Jeremy Huggett

2 Digital Haystacks: Open Data and the Transformation of Archaeological Knowledge

> "There is a great need for theorization precisely when emerging configurations of data might seem to make concepts superfluous to underscore that there is no Archimedean point of pure data outside conceptual worlds. Data always has theoretical enframings that are its condition of making …"(Boellstorff, 2013).

2.1 Introduction

Since the mid-1990s the development of online access to archaeological information has been revolutionary. Easy availability of data has changed the starting point for archaeological enquiry and the openness, quantity, range and scope of online digital data has long since passed a tipping point when online access became useful, even essential. However, this transformative access to archaeological data has not itself been examined in a critical manner. Access is good, exploitation is an essential component of preservation, openness is desirable, comparability is a requirement, but what are the implications for archaeological research of this flow – some would say deluge – of information? Lucas has recently pointed to the way archaeological reality can change as a consequence of intervention: as archaeologists change their mode of intervention so reality shifts and interpretations change (Lucas, 2012, p. 216). If this is true of archaeological practice, to what extent might the change in our relationship to data – the move from traditional modes of creation and access to digitally-enhanced methods – represent a potential paradigm shift in our archaeological reality, or place limits on future changes? As more data are 'born digital' with access to them open to an increasingly wide audience, is it realistic to assume that archaeological knowledge itself remains unchanged in the process? How does our relationship with archaeological data change as the observations, measurements, uncertainties, ambiguities, interpretations and values encapsulated within our datasets are increasingly subject to scrutiny, comparison, and re-use? What are the implications of increasing access to increasing quantities of data drawn from different sources which are more or less open, more or less standardised, and increasingly reliant on search tools with greater degrees of automation and linkage? Given the fundamental – and frequently contested – nature of archaeological data, it is surprising that the implications of open access to those data remain largely uncontested. Instead, archaeology's digital haystack repre-

Jeremy Huggett: University of Glasgow, Glasgow, UK

sents a largely unexplored set of practices mixing old and new in the creation of new infrastructures which transform the packaging, presentation, and analysis of the past. Examining this entails revisiting the notion of the 'archaeological record' within the context of the new technological frameworks, and considering the consequences of this digital data intervention.

2.2 Openness and Access

Open archaeology has been a concept receiving increasing attention in recent years, most evidently in an issue of World Archaeology which sought to extend awareness of the implications of open approaches to a wider archaeological audience (Lake, 2012, p. 471). As Lake observes, and as reflected in that issue and this volume, openness can cover the use and reuse of software, publications, creative works, and data, although within the archaeological debate attention has until recently focussed extensively, though not exclusively, on publication.

The most common starting point for considering 'openness' is the Open Definition: "A piece of data or content is open if anyone is free to use, reuse, and redistribute it – subject only, at most, to the requirement to attribute and/or share-alike." (Open Definition, 2014). Archaeology may seem to be well-served with free access to archaeological data via organisations such as the Archaeology Data Service in the UK, tDAR and Open Context (USA), DANS (Netherlands), as well as national heritage organisations (for example, Royal Commission on the Ancient and Historical Monuments of Scotland, English Heritage) and regional Historic Environment Records. However, with some exceptions, much of this data is only partially 'open', leaving Kansa to suggest that openness remains largely at the margins of archaeological practice (2012, p. 499). In part, this is a consequence of distinctions between different levels of 'open access data' and 'open data'. For example, a hierarchy can be defined in increasing order of 'openness':

1. Open access data which provides online access to view datasets, limited only by a presumption of Internet access and the requirement for a modern web browser. Use of the data beyond viewing and searching online is restricted (commonly seen with most Historic Environment Records, National Monuments data and including commercial organisations such as CyArk etc.). A variant of this approach enables a map to be created on demand within desktop GIS software. This generally entails access to Web Mapping Services (WMS) which provide a graphical image as output, with limited functionality beyond the image itself. These are typically available for National Monuments data accessed via open government websites such as data.gov.uk.
2. Open access data which returns summary geographical information as a downloadable output of a search query or via Web Feature Services (WFS). This can then be further analysed using GIS software as if the data were held locally. For

example, the Archaeology Data Service's ArchSearch has download functionality for registered users, and Historic Scotland/RCAHMS's PastMap similarly enables summary location data to be accessed via downloadable comma-separated values files. Currently most WFS feeds in archaeology are used internally within organisations, or to create interoperable services from multiple feeds (resources such as PastMap itself, and Scotland's Places) but are not accessible more widely (for example, McKeague et al. (2012)). Leaving technical issues aside, in part this seems to arise out of concern to limit bulk downloads of data: hence downloads from ArchSearch or PastMap are restricted to one or two hundred records at a time, for example.

3. Open access data consisting of entire datasets which can be downloaded but where restrictions apply to the use and reuse of data and hence is not truly open data in the technical sense. For example, the Archaeology Data Service Common Access Agreement (Archaeological Data Service, n.d.) specifies that the data should only be used for teaching, learning, and research purposes, although the definition of 'research' is drawn very broadly such that it includes commercial funding, and the primary condition is that the results are placed in the public domain. In other cases, the restriction is more of a 'health-warning': for instance, the PastMap terms and conditions specify that the data provided is intended for information only and that professional advice should be sought to properly interpret it, emphasising the need to understand its limitations (PastMap, 2013). On the other hand, English Heritage's Heritage Gateway applies strict copyright restrictions to data accessed and downloaded from the site (Heritage Gateway, 2007).
4. Open data which has no exclusions or restrictions on use, and conforms to the Open Definition or the most permissive Creative Commons licenses. In general these datasets relate to specific projects, sites, or collections. For example, in the United States both Open Context and tDAR organisations use the Creative Commons CC-BY licence which enables the data to be shared and reworked, simply requiring attribution or citation of the original work. As Kansa points out, certain datasets within the Archaeology Data Service collections are now also governed by the CC-BY license rather than the standard terms and conditions (Kansa, 2012, p. 507).

