The Death of Data

written by Jeremy Huggett | 08/11/2019



"Dead Data" by Stinging Eyes CC BY-SA 2.0

Yesterday was World Digital Preservation Day and saw the publication of the Digital Preservation Coalition's Bitlist – their global list of Digitally Endangered Species. Interestingly, under their 'Practically Extinct' category ("when the few known examples are inaccessible by most practical means and methods") sits *Unpublished Research Data*, which they define as

> "research data which has not been shared or published by any means and is thus in contravention of the 'FAIR' principles which require data to be Findable Accessible, Interoperable and Reusable".

Although the DPC jury hopes that this is a small group, I rather suspect that there is an unseen mountain of unpublished research data in archaeology (and in the interest of full disclosure: reader, I have some).

This crossed my screen at the same time as a paper published in the Harvard Data Science Review by Stephen Stigler: 'Data Have a Limited Shelf Life', in which he argues that data, unlike wines, do not improve with age. He suggests that old data are "Often … no more than decoration; sometimes they may be misleading in ways that cannot easily be discovered", while emphasising this is not the same as saying they have no value. Using three examples of old statistical data, he shows how misleading and incomplete they can be if their full background is not known. In each case, the data were selected from a prior source, not always accurately referenced if at all. In some instances, uncovering the original data flagged problems with the sample that had been taken, in others it revealed a greater breadth and depth of information which had gone un-used because the particular research question had stripped them away.

> "There was in each case a large degree of selection of data that was unreported. There are errors of transcription and commission. There are doubts about important aspects that cannot be answered at this point in time. And there is the unescapable conclusion that each data set had become only ornamental at an early stage of its history."

He points to both gains and losses for the data as a consequence and concludes "it is not possible to view them in the same way as before. Scientifically, they are dead data."

Stigler goes on to remind us that if data can die of old age, it can also die alongside the person who collected them, especially when they lack associated and adequate documentation. This absence is also highlighted in the DPC's 'Practically Extinct' category and something we are familiar with in archaeology, not least in the shape of attempting to deal with archives post-mortem (of individuals) and post-closure (of organisations) found in garages, sheds or their equivalent.

Stigler challenges the idea that data is only improved with more data, suggesting that:

> "The life span of data usually has an ending, and it is much earlier than generally realized. That end comes when people stop using the data in favor of later, different data for the same or similar goals, or when they become irrelevant … They die when they do not record what later researchers want, or when the trust in them has diminished, or when the questions they can answer are no longer asked, or when new, better data supersede them."

As he points out, this has major implications for big data built from collections of old data, which are as susceptible as any data to selection bias and sampling problems but their size militates against either recognising or dealing with such issues. Or if they are dealt with, they risk becoming what he calls 'zombie data', where old data is massaged to bring them back to life but in the process generates as many problems as it (might) solve.

Where does this leave us with archaeological data? Leaving aside the observation that given the focus of most archaeological data, our data are 'dead' from the outset, I don't think that Stigler's image of data death entirely fits the situation we find ourselves in. That said, the idea that datasets can die (and in some cases can be revivified but create zombie datasets in the process) presents some major challenges to archaeological data. Several commentators have observed that we tend to generate new data rather than reanalyse old data, partly as a result of the always-present threat to archaeology in the ground and the demand for preservation by record, and partly as a result of the complexities of dealing with old data, something Stigler also recognises. We certainly don't recognise the 'death' of data other than in the sense of its total loss – instead we collect more and more of it but see this as an additive practice rather than the replacement of deceased datasets, holding onto the belief that old data will, sometime, somewhere, be useful. Or as Stigler puts it, "Different people with different questions to answer could ask different questions. In that sense data could be dead to one user but not another".

And of course, the one area seeing a growing reuse of older archaeological data is through their amalgamation into big(ger) data to ask questions that were never the basis for the original data. As a result, we can all-too easily lose sight of the way that those original datasets were shaped by their

original research questions, by the limitations of their circumstances of capture, by the choices and selections made at the time by those responsible for their capture, and the consequences that may flow from wrenching such data – or elements of those data – from their original context. Indeed, the importance of context is something Stigler recognises in his conclusion: its significance in shaping the selection and form of the data, along with the temporal, political and social conditions of their creation.

Is it possible to take all this into account when we deal with old data? Are we really prepared to revisit the circumstances of data creation and build those conditions into our analyses? Capturing this information via meta/paradata is one thing, but how are these descriptive categorisations transformed into truly resurrected rather than zombified data (something that work by the SLO-Data project, for example, is currently looking at (e.g. Yakel and Whitcher Kansa 2019; Yakel et al. 2019; Faniel et al. 2018))? Or are we content to largely ignore these questions as has primarily been the case to date? *Are* old archaeological data destined for death regardless of their preservation in archives (or garages and sheds), and is this why the evidence for their reuse so limited?

References

Faniel, I., A. Austin, E. Kansa, S. Whitcher Kansa, P. France, J. Jacobs, R. Boytner and E. Yakel 2018. Beyond the Archive: Bridging Data Creation and Reuse in Archaeology. *Advances in Archaeological Practice*, 6(2):105-116. DOI: 10.1017/aap.2018.2 (also available at https://www.oclc.org/research/publications/2018/beyond-the-archive-data-creation-reuse.html)

Stigler, S. 2019. Data Have a Limited Shelf Life. *Harvard Data Science Review* 1(2). https://doi.org/10.1162/99608f92.f9a1e510

Yakel, E. and S. Whitcher Kansa 2019. Designing, Timing, and Determining the Feasibility of Curatorial Interventions to Support Data Reuse. OCLC Works in Progress Webinar, https://www.oclc.org/research/events/2019/103119-designing-timing-determining-curatorial-interventions-data-reuse.html

Yakel, E., I. Faniel and Z. Maiorana 2019. Virtuous and vicious circles in the data life-cycle. *Information Research*, 24(2). http://InformationR.net/ir/24-2/paper821.html.