How (and why) I use GLOSS to write XHTML+MathML

by
Richard Kaye (School of Mathematics, University of Birmingham, email: r.w.kaye@bham.ac.uk)

PLEASE NOTE: *This page only contains a brief abstract. To read the full article please goto Richard Kaye's web pages at:*

http://web.mat.bham.ac.uk/R.W.Kaye/papers/glossintro/

Abstract

‘A discerning friend of mine,’ said Don Quixote, ‘was of opinion that no one ought to waste labour in glossing verses; and the reason he gave was that the gloss can never come up to the text, and that often or most frequently it wanders away from the meaning and purpose aimed at in the glossed lines; and besides, that the laws of the gloss were too strict, as they did not allow interrogations, nor “said he,” nor “I say,” nor turning verbs into nouns, or altering the construction, not to speak of other restrictions and limitations that fetter gloss-writers, as you no doubt know.’

from Don Quixote,
by Miguel de Cervantes,
Translated by John Ormsby

XML is a general format for exchange of information between computer systems. It was originally devised as a ‘light’ version of SGML intended to present complex structured data containing the meaning or other information suggesting possible rendition of each individual part. The main idea behind XML (and SGML) is to present sections of text (or other data) with some particular purpose wrapped with markers typed using angle-brackets. Thus italic text may be marked-up as <italic>emphasised text</italic>. XML itself does not say which tag names (such as <italic> and its matching close-tag </italic> in this example) should be used, but there are many XML applications that do specify vocabularies of tags and mark-up.

One of these applications is the ubiquitous HTML in which almost all web pages are written. HTML is quite a rich vocabulary, not just for italic and other font choices, but also for paragraphs, sections and headings. It is possible to write short papers in HTML.

Actually, the previous paragraph contains a small lie. Historically, HTML actually precedes XML by a few years and is strictly speaking an SGML application. XML is somewhat stricter than SGML, and the XML version of HTML is called XHTML. To write XHTML you need to be more careful that you meet the additional rigours that XML imposes, but XHTML contains the same markup tags. Apart from one or two minor differences, XHTML is the same as HTML. However XHTML pages, because they conform to the XML standards, can be read and processed by a vast and increasing array of useful programs that are available for XML, such as stylesheet languages that can easily impose or change visual styles of presentation.

The down-side for using (X)HTML has been, until recently, lack of any support for writing mathematics. Things are changing here, with the adoption of XML-applications for mathematics such as MathML in some of the mainstream browsers. In particular, the Firefox browser has

built-in MathML-support. For the first time, it seems possible to write mathematical papers in XHTML and publish them directly on the web. Then XML stylesheets can convert the paper to PDF, TeX, or some other format. The mathematical content of the paper can be copied and pasted into a computer algebra system, or (for visually-impaired readers perhaps) the computer can read the paper and its mathematics aloud.

The problem is—and this will come as no surprise to readers here—that mathematics is complicated. The goal is to make the mathematics as accessible as possible to a range of present-day and future computer programs, and when one analyses the markup required for even a very simple mathematical expression, a great amount is required.

Presentation-MathML is one XML-vocabulary for the presentation of mathematics. Its main aim is to describe how mathematics should appear on the page......

In its basic form, GLOSS is a text-to-XML convertor. It enables you to write any XML (including MathML, XHTML, etc.) in plain text format, saving considerably better than 50% of the time and effort. For more specialist applications, the savings can be much more than this. GLOSS uses a syntax based on indentation, like the computer language Python, but unlike TeX and LaTeX. Plain characters and text are delimited by square brackets. (This choice was made as square brackets rarely occur in text, and are easily accessible on most keyboards.) Everything else in the input is a ‘word’ or ‘token’ or ‘command’, usually producing an XML element with the same name.......