The data deluge
Andrew Bevan

doi:10.15184/aqy.2015.102

**Introduction**
Archaeology has wandered into exciting but daunting territory. It faces floods of new
evidence about the human past that are largely digital, frequently spatial, increasingly
open and often remotely-sensed. The resulting terrain is littered, both with data that are
wholly new and data that were long known about but previously considered junk. This
paper offers an overview of this diluvian information landscape and aims to foster
debate about its wider disciplinary impact. In particular, I would argue that its
consequences: a) go well beyond the raw challenges of digital data archiving or
manipulation and should reconfigure our analytical agendas; b) can legitimately be read
for both utopian and dystopian disciplinary futures; and c) re-expose some enduring
tensions between archaeological empiricism, comparison and theory-building.

**Overground, underground, systematic and serendipitous**
It has become popular to speak of a ‘big data’ phenomenon in the social and natural
sciences, as well as to a lesser extent in the humanities. Narrowly defined, the term ‘big
data’ refers to those structured or unstructured digital datasets that are so colossal
(often hundreds of terabytes or more) that they present unusual contemporary
challenges with regard to everything from basic storage to read/write/search functions,
analysis and visualisation. In truth, only a few kinds of archaeological information even
approach this level of challenge: certain kinds of remote-sensing or ancient DNA
datasets undoubtedly do, and in the future so too might certain kinds of automatically
gathered evidence about public engagement (for example, online user statistics and
geo-social networking related to heritage experiences). While most archaeological
evidence may not typically be quite so vast, it is nonetheless at least as rich and
challenging in other ways and the quantities are rising sharply (see Figure 1).

One major source of new information is remote sensing and there are at least six key
developments that, together, have ushered in a remote-sensing revolution over the last
decade.

- First, since the advent of high spatial resolution (i.e. ≤1m) commercial
  satellites and the freeing up of global positioning systems (GPS) at the end
  of the 1990s, archaeologists have increasingly benefited from a combination
  of viable base maps and location-aware fieldwork (e.g. where surveyors use
devices that record where they are working in the landscape to a horizontal
spatial accuracy of at least ±5–10m). This has largely removed the awkward
positional uncertainties and paper mapping of previous decades. Earth
viewers (e.g. Google Earth), on-board handheld GPS base maps and
smartphone applications have become part of routine practice.
- Second, also worth noting is the sharply increased global geographic
  coverage of this imagery. That is partly been driven by the gradual
accumulation of satellite-borne datasets collected for non-archaeological
purposes (QuickBird, Ikonos, WorldView, Landsat, ASTER), but also by
enlightened re-dissemination of historical imagery (e.g. declassified
CORONA or wartime aerial photos; Cowley et al. 2010; Casana & Cothren 2013). Terrestrial geophysics can now also declare itself a truly landscape-scale set of methods. Towed or pushed multi-sensor systems can cover many hectares of ground quickly, especially across the world’s ploughed and open agricultural landscapes (Keay et al. 2009; White et al. 2013; Neubauer et al. 2014).

• Third, several of our airborne and ground-based techniques offer actively sensed data (where readings are taken based on a purpose-generated laser, radar pulse, electrical current, etc., rather than passively from an existing source such as electromagnetic reflectance from sunlight) and new kinds of 3D depth information. For example, the advent of ground-penetrating radar (GPR), plus theoretical and software advances allowing GPR profiles to be interpolated as 2D plan views or 3D image cubes, has made this the most volumetric and geo-archaeologically sophisticated of remote-sensing approaches (Sala & Linford 2012; Conyers et al. 2013; Herrmann 2013). At a larger scale, both synthetic aperture radar (SAR) and airborne laser-scanning (LiDAR) now offer excellent cloud-, (and in the latter case) woodland- or rainforest-penetrating ways to collect topographic information at both coarse and fine spatial scales (Devereux et al. 2008; Chase et al. 2011). This opens up some of the hitherto most poorly known parts of the world to rapid survey even if ground-truthing (e.g. for site date and function) will remain a key step.

