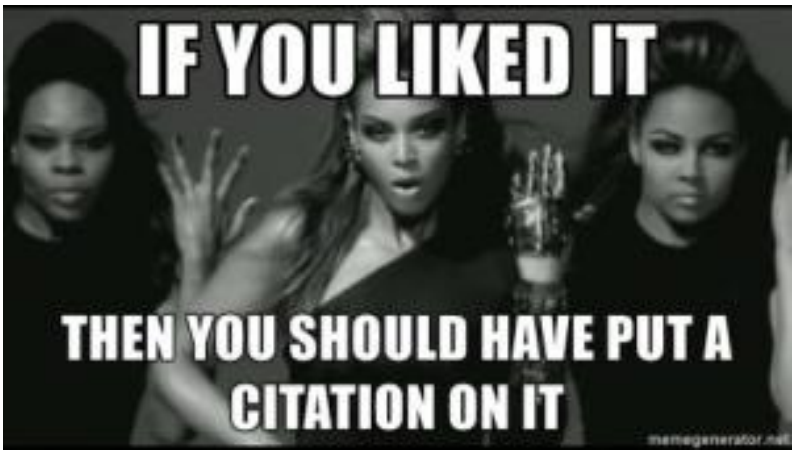
Citing Data Reuse

written by Jeremy Huggett | 23/05/2017



I've commented here and here about the question of data reuse (or more accurately, the lack of it) and the implications for archaeological digital repositories. It's frequently argued that the key incentive for making data available for reuse is providing credit through citation. So how's that going? I've not seen any attempt to actually quantify this, so out of curiosity I thought I'd have a go.

A logical starting point is Thomson Reuters Data Citation Index – according to its owners (it's a licensed rather than public resource), this indexes the contents of a large number of the world's leading data repositories, and, on checking, the UK's Archaeology Data Service (ADS) appears among them. So far so good.

Things don't start too well, however. A simple search on *TOPIC: (archaeology)* refined by *COUNTRIES/TERRITORIES: (UNITED KINGDOM)* gives 6 results. Oh. One of these is the ADS itself (as repository), plus 3 software resources and 2 data studies (both in the UK Data Archive – nothing from the ADS). Things aren't much better if the country is switched to the USA: of 223 results, 220 are mostly NAGPRA-related datasets deposited by one individual, 2 are software linked to publications, and 1 is the data repository OpenContext. Something's not right in relation to the Countries filter …

Rather more successfully, using *TOPIC: (archaeology) AND DOCUMENT TYPES: (Data study)* and refined by *SOURCE TITLES: (ARCHAEOLOGY DATA SERVICE)* (OpenContext, tDAR and other archaeological repositories don't appear as options) generates 476 results which usefully link back to the ADS archives themselves via their DOIs. The number seems rather small given that the Data Citation Index claims to be fully indexing repositories (the ADS website suggests it holds 1072 project archives, for instance), but leaving that on one side, let's look at the citation figures provided in the results.

Overall, 56 of the 476 data studies have apparently been cited across the Web of Science databases. The most cited is the Wharram Percy archive (13 times), then we have the Lyonesse Project (a study of the coastal and marine environment of the Isles of Scilly) (4 times), the CBA's Archaeological Site Index to Radiocarbon Dates from Great Britain and Ireland (4 times), Norwich Castle Mall (4 times), Newport Medieval Ship (3 times), Christchurch Spitalfields burial crypt (3 times), plus a further 6 datasets with 2 citations and 44 with 1 citation. Just to state the obvious,

that means 420 datasets received no citation according to the Data Citation Index.

Digging a little deeper into the citations throws up some problems. Ironically, the citations provided for the citations cited (!) are entirely inadequate, and leave one guessing (or using Google) to try and pin a publication down. It is also apparent that in most cases the citations are by the same authors who were responsible for the datasets themselves – indeed, in many instances they are citations to the datasets from within the original published reports. In some cases, the citations seem to possess a bizarre crystal ball-gazing capacity – for instance, the three of the four citations for the Lyonesse Project were published before the project commenced, in one case 35 years earlier … The citations are also evidently far from complete – for instance, the Newham Museum Archaeology archive is shown as having one citation but a quick Google Scholar search indicates at least four publications by members of the ADS in which the archive is used as a salutary case study of data preservation.

