Citations in Scholarly Markdown

Martin Fenner

June 19, 2013

In the comments on Monday's blog post about the Markdown for Science workshop, Carl Boettiger had some good arguments against the proposal for how to do citations that we came up with during the workshop. As this is a complex topic, I decided to write this blog post.

Citations of the scholarly literature are an essential part of scholarly texts and therefore have to be supported by scholarly markdown. Both the Pandoc and Multimarkdown flavors of markdown support citations, using a bibtex file that contains citations, placeholders for citekeys – [@smith04] for Pandoc and [#smith04] for Multimarkdown – and the Citation Style Language for citation formatting (Pandoc). A very reasonable approach would therefore be to use this functionality, with a preference for Pandoc because of the Citation Style Language support. All reference managers can export to the bibtex format, and some of them (e.g. Papers) make it very easy to copy and paste citekeys.

Ten days after the workshop I'm not so sure anymore this is the best approach. For four reasons:

1. **YFNS**. This approach failed the YFNS (your friendly neighborhood scientist) test. We came up with this term during the workshop and it means that our ideas about authoring should make sense to the workflow of the average scientist. I thought that using citekeys is a good idea, but my wife (my YFNS) tells me that she never uses citekeys because there are just too many [@smith04], and it is too easy get out of sync with the reference manager. She therefore prefers to put the complete reference information into the text while writing.
2. **Snippets**. As I said previously, I think that scholarly markdown has great potential not so much for writing full papers, but for all the little scientific documents we write on a daily basis. For this reason the citation information should ideally be embedded in the document if it is short, and that is difficult with bibtex (which is not human-readable).
3. **Citations as links**. Carl Boettiger reminded me that I wrote a blog post in 2010 stating that citations are nothing else than links, and that we should treat them accordingly. He has written a tool (knitcitations) for R that does just that, and Phil Lord and colleagues have written a similar

tool (kcite) for Wordpress. In 2010 I wrote a tool for Wordpress (Link to Link) that takes a different approach but also treats citations as links. All that we need is the DOI (or URL) for the article.
4. **Vendor lock-in**. Although a number of excellent reference managers are available now, users are still limited in their choices because everyone has to use the same reference manager when multiple authors work on the same document. This has always annoyed me. It would no longer be the case if we embed the citation information in the document in a standard format.

Part of the motivation for using scholarly markdown is that we can come up with best practices that make sense for digital content and don't need to support conventions from an era when articles were still printed on paper. Reference information in the form of volumes and pages, and 1000s of citation styles certainly have outlived their purpose. Citation styles are a particular pain point, as they are nothing more than a visual representation of a citation - we should care much more about the machine-readable metadata, in particular the DOI or other identifier.

The best practice for scholarly markdown could therefore be to treat citations as links, using DOIs or other standard identifiers (PMID, ArXiV, etc.) where possible. Because we typically want to list the citations as references at the end of the document, reference-style links should be preferred over inline links. From the markdown syntax documentation:

```
This is [an example][id] reference-style link.

This is [an example](http://example.com/ "Title") inline link.
[id]: http://example.com/   "Optional Title Here"
```

It might be tempting to use sequential numbers as id for the reference-style links, but the order of links can of course change during writing. It may make sense to think of the id in reference-style links as a citekey, and people should be free use that functionality of their reference manager. The citekey is used to link to the reference list at the bottom of the document, different from linking to the citekey in a separate bibtex file.

All of the above can be done in any text editor. This also includes the text editor that scholars spend most of their time with - their email program. Reference-style citations in an email are very readable, and also actionable since they are links and not text with bibliographic information.

One problem with this approach is of course that all links are inline in the resulting HTML, without a references section at the end of the document. This may be fine, as we can provide citation information in the title attribute, available upon hovering over the link (try hovering over this link, the journal eLife is doing something similar). The markdown could look like this (using the *Vancouver* citation style):

```
[@Ioannidis2005]: http://dx.doi.org/10.1371/journal.pmed.0020124 "Ioannidis JPA. Why Most P
```

The title attribute now of course uses a citation style, but this is optional information and can easily be reformatted as we have the DOI.

Or we break away from standard markdown and display reference-style links at the end of the document - similar to footnotes, which are also not part of standard markdown. But this is just a display issue that can be solved, and the solution might look different depending on whether the output is HTML, PDF or XML. This document for example contains 14 reference-style citations.

There is obviously a need for tools that make adding citations to scholarly markdown easier. This could be accomplished by relatively small changes to existing reference managers (enabling copy/paste of citations in reference-style markdown format), or by tools similar to the knitcitations and kcite mentioned above.