Data Citation Reprised

written by Jeremy Huggett | 30/01/2018

CC0 by Tama66 via Pixabay

So here's a thing. A while ago, I asked whether there was any way to quantify the extent to which archaeologists were citing their reuse of data. I used the Thomson Reuters/Clarivate Analytics Data Citation Index (DCI) as a starting point, but it didn't go too well … Back then, the DCI indicated that 56 of the 476 data studies derived from the UK's Archaeology Data Service repository had apparently been cited elsewhere in the Web of Science databases (the figure is currently 58 out of 515). But I also found that the citations themselves were problematic: the citation of the published paper/volume was frequently incomplete or abbreviated, many appeared to be self-citations from within interim or final reports, in some cases the citations preceded the dates of the project being referenced, and in many instances it was possible to demonstrate that the data had been cited (in some form or other) but this had not been captured in the DCI. At that point I concluded that the DCI was of little value at present. So what was going on?

The devil finds work for idle hands, and so I recently decided to dig deeper. According to the Data Citation Index Handbook, the citations reported in the DCI are in fact not just derived from the range of Web of Science databases, but also from references captured and provided by the repositories themselves. This explains why most of the DCI citations found in my search didn't link through to one or other of the Web of Science databases – they were actually provided by the ADS themselves.

Sure enough, the 13 citations for Wharram Percy in the DCI are matched by the *Associated Publications* entries in the ADS metadata describing that data collection, and, unlike the DCI variants, these helpfully have full references. Things are slightly less straightforward with the next data collection in the DCI list of most-cited ADS data collections: the Lyonesse Project from the Scilly Isles. The DCI reports 4 citations and the same four appear in the ADS metadata along with a fifth – the 1756 volume on the ancient Scilly Isles by W. Borlase. Presumably this doesn't appear in the DCI

because it falls outside of some date bounds (and certainly can't be a data citation in any case)? Things go rapidly downhill with the next in the list: the Castle Mall, Norwich with 4 citations in the DCI. These are all shown as anonymous, but strangely the ADS metadata records authors for these. However, the format of the references in the ADS metadata is different – unlike the previous examples, the authors are preceded by the title of the publication, and this seems to have thrown off the DCI citation. The DCI citations also include dates of publication where none were provided in the ADS metadata – for instance the volume '*Norwich Castle: Excavations and Historical Survey 1987–98*' is shown in the DCI data as published in 1987 (picking the date up from the title?) whereas it was actually published in 2009 as East Anglian Archaeology Report 132 … I won't go on!

So other than confirming my initial conclusion that the DCI data weren't a useful means of evaluating the reuse and citation of archaeological digital data, what else can we take from this? There seems to be a combination of issues giving rise to this situation:

- The original metadata is collected at different times (and likely by different people) and is inconsistently recorded (e.g. the different formats used in references leading to 'anonymous' authorship recorded in the DCI)
- The metadata is incomplete and incorrect assumptions are being made in its parsing between systems (e.g. the invalid creation of dates of publication where none existed)

Automated ingest of metadata from the ADS into the DCI therefore appears to be problematic, and perhaps serves as a warning of the complexities of data mapping and serving up digital information into different aggregation systems: the basis of many of the online tools that we increasingly rely on.

But furthermore, from the perspective of evaluating levels of data citation, there is a fundamental error in mapping the ADS *Associated Publication* category onto DCI citations which are subsequently counted in their *Times Cited* category, since the ADS *Associated Publication* metadata as used has little or nothing to do with data citation as such. Consequently, not only are the DCI data not a useful means of evaluating the reuse and citation of archaeological digital data, worse, they are actually misleading as things currently stand.

So how can we best gauge data citation levels in archaeology? An alternative approach is to use systems which track the use of persistent identifiers such as DOIs to uniquely identify digital objects (whether publications or data collections). An example of this is the beta Events service provided through DataCite and Crossref. The flaw here is that DOIs are, as yet, inconsistently used: research data are often uncited, or if referenced at all, are often mentioned as the title of the dataset in the acknowledgements rather than using the DOI. Archaeology is not alone in this: see, for example, Mooney and Newton 2012; Park and Wolfram 2017; Peters *et al*. 2016.

Frustratingly, it remains problematic to evaluate the level of data reuse in archaeology, and such tools as we have are at best inadequate in enabling us to make any realistic assessment of this. In the absence of any clear requirement on the part of publishers and professional bodies to properly reference our use of data (whether through DOIs or some other alternative), it looks as if this will continue to be the case for some time to come. As I suggested at the end of the original piece, the most useful approach could likely be for the data repositories themselves to track and record the reuse of the datasets they provide access to. By way of compensation, in publicising this

information (not burying it in metadata!) they could give their invaluable work in data preservation and data sharing a much higher profile, and – who knows? – maybe in the process secure a more secure financial base into the future.

References

Mooney, H. and Newton, M. 2012 'The Anatomy of a Data Citation: Discovery, Reuse, and Credit', *Journal of Librarianship and Scholarly Communication* 1(1), eP1035. https://dx.doi.org/10.7710/2162-3309.1035

Park, H. and Wolfram, D. 2017 'An Examination of Research Data Sharing and Re-use: Implications for Data Citation Practice', *Scientometrics* 111, 433-461. https://dx.doi.org/10.1007/s11192-017-2240-2

Peters, I., Kraker, P., Lex, E., Gumpenberger, C. and Gorraiz, J. 2016 'Research Data Explored: An Extended Analysis of Citations and Altmetrics', *Scientometrics* 107, 723-744. https://dx.doi.org/10.1007/s11192-016-1887-4