• Fourth, we are about to become far more obsessed with deep spectral resolution. As imaging, and especially ‘hyper’-spectral imagery (sometimes also called imaging spectroscopy, where the electromagnetic spectrum is sampled more or less continuously rather than with gaps), becomes increasingly captured by unmanned aerial vehicles and plane-mounted systems. We are starting to be able to recover the kinds of spectral signatures for archaeological phenomena that have long been anticipated but largely unachievable due to the coarseness of much multispectral satellite coverage to date.

• A fifth aspect is that we can now undertake meaningful time series analysis, either by systematically collecting new diurnal, seasonal or inter-annual data (e.g. Poirier et al. 2013; Casana et al. 2014) or by fusing existing evidence in new ways (e.g. Menze & Ur 2012). As archives of historical aerial photographs and declassified Cold War imagery (see above) also come online, such analysis can be extended backwards to the early years of the twentieth century (albeit only for certain temporal windows, where serendipitous coverage permits, and typically only for qualitative visual comparison).

• Sixth, the last 10–15 years have seen plummeting costs in, and an overall democratisation of, remote sensing, at least for certain applications. As noted above, archaeological findspots can now often be located to within 5–10m by anyone with a handheld GPS or by viewing the locations on Google Earth. This enables less technically minded archaeologists and members of the public to contribute, and allows less well-resourced projects to develop modest remote-sensing programmes. The cost of high-resolution satellite imagery has dropped rapidly as the number of commercial platforms has diversified. Furthermore, structure-from-motion/multi-view stereo (SfM
and more broadly ‘computer vision’) methods provide an ultra-low-cost method for 3D data capture, merely by taking a sufficient number of overlapping images with a standard digital camera (Ducke et al. 2011; Green et al. 2014). Many of these low cost remote-sensing methods are also increasingly scale-agnostic in their application, with new opportunities to document and analyse not just sites and landscapes, but also artefacts (e.g. Bevan et al. 2014). At this ‘artefact-scale’, a further major advance has been the use of portable X-ray fluorescence (pXRF; Milić 2014) to achieve a rough impression of surface material composition. While it is worth sounding a note of caution about what SfM and pXRF can and cannot achieve (as well as stressing the continuing need for specialist input), both methods alleviate a whole series of traditional operational bottlenecks (e.g. needing to bring the object to the equipment), thereby allowing for much larger sampling programmes.

The above review offers a quick sense of the sheer pace of change in remote-sensing applications, both in situ and in the storeroom, but it would be a mistake to assume that such technologies were exclusively responsible for our current floods of archaeological information. For example, while the enormous threat posed to archaeology by modern construction is a story many decades old in certain Western countries (e.g. Webley et al. 2012), what has changed more recently is access to ‘grey literature’ reports from developer-led archaeological mitigation (Fulford & Holbrook 2011). This is in step with, but less widely known than, the massively expanded access provided by journal digitisation. As methods for digitising, searching, summarising and mining grey reports have advanced (Haselgrove & Moore 2007; Jeffrey et al. 2009; Vlachidis & Tudhope 2011), the sheer scope of fresh evidence that they represent has also become clear. In addition, while the gold standard of regional-scale digital inventories in the UK has long been the ‘Sites and Monuments Record’ (aka Historic Environment Record)—and these inventories continue to be important—we can now also point to a range of other synthetic georeferenced datasets for large areas, covering specific categories of archaeobotanical, metallurgical, zooarchaeological and landscape evidence (e.g. Bland 2005; Shennan & Conolly 2007; Conolly et al. 2012; Cooper & Green 2015). Increasingly, other old legacy datasets lurking in hard copy can be ‘crowdsourced’ via a huge group of interested volunteers (who each assist with individually small but collectively significant data collection tasks online, see below) and georeferenced via semi-automatic place-name look-up (‘geocoding’). Further categories of existing data are finding new and unanticipated onward uses: perhaps the most important (albeit still contentious) example is the treatment of tens of thousands of radiocarbon dates ‘as data’, that is, as a proxy for aggregate quantities of human activity and therefore potentially as evidence for highs and lows in human population across time and space (e.g. Shennan et al. 2013). Likewise, there are also vast quantities of genetic data (ancient and modern, from humans, plants, animals and vestigial in ‘dirt’) as well as multiple long-term climate proxies (lake cores, speleothems, tree-rings) that are increasingly being shared digitally and tuned to address questions of mainstream archaeological interest.

