Big Data Analytics

written by Jeremy Huggett | 08/02/2015

It was only a matter of time before a ‘big data’ company latched onto archaeology for commercial purposes. Reported in a *New Scientist* article last week (with an unfortunate focus on ‘treasure’), a UK data analytics start-up called Democrata is incorporating archaeological data into a system to allow engineering and construction firms to predict the likelihood of encountering archaeological remains. This, of course, is what local authority archaeologists do, along with environmental impact assessments undertaken by commercial archaeology units. But this isn’t (yet) an argument about a potential threat to archaeological jobs.

The ability to use online resources such as Heritage Gateway in England, Canmore/Pastmap in Scotland, Coflein in Wales, and the Archaeology Data Service, has been growing in recent years, but all these resources hedge their use for commercial purposes by emphasising the need for professional archaeological advice in conjunction with the data. This same data is presumably the basis for the Democrata system. I say presumably because the *New Scientist* article is vague, describing it as “documents from government departments such as the Forestry Commission, English Heritage and Land Registry” and including “‘grey literature’, the massive set of unpublished reports written by contractors every year” (the Archaeology Data Service holds over 30,000 of these at the last count through a project supported by, among others, English Heritage). Nor is the Democrata website any guide – for a start-up company set up in July 2014, their website is remarkably uninformative.

It’s safe to assume that archaeological data are just one small, very minor, aspect of Democrata’s business (for instance, they’re reported to have won the Open Innovation prize from the Science & Technology Facilities Council, and incubator funding from the European Space Agency, which would suggest a primary focus other than archaeology). It is a concern, however, that public archaeological data is being incorporated into a commercial system from which it seems improbable that archaeology itself will be able to benefit in terms of knowledge of enhanced data processing and the like. But even that isn’t really the issue.

The issue is with the data (and the ends to which they are being put). The Archaeology Data Service, in conjunction with English Heritage, the University of Glamorgan and others, has for some years been working on aspects of automated classification and natural language processing of these archaeological datasets through projects such as SENESCHAL, STAR, STELLAR, and Archaeotools. These and other projects are developing tools that are part and parcel of ‘big data’ processing, and a lot of valuable work has been done largely focusing on extracting what/when/where information. But one thing remains clear – to automate the extraction and processing of information from these resources sensibly is far from trivial when those resources weren’t created with this kind of use in mind. The levels of accuracy are improving, but – as anyone who has tried computer dictation has experienced – even 99% accuracy leaves a lot of manual tweaking required when the volumes of data are large. Furthermore, embedded in these systems are content standards, documentation standards, and ontology standards in an ill-defined web of relationships and dependencies, about which we know and understand too little (Huggett 2012), and into which these data are

automatically fitted to make them useful. Couple that with the known challenges of archaeological predictive modelling within GIS. The location factors reported as being employed by Democrata seem very familiar – proximity to water, to mineral resources, to known sites of religious significance, etc. – and haven't as a rule dramatically improved archaeological predictive models. These are key ongoing areas of archaeological research and development – and yet here we have a company without apparent archaeological expertise offering a commercial service predicting the whereabouts of archaeological remains on the strength of their access to and processing of these datasets.

Henry Chapman, senior lecturer in Archaeology and Visualisation at Birmingham University, is quoted by *New Scientist* as expressing concern that the Democrata tool might limit future archaeological discoveries by reducing the number of excavations that take place in advance of development – which, after all, is Democrata's selling point to would-be developers. However, given the nature of the data and the challenges in their processing, whether a developer could (should?) sensibly or legitimately rely on such information remains open to question.

It would be nice to think that a collaborative archaeological research project might be part of Democrata's forward plans since, after all, they will be profiting from publicly funded data and resources.

References

Huggett, J. 2012 'Lost in information? Ways of knowing and modes of representation in e-archaeology', *World Archaeology* 44 (4), 538-552.

Rutkin, A. 2015 'Data archaeology helps builders avoid buried treasure', *New Scientist* 3006, January 2015.
http://www.newscientist.com/article/mg22530062.400-data-archaeology-helps-builders-avoid-buried-treasure.html