So, as a means of evaluating the levels of reuse and citation of archaeological digital data, it seems clear that the Digital Citation Index is, regretfully, of little value at the moment. If it serves any purpose, it does underline the importance of the context of a citation, and emphasises that the possession of a citation is no measure of the actual reuse of a dataset.

By way of comparison, we could look at DataCite, a publicly accessible resource designed to locate, identify, and cite research data though the use of persistent DOIs. Plugging 'archaeology' into their search tool returns 89,450 works: the majority (42,628) from the ADS with 29,697 from DANS-EASY and 730 from tDAR (for some reason, OpenContext does not figure large in this database). Most of the ADS records (41,460) are classified as 'texts' with 929 as 'datasets' and 236 as 'collections'. The ADS 'texts' relates to their grey literature library (currently 41,990 reports) allowing for some lag in updating, although the distinction between 'collections' and 'datasets' is rather less clear – most of the 'collections' appear to be unpublished fieldwork reports (i.e. grey literature and yet not 'texts'), although 'datasets' are still, thankfully, datasets. However, unlike the Data Citation Index, DataCite does not (yet) include citation information: their Event Data service is designed to detect when an item's DOI has been saved, liked, shared, referenced or commented upon, but is not yet available (currently in beta as of May 2017 via an API).

Still early days in terms of data citation indexes, then – and some signs that, when available, their insensitivity to context makes them of limited value as a means of assessing data reuse. Which means, for now at least, we still have to fall back to Google – and that entails a laborious one-by-one examination of each dataset.

To take one example not entirely at random: the Archaeology Data Service holds the data archive for the Anglo-Saxon cemetery at Cleatham, Lincolnshire. Back in 2009, I reviewed the published excavation report (Leahy 2007) for the *American Journal of Archaeology* and in the process, commented that the limited statistical analysis in the report was offset by the fact that the data had been archived with the ADS which made a more systematic GIS-based analysis easily within grasp. So has anyone done so?

The usage statistics page for Cleatham shows that it has currently seen 1763 unique visits since May 2011, with 1025 file downloads and 9903 page views. Focusing on the file downloads as a more likely useful statistic to indicate reuse, the archive itself consists of 12 database files, 2 shapefiles,

and 2022 images (the images are also downloadable in 10 zip files). So, a full download of the least number of files would consist of 24 files in all. Assuming a worst-case scenario, that the 1025 file downloads consisted of individual files, that could equate to just under 43 complete archive downloads. A very reasonable figure over six years, one might think. But is there any evidence for their reuse? Not according to the Data Citation Index.

Searching Google using the Cleatham archive DOI simply finds links to the ADS and the DataCite records, which might suggest that people are only gradually starting to use DOIs as a means of reference. Lots of papers and books refer to Cleatham typically as part of broader regional studies, but many do not cite the original excavation report let alone the data archive. Other work relies on the Cleatham material archive, and whilst it might seem more than likely that the digital archive would have been of use, there is no unequivocal evidence of its application. Trying to whittle the results down to distinguish a reference to the archive rather than the original published report turned up a total of three published papers and one PhD thesis which explicitly reference the Cleatham archive. The PhD and two of the papers are research by Kirsty Squires (2011, 2012, 2013) which included a substantial analysis of the Cleatham material and there is direct evidence of archive use in, for instance, the published GIS plots illustrating aspects of her analyses. The remaining paper by Ruth Nugent and Howard Williams (2012) cites the archive but from the in-text references uses it primarily as the basis for the reuse (redrawing) of a number of illustrations. So, since its release in 2007, the Cleatham archive has been reused in two distinct studies, and was fundamental to one of them. Is this good? We don't know, because we've little to compare it with, but it perhaps doesn't seem unreasonable for a ten-year period. Is this an accurate picture? Who knows, but almost certainly not. And we certainly can't do this kind of search for every individual dataset …