Much archaeological data therefore is not truly 'open', and recent papers on open data in archaeology tend to focus on the desirability of increasing openness and the restrictions and impediments to achieving it (for example, Beale 2012; Beck and Neylon 2012; Bevan 2012*b*; Kansa 2012). These are not new issues: for example, in a discussion of copyright and archaeological data in 1997 Carson asked: "Who owns the right to reproduce raw data? Who owns the right to publish a manipulated version of that data? And who owns the right to produce second-generation items, such as models, from that data?" (Carson, 1996, p. 291). The ethical responsibility of archaeologists to make their data available is frequently cited: for example, Carson argues that:

> "Archaeologists, like other scientists, have an ethical obligation to publish, and to allow others to critique, their findings. Publishing data sets in machine-readable form is the ultimate expression of this obligation, in that others are free to analyze the basis of an archaeologist's findings and come to their own conclusions." (Carson, 1996, p. 316).

Kansa puts the case more strongly, arguing that "the discipline should not continue to tolerate the personal, self-aggrandizing appropriation of cultural heritage that comes with data hoarding" (2012, p. 507) and goes on to say:

> "Failure to incentivize greater data transparency would demonstrate an egregious failure of leadership and utter dysfunction in a discipline supposedly devoted toward building and preserving knowledge of the past." (2012, p. 507).

Most professional archaeology codes of practice emphasise this link between the stewardship of the past and the requirement to report and publish and to preserve the records made, including computer data. For example, the Institute for Archaeologists in the UK specifies that the results of archaeological work should be made available with reasonable dispatch (Institute for Archaeologists, 2013, Principle 4) and establishes that this includes the analysis and publication of data (Institute for Archaeologists, 2013, 4.4). In the light of this it would be tempting to ask why more open data is not available. One reason may be that the ethical codes emphasise that rights of primacy exist: in the case of both the IfA and the European Association of Archaeologists this persists for up to ten years (Institute for Archaeologists 2013, 4.4; European Association of Archaeologists (1997, 2.7)), although the Archaeological Institute of America, the Society for American Archaeology, and the Canadian Archaeological Association, for example, only specify the need to make results available in a timely fashion and to make evidence available to others within a reasonable time (of America 2008, I.4; Society for American Archaeology (1996, 5); Canadian Archaeological Association (n.d.)). Consequently rights of primacy may restrict access to data and, without enforcement, the timescales specified may be stretched: indeed, there is a long and unfortunate history of archaeological archive data being retained by an individual for a lifetime. In such a context, Kansa's expostulation is understandable.

One issue regularly raised in relation to open archaeological data is that they frequently include spatial information which might facilitate looting (for example, Bevan 2012*b*, p. 7–8; Kansa 2012, p. 508–509). Degrading the quality of spatial data and making full resolution data available only to 'approved' users are approaches that have been adopted, but restricting access like this flies in the face of open data requirements. Other common arguments about the limits to open data relate to authority and the risk of reducing confidence 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 lo-

cational information are likely to be more problematic for would-be users than the occasional rogue item.

2.3 Openness and Reuse

In the light of the pressures for access to open data it is perhaps worth emphasising that there has been no empirical study of the demand for open data in archaeology. This means that, to a large extent, the level of demand remains undemonstrated and unquantified. However, a recent study of the Archaeology Data Service sought to evaluate and quantify the 'value' of online access to data (Beagrie and Houghton, 2013). It employs a range of approaches to assessing value: for example, investment value (amount invested in the services), use value (amount spent by users to access the service), contingent value (for instance, how much people would be willing to pay). In combination these give rise to the net economic value (the difference between the willingness to pay and the cost of obtaining the service minus the investment value) (Beagrie and Houghton, 2013, figure 4.1). On this basis, the investment value of the Archaeology Data Service was calculated to be about £1.2m per annum, made up of £698,000 from funders or sponsors and around £465,000 indirectly contributed by depositors (Beagrie and Houghton, 2013, p. 35). Direct use value to the user community was estimated to be about £1.4m per annum (Beagrie and Houghton, 2013, p. 35) but the efficiency impacts were estimated to be anywhere between £13m and £58m per annum (Beagrie and Houghton, 2013, p. 40). Research efficiency gains were equivalent to around 7 hours per week as a consequence of access to ADS data (Beagrie and Houghton, 2013, p. 39). Interestingly, there were objections to the survey's use of questions about willingness to pay for the service and how much people would be willing to accept in return for giving up the service, and 6-9% of respondents refused to estimate this, arguing that access and data should be free (Beagrie and Houghton, 2013, p. 36–37). The results show that the value of access to data is considerable – however, as with everything 'open', the challenge is to make openness sustainable financially.

The extent to which open access data is actually used also remains largely unquantified. Ironically, access to data about access to open archaeological data is often not directly accessible; however the Archaeology Data Service website provides statistics for a variety of metrics and, as one of the longest-established providers of a broad range of archaeological data, could reasonably be viewed as representative. Web metrics are notoriously difficult to disentangle and interpret, but the evidence suggests a surprisingly high number of downloads relative to visits to the site (Figure 2.1). Much of this relates to downloads of PDF files from the large collections of unpublished grey literature reports and back-issues of journals and other volumes (Green pers comm – Figure 2.2), rather than downloads of specific datasets.

