CSL is more than citation styles

Martin Fenner, Gobbledygook

August 8, 2013

According to the description on the Citation Style Language (CSL) website, CSL *is an open XML-based language to describe the formatting of citations and bibliographies*. We use reference managers such as **Zotero**, **Mendeley**, or **Papers** to format our references in manuscripts we submit for publication, and underneath a CSL processor such as Citeproc-js - together with a CSL file for a particular citation style - is doing the work:

When the journal article is accepted the publisher takes the text with the formatted text citation and turns it into XML, a process that is error-prone and takes time:

It is not hard to see that something is very wrong here:

- Authors are required to use a specific citation style (there are probably about 1,000 different citation styles and many more dependent styles) even though the publisher doesn't directly use the formatted text. The publisher eLife accepts references in any format.
- Turning structured information into plain text and back into structured XML is always a bad idea. Kaveh Bazargan is a typesetter who has gone on record for saying that we should stop this nonsense and put him out of business.

It is also obvious how the ideal workflow should look like:

We go from structured content to structured content, and never use citations formatted as text as intermediary steps in the workflow.

What is surprising is that this is an ideal workflow and not something that publishers actually do. Most journal author instructions don't even mention CSL styles (I work for PLOS and they are no exception). There are some issues to be solved, but they are all minor:

- The Citeproc JSON citation format isn't really an official standard, but rather something invented for the most popular CSL processor, Citeproc-js.
- People like to fight over standards, and there are always people you prefer bibtex, RIS, MODS or BibJSON over Citeproc JSON, or want authors to to use JATS XML.

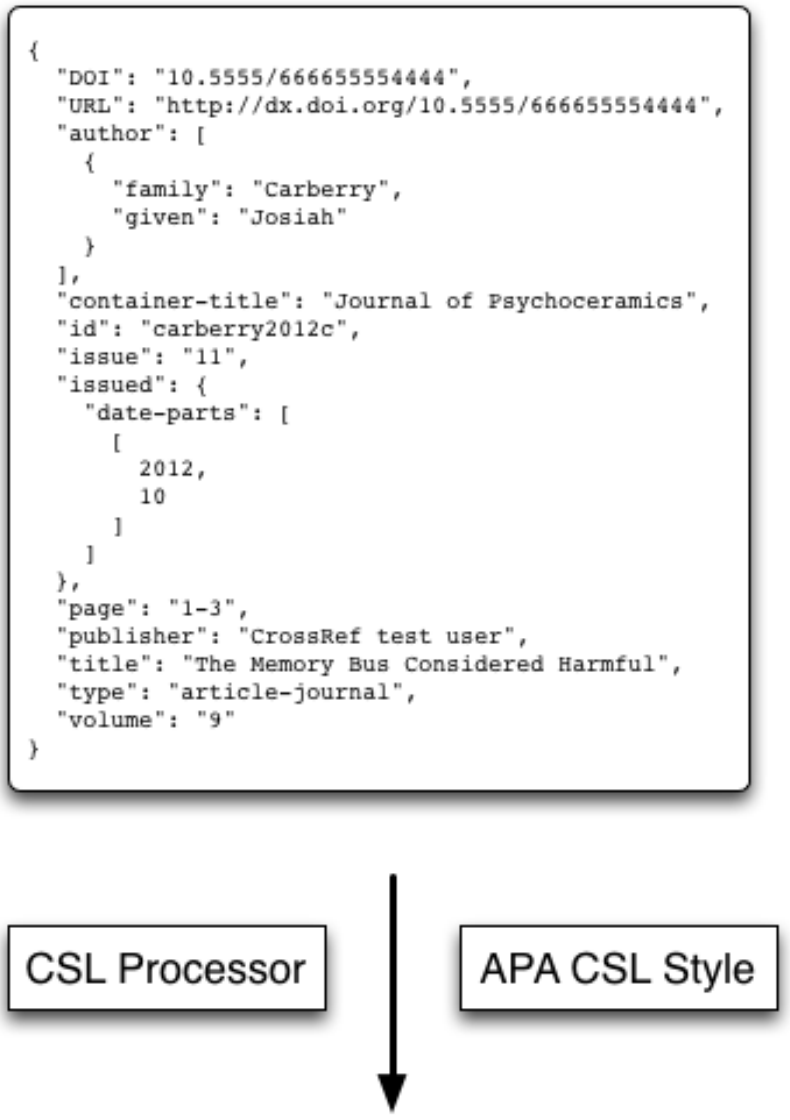
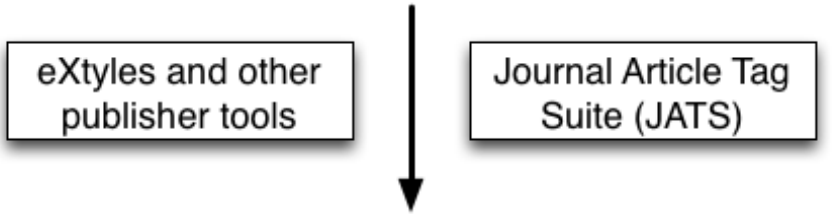


