Dark Data

written by Jeremy Huggett | 22/04/2020

There are quite a few metaphors associated with archaeological data, many of which relate to its apparent mystery. For example, Gavin Lucas has described the archaeological record as being "haunted by absences" created by decay and destruction (Lucas 2012, 178). In a similar vein, Alison Wylie has described archaeological data as "shadowy" and that archaeology is defined "by the challenges of working with gaps and absences in its primary data" (Wylie 2017, 204). In a special issue of the *Science, Technology, & Human Values* journal on 'Data Shadows', Leonelli *et al.* describe data in terms of its presence, but also in terms of its unavailability, inaccessibility, or its absence, defining absence as a descriptor of how "data are missing, incomplete, unreliable, ignored, unwanted, or untagged" (Leonelli *et al*. 2017, 192). As Chris Chippendale described it,

> Archaeology is plagued in many an instance with poorly defined variables (usually thought of as 'data') drawn from ill-understood populations, and with uncertain articulations between the entities whose logical relations we seek to understand. (2000, 611)

So far, so well understood. We appreciate that archaeological data is incomplete, that in many respects it is more absent than present and so we make the best of what data we do have and use it to build our conclusions about the past.

But what does this actually mean? In Gavin Lucas's discussion of the archaeological record, he cites a list formulated by Michael Collins (1975, 27) on sources of bias in our data:

1. Not all behaviour results in patterned material culture.
2. Of those that do, not all can enter the archaeological record.
3. Of those that do, not all will enter the archaeological record.
4. Of those that do, not all will be preserved.
5. Of those that do, not all survive indefinitely.
6. Of those that do, not all will exposed by the archaeologist.
7. Of those that do, not all will be identified and/or recognized by the archaeologist.

Collins wasn't without his critics: Binford, for example, complained that this was of no relevance to a discussion of sampling and his "… arguments are more relevant to the questions of whether we should even attempt to use archaeological facts in evaluating our ideas about past dynamics or whether it is worthwhile even to do archaeology of any kind!" (Binford 1975, 254). But as Lucas says, the list is important because it links the incompleteness of the archaeological record with the

incompleteness of archiving and/or collecting (Lucas 2012, 64).

So in fact we can extend this list further:

8. Of those that do, not all can be recorded by the archaeologist
9. Of those that do, not all will be recorded by the archaeologist
10. Of those that do, not all will be recorded in the same way by different archaeologists
11. Of those that do, not all will be deposited in an archive by the archaeologist

And of course, in seeking to re-use such data, the challenge is how we reconstruct these sets of decisions surrounding the collection of data so that we can subsequently use the data with confidence. What are the implications of data we know are missing, data we don't know we're missing, data which might have existed but hasn't survived, data which might have existed but hasn't been collected, data which we don't know is selective or the reasons for its selection, and data we don't know is unreliable because of issues of measurement, accuracy, uncertainty, ambiguity and bias? What are the consequences of these absences for knowledge creation?

The increasingly common response to this is to use meta- or paradata: data concerning the context within which the data were captured, and the processes and interpretations that were applied in their creation. For example, the London Charter talks about documentation of "the evaluative, analytical, deductive, interpretative and creative decisions", and more recently paradata has been characterised as "detailed information about the excavation of the remains, the analyst's training and expertise, where analysis took place, which methods and reference materials were used, how the dataset was modified, etc." (Kansa *et al*. 2020, 45). However, as soon as that paradata is codified or structured (e.g. Doerr *et al*. 2014) it seems to focus almost inevitably on the technical aspects of the data processes, pushing the non-technological, more human-centred decisions and actions into the background. This might reasonably be expected to affect our approaches to the data and consequently skew results in unforeseen and unrecognised ways.

