Delving into Data Reuse

written by Jeremy Huggett | 10/10/2019

Given the years, the money, expertise and energy we've spent on creating and managing archaeological data archives, the relative lack of evidence of reuse is a problem. Making our data open and available doesn't equate to reusing it, nor does making it accessible necessarily correspond to making it usable. But if we're not reusing data, how can we justify these resources? In their reflections on large-scale online research infrastructures Holly Wright and Julian Richards (2018) have recently suggested that we need to understand how to optimize archives and their interfaces in order to maximize the use and reuse of archaeological data, and explore how archaeological archives can better respond to user needs alongside ways to document and understand both quantitative and qualitative reuse.

However, I would argue that all these kinds of issues (alongside those of citation, recognition, training, etc.) while not resolved are at least known and mostly acknowledged. The real challenges to data reuse lie elsewhere and entail a much deeper understanding and appreciation of what reuse entails: issues associated with the re-presentation and interpretation of old data, the nature and purpose of reuse, and the opportunities and risks presented by reuse. Such questions are not specific to digital data; however, digital data change the terms of engagement with their near-instant access, volume, and flexibility, and their potentially transformative effects on the practice of archaeology now and in the future.

So what are these 'deeper' challenges for reuse? I offer six suggestions …

1. The performative nature of archaeological data

It might seem obvious to say that data do not create themselves but are created by us – or in fact mostly by others, separated from us in space and time. But there is an ambiguity in our approach to data, where it is seen as both 'given' (atomic fragments) and 'made' (created or reconstructed). There is a long history of seeing data as 'givens', as raw materials, the stuff of archaeology, existing independent of the archaeologist, waiting to be discovered, and when brought into the light, used to create information and build knowledge in the classic DIKW pyramid. On the other hand, we can see data as a result of acts of performative creation and not 'given' at all. Instead, based on our experience, research objectives and so on, we articulate our knowledge to recognise, identify and categorise information, and that information is atomised within a digital environment to create data. Data in these terms are not raw at all, but embedded with theory, process, and purpose, presenting a major challenge in terms of reusability. However, in a Big Data environment, the earlier perception

of data as ‘raw’, independent of the relations and contexts that gives rise to them, has become increasingly resurgent. How we perceive the origination of data should fundamentally affect our subsequent handling and use of it, and the challenge lies in recognising this and, I would argue, ensuring our data systems capture its performativity alongside what might otherwise be seen as the ‘raw’ data.

2. The incomplete and imperfect nature of archaeological data

I’m certainly not the first to suggest that archaeological data are messy, complicated by their partial, fragmentary, interpretative nature. However, it should alert us to the consequent complexity of handling such data whereas we are increasingly seeing the growing perception that data are relatively straightforward and unproblematic, with any issues of reliability or quality capable of being overcome by virtue of their quantity. This is a frequent claim of Big Data proponents, and intuitively the bigger a sample is the better the outcome is likely to be. However, some studies have shown that the impact of poor-quality data can increase rather than reduce as dataset size increases (e.g. Succi and Coveney 2019; Woodall et al. 2014) which, given the nature of archaeological data, should act as a red flag. The challenge here is to find a means of capturing the data shadows, to properly record the set of preunderstandings (Wylie 2017) that structure our data, recognising that much of our data are recorded over time, by different people with different practical and theoretical agendas. The metadata we are accustomed to creating and using falls far short in this respect, concentrating as it does on discovery rather than reuse.

3. The operationalisation of archaeological data

We often seem to assume that data are self-evident but in reality, are often unclear about how the data have been identified and characterised, what actions were taken in their recording, and how they were subsequently prepared for deposit within our archives. This lack of transparency over the manipulations that data typically undergo makes it difficult to evaluate the decisions taken to address the different recording conventions, data formats, data models, encountered within our data and to resolve the host of anomalies within the data themselves. Recognising and recording these kinds of issues represents a major challenge to our current digital practice and we tend to have a somewhat improvisational approach to paradata at present. This becomes increasingly significant with the increasing tendency to amalgamate our datasets into larger datasets, if not ‘Big Data’, combining data from multiple sources into massified datasets for analytical purposes.

Another challenge linked to the operationalisation of our data lies with the very data infrastructures themselves. These represent new data gatekeepers and although they have been largely built and driven by digitally knowledgeable archaeologists, we are barely beginning to understand the predispositions of these systems. It is not just the data that are situated but the data infrastructures themselves are also situated culturally, socially, politically, technologically and spatially, and consequently risk the creation of ‘filter bubbles’ which prioritise certain kinds of data retrieval and use through the design of their search tools and the structuring of their data. So an additional challenge for operationalisation is a greater understanding and appreciation of the tools we are building to provide access to these data.

4. The unanticipated potential of archaeological data

One of the most challenging aspects of archaeological data is its almost infinite flexibility in terms of

potential applications. Much archaeological research entails using data as proxy for immaterial processes meaning that archaeological data have almost infinite potential applications as we come up with new proxies. In reality, though, the digital data we seek to reuse can constrain and limit subsequent analysis. Our databases are theoretically structured in ways that enable certain perspectives and disable others. So the challenge here, therefore, is to recognise that the very tools that are seen as enabling, that facilitate and simplify our reuse of data constrain our actions and reduce potential applications. They limit as well as enable, and this is often overlooked in the thrill of access and retrieval. Alongside the unanticipated potential of our data, in other words, are unanticipated constraints on our use of that data which we need to find ways to overcome that don't just entail ignoring them.

