From Markdown to JATS XML in one Step

Martin Fenner, Gobbledygook

December 12, 2013

The Journal Article Tag Suite (JATS) is a NISO standard that defines a set of XML elements and attributes for tagging journal articles. JATS is not only used for fulltext content at PubMed Central (and JATS has evolved from the NLM Archiving and Interchange Tag Suite originally developed for PubMed Central), but is also increasinly used by publishers.

For many publishers the *version of record* of an article is stored in XML, and other formats (currently HTML, PDF and increasingly ePub) are generated from this XML. Unfortunately the process of converting author-submitted manuscripts into JATS-compliant XML is time-consuming and costly, and this is a problem in particular for small publishers.

In a recent blog post (The Grammar of Scholarly Communication) I argued that publishers should accept manuscripts in any reasonable file format, including Microsoft Word, Open Office, LaTeX, Markdown, HTML and PDF. Readers of this blog know that I am a big fan of markdown for scholarly documents, but I am of course well aware that at the end of the day these documents have to be converted into JATS.

As a small step towards that goal I have today released the first public version of pandoc-jats, a custom writer for Pandoc that converts markdown documents into JATS XML with a single command, e.g.

```
pandoc -f example.md --filter pandoc-citeproc --bibliography=example.bib --csl=apa.csl -t JA
```

Please see the pandoc-jats Github repository for more detailed information, but using this custom writer is as simple as downloading a single `JATS.lua`file. The big challenge is of course to make this custom writer work with as many documents as possible, and that will be my job the next few weeks. Two example JATS documents are below (both markdown versions of scholarly articles and posted on this blog as HTML):

- Nine simple ways to make it easier to (re)use your data (HTML, JATS)
- What Can Article Level Metrics Do for You? (HTML, JATS)

Both JATS files were validated against the JATS DTD and XSD and showed no errors with the NLM XML StyleChecker - using the excellent jats-conversion

conversion and validation tools written by Alf Eaton. Markdown is actually a nice file format to convert to XML - in contrast to HTML authors can't for example put closing tags at the wrong places. And a Pandoc custom writer written in the Lua scripting language is an interesting alternative to XSLT transformations, the more common way to create JATS XML. The custom writer has not been tested with other Pandoc input formats besides markdown, of particular interest are of course HTML and LaTeX - Microsoft Word .docx is unfortunately only a Pandoc output format.

This is the first public release and there is of course a lot of room for improvement. Many elements and attributes are not yet supported - although ORCID author identifiers are of course included. Please help me improve this tool using the Github Issue Tracker.