The Archaeology Data Service download statistics do not differentiate between PDF and other file types, so estimating usage of datasets is not straightforward. How-

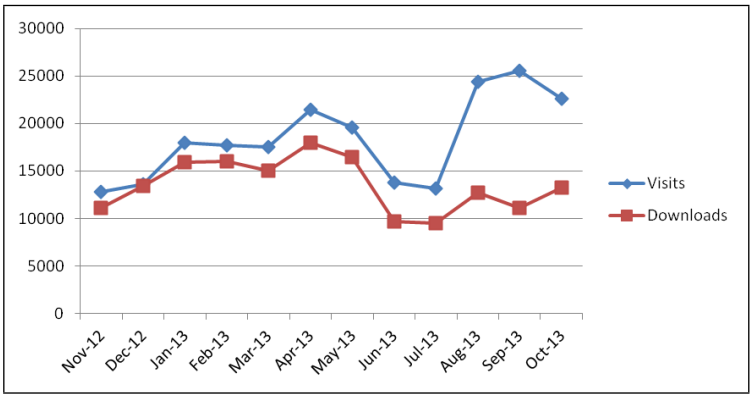

Figure 2.1: Archaeology Data Service access statistics

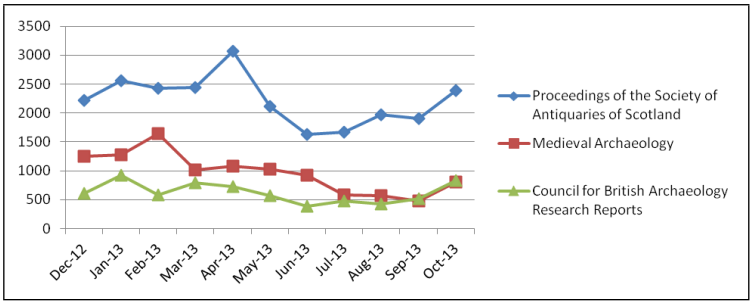

Figure 2.2: ADS access statistics - examples of PDF downloads

ever, Figure 2.3 provides an approximate comparison to Figure 2.2 based on examples of field projects which include downloadable data, simply to demonstrate the order of magnitude difference between field data downloads and PDF downloads. The reason for this difference may be simply that the majority of PDFs relate to free access to back issues of journals and volumes that would otherwise require subscription or purchase, or access to grey literature about excavated sites that would be costly in time and effort to acquire otherwise (for example, Bradley 2006, 7–8), while the field datasets require a very specific level of interest and, to some extent, expertise. Clearly, there is much more to be gained from a deeper and more nuanced analysis of these kinds of access data.

Issues with open data (and non-open data, for that matter) really come to the fore only when those data are put to analytical use. Detailed accounts of data reuse are as yet rare, and those reports there are tend to stress the positive outcomes and minimise the efforts entailed in achieving them. For example, Bevan (2012*a*) demonstrates the potential benefits from the examination of several large scale georeferenced inventories and how built-in data biases might be overcome, but apart from reference to an

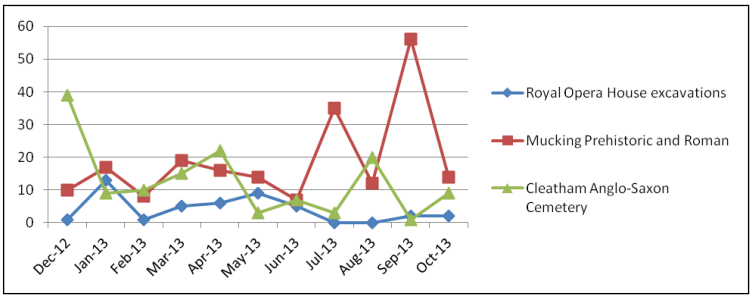

Figure 2.3: ADS access statistics - examples of fieldwork data downloads

"intensive effort of cross-checking and problem-flagging" (2012*a*, p. 493) there is no information provided about any data-cleansing and manipulation that may have been required in advance of analysis. An earlier study using some of the same data provides a clearer indication of the kind of work that can be required to make data usable. The Viking and Anglo-Saxon Landscape and Economy of England (VASLE) project combined data from the Portable Antiquities Scheme database with the Early Medieval Corpus of Coin Finds and an extensive data cleansing exercise was required (Naylor and Richards, 2005; Richards et al., 2008, 2009) to resolve issues of comparability, compatibility, and standardisation of classifications across the two datasets. For example, many dates in the Portable Antiquities Scheme database were only recorded at a generic level, while different recorders classified the same kinds of artefacts under different headings. The level of effort entailed underlines not so much the complexity of the data but the complexity of the task. Unsurprisingly, the researchers concluded that

> "Re-use of data requires a close understanding of the context of data collection and of the vocabulary used to describe the observations. The archaeologist of tomorrow needs training not so much in methods of data collection, but in data analysis and re-use." (Naylor and Richards, 2005, p. 90).

Similar conclusions are reached in a recent study which interviewed a sample of archaeologists about their experience of reusing data and reported that the lack of context was a persistent problem (Faniel et al., 2013). This arose for a variety of reasons, including the variability of archaeologists and their recording procedures which ranged from the meticulous to the careless (Faniel et al., 2013, p. 298). As a consequence, they identify a series of gaps in current archaeological data standards such as the need to capture the range of methodological procedures undertaken during excavation or survey, including specifications of instruments, information about how the data were collected, the strategy adopted, etc. (Faniel et al., 2013, p. 302). However, despite the problems encountered, they note that archaeologists still reused data, sometimes finding alternative means of recovering context – or, presumably, either making

assumptions about context or ignoring it altogether. Archaeologists are not unique in this respect. For example, an examination of the reuse of open government data drew attention to the lack of contextual metadata, poor documentation, variability of data quality, conflicting data definitions, and a host of other practical impediments to reuse (Zuiderwijk et al., 2012, p. 162–164). What this highlights is one of the key paradoxes that lies behind open data: increasing access to increasing amounts of data has to be set against greater distance from that data and a growing disconnect between the data and knowledge about that data (Huggett, Forthcoming).

2.4 Approaches to Open Data

One of the problems of open data is that archaeologists are only just starting to consider the issues surrounding open access to archaeological data. Most discussions focus on the desirability of openness, the ethical responsibility to be open, and what benefits might accrue from open access to data for both archaeology, archaeologists, and the wider community. The very diversity of archaeology – its coverage, scope, quantity and range of data sources, multiplicity of practices, and limited standardisation – is often seen as an attraction for e-science studies (e.g. Faniel et al. 2013, p. 295–296; Jeffrey et al. 2009, p. 2515; Richards et al. 2011, p. 42), but the technological responses to this diversity tend to focus on deconstructing archaeological information into semantic structures as a means of managing and controlling the data, a process which itself is not without issues (Huggett, 2012). However, this diversity of archaeological data is not what makes them really distinctive: what is particularly characteristic about archaeological data is their time dimension (what Arbesman 2013) has termed 'long data' in contrast to 'big data') and the peculiarly destructive nature of much of their collection methodology. Individually, neither is especially unique – geology deals with especially 'long data', for instance – but in combination, it makes for an especially challenging prospect for open data. This is because of the conceptual approach to open data and its subsequent reuse.

