Data Detachment

written by Jeremy Huggett | 20/10/2022

‘Data detachment’ via Craiyon

A couple of interesting but unrelated articles around the subject of humanities digital data recently appeared: a guest post in *The Scholarly Kitchen* by Chris Houghton on data and digital humanities, and an *Aeon* essay by Claire Lemercier and Clair Zalc on historical data analysis.

Houghton’s article emphasises the benefits of mass digitisation and large-scale analysis in the context of the increasing availability of digital data resources provided through digital archives and others. According to Houghton, “The more databases and sources available to the scholar, the more power they will have to ask new questions, discover previously unknown trends, or simply strengthen an argument by adding more proof.” (Houghton 2022). The challenge he highlights is that although digital archives increasingly provide access to large bodies of data, the work entailed in exploring, refining, checking, and cleaning the data for subsequent analysis can be considerable.

> An academic who runs a large digital humanities research group explained to me recently, “*You can spend 80 percent of your time curating and cleaning the data, and another 80 percent of your time creating exploratory tools to understand it*.” … the more data sources and data formats there are, the more complex this process becomes. (Houghton 2022).

Houghton sees this as a significant problem, and a key hurdle that he identifies is the need for technical skill alongside disciplinary knowledge (and its implicit absence). Rather than addressing the skill shortage, his solution is instead to point to the environments provided by Gale (his employer) and others which seek to replace the requirement for coding skills with user-friendly interfaces.

> Tools like these can seriously streamline the workflow of collecting, curating, cleaning, and analyzing huge sets of data. Instead of researchers spending 80 percent of their time on pre-processing tasks, using these tools radically reduces the time spent on them, bringing it down from months to days, if not hours. (Houghton 2022).

The requirement or otherwise to have coding abilities has long been debated within digital humanities, and indeed Lemercier and Zalc (2022) point to the resurgence of "the historian-as-programmer" associated with the growth in digital and 'big' data. Personally, I've always been in favour of possessing some coding knowledge as it provides useful insights into how computer systems and software operate as well as being a desirable transferable skill. However, Houghton sees coding requirements as a problem (perhaps because such skills are still relatively rare due to lack of training opportunities), and his solution is the insertion of user-friendly tools between the historian/humanist/archaeologist and their data. At the risk of seeming perilously close to an argument that to be worth anything an analysis must necessarily be 'hard', my concern with this approach is that this necessarily introduces separation or distance between the analyst and their data: as I've suggested elsewhere, there's a paradox at work here in which quantity, convenience, and availability makes data more accessible at the same time as it becomes more remote (Huggett 2022, 276).

What Houghton characterises as a problem is arguably a desirable requirement, so that his removal of the problem itself becomes the problem. For example, data cleaning is all too often seen as a chore,

> … implying that heterogeneity in the sources is a problem to be solved – and that solving it is a subaltern task. For us, on the contrary, building data from sources and creating categories that do not erase all complications is not just the longest and most complicated stage of research; it is also the most interesting. (Lemercier and Zalc 2022).

As Lemercier and Zalc emphasise, dirty data is itself of value, through its absences, outliers and downright weirdness. By their very nature, friendly interfaces insert themselves between us and our dirty data and may thereby encourage a more passive rather than active approach to the data, with the data becoming little more than tokens to be shuffled, reordered, and rearranged. Handling data at arms-length in this way is associated with a remoteness or detachment which might be mistakenly linked to objectivity but risks becoming a derogation of responsibility, engagement, familiarity, and awareness. As Lemercier and Zalc observe, using new data analysis techniques (and they specifically reference 'big' data approaches) can make source criticism seem optional, the provenance and construction of the data being accepted without question.

Although archaeological data starts out as primary (derived from the direct observation, recognition, and recording of archaeological evidence), a large proportion of archaeological data analysis is essentially secondary, using prior data made available for reuse and combining and reworking it for different purposes, while some is tertiary in that it is entirely reliant on these reworked datasets. Consequently, much archaeological analysis is already distanced from the original material record, so introducing software environments that may effectively increase this remoteness may seem open to risk. They may encourage us to work with data from a high-level, somewhat disconnected

perspective which makes it difficult to track back through the sets of transformations, modifications, and re-articulations which in combination represent the data journeys from initial discovery to their reuse in knowledge creation.

To have confidence in our analyses and results we need to be able to trace these data journeys and be able to confirm the data's reliability. One of the key features of digital archaeology over recent decades has been the construction of infrastructures for storing, archiving, organising, and making data available, as well as systems to streamline and simplify data processing and analysis. Working on and accepting data at infrastructural or environmental level is to operate at at least one remove, taking the data as delivered for granted unless it is possible to drill down into the original datasets and examine their origins and transformations. This requires what Paul Edwards calls 'infrastructural inversion', where the infrastructure is turned upside down to examine "the parts you don't normally think about precisely because they have become standard, routine, transparent, invisible" (Edwards 2010, 20). This is not easy to do and, as things stand, not made easy by existing infrastructures which are often opaque about how they function (e.g. Huggett 2022, 284-5). Nor are our data organised or well-equipped to support this tracking back and – other than in terms of accessibility – digital data are not dramatically easier to work with than older analog data, so in this sense Houghton is correct to highlight the labour involved, even though there is value and significance in that effort. Periodically reminding ourselves that our data are not simple and that they should be used in full knowledge of their origins, the circumstances of their discovery, their recording strategies and standards, their interpretative modifications, and the nature of any relationships with other data, is always a valuable precursor to analysis.

References

Edwards, P. (2010) *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming*. Cambridge, Mass: MIT Press.

Houghton, C. 2022 'Three Challenges (and Solutions) to Expand Digital Humanities', *The Scholarly Kitchen* (September 26, 2022). Available at: https://scholarlykitchen.sspnet.org/2022/09/26/guest-post-three-challenges-and-solutions-to-expand-digital-humanities/

Huggett, J. (2022) 'Data Legacies, Epistemic Anxieties, and Digital Imaginaries in Archaeology', *Digital*, 2(2), pp. 267–295. Available at: https://doi.org/10.3390/digital2020016

Lemercier, C. and Zalc, C. (2022) 'History by Numbers: Is history a matter of individual agency and action, or of finding and quantifying underpinning structures and patterns?' *Aeon* (September 2, 2022). Available at: https://aeon.co/essays/historical-data-is-not-a-kitten-its-a-sabre-toothed-tiger