Creating Webs of Math using LaTeX

Bruce R. Miller[*]

National Institute of Standards and Technology

My colleagues Dan Lozier, Abdou Youssef and I organized the mini-symposium 'Math on the Web: Content Development and Implementation' with the obvious goal of showcasing our recent progress in this field. But more than that, we were motivated by the desire to show what is becoming possible, and even more, to motivate authors to think about their role in the process. To take full advantage of the technologies that are available or are becoming available, writers of scientific material, particularly that involving mathematics, should adapt their approach to composition — emphasis on semantic rather than presentation markup has the potential for great payoff when electronic delivery is considered. In this context, I describe LaTeXML, a tool developed for converting LaTeX to XML, including MathML.

1 Background

The Digital Library of Mathematical Functions (DLMF) project was established to develop a new edition of Abramowitz and Stegun's 'Handbook of Mathematical Functions'. Clearly this is a work heavy in mathematical content — in fact, it has relatively little text. A goal of the project was to produce a book in the finest traditions of mathematical typesetting. There was as well the goal to put the work online, taking advantage of all the possibilities of the web for finding the material, presenting it, navigating through it, reusing it in applications and making the material accessible.

Virtually all authors and editors involved in the project were both familiar and comfortable with LaTeX. The high quality of its typesetting and its extensibility stood also in it's favor. Thus, the choice of LaTeX for authoring was an easy one. The only technical issue was how to get it on the web.

Although tools existed for converting LaTeX to HTML, none were quite up to the task as we envisioned it. A critical need was the extensibility and customizations necessary to preserve and enhance the semantic information present in the documents. In particular, faithful conversion of the mathematics to MathML needs more information than is available in LaTeX markup as typically written. Converting to an intermediate XML allows for a variety of analyses and data extraction to take place, such as cross-referencing and search indexing. Standard XML processing tools can then be used to convert the results into a variety of forms, such as HTML, needed for the web.

We therefore felt compelled to develop a tool, LaTeXML, to accomplish this task; Other authors and developers finding themselves in a similar situation may find LaTeXML useful.

2 Webs of Math

A document typically has many cross-references such as references to equations, citations to a bibliography, or even links from a bibliography to a full copy of a cited paper in some external archive. Tools that convert to online forms trivially capture these connections as hyperlinks that can be clicked and followed to the referenced point.

But less obvious 'webs' can significantly enhance the usefulness of the resource; connections from a symbol to the place where it is defined, or places where it is used. Interconnected indexes, keyword and glossary lists are other such features. Even less obvious webs become feasible only in electronic contexts. Layers of detail, such as full proofs or other annotations, can be normally hidden, but revealed on request. Alternative representations of the information can be offered, such as reusable standard forms of formula that can be downloaded and inserted into a computer algebra system, for example.

A recurring theme becomes apparent when one imagines implementing features such as these: The more an author concentrates on meaning rather than appearance, the more possibilities arise.

3 LaTeXml

The LaTeXML tool was developed to fill these needs. It is a tool for transforming LaTeX into an XML format closely modelling LaTeX's document structure. It has the goals of mimicking TeX's behavior as closely as possible, to be loss-less (not losing any information embedded in the authors markup), but extensible and adaptable.

[*] Email: bruce.miller@nist.gov

In the limited space available here, I'll focus on some issues of processing mathematics. While LaTeX provides more semantically oriented markup at the document-level, eg. sectioning commands, than does plain TeX, the almost exclusive focus on the presentation of mathematics, rather than content, is retained. Typical LaTeX markup of mathematics is notoriously ambiguous, resisting attempts for machine generation of, say, Content MathML. Indeed, and ironically, even the generation of high quality Presentation MathML requires a certain level of parsing.

The approach with LaTeXML, then, is to encourage higher-level markup, both functional and declarative, and to use this information to guide the math parser, and thus improve the fidelity of the conversion.

Consider a somewhat contrived example: $|a|b|c|$, likely marked up as `$|a|b|c|$`. While the human reader may possibly understand it in context, a machine parser will be baffled. A trivial macro `\abs{x}` for TeX, with the corresponding semantic mapping for LaTeXML, solves the problem handily, however. The author simply writes

`\abs{a \abs{b} c}` *or* `\abs{a} b \abs{c}`,

according to the intended meaning, and the result becomes understandable by machine.

Another example relates to the ambiguities of function application, say $f(x)$: Is f a function? Does the juxtaposition correspond to multiplication or function application? Here, LaTeXML defines markup such as

`\lxDeclare[role=FUNCTION]{$f$}`

to declare that f is used as a function, thus assisting the math parser. Additionally, this markup is taken to indicate its location as the 'definition' of f, providing the capability to link uses of symbols to their definitions.

Carrying these examples a step further to the special functions as encountered in the DLMF, notice that symbols like F and J tend to be heavily used and can stand for several distinct functions. Again, a human reader *may* be able in many cases to guess which is intended, but the machine has more difficulty. And again, the appropriate markup can resolve these problems, and in fact be easier to type than the traditional ambiguous markup.

For example, we define macros for special functions like

`\HyperpFq{p}{q}` $\Rightarrow {}_{p}F_{q}$.

Introducing @ to stand for the notion of *evaluating at* or *apply* (as you prefer) allows us to write

`\HyperpFq{p}{q}@{a}{b}{z}` $\Rightarrow {}_{p}F_{q}(a;b;z)$.

An alternative notation can be written as

`\HyperpFq{p}{q}@@{a}{b}{z}` $\Rightarrow {}_{p}F_{q}\left(\genfrac{}{}{0pt}{}{a}{b};z\right)$,

which is certainly more palatable to type than

`\sideset{_p}{_q}{\mathop{F}}\left(\genfrac{}{}{0pt}{}{a}{b};z\right)`.

More important, it has established unequivocally exactly *which* F is involved, namely the hypergeometric function, and what arguments it is being applied to. With a little extra markup (as described above), the system is able to establish a direct link from any use of F to the definition of the intended F.

4 Caveats on Leveraging

Of course, any extra markup requires an extra effort to remember and to include. Thus one is tempted to try to leverage anything in the document as far as possible. For example, indexing data as typically included in LaTeX, would seem to be a rich source of search data and for classification keywords. One must be careful, however.

While such indexing data, appropriately 'stemmed' does indeed help feed a search engine, the phrasing needs of a readable, high quality index are in fact quite different from the kind of controlled vocabulary needed by a classification system. Thus, short of specialized natural language processing, it is left to the author to duplicate information in slightly different forms.

5 Conclusion

Whatever technology one chooses to use for authoring mathematical material, we claim there are a number of very useful concepts and metadata that an author should explicitly embed in those documents that typically are not included. It is difficult to imagine an automated system in the near future correctly making these inferences. We encourage authors to get used to the idea of enriching their material with semantics and hidden metadata, in the expectation that the promises of extra features and capabilities will make this effort worthwhile.

For those authors that choose to use LaTeX, we offer our tool LaTeXML as one that may help them meet these goals. It is not perfect, but under active development — and available at `http://dlmf.nist.gov/LaTeXML/`.