For example, in a recent definition of what constitutes archaeological open data, Anichini and Gattiglia (2012, p. 54) follow the Italian Association for Open Government (Belisario et al., 2011, p. 11–12) in defining archaeological digital open data as being complete (capable of being exported, used, integrated and aggregated with other data, and including information about their creation), primary ('raw' data capable of integration with other data), timely (available), accessible (free, subject only to costs associated with Internet access), machine-readable (capable of automated processing), non-proprietary (free from licenses that limit their access, use or reuse), reusable, searchable (through catalogues and search engines), and permanent. Unsurprisingly, these do not greatly differ from other open data definitions such as that provided by Open Definition (n.d.).

Although such a characterisation may seem fairly uncontroversial, the concept of

the completeness and primacy of the data is problematic from an archaeological perspective since it loses sight of what these data actually are. Completeness and availability may imply that the data are finished and ready for reuse. Beale's reminder that "data must first be prepared and then care taken to identify the moment when we are no longer preparing them for release, but are in fact working on them" (2012, p. 623) distinguishes between data preparation and subsequent analysis but in the process implies the existence of a form of basic un-worked data – 'raw' data in Anichini and Gattiglia's characterisation – which is seen as providing the building blocks for archaeological knowledge. Bevan (2012*b*, p. 6) is suspicious of the use of 'raw data' as a term for something which has a clear interpretative component, but sees the actual problems for open data lying at a higher level with its spatial content (2012*b*, p. 7).

Kansa et al. observe that primary archaeological data have received little theoretical attention while recognising their importance in the production of archaeological knowledge (2010, p. 303). Although it is true that primary data have not been critically discussed within an archaeological informatics context, the nature of archaeological data has been a focus of much debate over the years, recognising that these data are situated, contingent, incomplete, and theory-laden (for example, Patrik 1985; Binford 1987; Barrett 1988; Hodder 1999; Chippindale 2000; Lucas 2001; Lucas 2012). An exception in this regard is Llobera's discussion of data within the context of defining the basis of an Archaeological Information Science (Llobera, 2011), although much of his concern lies with data representation and data structures:

> "…the topic of data representation within archaeology has not received as much attention as it should, especially in the light of the pivotal role it has in the production of archaeological knowledge and its potential to precipitate different interpretations. The consequences of this oversight become deeper and more far-reaching the moment information systems are adopted. It is all too easy for the user to forget that he/she is subscribing to a particular form of data representation." (Llobera, 2011, p. 213–214).

Llobera suggests that archaeologists have generally been concerned with the choice of which data to collect based on prior research questions, rather than the form in which the data are collected (Llobera, 2011, p. 214). Although he recognises that recording data is subject to the theoretical orientation and the goals of the researcher, he argues that the data structures used to contain these archaeological observations are not themselves interpretative and hence:

> "The fact that they organize observations explicitly and that their manipulation is done via a set of operations defined a priori provides transparency and flexibility. Indeed, it is the marriage between data and purpose that make them so powerful and appealing." (Llobera, 2011, p. 215).

Although Llobera's focus on the significance of data structures and their ability to support new forms of archaeological investigation is important, it largely sidelines the origins and nature of data themselves: they become 'reasoning artefacts' that contribute to analysis and interpretation (Llobera, 2011, p. 214). How the data are structured is

without doubt crucial to their analysis and any subsequent reuse. However, the significance of the data themselves is equally profound, if not more so.

2.5 From Data to Knowledge?

Within the field of archaeological informatics, responses to issues raised concerning data tend to emphasise structural and organisational approaches and solutions – while the data may be recognised as essentially interpretive, the implications of this are generally left for others to deal with. Consequently a term like 'raw data' is frequently used without reflection and indeed, the term 'data' itself is often open to confusion. In the last things seemed much simpler. For example, Trigger (1998, p. 3) identifies Glyn Daniel, Stuart Piggott and Christopher Hawkes as drawing a clear distinction between facts and interpretations – archaeological data were facts and constituted the core of the discipline, while interpretations were transient and changing. Accordingly the archaeological record was seen to become 'better' as a result of the collection of more data and the development of better techniques for interpreting these data (Trigger, 1998, p. 22). A similar view is held within Information Systems studies, where data are often seen as facts – the raw materials captured within data structures for creating information (for example, Räsänen and Nyce 2013, p. 656), and in the context of 'big data' large datasets are seen increasingly as providing significant opportunities to create new knowledge. In much the same way, the knowledge management industry is predicated on refining data into knowledge (for example, Tuomi 1999, p. 103; Weinberger 2011, p. 2–3).

Superficially, data are not complex. For example, the Royal Society recently defined data as "Numbers, characters or images that designate an attribute of a phenomenon", and as

> "Qualitative or quantitative statements or numbers that are (or assumed to be) factual. Data may be raw or primary data (e.g. direct from measurement), or derivative of primary data, but are not yet the product of analysis or interpretation other than calculation." (Royal Society, 2012, p. 12).

However, this immediately introduces two types of data – 'raw' and 'derived' – and a corresponding contradiction: on the one hand derived data are calculated from other data (for example, average rainfall); on the other hand, calculated data are seen as information (for example, the numbers generated by a survey instrument are data used to calculate the height of a feature which is classified as information) (Royal Society, 2012, p. 14). Not surprisingly, the Report admits that there is sometimes confusion, with data, information, and knowledge being used as overlapping concepts.