5. The insertion of non-human actors in archaeological data

One of the characteristics of our digital era is the increasing insertion of non-human actors in the process of data collection. Many of our digital devices capture data and are increasingly automated, with less need for human intervention. Data becomes less something that arises out of our observations and engagement and more something that is automatically captured on our behalf. The trap is to assume that, because such data are created by machine, they are somehow objective and more reliable than any human equivalent whereas machine-generated digital data obscures the role of human decisions in its creation. So the challenge with such automated data collection is to ensure that knowledge of the human decisions behind the collection, along with the mixture of human-derived and technical constraints of the instrumentation are appropriately captured and recorded alongside the data themselves.

6. The environmental impact of archaeological data

Back in 2016 I wrote a post on 'Digital Data Realities' which generated a lot of interest, not least because I openly claimed that we weren't using our archived archaeological data. Interestingly, though, one aspect of that post which no-one picked up on was the question of the environmental cost of our emphasis on immediate online access to data which we weren't actually using. For example, Keith Pendergrass *et al.* (2019) have argued that we need a

> … renewed emphasis on critical *appraisal*; reduce the resource-intensity of digital storage and management by rethinking digital *permanence*; and meet user needs in different ways by challenging assumptions about the *availability* of digital content and the need for 'always-on' digital access infrastructure." (2019: 181).

This is not something we seem to have explicitly addressed as archaeologists, although digital archivists will be all-too-aware of these issues. Do all digital data require the same level of preservation and migration? Do all data have to be held at the highest resolution possible? Are there levels of acceptable loss – as Pendergrass *et al*. point out, a random bit flip in an uncompressed TIF graphic file will be less significant than in a JPG file. Do we always need to keep storage media powered up to ensure immediate access, or can we conceive of different levels of speed of access according to levels of demand, for instance? The challenge here is to incorporate the environmental costs associated with our demands for digital data more explicitly into our data infrastructures.

Et enfin

Alongside our 'known unknowns' for data reuse (more data, more data citation, greater digital data literacy, better interfaces, enhanced access, etc.), these six interrelated challenges represent some 'unknown unknowns', emphasising the need to look deeper at the nature of our data and how that nature is properly represented and captured alongside the data themselves. In short, we need to:

1. capture the performativity of data and its processes of creation;
2. capture the data shadows and the set of preunderstandings associated with the structuring of data;
3. appreciate the predispositions of our data and our repositories;
4. understand the constraints on our tools and our data which limit their potential;
5. capture the human decisions embedded within automated data devices;
6. incorporate the environmental costs in the consideration of our digital data

These challenges will affect our reuse of data, whether we choose to recognise it or not. And because they are not primarily technical issues, they will not be resolved unless we choose to address them.

[This post is part of a presentation given to the Mémoires Archéologues et des Sites Archéologiques (MASA) seminar on 'La réutilisation des données archéologiques: dépasser les frontières?' at the University of Nanterre on 19th Sept 2019. Thanks to Xavier Rodier, Emmanuelle Bryas, Nathalie Le Tellier-Becquart and colleagues for their invitation and hospitality]

References

Pendergrass, K., W. Sampson, T. Walsh and L. Alagna (2019) 'Toward Environmentally Sustainable Digital Preservation'. *The American Archivist* 82(1), 165-206. https://doi.org/10.17723/0360-9081-82.1.165

Huggett, J. 2016. 'Digital Data Realities'. *Introspective Digital Archaeology* Jun 29. http://introspectivedigitalarchaeology.com/2016/06/29/digital-data-realities/

Huggett, J. 2016. 'Gatekeepers in Digital Archaeology'. *Introspective Digital Archaeology* Nov 11. http://introspectivedigitalarchaeology.com/2016/11/11/gatekeepers-in-digital-archaeology/

Huggett, J. 2017. 'The Idle Archive?'. *Introspective Digital Archaeology* Feb 10. http://introspectivedigitalarchaeology.com/2017/02/10/the-idle-archive/

Huggett, J. 2017. 'Digital Proxies'. *Introspective Digital Archaeology* Nov 28. http://introspectivedigitalarchaeology.com/2017/11/28/digital-proxies/

Huggett, J. 2018. 'Digital Data Relations'. *Introspective Digital Archaeology* Jun 28. http://introspectivedigitalarchaeology.com/2018/06/28/digital-data-relations/

Succi, S. and P. Coveney 2019. "Big data: the end of the scientific method?" *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 377. https://doi.org/10.1098/rsta.2018.0145

Woodall, P., A. Borek, J. Gao, M. Oberhofer and A. Koronios. 2014. 'An Investigation of How Data

Quality is Affected by Dataset Size in the Context of Big Data Analytics'. *Proceedings of the 19th International Conference on Information Quality (ICIQ-2014)*, pp. 24-33. https://ualr.edu/informationquality/iciq-proceedings/iciq-2014/.

Wright, H. and J. Richards. 2018. 'Reflections on Collaborative Archaeology and Large-Scale Online Research Infrastructures'. *Journal of Field Archaeology* 43(sup 1), S60-S67. https://doi.org/10.1080/00934690.2018.1511960

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