Essentially we're missing what might be characterised as the 'marginalia' of archaeological data capture – the notes, discussions, comments made by the participants during the course of data discovery, collection, and recording. It's not simply that archaeology begins at Ian Hodder's trowel's edge, but it is also situated in the conversations at the edge of the trench, as Colleen Morgan and Holly Wright (2018, 146) have argued. Marginalia are said to exist outside the bounds of the parameters of a study, and may simultaneously indicate misunderstanding and miscommunication, flag difference and disagreement: things we might expect to appear in analog records to some degree (such as in field notebooks or on plan/section drawings), but which on the whole are not party to digital data. For example, Sara McClelland (2016, 160) suggests that marginalia disrupt and challenge assumptions about research processes, conceptual definitions, and issues of measurement and analysis; they provide a means by which the original participants interrupt or disrupt subsequent researchers' expectations. In particular they shine a light on intuitive knowledge and know-how, things that are rooted in experience and practice, but which are difficult to communicate. Aspects of these marginalia may be captured in the more unstructured narrative forms of paradata, but only if the hurdle of capturing these assumptions and decisions is overcome in the first place. More often, the difficulty of articulating these kinds of data means that they remain unexpressed.

These constitute archaeology's 'dark data' – what David Hand (2020) defines as data that are effectively concealed from us, which means we are at risk of misunderstanding, drawing incorrect conclusions – things we should surely be concerned about. Hand defines a taxonomy of dark data (2020, 291ff): some fifteen characteristics which highlight the dangers and issues associated with data, and much of what I've been writing about here falls within these categories.

Isto Huvila has recently characterised paradata as a 'wicked' problem: it's "the practical impossibility to document and keep everything and the difficulty of determining how to capture just enough". So how much is enough? Where do we draw a pragmatic line between some kind of theoretical 'completeness' of data knowledge and ignorance of the unknown unknowns of our data? And in the end, would we use this information anyway?

[Some of these thoughts about paradata formed part of a presentation to the SEADDA Working Group 4 Exploratory Workshop on the 'Use and Re-use of Archaeological Data', held online on 31st March – 2nd April 2020, and I'm grateful to Holly Wright and her colleagues for the invitation to speak.]

References

Binford, L. 1975. 'Sampling, Judgement, and the Archaeological Record'. In J. Mueller (ed.) *Sampling in Archaeology.* Tucson: University of Arizona Press, pp. 251-257.

Chippindale, C. 2000. 'Capta and data: On the true nature of archaeological information', *American Antiquity* 65(4), 605–612. https://doi.org/10.2307/2694418

Collins, M. 1975. 'Sources of Bias in Processual Data: An Appraisal'. In J. Mueller (ed.) *Sampling in Archaeology.* Tucson: University of Arizona Press, pp. 26–32.

Doerr, M., Chrysakis, I., Axaridou, A., Theodoridou, M., Georgis, C. and Maravelakis, E. 2014 'A framework for maintaining provenance information of cultural heritage 3D-models', *25th International Conference of Electronic Visualisation and the Arts 2014*. https://doi.org/10.14236/ewic/eva2014.63

Hand, D. 2020 *Dark Data: Why What You Don't Know Matters.* Oxford: Princeton University Press.

Kansa, S., Atici, L., Kansa, E. and Meadow, R. 2020 'Archaeological Analysis in the Information Age: Guidelines for Maximizing the Reach, Comprehensiveness, and Longevity of Data', *Advances in Archaeological Practice* 8(1), 40-52. https://doi.org/10.1017/aap.2019.36

Leonelli, S., Rappert, B. and Davies, G. 2017, Data Shadows: Knowledge, Openness, and Absence, Science, Technology, & Human Values 42 (2), 191-202. https://doi.org/10.1177/0162243916687039

Lucas, G. 2012. *Understanding the Archaeological Record*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511845772

McClelland, S. 2016. 'Speaking Back From the Margins: Participant *Marginalia* in Survey and Interview Research', *Qualitative Psychology* 3(2), 159-165. DOI: https://doi.org/10.1037/qup0000061

Morgan, C. and Wright, H. 2018. 'Pencils and Pixels: Drawing and Digital Media in Archaeological Field Recording', *Journal of Field Archaeology* 43(2), 136-151.

https://doi.org/10.1080/00934690.2018.1428488

Wylie, A. 2017. How Archaeological Evidence Bites Back: Strategies for Putting Old Data to Work in New Ways, Science, Technology, & Human Values 42 (2), 203 – 225.
https://doi.org/10.1177/0162243916671200