A final point to note, but in many ways perhaps the most important, is the fact that almost all of the above categories of information are increasingly released into a public commons under liberal data licenses that encourage rather than prohibit ‘downstream’ re-use (e.g. Creative Commons’ CC0, CC-BY). They are also increasingly the subject of peer-reviewed ‘data papers’ in publications such as Internet Archaeology, Journal of
Open Archaeology Data or Open Context, in which the strengths, weaknesses, biographies and re-use potential of this information is discussed (see also Huggett 2014; Kansa et al. 2014). This widening access and broader contextualisation is occurring not only ‘top-down’ with respect to the outputs from large-scale data collection exercises, but also ‘bottom-up’ by unlocking the potential of many smaller scale datasets that previously circulated in rather eccentric, narrow and conditional networks of colleague-to-colleague gifting (see below).

**Mining and mashing; cases and controls**

One obvious result of the data deluge is that, at least in certain parts of the world, we cannot in all good conscience claim “we don’t yet have enough evidence” or that we should “wait till the evidence is in”. The question then becomes how best to mine, mix and otherwise analyse a potential embarrassment of riches. It is not possible to do justice here to the full range of techniques that can be applied to such datasets, but some general points are worth making. First, there should be no shame in espousing empirical, often inductive, classification methods that continue to unpick the large-scale ‘systematics’ of this growing archaeological evidence. Even if this heralds, for some, an unwelcome return to culture-historical topics such as population-scale effects, continental-scale comparison and deep artefact traditions (Dunnell 1971; Roberts & Vander Linden 2011; Shennan 2011; Bevan 2014). In fact, there remains a clear role here for archaeology to contribute to more general debates about how best to collate, split or connect complex spatio-temporal datasets (Womble 1951; Barbujani et al. 1989; Huson & Bryant 2006).

Second, however dense it becomes, archaeological evidence will always remain patchy, with levels of uncertainty and variable expert opinions that are hugely challenging. In certain instances, the diluvian arrival of new data is a good reason to favour Bayesian inferential approaches that provide a formal framework both for expressing the strengths and weaknesses of existing knowledge, and then for incorporating new updates to it (Buck et al. 1996; Ortman et al. 2007; Fernández-López de Pablo & Barton 2015). Moreover, in addition to standard Bayesian methods, there is now also a range of further related approaches to data uncertainty (Beaumont 2010; Crema 2012; Bevan et al. 2013; Ducke 2014).

Third, we need to be clearer about analytical objectives. Is this growing body of evidence to be harnessed for ‘pure prediction’ or for wider behavioural, processual and interpretative insight? In certain heritage management situations, hyper-automated pure prediction may be both necessary and sufficient, for example to anticipate the extent of archaeological mitigation needed due to a proposed commercial development, to predict likely conservations risks to museum objects or to target coarse-grained public engagement strategies in cost-efficient ways. However, in the vast majority of archaeological situations, we do wish to interpret what the evidence is saying about past behaviours or processes. For example, we may wish to use archaeological settlement and field-system evidence not simply to anticipate where further examples of such archaeology might exist for protection purposes, but also to understand changing social logics influencing where people live in the landscape and how they organise it. It may also be of interest to know how these logics vary both regionally in the same time period and within the same region over different time periods. In such instances we need to: a) develop methods that are more transparent and grapple better with change detection, comparative distinctions and cause-effect wherever possible (bearing in mind the likelihood of equi- and multi-final outcomes, but not taking these as an excuse.
for never attempting formal analysis); b) recognise that both bottom-up agencies and patterning on a larger scale can be interesting; and c) balance more data-led treatments of our evidence with a continuing attention to behavioural theory and close, contextual archaeology.

More generally, faced with bewilderingly diverse but patchy sets of new archaeological information, we should be careful to avoid two kinds of fallacy: a) to presume that analytical confusion in the present necessarily implies the existence of behavioural complexity in the past; and b) to assume that real behavioural complexity in the lived past necessarily requires similar complexity in our present-day modelling. Put a different way, there is a risk that we explain away poorly constructed analyses and awkwardly complex analytical results by claiming they are due to a past reality that was behaviourally complex and irreducibly entangled. Even when we do think that a past situation was not straightforward, the fitting of complex models to complex data (just because we now can) will not necessarily lead to a more sophisticated understanding (e.g. Box 1976). Simplicity of approach remains a great virtue.

