural information:

Article Structure Articles are a list of undifferentiated definitions and theorems, with no thematic sections. For larger articles this can make the table of contents, and the article itself, unnavigable. A limited amount of information is potentially available from Mizar source comments, currently ignored by the XML output.

Commentary There is a lack of glue content that provides examples, motivation and significance, with no indication of the role or importance each part has in the theory. MML articles from the JFM could use the journal's prose abstracts, but this is not sufficient.

Naming Definitions are named after the defined symbol, which is not always clear enough for searching and browsing. Theorems are labeled with a version stripped of quantifiers and preconditions, and this is often too lengthy or obscure. The most important (if not all) theorems and definitions need to be explicitly named, including shorter names for menus.

Proof Structure Proofs can be very large, deeply nested structures, with no clues given as to how they should be read. Presentation needs to deal with structuring the proof into understandable and navigable sections. Retaining the hierarchical structure may aid navigation but could also have its own problems (Cairns and Gow, 2003).

Search A full-text search index is not an effective way to retrieve information in the library: finding occurrences of particular symbols is probably insufficient for our users. A more specialised search feature needs to be developed.

4 Further work

Continued development of the library will iron out some problems: some form of NL generation and an specialised search facility are badly needed. Mizar's XML output also needs to be a more faithful representation of the source, including comments and variable names.

We anticipate that the more general issues identified so far need to be addressed by improvements in formalisation practice and presentation techniques. Formalised mathematics could provide more structured and annotated content — unfortunately, there is little motivation for authors to do this given that most articles are primarily addressed to machine rather than human readers. On the presentation side, better use needs to be made of structural information that is present or can be inferred. We hope further development of the library will lead to more specific requirements for formal mathematics and its presentation.

References

G. Bancerek. Automatic translation in formalized mathematics. *Mechanized Math. and Its Applications*, 5(2):19–31, 2006.

J. Borwein and W. Farmer, editors. *Proc. 5th Int. Conf. on Mathematical Knowledge Management*, 2006. Springer.

P. Cairns and J. Gow. A theoretical analysis of hierarchical proofs. In A. Asperti, B. Buchberger, and J. Davenport, editors, *2nd Int. Conf. Mathematical Knowledge Management*, pages 175–187. Springer, 2003. LNCS 2594.

A. Holland-Minkley, R. Barzilay, and R. L. Constable. Verbalization of high-level formal proofs. In *Proc. 16th National Conference on Artificial Intelligence (AAAI-99)*, pages 277–284. AAAI Press, 1999.

H. Horacek. Presenting proofs in a human-oriented way. In Harald Ganzinger, editor, *Proc. 16th Int. Conf. on Automated Deduction (CADE-16)*, pages 142–156. Springer, 1999. LNCS 1632.

I. H. Witten, R. J. McNab, S. J. Boddie, and D. Bainbridge. Greenstone: A comprehensive open-source digital library software system. In *Proc. ACM Digital Libraries, San Antonio, TX*, pages 113–121, 2000.

C. Zinn. *Understanding Informal Mathematical Discourse*. PhD thesis, Arbeitsberichter des Instituts für Informatik, Friedrich-Alexander-Universität, 2004.