Mel Chua » Blog Archive » A cool idea that failed: you can't reverse-engineer a paper for open access

http://blog.melchua.com/2012/05/03/a-cool-idea-that-failed-you-cant-reverse-engineer-a-paper-for-open-access/ May 3, 2012
May 3, 2012 – 8:42 pm

One of the things I tried out as part of my independent study on open access this semester was the idea of reverse-engineering a publication. This isn't about hacking code; it's about hacking copyright. And as it turns out, it doesn't work.

Here's the setup: imagine you're a researcher and you've written a great paper that's published in a prestigious journal. You beam with pride! Life is fantastic. And then you find out about the open access citation advantage, realize your publisher allows archiving of preprints, and think that life is about to get even better.

There's just one problem. You can't find your preprint version (the final edited version you send to the publisher, usually a plain Word or LaTeX document). You only have the final copy PDF with all the branding and pretty-print formatting on it – the version that got published in the journal. Somehow, in the frenzy of hard drive clean-up that accompanied your "I am done with this paper forever!" project completion celebration, you… you lost the file.

But wait… the final print version is identical to the text you sent in, right? All the publisher did was add formatting. So if you could just grab the text from the final print version and throw it *back* into a Word document, that would be identical to the preprint, and you could post that. A preprint is just the end publisher content there without the end publisher formatting. Right?

Wrong. The problem here isn't technical, it's legal. I actually took a print pdf and "reverse engineered" it into a LibreOffice document, and it looked *fantastic* — I did the process by hand, but it would be easily automatable, so the software portion of the problem is trivial. I talked with Donna Ferullo, Purdue's copyright librarian, and the copyright portion of the problem is, unfortunately, a blocker bug. The crux of it the matter is that we don't know what value the publisher added before printing. Okay, this probably is "not much other than formatting," but still… it's legal grey. So we hit a hard wall on that, but at least we learned something.

I promised to write something up about this since I don't think the reverse-engineering idea has been broached before, and it's at least good for others to know that it's a dead-end — so here it is.