A

Why scholars should write in Markdown

Stuart M. Shieber

With few exceptions, scholars would be better off writing their papers in a lightweight markup format called Markdown, rather than using a word-processing program like Microsoft Word. This post explains why, and reveals a hidden agenda as well.[1]

MICROSOFT WORD IS NOT APPROPRIATE FOR SCHOLARLY ARTICLE PRODUCTION

Before turning to lightweight markup, I review the problems with Microsoft Word as the lingua franca for producing scholarly articles. This ground has been heavily covered. (Here's a recent example.) The problems include:

Substantial learning curve. Microsoft Word is a complicated program that is difficult to use well.

Appearance versus structure. Word-processing programs like Word conflate composition with typesetting. They work by having you specify how a document should look, not how it is structured. A classic example is section headings. In a typical markup language, you specify that something is a heading by marking it as a heading. In a word-processing program you might specify that something is a heading by increasing the font size and making it bold. Yes, Word has "paragraph styles", and some people sometimes use them more or less properly, if you can figure out how. But most people don't, or don't do so consistently, and the resultant chaos has been well documented. It has led to a whole industry of people who specialize in massaging Word files into some semblance of consistency.

Backwards compatibility. Word-processing program file formats have a tendency to change. Word itself has gone through multiple incompatible file formats in the last decades, one every couple of years. Over time, you have to keep up with the latest version of the software to do anything at all with a new document, but updating your software may well mean that old documents are no longer identically rendered. With Markdown, no software is necessary to read documents. They are just plain text files with relatively intuitive markings, and the underlying file format (UTF-8 née ASCII) is backward compatible to 1963. Further, typesetting documents in Markdown to get the "nice" version is based on free and open-source software (markdown, pandoc) and built on other longstanding open source standards (LaTeX, BibTeX).

Poor typesetting. Microsoft Word does a generally poor job of typesetting, as exemplified by hyphenation, kerning, mathematical typesetting. This shouldn't be surprising, since the whole premise of a word-processing program means that the same interface must handle both the specification and typesetting in real-time, a recipe for having to make compromises.

[1] Many of the ideas in this post are not new. Complaints about WYSIWYG word-processing programs have a long history. Here's a particularly trenchant diatribe pointing out the superiority of disentangling composition from typesetting. The idea of "scholarly Markdown" as the solution is also not new. See this post or this one for similar proposals. I go further in viewing certain current versions of Markdown (as implemented in Pandoc) as practical already for scholarly article production purposes, though I support coordinated efforts that could lead to improved lightweight markup formats for scholarly applications.

Lock-in. Because Microsoft Word's file format is effectively proprietary, users are locked in to a single software provider for any and all functionality. The file formats are so complicated that alternative implementations are effectively impossible.

LIGHTWEIGHT MARKUP IS THE SOLUTION

The solution is to use a markup format that allows *specification* of the document (providing its logical structure) separate from the *typesetting* of that document. Your document is specified – that is, generated and stored – as straight text. Any formatting issues are handled not by changing the formatting directly via a graphical user interface but by specifying the formatting textually using a specific textual notation. For instance, in the HTML markup language, a word or phrase that should be *emphasized* is textually indicated by surrounding it with .... HTML and other powerful markup formats like LaTeX and various XML formats carry relatively large overheads. They are complex to learn and difficult to read. (Typing raw XML is nobody's idea of fun.) Ideally, we would want a markup format to be *lightweight*, that is, simple, portable, and human-readable even in its raw state.

Markdown is just such a lightweight markup language. In Markdown, emphasis is textually indicated by surrounding the phrase with asterisks, as is familiar from email conventions, for example, *lightweight*. See, that wasn't so hard. Here's another example: A bulleted list is indicated by prepending each item on a separate line with an asterisk, like this:

```
* First item
* Second item
```

which specifies the list

— First item
— Second item

Because specification and typesetting are separated, software is needed to convert from one to the other, to typeset the specified document. For reasons that will become clear later, I recommend the open-source software pandoc. Generally, scholars will want to convert their documents to PDF (though pandoc can convert to a huge variety of other formats). To convert file.md (the Markdown-format specification file) to PDF, the command

```
pandoc file.md -o file.pdf
```

suffices. Alternatively, there are many editing programs that allow entering, editing, and typesetting Markdown. I sometimes use Byword. In fact, I'm using it now.