One outcome of this more-or-less commonsense technical approach to data is a view of data as sitting at the bottom of a hierarchy which moves from data through

information to knowledge (and in some models, to wisdom beyond that). This data-information-knowledge (-wisdom) pyramid (for example, Weinberger 2011, p. 1–5) essentially sees the acquisition of knowledge (or wisdom) as constructed from a series of building blocks: data are used to create information, information combined to generate knowledge. For instance,

> "Data are seen as raw materials for information. Data become information when it is structured and arranged in a particular context or relations set. Information is talked about as though it has a meaning, but no (appended) judgments. It is commonly thought that knowledge contains meaning and judgment and beliefs and commitment regarding a particular action." (Räsänen and Nyce, 2013, p. 659).

From an archaeological perspective, Darvill has expressed concern that such a structure is destabilised by the generation of vast amounts of archaeological data which remain to be turned into information or knowledge (Darvill, 2007, p. 445): an archaeological digital data mountain which increasingly we struggle to deal with (Huggett, 2000, p. 15–16) but which in a world of 'big data' appears much more amenable. When presented with access to these large quantities of data, the data-information-knowledge approach seems self-evident: we are faced with data which we seek to make sense of and ultimately use to draw conclusions about aspects of the past. This is one of the key benefits identified for open data – the provision of access to fundamental building blocks which will enable us to create new knowledge which would otherwise be much harder – or impossible – to do.

However, this outwardly logical approach disguises a hidden technological agenda: as Weinberger observes, this image of knowledge creation as a pyramid with increasingly fine filters being applied at each level is associated with an Information Age "which has been all about filtering noise, reducing the flow to what is clean, clear and manageable. Knowledge is more creative, messier, harder won, and far more discontinuous" (Weinberger, 2010). One might equally add that information and data are just as messy and creative in nature.

The issue lies with the fundamental nature of data. For example, Borgman points out that data carry very little information in and of themselves: "Data are subject to interpretation; their status as facts or evidence is determined by the people who produce, manage, and use those data." (Borgman, 2007, p. 121). Data have no value – indeed, data do not exist – without some degree of interpretation. In archaeological terms, data are contemporary observations about attributes we consider to have some value in understanding past activities – they are the result of the archaeologist's judgements at the time as to what might be worthy of recording: "all archaeological data are generated by us in our terms" (Binford, 1987, p. 393). The kind of data collected from a given assemblage will vary between individuals depending on a variety of factors including recovery methods and research questions (for example, Atici et al. 2013, p. 665). A perspective of data as 'raw' in the sense of being uncontaminated by methodological and theoretical biases and therefore more likely to result in an accu-

rate outcome (Carson, 1996, p. 316) is therefore a simplistic view of what constitutes data. Indeed, some would claim that 'data' is a misleading word to use in the first place. Both Chippindale (2000) and Drucker (2011) have independently argued that 'capta' is a more appropriate term. Drucker emphasises that 'capta' are taken actively, whereas data are assumed to be a given that can be recorded and observed (Drucker, 2011, p. 3) . Chippindale proposes that data as 'capta' are "things we have ventured forth in search of and captured", with all the associated connotations of hunting and gathering, danger, uncertainty and risk (Chippindale, 2000, p. 605). Both emphasise the creative aspect of data (or capta), that they are partial, selective, and change through time. Data/capta rely on prior knowledge and experience: to capture data requires recognition, identification, and classification in order to be recorded in the first place. Additionally, data may not be easily described and hence receive a decreasing amount of attention, they may not break up into natural units so are highly dependent on the level of analysis applied at the time, and they may not be considered worthy of recognition or capable of capture (Bowker, 2005, p. 141–144). As a result,

> "…we are producing a set of models of the world that – despite its avowed historicity – is constraining us generally to converge on descriptions of the world in terms of repeatable entities, not because the world is so, but because this is the nature of our manipulable data structures" (Bowker, 2005, p. 146).

Data and datasets are therefore of their place and time: they are constrained by the conditions of their creation, all the more so if the question of when data become data is considered. As Borgman points out, in some circumstances data may not be considered to be data until they are cleaned and verified – and how much cleaning and verification is required before they are considered usable data is a question of judgement (Borgman, 2007, p. 183). This is a constant issue for digital archives: the distinction between processed and unprocessed data, and how much processing is 'enough'. So what one person considers data might not be recognised as such by another, in terms both of what is captured and what is not, as well as the extent to which it has been processed. As an example of the problem, Chippendale cites the case of recording rock art where effort went into removing the natural elements from the data, overlooking that the natural features may have been an integral aspect of the art which subsequently required the works to be re-recorded (Chippindale, 2000, p. 608). Of course, the rock art was still there to be re-recorded, which cannot be said for the objects of much archaeological data.

2.6 From Knowledge to Data?

The simple perception of data as the base constituents for the construction of information and knowledge may seem attractive and logical when faced with a technological

infrastructure consisting of large quantities of data, but it misrepresents the situation and as a result reuse risks misuse. Making sense of data in computer systems is not a straightforward process:

> "Someone has articulated knowledge using languages and conceptual systems available and – in the case of a computer database – represented the articulated knowledge using a predefined conceptual schema. Someone else then accesses these data and tries to recover their potential meaning." (Tuomi, 1999, p. 111).

In order to make sense, the data-information-knowledge model should actually be reversed (for example, Knox 2007; Tuomi 1999) such that data are seen to emerge only as a consequence of knowledge and information; in other words, data come into existence in the first place through human engagement. This is all the more true in the context of digital data: "Data can emerge only if a meaning structure, or semantics, is first fixed and then used to represent information" (Tuomi, 1999, p. 107). Tuomi argues that knowledge has to be articulated in order to become information which can be represented; in order for it to be represented in a digital environment, information needs to be broken down into atomic elements, or data (Tuomi, 1999, p. 107) – a situation familiar to anyone who has constructed a database from scratch. The problem here is that the knowledge that is articulated and atomised is by definition explicit and more easily represented and communicated than contextual tacit knowledge. Tacit knowledge is more easily displayed or exemplified as practice rather than transmitted (Duguid, 2005, p. 113) and therefore tends to be more or less invisible in a digital data environment. As Borgman observes, "The effort required to explain one's research records adequately increases as a function of the distance between data originators and users" (2007, p. 167). The data are therefore accessed in a largely de-contextualised state, and the increasing development of automated processing techniques associated with 'big data' exacerbates this situation still further.

