Manifests and Reference Lists

Martin Fenner, Gobbledygook

February 5, 2015

Last month at the Force15 conference in Oxford Ian Mulvany and I ran a workshop on data citation support in reference managers. The report of that workshop isn't done yet, but I can say that it was a success - we now have a pretty good idea what the problems are and what needs to be done to fix them. The short summary of the workshop is in this slidedeck of the presentation that summarized the workshop for the other Force15 attendees.

The whole idea of the workshop was to treat data citation as similar as possible to the citation of journal articles, i.e. to allow authors to use the same tools (reference managers) and conventions (citation styles). Putting a data citation into a reference list makes it easier to find that data citation because reference lists contain more metadata, are more structured, and more accessible than data citations in the form of identifiers or links within the body text of the article.

But I have to admit that there is one problem with reference lists: although there is always some self-citation, reference lists usually contain references to articles (and other resources) created by other people and before the article was published. It feels a little bit odd to put a dataset created by the same group of people and published at the same time into the reference list. And although we could use a separate reference list or highlight the data associated with the article in some other way, what we really want is something slightly different, a manifest file.

The journal article has been a (mainly) textual document for many centuries not because this is the essence of science communication, but rather because there was no practical way to include all the other information (raw data, tools used for experiments, etc.). Very few of these limitations remain with the digital journal article that we have since the 1990s, but we have for the most part failed to change the format other than going from paper to PDF. One of many examples: figures in publications typically still are has limited as they were decades ago with no way to see the data underlying the figure, options for selecting what data points are shown, or animation for time-based information.

So what we really care about is the sum of artifacts and resources that together make what Carol Goble and others call research object (Bechhofer et al., 2010), the journal article is an important part, but clearly doesn't include everthing

that is needed to understand and reproduce the work. Reference lists can help with linking to some of the resources not included in the article text, but they typically don't link to supplementary information or other places where the underlying data are made available, or to the figures of the article. Although some publishers provide navigation tools for readers to get to this information, what we really need is a machine-readable list of all the resources used in an article.

As it happens, this is exactly what the ePub format for electronic books is doing, as every ePub must include a manifest file that lists all the files that are part of the publication, defined in the Open Packaging Format (OPF). I need to do more research to figure out how to do this with JATS, the standard for scholarly articles, and how to generate something similar to the manifest file when using different formats, e.g. html or markdown. This has to be linked to some of the information we are collecting already, e.g. described in JATS (Beck, 2011), or the `relatedIdentifier` in the DataCite metadata (Starr, 2014).

References

Bechhofer, S., Bechhofer, S., De Roure, D., Gamble, M., Goble, C., & Buchan, I. (2010). Research Objects: Towards Exchange and Reuse of Digital Knowledge. *Nature Precedings*, (713). https://doi.org/10.1038/npre.2010.4626.1

Beck, J. (2011). NISO Z39.96 The Journal Article Tag Suite (JATS): What Happened to the NLM DTDs? *The journal of electronic publishing : JEP*, *14*(1). https://doi.org/10.3998/3336451.0014.106

Starr, J. (2014). DataCite Metadata Schema for the Publication and Citation of Research Data, 1–38. https://doi.org/10.5438/0010