Metadata in Scholarly Markdown

Martin Fenner, Gobbledygook

June 29, 2013

Scholarly documents often need metadata that describe them: typically author(s), title and location (DOI or URL), but possibly many other things. For some metadata it makes sense to store them in the document text, e.g. as is typically done for citations. The problem is that this can make it hard to make the metadata machine-readable. The worst place for metadata is of course outside of the document, and unfortunately that it is the most common way of doing this. Two examples:

- Manuscript submission. Papers submitted to scholarly journals contain the metadata in the text, but authors are required to enter the information again into a webform. You can add metadata (property information) to Microsoft Word documents, but it seems that nobody is doing it.
- PDFs and image files. Even though we have at least one good standard with XMP to store metadata in these documents, it is not a common practice. Information about these documents is therefore stored somewhere else and doesn't automatically travel with them.

The best place for metadata is the document itself, and the metadata should be stored in machine-readable format. Another requirement is flexibility in what we can store, and we shouldn't limit ourselves to a predefined list. Pandoc for example allows only three attributes in the title block:

```
% title
% author(s) (separated by semicolons)
% date
```

For Scholarly Markdown we have another requirement: the metadata should be writeable and readable by humans. YAML is the perfect format for this. JSON is closely related to YAML (and is in fact a subset of YAML 1.2), but YAML can also be written with whitespace instead of curly braces. The static website generator Jekyll - which I use to parse the markdown for this blog into HTML - uses YAML at the beginning of markdown documents to store metadata, and we can easily extend this functionality. Carl Boettinger posted a comment yesterday saying that YAML support is on the Pandoc development roadmap.

Below is the YAML for (Ethan P. White, 2013), where I reposted a paper written

in markdown:

```
---
layout: post
title: "Nine simple ways to make it easier to (re)use your data"
tags: [example, citation]
authors:
 - name: Ethan P. White
   orcid: 0000-0001-6728-7745
   affiliation: Dept. of Biology and the Ecology Center, Utah State University, Logan, UT, U
 - name: Elita Baldrige
   orcid: 0000-0003-1639-5951
   affiliation: Dept. of Biology and the Ecology Center, Utah State University, Logan, UT, U
 - name: Zachary T. Brym
   affiliation: Dept. of Biology and the Ecology Center, Utah State University, Logan, UT, U
 - name: Kenneth J. Locey
   affiliation: Dept. of Biology, Utah State University, Logan, UT, USA, 84341
 - name: Daniel J. McGlinn
   affiliation: Dept. of Biology and the Ecology Center, Utah State University, Logan, UT, U
 - name: Sarah R. Supp
   affiliation: Dept. of Biology and the Ecology Center, Utah State University, Logan, UT, U
---
```

In JSON the same information would look like this (and Jekyll is able to parse it, since JSON is a subset of YAML 1.2):

```
---
{
  "layout": "post",
  "title": "Nine simple ways to make it easier to (re)use your data",
  "tags": [
    "example",
    "citation"
  ],
  "authors": [
    {
      "name": "Ethan P. White",
      "orcid": "0000-0001-6728-7745",
      "affiliation": "Dept. of Biology and the Ecology Center, Utah State University, Logan,
    },
    {
      "name": "Elita Baldrige",
      "orcid": "0000-0003-1639-5951",
      "affiliation": "Dept. of Biology and the Ecology Center, Utah State University, Logan,
    },
    {
      "name": "Zachary T. Brym",
```

```
      "affiliation": "Dept. of Biology and the Ecology Center, Utah State University, Logan,
    },
    {
      "name": "Kenneth J. Locey",
      "affiliation": "Dept. of Biology, Utah State University, Logan, UT, USA, 84341"
    },
    {
      "name": "Daniel J. McGlinn",
      "affiliation": "Dept. of Biology and the Ecology Center, Utah State University, Logan,
    },
    {
      "name": "Sarah R. Supp",
      "affiliation": "Dept. of Biology and the Ecology Center, Utah State University, Logan,
    }
  ]
}
---
```

You can see that the author information required for manuscript submission can easily be written in YAML (email addresses were removed to protect privacy). JSON is also possible for people where this is a better fit into their workflow, but it is more difficult to write for humans because of the curly braces, and because all strings need to be in double quotes.

Once the ORCID Registry adds affiliation information, we no longer need to provide email and affiliation when submitting manuscripts. I have stored my own name, orcid, email and affiliation in my site configuration file so that I don't have to provide this info for every blog post.

In this blog markdown files are currently only processed to HTML, and I store the metadata in HTML `meta` tags in a format used by many sites and services, including Google Scholar - look at the source code of Ethan P. White et al. (2013) for an example. These metadata are also understood by the Greycite service built by Phil Lord and Lindsay Marshall (2012) that generates citation information for weblinks, adding important metadata such as title, authors and publication_date so that we can properly cite our blog post (Ethan P. White, 2013).

And I use the metadata to link the author names to their ORCID profile (if they have an ORCID) or email address, with the affiliation visible when you hover over the name. My own name is linked to the About page of this site, but with a little development effort I could automatically add all my publications (and other works) in my ORCID profile to that page.

Metadata are important, and Scholarly Markdown makes it easy to embed them.

Update 06/30/13: added JSON example to demonstrate the differences to YAML, and to show that Jekyll also works with JSON (used in this blog post, and tested with the examples above which produce identical HTML output). Also added

two references, using the embedded HTML metadata and the Greycite service to generate citations in bibtex.

References

Ethan P. White, Z. T. B., Elita Baldrige. (2013). Nine simple ways to make it easier to (re)use your data. *Gobbledygook*. Retrieved from http://blog.martinfenner.org/2013/06/25/nine-simple-ways-to-make-it-easier-to-reuse-your-data

Lord, P., & Marshall, L. (2012). Greycite: Citing the web. *An Exercise in Irrelevance*. Retrieved from http://www.russet.org.uk/blog/2071