As far as the data user is concerned, making sense of the data relies to a considerable extent on their own tacit knowledge and – as Tuomi emphasises – ultimately requires trust in the data originator, since the data-information-knowledge of the end user only emerges as a consequence of their understanding of the knowledge-information-data disarticulation by the original creator ((Tuomi, 1999, p. 112). If, as Gramsch argues, we also need to be able to scrutinise what the data might reveal beyond the originator's intentions (Gramsch, 2011, p. 62), the significance of knowledge about the whole data lifecycle, including the original knowledge-information-data process, the circumstances of collection, and the contextual and tacit information associated with it, becomes greater still. The alternative risks data being wrenched from context, arguments separated from evidence, interpretations transformed into 'facts', explicit knowledge separated from tacit knowledge, and, in the context of digital data processing, push-button solutions substituted for knowledgeable actions (Huggett, 2004*a*,*b*).

The concern, therefore, is that the combination of access to data and distance from understanding the nature of those data in many respects reinforces Postman's prediction, that:

> "…the tie between information and human purpose has been severed, i.e., information appears indiscriminately, directed at no one in particular, in enormous volume and at high speeds, and disconnected from theory, meaning, or purpose." (Postman, 1993, p. 70).

This is all the more prescient given the development of 'big data' and Chris Anderson's famous claim that the new 'Petabyte Age' :

> "…calls for an entirely different approach, one that requires us to lose the tether of data as something that can be visualized in its totality. It forces us to view data mathematically first and establish a context for it later … We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot." (Anderson, 2008).

Delivering data in increasingly large amounts but without accompanying awareness about the theories, purposes and processes which lie behind those data means that the data arrive at the would-be user contextless and consequently open to misunderstanding, misconception, misapplication, and misinterpretation.

2.7 Putting the 'Capta' Back into Data?

The expansion in access to increasing volumes of open archaeological data in many respects presages the arrival of a new archaeological 'record'. In 2005, for example, Naylor and Richards predicted that researchers will be increasingly expected to use existing data and will need to justify primary data collection in future (2005, p. 90). More recently, Beck and Neylon suggested that access to dynamic open archaeology data may question the orthodoxy of excavation (2012, p. 494). The risk identified here is that we may get caught up in this brave new technological world of data and lose sight of the underlying issues in the thrill of enhanced access. For instance, Gitelman and Jackson warn that a shared sense of starting with the data

> "…often leads to an unnoticed assumption that data are transparent, that information is self-evident, the fundamental stuff of truth itself. If we're not careful, in other words, our zeal for more and more data can become a faith in their neutrality and autonomy, their objectivity." (Gitelman, 2013, p. 2–3).

Archaeological debates about open data may not fall into this trap and certainly cannot be characterised as excessively utopian in outlook. However, focussing on structures and organisation rather than the data, emphasising their access and delivery,

pays relatively less attention to what the access is to, what the delivery is of, and what the consequences of such access and delivery might be.

Lucas characterises archaeological intervention and the consequent creation of a record as a combination of re-materialisation and de-materialisation: re-materialisation in the sense of creating new interpretative objects from the old (sherds, flakes etc.) and the new (photographs, drawings, descriptions etc.), and de-materialisation in the conversion of the physical (excavation trench) into the paper records, photographs, finds and so on (Lucas, 2012, p. 258–259). This is reminiscent of the classic view of information technology as bringing about a shift from atoms (the material world) to bits (the digital world) (Negroponte, 1996, p. 11ff). The introduction of a digital dimension to the archaeological record can be seen as an additional level of de-materialisation, further removing the original objects of record from the interpretative traditional record. The digital record is therefore distanced from the objects lying behind those data just as access to digital data is distanced from the conditions of creation of those data.

Solutions to this distancing are available; however they entail adding new data structures which attempt to capture missing contextual information in the form of elaborated metadata and ontologies. As this effectively applies more technology to a problem created by the technology in the first place, it is not necessarily a robust methodology (Tuomi 1999, p. 110, Bowker 2005, p. 126), even assuming the information can be adequately captured and represented in the first place. For example, the London Charter is frequently cited as an example of the attempt to document computer-based visualisation of cultural heritage by incorporating information about the interpretative decisions made in the course of creating a visualisation. Hence:

> "Documentation of the evaluative, analytical, deductive, interpretative and creative decisions made in the course of computer-based visualisation should be disseminated in such a way that the relationship between research sources, implicit knowledge, explicit reasoning, and visualisation-based outcomes can be understood." (Charter, 2009, p. 8–9).

This is undertaken through the capture of provenance metadata (or paradata) (for example, Baker 2012; Mudge 2012), which contrasts with the more typical metadata currently used by organisations such as the Archaeology Data Service which focuses on issues of authorship, rights, and sources, and carries only limited descriptive information and nothing relating to process or derivation. To a large extent this provenance metadata remains vapourware, with little or no implementation to date. That said, there are technically-derived provenance metadata available which are captured automatically: for instance, ESRI's ArcGIS system captures information about derivable properties of the data and some information about geoprocessing techniques applied to the data without user intervention. Similarly, the EXIF metadata automatically captured by many digital cameras includes information about settings used in the creation of the photograph. If it were feasible, the availability of this kind of contextual

metadata would offer the prospect of providing a better understanding of some of the collection processes and circumstances that lie behind the data themselves, as well as potentially improving appreciation of the authority and reliability of the data.