Markup languages range from the simple to the complex. I argue for Markdown for four reasons:

1. Basic Markdown, sufficient for the vast majority of non-mathematical scholarly writing, is dead simple to learn and remember, because the markup notations were designed to mimic the kinds of textual conventions that people are used to – asterisks for emphasis and for indicating bulleted items, for instance. The coverage of this basic part of Markdown includes: emphasis, section structure, block quotes, bulleted and numbered lists, simple tables, and footnotes.
2. Markdown is designed to be readable and the specified format understandable even in its plain text form, unlike heavier weight markup languages such as HTML.

3. Markdown is well supported by a large ecology of software systems for entering, previewing, converting, typesetting, and collaboratively editing documents.
4. Simple things are simple. More complicated things are more complicated, but not impossible. The extensions to Markdown provided by pandoc cover more or less the rest of what anyone might need for scholarly documents, including links, cross-references, figures, citations and bibliographies (via BibTeX), mathematical typesetting (via LaTeX), and much more.
   For instance, this equation (the Cauchy-Schwarz inequality) will typeset well in generated PDF files, and even in HTML pages using the wonderful MathJax library.

$$\left(\sum_{k=1}^{n} a_k b_k\right)^2 \leq \left(\sum_{k=1}^{n} a_k^2\right)\left(\sum_{k=1}^{n} b_k^2\right)$$

   (Pandoc also provides some extensions that simplify and extend the basic Markdown in quite nice ways, for instance, definition lists, strikeout text, a simpler notation for tables.)

Above, I claimed that scholars should use Markdown "with few exceptions". The exceptions are:

1. The document requires nontrivial mathematical typesetting. In that case, you're probably better off using LaTeX. Anyone writing a lot of mathematics has given up word processors long ago and ought to know LaTeX anyway. Still, I'll often do a first draft in Markdown with LaTeX for the math-y bits. Pandoc allows LaTeX to be included within a Markdown file (as I've done above), and preserves the LaTeX markup when converting the Markdown to LaTeX. From there, it can be typeset with LaTeX. Microsoft Word would certainly not be appropriate for this case.
2. The document requires typesetting with highly refined or specialized aspects. I'd probably go with LaTeX here too, though desktop publishing software (InDesign) is also appropriate if there's little or no mathematical typesetting required. Microsoft Word would not be appropriate for this case either.

Some have proposed that we need a special lightweight markup language for scholars. But Markdown is sufficiently close, and has such a strong community of support and software infrastructure, that it is more than sufficient for the time being. Further development would of course be helpful, so long as the urge to add "features" doesn't overwhelm its core simplicity.

THE HIDDEN AGENDA

I have a hidden agenda. Markdown is sufficient for the bulk of cases of composing scholarly articles, and simple enough to learn that academics might actually use it. Markdown documents are also typesettable according to a separate specification of document style, and retargetable to multiple output formats (PDF, HTML, etc.).[2] Thus, Markdown could be used as the production file format for scholarly journals, which would eliminate the need for

[2] As an example, I've used this very blog post. Starting with the Markdown source file (which I've attached to this post), I first generated HTML output for copying into the blog using the command

```
pandoc -S --mathjax --base-header-level=3 markdownpost.md -o markdownpost.html
```

A nicely typeset version using the American Mathematical Society's journal article document style can be generated with

converting between the authors' manuscript version and the publishers internal format, with all the concomitant errors that process is prone to produce.

In computer science, we have by now moved almost completely to a system in which authors provide articles in LaTeX so that no retyping or recomposition of the articles needs to be done for the publisher's typesetting system. Publishers just apply their LaTeX style files to our articles. The result has been a dramatic improvement in correctness and efficiency. (It is in part due to such an efficient production process that the cost of running a high-end computer science journal can be so astoundingly low.)

Even better, there is a new breed of collaborative web-based document editing tools being developed that use Markdown as their core file format, tools like Draft and Authorea. They provide multi-author editing, versioning, version comparison, and merging. These tools could constitute the system by which scholarly articles are written, collaborated on, revised, copyedited, and moved to the journal production process, generating efficiencies for a huge range of journals, efficiencies that we've enjoyed in computer science and mathematics for years.

As Rob Walsh of ScholasticaHQ says, "One of the biggest bottlenecks in Open Access publishing is typesetting. It shouldn't be." A production ecology built around Markdown could be the solution.

`pandoc markdownpost.md -V documentclass:amsart -o markdownpost-amsart.pdf`

To target the style of ACM transactions instead, the following command suffices:

`pandoc markdownpost.md -V documentclass:acmsmall -o markdownpost-acmsmall.pdf`

Both PDF versions are also attached to this post.