Finally, if we wish to make sense of the ever increasing, but still very uneven evidence, then we should follow the lead of subjects such as epidemiology or ecology and think far more in terms of cases-and-controls. This is one of the most robust approaches developed so far for evidence-based inference in situations where fresh, randomised experiments are impossible (Woodman & Woodward 2002). Typically, ‘cases’ refer to those observations that positively exhibit a property of interest (e.g. pottery of a certain type found at different sites) while ‘controls’ refer to those observations that do not (e.g. other types of pottery at the same sites). The idea can be further generalised and mapped as spatial surface via the notion of a continuous ‘relative risk’ of one kind of observation or measurement vs another (e.g. Kelsall & Diggle 1995; Hazelton & Davies 2009). An example from the spatial analysis of metal detecting data is the comparison of 2D kernel densities of one kind of metal find (e.g. Iron Age coins as the cases) with another (e.g. coins of other types as the controls) or with metal detector finds overall (Bevan 2012). More broadly, such compare-and-contrast approaches to archaeological data in high volumes are invariably much stronger strategies than single variable discussions, as recent work in multi-method terrestrial geophysics also makes clear (Kvamme et al. 2006). The rationale in all of these situations is what we might call ‘adversarial’, a term here borrowed very loosely from journalism (e.g. Hanitzsch 2007: 373) to place emphasis on the need to create and compare countervailing voices or datasets when faced with perceived biases in investigation. Hence the adversarial archaeological researcher seeks to mitigate investigative biases by juxtaposing one subset of evidence with the rest of a parent sample, with another distinct subset that has similar recovery conditions, or with other evidence known to be subject to quite different recovery conditions. Some archaeologists have long adopted this comparative, contrastive approach as a matter of course (see Smith & Peregrine 2012), but there are good reasons to make it more practically and theoretically explicit. In certain instances, this shift in emphasis from an obsession with one category of evidence to a strategic juxtaposition of several, may present unexpected challenges to current record-keeping practice. For example, if we are interested in modelling the locational preferences exhibited by Iron Age settlements recorded in a sites and monuments record, it may well be useful to recruit not just evidence about these Iron Age cases, but also about settlement locations in any other archaeological period in the study region, or alternatively all known archaeological findspots in the SMR. The latter two categories of ‘control’ give us a sharper way to handle biases in our record, rather than simply
assuming the existing data is representative. However, asking for all SMR records because you are interested in one category of evidence is not, in many operational contexts worldwide, something that is currently easy to do in terms of permissions, or currently supported by adequate financial resourcing at the level of the record-keeping body itself.

Citizens and scientists
As briefly mentioned above, archaeologists are also crossing a major watershed regarding how they disseminate information (Lake 2012), with a dramatically increased emphasis on open-access and third-party licensing in various flavours (traditional publications, software source code, raw data). A related opportunity is provided by our growing ability to collect and discuss archaeological evidence in more public-facing and participatory ways (e.g. Kansa et al. 2011; Bonacchi 2012). Archaeological ‘citizen science’ has a long history offline involving volunteer archaeological and historical societies (especially, but not exclusively, in the UK), academic community collaborations (e.g. most recently, The Survey of Hillforts n.d.) and its fair share of contentious encounters (e.g. over metal detecting). But in many ways both the pros (predominantly) and cons (more rarely in my view) are now being turbo-charged by online coordination and modern digital documentation. For example, perhaps the earliest and most enduring example of massive online crowd-sourced documentation in archaeology is the UK-based Megalithic Portal (n.d. in operation since 2001). Over the last decade, vocational archaeologists have greeted it with an interesting mixture of enthusiasm (e.g. as a huge, community-driven database) and dismissiveness (e.g. about perceived poor data quality or about its inclusion of fringe archaeologies). Overall I would suggest that many or most would now consider it to be an extremely successful venture. Regardless and put simply, the archaeological record is a massive, spatially scattered, constantly threatened, and rapidly dwindling resource, and as a subject we have traditionally only received the support of rather small amounts of money. So there are compelling reasons to share responsibility for data collection, and where possible analysis and interpretation, beyond a group of traditional specialists (see also Bevan et al. 2014). Overall, ‘volunteer sensing’ is both a trend we cannot avoid (given metal detectors and earth viewers) and one we should be enthusiastic about (in certain circumstances and if conducted responsibly). There are many recent examples of effective collaboration, including the UK Portable Antiquities Scheme (n.d.), community-based geophysics projects (Geophysics and the Landscape of Hertfordshire n.d.) and SfM-led 3D modelling of archaeological features (MicroPasts n.d.; HeritageTogether n.d.; ACCORD n.d.).