Provenance metadata can, therefore, be seen as a means of addressing the lack of contextual information typically associated with digital data, the absence of which ought to present significant issues when those data are situated, contingent, and incomplete. On the other hand, provenance metadata also increases the data load associated with any given dataset, especially since it cannot necessarily be assumed to exist simply at the collection level. For example, individual records or sets of records within an excavation database will be created by different people and individual contexts will be excavated using different methods; likewise a single individual might be associated with the creation of a GIS dataset but that dataset itself consists of multiple layers which have been created using various data sources and algorithms. It could be argued therefore that provenance metadata would be required at all levels of a given dataset, with significant implications for capturing and representing this information.

The creation of metadata – both supporting resource discovery and providing provenance or contextual information about data – essentially creates more data about data in a structured web of dependencies and relationships. Issues of identification, classification, atomisation, and standardisation are compounded in an environment which adds new data definitions to old. If the original data are perceived in some respects to have been squeezed into pre-defined pigeonholes in order to capture them, this is equally the case with metadata. In this way the technological solution offered by metadata can be seen to reinforce the issues it is intended to resolve. Additionally, it remains to be demonstrated that such contextual metadata would be either useful or used. While the metadata in common use currently is understood to have value as a finding aid, there is little evidence of provenance metadata use as yet or indeed a clear demonstration of how it would work. Provenance metadata may be theoretically valuable, but data users are more accustomed to resorting to textual documentation in order to understand the meaning of a particular data field or its contents, assuming such documentation exists in the first place (c.f. Faniel et al. 2013). Indeed, metadata is in many respects of more significance to computational tools than to the human agents themselves who simply receive the results of the computations as a consequence of a query. We commonly perceive knowledge as passing from one knowledge worker to another with data as the intermediary, whereas increasingly knowledge is handled via a program–data–program or data–program–data cycle with a minimum of human intervention (Bowler, n.d., p. 169–170).

2.8 Transforming Knowledge?

Computer software can be seen as protecting the human user by disguising the underlying complexities of a problem or task through inserting layers of opacity (for ex-

ample, Huggett 2004*a*, p. 83–84). Similarly, in a kind of utopian determinism, the expectation is that computer systems will resolve current limitations and remove restrictions in terms of access, processing, and storage of data. However, access to these data and the immediateness of their delivery can both overwhelm and isolate the data user from the moment of discovery and capture, with the de-contextualised knowledge-information-data process inserting distance between originator and user. Recognition of this is key to knowledgeable action: for example, Turkle has characterised computer systems as creating a seduction of simulation, in which we become accustomed to manipulating a system whose core assumptions we do not understand, hence leading to the abdication of authority to the simulation (Turkle, 1997, p. 36–42). Equally we may become accustomed to manipulating data whose core assumptions we no longer understand, abdicating authority and responsibility to the system which delivered those data in response to our query. This becomes all the more important when those data are removed from their original context and repurposed – in other words, the data may be purpose-laden, collected not so much with research in mind but resource-management (Huggett, 2004*a*) which brings different priorities and concerns to the fore. Indeed, in addition to being theory-laden and purpose-laden, data may also be process-laden, with aspects of their creation and subsequent modification embedded, often invisibly, within them. The operationalisation of data within a computer environment strips out the context of creation – or at the very least, increases the distance from it (and provenance metadata seems likely to provide a poor proxy at best).

Digital data structures can be seen as constraining subsequent action and analysis, an argument which goes back to the near prehistory of computer archaeology (for example, contributions to Cooper and Richards 1985) but has seen relatively little attention since. These largely unseen and potentially unrealised aspects of digital data are not dissimilar to discussions about the way that traditional context sheets work "to make the objects of archaeology comparable . . . by making the actions of the people that use them comparable" (Yarrow 2008, p. 123; Lucas 2001, p. 9). Although Yarrow suggests that context sheets are actually less restrictive than they might appear (2008, 130-2), it is not clear that the same can be said for data structures. The database is not neutral: data have to be structured in order to be represented, and the choice of representation carries different implications for the data. For example, hierarchical databases, where each item has a single parent, impose a detailed line of authority which required to be followed to retrieve any information (Bowker 2005, p. 130–131, Bowler n.d., p. 169). Relational databases separate the physical organisation of data in the computer and the representation of the data: each entity is identified by a record number, and – in theory – at any point the user can specify a set of relationships to produce a view that reflected those relationships, though in reality the range of relationships is more limited (Bowker, 2005, p. 131). The structure of object-oriented and object-relational databases means that basic operations can be redefined and reconfigured: "any structure can be evanescent providing we know the inputs or outputs of any objects within it" (Bowler, n.d., p. 169), but these are not yet the source of most

archaeological data, and even if they were, it remains to be seen how much control the data user is actually allowed. A database is therefore not an 'empty vessel' into which data can be poured – and if it were, the computer would be invisibly organising and making sense of the data which would make the process still less transparent than the traditional structures currently in use. However, there has still been relatively little attention paid within archaeology to the effects of structuring data for a database on the way that we think about that data, on the way we go about recording that data, the way in which we retrieve that data, and the way in which we subsequently analyse that data (Huggett, 2004*a*).

This become more important, not less, as we move into the disruptive realms of what Anderson has described as the "end of theory", in which he claims "'Correlation is enough.' … We can analyze the data without hypotheses about what it might show." (Anderson, 2008). Such proponents of 'big data' frequently adopt a fetishistic approach to the power of systems to overcome the limits of 'small data'. The sheer quantity of data is argued to make quality less significant, so that the size of the datasets will offset any problems associated with errors and inaccuracies in the data to the extent that "It isn't just that 'more trumps some', but that, in fact, sometimes 'more trumps better'." (Mayer-Schönberger and Cukier, 2013, p. 33). However, just because a dataset is large does not mean it is representative or unbiased, and methodological issues are even more important with large and disparate datasets (Boyd and Crawford, 2012, p. 669). Indeed, boyd and Crawford highlight the mythological aspects of 'big data': specifically that large datasets somehow offer a higher form of intelligence and knowledge that can generate insights that were previously impossible, with the aura of truth, objectivity, and accuracy (2012, p. 663). 'Big data' explicitly adopts the data-information-knowledge model, with the 'bigness' of data seen as requiring it to be collected prior to interpretation (Boellstorff, 2013), and in the process presumes that knowledge can be generated in a theoretical vacuum. This may be true in the sense of data automatically captured through instruments, sensors, click-throughs, and the like, but even the creation of a device (whether hardware or software) has knowledge fixed into it, since what it records is designed into the system (Tuomi, 1999, p. 108–109). In reality, 'big data' always entails 'big theory', whether or not this is recognised (Boellstorff, 2013). Losing sight of these issues risks what Carr (2013) has identified as automation complacency and automation bias, lulling the user into a false sense of security and certainty such that we fail to recognise errors and shortcomings as computers increasingly mediate our understanding (Carr, 2010).

