chem-bla-ics

No, PDFs really do suck!

Egon Willighagen

Published June 17, 2009

Citation
Willighagen, E. (2009). No, PDFs really do suck!. In *chem-bla-ics*. chem-bla-ics. https://doi.org/10.59350/dv8xh-5dk63

Keywords
Publishing

Abstract
A typical blog by Peter MR made (again), The ICE-man: Scholary HTML not PDF, the point of why PDF is to data what a hamburger is to a cow, in reply to a blog by Peter SF, Scholarly HTML.

A typical blog by Peter MR made (again), The ICE-man: Scholary HTML not PDF, the point of why PDF is to data what a hamburger is to a cow, in reply to a blog by Peter SF, Scholarly HTML.

This lead to a discussion on FriendFeed. A couple of misconceptions:

"But how are we going to cite without paaaaaaaaaaage nuuuuuuuuuumbers?"
We don't. Many online-only journals can do without; there is DOI. And if that is not enough, the legal business has means of identifying paragraphs, etc, which should provide us with all the methods we could possibly need in science.

Typesetting of PDFs, in most journals, is superior than HTML, which is why I prefer to read a PDF version if it is available. It is nicer to the eyes.
Ummm... this is supposed to be Science, not a California Glossy. It seems that pretty looks is causing major body count in the States. Otherwise, HTML+CSS can likely beat any pretty looks of PDF, or at least match it.

As I seem to be the only physicist/mathematician who comments on these sort of things, I feel like a broken record, but math support in browsers currently sucks extremely badly and this is a primary reason why we will continue to use PDF for quite some time.
HTML+MathML is well established, and default FireFox browsers have no problem showing mathematical equations. For years, the Blue Obelisk QSAR descriptor ontology has been using such a set up for years. If you use TeX to author your equations, you can convert it to HTML too.

We can mine the data from the PDF text. Theoretically, yes. Practically, it is money down the drain. PDF is particularly nasty here, as it breaks words at the end of a line, and even can make words consist of unlinked series of characters positioned at (x,y). PDF, however, can contains a lot of metadata, but that is merely a hack, and unneeded workaround. Worse, hardly used regarding chemistry. PDF can contain PNG images which can contain CML; the tools are there, but not used, and there are more efficient technologies anyway.

I, for one, agree with Peter on PDF: it really suck as scientific communication medium.