Citeproc YAML for bibliographies

Martin Fenner, Gobbledygook

July 30, 2013

The standard local file formats for bibliographic data are probably bibtex and RIS. They have been around for a long time, and are supported by all reference managers and many other tools and services. Unfortunately these formats are far from perfect:

- neither bibtex nor RIS use a web-friendly data interchange format such as XML or JSON, which makes it harder to work with these formats
- bibtex - and to a lesser extend RIS - don't support all entry types that we need, e.g. datasets, or new standards such as ORCID author identifiers
- bibtex stores all authors in a single field, which makes author names hard to parse

bibtex

```
@article{fenner2012a,
  title = {One-click science marketing},
  volume = {11},
  url = {http://dx.doi.org/10.1038/nmat3283},
  doi = {10.1038/nmat3283},
  number = {4},
  journal = {Nature Materials},
  publisher = {Nature Publishing Group},
  author = {Fenner, Martin},
  year = {2012},
  month = {mar},
  pages = {261-263}
}
```

One obvious solution would be to store bibliographic data in XML or JSON. These formats have very good support in all programming languages, and they are the formats used by APIs on the web. There have been some efforts to standardize these formats for bibliographic data, e.g. BibJSON, MODS, BibTeX XML or Endnote XML.

BibTeX XML

```
<bibtex:entry id='fenner2012a'>
  <bibtex:article>
    <bibtex:title>One-click science marketing</bibtex:title>
    <bibtex:volume>11</bibtex:volume>
    <bibtex:url>http://dx.doi.org/10.1038/nmat3283</bibtex:url>
    <bibtex:doi>10.1038/nmat3283</bibtex:doi>
    <bibtex:number>4</bibtex:number>
    <bibtex:journal>Nature Materials</bibtex:journal>
    <bibtex:publisher>Nature Publishing Group</bibtex:publisher>
    <bibtex:person>
      <bibtex:first>Martin</bibtex:first>
      <bibtex:last>Fenner</bibtex:last>
    <bibtex:person><bibtex:author/>
    <bibtex:year>2012</bibtex:year>
    <bibtex:month>mar</bibtex:month>
    <bibtex:pages>261-263</bibtex:pages>
  </bibtex:article>
</bibtex:entry>
```

My problem with these formats is that they are made for computers talking to each other and not humans. I personally think that a file with bibliographic data should be human-readable, similar to why I like markdown for writing scientific documents.

When you have too many standards and are not happy with any of them, you of course create a new standard.

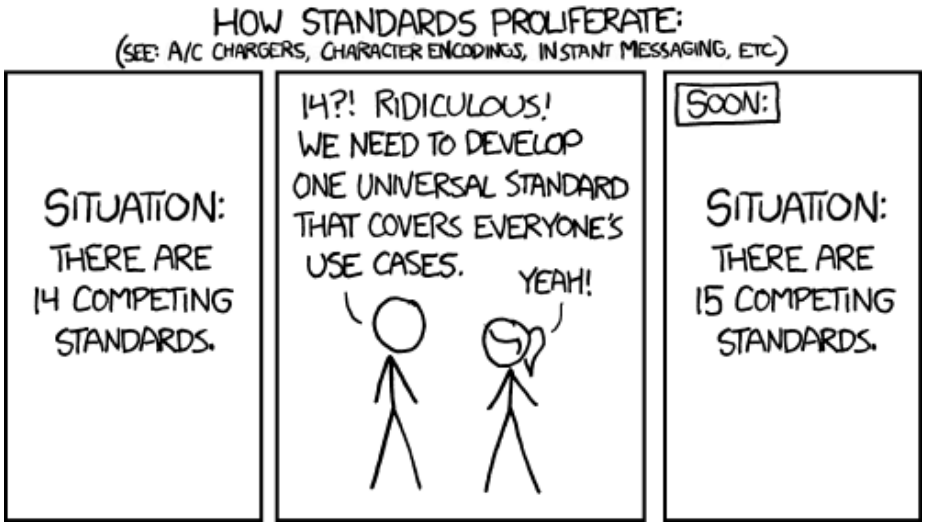

Figure 1: **How Standards Proliferate**. Taken from http://xkcd.com/927/

My suggestion for a new bibliographic file format is twofold: a) use YAML for data serialization and b) use CSL as data format. YAML is a data format popular with Ruby Developers and is described on the YAML website as

> YAML is a human friendly data serialization standard for all programming languages.

Something that not may people seem to know is that YAML is a superset of JSON and that every JSON file is also a valid YAML file. The main difference is the better human readability of YAML.

Citation Style Language is described on the CSL website as

> CSL is an open XML-based language to describe the formatting of citations and bibliographies.

Although some commercial applications still use proprietary citation styles, CSL has become the de facto standard, and is used by the reference managers **Zotero**, **Mendeley**, **Papers**, and others. This blog uses CSL via Pandoc and the citeproc-hs library. CSL processors need bibliographic data in a standard format. The popular Citeproc-js Javascript CSL processor by Frank Bennett for example uses JSON, but we might as well use YAML:

Citeproc YAML

```
- title: One-click science marketing
  volume: '11'
  URL: http://dx.doi.org/10.1038/nmat3283
  DOI: 10.1038/nmat3283
  issue: '4'
  container-title: Nature Materials
  publisher: Nature Publishing Group
  author:
  - family: Fenner
    given: Martin
    orcid: 0000-0003-1419-2405
  page: 261-263
  id: fenner2012a
  type: article-journal
  issued:
    date-parts:
      - 2012
      - 3
```

I hope you agree that this format is not only structured and can be understood by computers, but is also very readable by humans. You may have noticed that I have inserted my ORCID, something that is very difficult to do with bibtex where all authors are stored in one text string (see above).

Careful readers of this blog will of course remember that I have written about using YAML to store metadata about a blog post. We could now add bibliographic information to these metadata, either in the YAML frontmatter (if it is a Jekyll blog), or in a separate file. It should be straightforward to adapt the existing CSL processors to understand YAML since YAML and JSON are so similar. To get started with some Citeproc YAML, use the new (and still experimental) **ORCID Feed** Webservice with your ORCID and specify the `yml` format, e.g. http://feed.labs.orcid-eu.org/0000-0003-1419-2405.yml for my publications.