Figure 1: Citation processing during manuscript writing

Carberry, J. (2012). The Memory Bus Considered Harmful. *Journal of Psychoceramics*, 9(11), 1-3. doi:10.5555/666655554444




```
<ref id="random_id">
  <label>69</label>
  <mixed-citation publication-type="journal"
xlink:type="simple"><name name-style="western"><surname>Carberry
</surname><given-names>Josiah</given-names></name> (<year>2012</
year>) <article-title>The Memory Bus Considered Harmful</article-
title>. <source>Journal of Psychoceramics</source> <volume>9</
volume>: <fpage>1</fpage>–<lpage>3</lpage>.
  <pub-id pub-id-type="doi">10.5555/666655554444</pub-id></mixed-
citation>
</ref>
```

Figure 2: Citation processing by the publisher

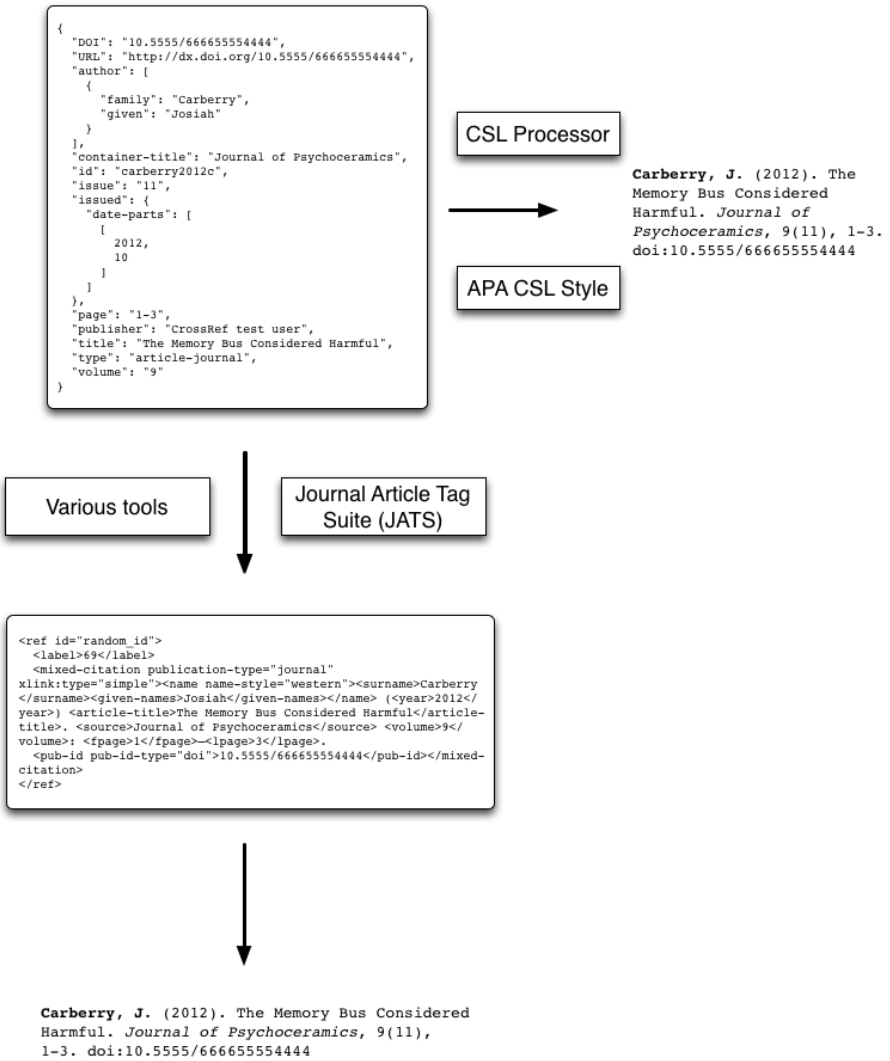


Figure 3: Ideal workflow of citation processing

I would really like to push Citeproc JSON as a standard bibliographic exchange format for authors. There are several things I like about Citeproc JSON:

- It is the native format to format citations, so it is used internally by many reference managers anyway.
- Citeproc JSON is really good in handling all the possible variations of author names. Putting all authors into a single text field as in bibtex requires a lot of trickery to get it right.
- JSON is a standard serialization format and there are a kinds of libraries in different programming languages to do things like searching, sorting or finding of duplicates. And JSON is easily extensible, e.g. if we would want to add ORCID identifiers for authors.

I have five suggestions to move forward:

- Make a specification for Citeproc JSON that is as clear as the CSL specification.
- Consider extending the specification to include content other than citations. Ideally we should be able to add arbitrary metadata about a manuscript.
- Consider other serialization formats besides JSON. I particularly like YAML as it is very similar to JSON, but human-readable, but other people might prefer XML. It is relatively easy to transform data between these serialization formats, in particular between JSON and YAML. In my About page I only need the js-yaml library and one extra line of code to use Citeproc YAML instead of Citeproc JSON (in the d3.js visualization).
- Add Citeproc JSON (and YAML) support to reference managers. Zotero is already doing this, but it should be an easy to add feature if the reference manager is already using CSL internally (Mendeley and Papers).
- Push publishers to accept Citeproc JSON with manuscript submissions.