So where does all this leave us? Presumably a measure of successful data reuse is the number of other people picking up the data and doing something with it. On that basis, the Cleatham example is a success of sorts given it's only been available for ten years with a PhD taking three years plus while publications can easily take a couple of years to appear in press. However, Cleatham wasn't picked up by the Data Citation Index, and the insensitivity of citation indexes to context appears to make them of limited value as a means of directly assessing data reuse. And, of course, none of this allows for the likelihood that all such calculations are likely to substantially underestimate the actual level of reuse, not just in terms of missing citations. For instance, reuse as a teaching dataset or in student projects would not likely be revealed through a citation search and yet are perfectly valid examples of reuse (for example, 5 of 22 independent student projects on my GIS course this year used ADS datasets – none of them Cleatham!).

It seems unlikely that general citation tools are going to be capable of doing much more than flagging citations that may or may not represent reuse, given the need for a more nuanced contextual assessment illustrated by the Cleatham example above. Equally, citation counts are a poor measure for actual reuse – we need some other, more subtle, measure since reuse itself can be defined according to various criteria. It may be that this is something we have to look to our repositories to provide along with the data, both for the benefit of researchers – depositors and users alike – and also out of their own self-interest. For example, the ADS already captures some relevant information in its record metadata – the metadata for Squires' thesis (2011) is linked (via 'Associated Publication') to relevant publications and to the original excavation report as well as to the original archive (as 'Related Collection'). Likewise, the Cleatham archive metadata is cross-

linked to Squires' thesis (via 'Related Collection'), although not to the other relevant publications. Clearly such metadata needs to be maintained and updated as new items become available and, at the same time, these metadata need to be exposed more clearly to users in order to highlight the reuse of the datasets. Both requirements place greater burdens on repositories, but the combination of their access to automated citation systems and their need to quantify reuse in order to demonstrate the value of their resources might enable the means to be found to make this both technically achievable as well as more generally useful.

In the end, though, whether or not citations are valuable in gauging reuse, the evidence suggests we're still inconsistent in referencing our use of datasets, which ultimately makes it an ethical matter beyond any practical need to calculate reuse.

References

Leahy, K. 2007. *"Interrupting the Pots": The Excavation of Cleatham Anglo-Saxon Cemetery, North Lincolnshire*. Council for British Archaeology, York.

Leahy, K. 2014. *The Excavation of the Cleatham Anglo-Saxon Cemetery, North Lincolnshire* [data-set]. York: Archaeology Data Service [distributor] https://doi.org/10.5284/1000011

Nugent, R. and Williams, H. 2012. Sighted surfaces. Ocular Agency in early Anglo-Saxon cremation burials. In: I.-M. Back Danielsson, F. Fahlander and Y. Sjöstrand (eds.) *Encountering images: materialities, perceptions, relations*. Stockholm Studies in Archaeology, 57. (Stockholm: Stockholm University), 187–208.
http://www.mikroarkeologi.se/publications/encounteringimagery/11.Howard_Ruth.pdf

Squires, K. 2011. *An Osteological Analysis and Social Investigation of the Cremation Rite at the Cemeteries of Elsham and Cleatham, North Lincolnshire.* PhD Thesis, University of Sheffield
https://doi.org/10.5284/1029431

Squires, K. 2012. Populating the Pots: The Demography of the Early Anglo-Saxon Cemeteries at Elsham and Cleatham, North Lincolnshire. *Archaeological Journal* 169 (1), 312-342.
https://doi.org/10.1080/00665983.2012.11020917

Squires, K. 2013. Piecing Together Identity: A Social Investigation of Early Anglo-Saxon Cremation Practices. *Archaeological Journal* 170 (1), 154-200.
https://doi.org/10.1080/00665983.2013.11021004