Open Data quality

written by Jeremy Huggett | 11/11/2014

In relation to the Portable Antiquities Scheme (PAS) database, David Gill on his 'Looting Matters' blog has pondered "How far can we trust the information supplied with the reported objects? Are these largely reported or 'said to be' findspots?".

Spatial information is frequently cited as a problem in relation to open archaeological data – but the focus tends to be on the risks it poses for looting (for example, Bevan 2012, 7-8; Kansa 2012, 508-9).

While this kind of problem is not to be underestimated, the solutions – degrading the quality of spatial data and/or making full resolution data available only to 'approved' users fly in the face of open data requirements. In relation to the data themselves, more important issues are arguably related to the authority of the data and the risk of reducing confidence in a data source as a consequence of revealing discrepancies and errors in the data. With datasets consisting of millions of records in some cases, it would be surprising if errors did not creep in, especially as the data are increasingly manipulated by automated means. Whether this damages the authority of the data is open to question: arguably issues with the data such as different levels of precision of locational information are likely to be more problematic for would-be users than the occasional rogue item.

However, what the PAS example highlights is that the underlying concerns relate to more than 'just' the essential quality of the data themselves, but have to include a consideration of the recording that gave rise to those data, and this is where the real question with open data quality lies. Open data frequently lacks the context of recording, and it is all to easy to overlook the theory-laden, process-laden and purpose-laden nature of the data. In this instance, the risk clearly lies in a bias introduced by differential recording, for whatever reason that might have arisen. But how would an end-user, separated in time and space from the original discovery, and remote from the data source itself through the intervention of search engines, retrieval tools, etc., know whether or not this is the case or be able to disentangle the good from the rest?

One of the challenges here is for those who do seek to wrangle such datasets to make the process of cleaning and checking clear and transparent. All too often, it seems that those relatively few accounts of data reuse tend to stress the positive outcomes and minimise the efforts entailed in achieving them, presumably because the typical reader – including the digital archaeologist – is more interested in results than the nitty gritty detail of the resolution of problems within the data, despite the implications this might carry for the validity of the conclusions drawn.

References

Bevan, A. 2012 'Value, authority and the Open Society: some implications for digital and online archaeology', in C. Bonacchi (ed.) *Archaeology and Digital Communication: Towards Strategies of Public Engagement* (Archetype: London), 1-14.

Kansa, E. 2012 ‘Openness and archaeology’s information ecosystem’, *World Archaeology* 44 (4), 498-520.