Authoring Mathematical Knowledge

Bruce R. Miller*

November 25, 2003

1 Authoring Mathematical Knowledge

The Digital Library of Mathematical Functions project[1] is developing a successor to Abramowitz and Stegun's *Handbook of Mathematical Functions*. This new work will be published digitally as an annotated, searchable web site[1], and in print. An important issue is how to get the 'Mathematical Knowledge' from the authors into a form that can be 'Managed'. Given our context — author familiarity, need for quality printing — LaTeX is the natural choice for source documents. However, while LaTeX documents tend to be far more structured than those of TeX or other systems, automatic extraction of the knowledge for reuse in computational systems is non-trivial.

In this presentation, I describe the tool we are developing for this purpose, LaTeXML[2], the design issues we have dealt with, as well as some open issues. While the immediate motivation is the DLMF project, the goal is to develop a general purpose, extensible, tool.

To leverage author's familiarity, LaTeXML attempts to behave as much like LaTeX as possible. The input is tokenized, expanded and processed through a digestive tract that mimics Knuth's design. The important distinction is that the Stomach builds structured 'boxes' which preserve the original markup and associate it with the desired document schema. In conjunction with a document model extracted from the DTD (or eventually XSchema), the Intestines construct a document tree — the XML — from this stream of augmented boxes. Both the boxes and the constructed tree may have additional transformations applied in order to synthesize the desired structure and semantics, for example, to introduce ligatures or to parse math formulae. The system provides the means of defining the correspondence between document markup and the desired XML structure.

In the context of Mathematical Knowledge Management, two issues stand out as particularly important: (1) extracting the mathematical content from the decidedly presentation oriented math markup of LaTeX, and (2) determining the 'role' of the mathematics and what purpose it fulfills within the document.

*National Institute of Standards and Technology

[1] See `http://dlmf.nist.gov/`

2 Mathematical Content

Inferring the semantics of formula marked up in LaTeX is not possible, in general, without knowing what notations are used and what they mean. Juxtaposition of elements, for example, requires at least minimal type analysis to determine whether multiplication, function application or operator action is implied.

Given the relatively limited notations used in the field of special functions (mainly algebra and calculus), we have been able to make good progress for our application. It currently produces passable presentation MathML with content MathML within reach. OpenMath output will require more analysis.

Two approaches to teaching LaTeXML about notations and meaning have been developed so far. For best fidelity, extensions of LaTeX markup minimizes the ambiguity of the source document itself. For example, we define:

`\hyperpFq{p}{q}@{a_1,\ldots}{b_1,\ldots}{z}` $\Rightarrow$ ${}_pF_q\left(\begin{matrix}a_1,\dots\\ b_1,\dots\end{matrix};z\right)$

helping the author, while unambiguously representing to the parser an application of the hypergeometric function to arguments. Alternatively, external sets of patterns assert that certain markup sequences represent functions, etc.

Clearly, for the general community more application specific notations must be specified, and the grammar used for mathematics parsing must be extended. The mechanisms described get us, at least, part of the way to that goal.

3 Mathematical Significance

Assuming the success of the above techniques, one already has a considerable benefit: searchable content; presented on the web. However, true MKM needs meta information about interrelationships between document elements: which are proofs; which prove which; which are definitions; and so forth. OMDoc provides an example of the kind of richness that one needs for MKM. This level of sophistication first requires a concerted effort to develop the appropriate ontologies — we look to the MKM community for this — and a serious commitment to apply them to the documents in question.

Given the appropriate ontologies, however, the framework we have presented should make the task fairly easy, simply by adding the appropriate declarative markup to the documents. In its current state, LaTeXML processes indexing keywords, free-text annotations, validity constraints on formula.

References

[1] D. W. Lozier, 'NIST Digital Library of Mathematical Functions', *An. Math. and Artificial Intelligence* **38**, p. 105, 2003.

[2] B. R. Miller, A. Youssef, 'Technical aspects of the Digital Library of Mathematical Functions', *An. Math. and Artificial Intelligence* **38**, p. 121, 2003.