A more pernicious feature of the current growth phase in digital archaeological information is the risk that it will make us less, not more, equal as researchers. For example, it is the historically wealthy, computer-savvy, ploughzone-rich and Metallschock-ed parts of the world that are arguably best placed to collect high-quality remote-sensing datasets. Even when other parts of the world have useful coverages of this kind (e.g. because they have been under surveillance in military conflicts), it is often Western institutions that retain best access to such information, and it is often researchers who have grown-up (educationally speaking) in data-, resource- and technique-rich Western archaeological settings who will be academically best placed to exploit it. Likewise, the subsidising of open access arguably raises as many problems as it solves, opening up research to a wider group but often involving premiums paid to traditional publishers by wealthy institutions for super-charged academic exposure. Hence, while remaining optimistic overall about the impact of the data deluge, there are
good grounds for continuing concern about how such flows might recapitulate wider economic, political and knowledge asymmetries.

Final thoughts
In a sense, the metaphor of a single deluge of archaeological data implies too clear-cut a before-and-after moment, when in fact we will have to get used to chronic annual flooding. Regardless, the operating conditions for archaeological research are being transformed. For example, even if it will always remain crucial for archaeologists to frame arguments clearly, narrate them in an attractive way, contextualise closely and build theory, the need for more specific computational and numerical skills is also likely to grow further in importance alongside the growing complexity of our evidence and the scope of our explanatory ambitions (for a recent perspective on the latter, see Kintigh et al. 2014). More generally, the way we understand ‘reproducible research’ is likely to change (Stodden et al. 2014; rOpenSci n.d.) and require us to revisit which stages of the research process are given due credit and transparency as academic outcomes. Should it remain just the summary results of a research project that are so treated, or should data papers, collection protocols and analytical software also be peer-reviewed, published and given due intellectual credit? The answer to the latter questions is almost certainly yes, but the mechanisms by which this is done remain very new and not yet widely known. More generally, there is a pressing need to demonstrate that our information glut does indeed lead to better (rather than just more) archaeology and part of that challenge is related to the trade-off between archaeological empiricism and broader knowledge construction (to force a slightly unfair distinction). Matthew Johnson’s concern (2011) that archaeologists are prone to disciplinary episodes of hyper-empiricism is fair: the risk is that we will now enter a rather undirected phase of gathering ever greater amounts of evidence, whilst assuming that it otherwise largely speaks for itself. In many instances, a more adversarial analytical approach is called for, in which we harness increased flows of archaeological information to set up explicit forms of comparison-and-contrast. Ultimately, whilst we should avoid undermining the kinds of detailed, hard-won specialist knowledge that remain key to a proper understanding of the archaeological record, this comparative effort will also need to cross some persistent sub-disciplinary barricades, and work ‘ultra-longitudinally’ across different chronological periods; periods that are usually treated separately in the same study region, ‘trans-geographically’ across traditional regional or modern political borders, and ‘counter-artefactually’ across constellations rather than single classes of material culture.

References
http://dx.doi.org/10.2307/2992403
http://dx.doi.org/10.1146/annurev-ecolsys-102209-144621


UK Portable Antiquities Scheme n.d. Available at: https://finds.org.uk/ (accessed 29 April 2015).


Figure 1. Archaeological evidence in flood, top left to bottom right: people’s incidental spreadsheets as digital natives; finds from the UK Portable Antiquities Scheme; 3D modelling via structure-from-motion; digital ‘grey literature’ from developer-led archaeology; large-scale community-led geophysics; vehicle-towed GPR; radiocarbon dates ‘as data’; crowd-sourced archival transcription; ancient and modern DNA; georeferencing via earth viewers; open data; and digital repositories.