Status

I had hoped to polish this more before publishing it, but I can't seem to get caught up... there's so much new stuff all the time!

Some Background on SGML for the World-Wide Web

In late 1992 and early 1993, I did quite a bit of work on the HTML DTD while I was working at Convex in the online documentation group.

When I began, there was the LineMode browser and the NeXT implementation, and a few nodes in The Web describing HTML with some oblique references to SGML. I was not intimately familiar with SGML, but I was quite familiar with the problems of document interchange, and I was eager to apply some of my formal systems background to the problem.

On Formally Unconvertable Document Formats

My experience with document interchange led me to classify document formats using the essential distinction that some are "programmable" and some are not. Most widely used source forms are programmable: TeX, troff, postscript, and the like. On the other hand, there are several "static" formats: plain text, Microsoft RTF, FrameMaker MIF, GNU's TeXinfo,

The reason that this distinction is essential with respect to document interchange is that extracting information from documents in "programmable" document formats is equivalent to the halting problem. That is, it is arbitrarily difficult and cannot be automated in a general fashion.

For example, I conjecture that it is impossible to write a program that will extract the third word from a TeX document. It would be an easy task for 80% of the TeX documents out there -- just skip over some formatting stuff and grab the third bunch of characters surrounded by whitespace. But that "formatting stuff" might be a program that generates 100 words from the hypenation dictionary. So the simple lexical scan of the TeX source would find a word that is *not* third word of the document when printed.

This may seem like an obscure and unimportant problem, but I assure you that the problem of converting TeX tables to FrameMaker MIF is just as unsolvable.

So while "programmable" document formats have the advantage that features can be added on a per-document basis, they suffer the disadvantage that these features cannot be recovered by the machine and translated in an automated fashion.

TEST	TEST	TEST	TEST	zz
a	a	a	a	aa
a	a	a	aa	ze
a	a	a	a	zae