Archaeology may not yet be dealing in 'big data', but the foundations are being laid for doing so. Open data are implicated in this, as is the construction of new data infrastructures (for example, Niccolucci and Richards 2013), the creation of automated processes to align data of different types drawn from different sources (for instance, Jeffrey et al. 2009; May et al. 2010), and processes to automatically extract information from online publications and datasets (Byrne and Klein 2010; Vlachidis et al. 2010, for example). These, and projects like them, are challenging, innovative, and excit-

ing; however, all are based on automatic extraction, processing, and transformation of archaeological data and their results typically become the basis of the tools we use to access archaeological data in the future. One of the clearest examples of this is the faceted classification system developed by the Archaeotools project, which now sits beneath the ArchSearch browser used by the Archaeology Data Service as a primary means of accessing its data and resources (Jeffrey et al., 2009; Richards et al., 2011). The ARIADNE project seeks to integrate existing archaeological data infrastructures across Europe, and while there is no doubting that digital data across Europe are scattered amongst different silos and access is constrained by a lack of common standards and agreed metadata (Niccolucci and Richards, 2013, p. 85), the level of manipulation of data in order to achieve integration across these disparate datasets is likely to be considerable, and the data users potentially removed still further from the data as originated.

2.9 Open Data is for Sharing

None of this should deny the value, importance, and potential of open data in archaeology. When access to the Archaeology Data Service has reduced the time required for data acquisition and data processing for 79% of archaeologists surveyed, has improved the efficiency of archaeological research in the UK (JISC/Research Information Network, 2011, p. 34), and those efficiency impacts are valued at between £13m and £58m per year (Beagrie and Houghton, 2013, p. 40), the benefits seem unarguable. Instead, the concern is to recognise the implications of increasing access to data for users separated by space, and inevitably and increasingly time, from the data originators, and the effects of the ways in which the tools used seek to capture the consequences of interpretation, classification, and identification which remain largely tacit. The benefits for archaeology, in terms of an enhanced ability to access and use data, are predicated upon a clear understanding of those data as well of the level of control and authority implicit in their delivery. Indeed, as the tools formalising the information for delivery are increasingly automated, the status of the data user can become little more than a powerless consumer. Given the way that classification standards, information infrastructures, and the data themselves shape future practice, it is all the more important to reveal the forms, decisions and assumptions which underpin them rather than allow them to remain invisible. These classifications and standards are the means by which data from one time and place are linked to data from another, since they provide for the regularisation of the data, allowing them to be communicated between different contexts (for example, Bowker and Star 1999, p. 290; Huggett 2012).

The ease with which data are communicated within a technological environment is in marked contrast to earlier generations where data were held in notebooks and

card indexes and presented in the form of published reports. The benefits of this seem clear:

> "…sharing primary data allows us to better confront some of the biases in the data collection and analysis process, and to do more informed research, rather than simply taking the interpretive publication at face value." (Atici et al., 2013, p. 666).

Making individual datasets available for reuse is largely a matter of providing access and adequate documentation to provide the necessary theoretical and methodological background and explanation (for example, Atici et al. 2013, p. 677–679). In certain respects reusing such data presents similar challenges to reinterpreting traditional non-digital archives. This is not the case where the data have been made interoperable for the purposes of comparison and combination into large datasets. Linking data with other datasets is not a simple process: although semantic tools such as ontologies are used to provide mappings between the different datasets, these are in no sense absolute (Huggett, 2012, p. 543–545). These mappings may be carried out manually or increasingly automatically but "their methods require potentially contestable judgement calls" (Atici et al., 2013, p. 674), and these methods and judgement calls are not made explicit nor are they widely appreciated. As argued elsewhere (Huggett, 2012), little attention has been paid to the means by which data standards have been developed and implemented in order to achieve interoperability and communication – or at least, such as there has been is not in the public domain. In the process, the implications of the methods by which data become interoperable become lost in the face of engagement with these unified datasets which are, by definition, no longer primary and yet may be treated as if they are. Where these mappings are undertaken automatically, the data themselves are no more than tokens shunted around in a manner which reshapes and reformulates them within a technical environment. This is far removed from the eventual human agents who remain largely oblivious to the actions that have been undertaken in order to deliver the data to them.

For example, in the context of the thousands of mostly small-scale archaeological interventions undertaken across the UK and only available as grey literature, Fowler estimated that he was able to take account of less than five percent of the information gained over the past 20 years in attempting to write a work of archaeological synthesis (Fowler, 2001, p. 607). Similarly, Bradley's synthesis of British and Irish prehistory entailed four years of professorial research leave, plus the salary of a research assistant for three years (Bradley 2006, 10) in order to travel the country to seek out grey literature reports accumulated over 20 years. Now, however, there are over 22,000 grey literature reports in the Archaeology Data Service digital library, and more are added each month through the OASIS project in England and Scotland. Access to this resource clearly changes the nature of the task of synthesis, but if natural language techniques are applied to these reports in the search to gain comparability and interoperability of the data and the information codified within them (for example, Richards et al. 2011),

what would then be the nature of a synthesis that might be derived as a consequence of such technical intervention?

As Bevan (2012*a*, p. 493) has recently pointed out, the availability of large-scale datasets should shift our goalposts and enlarge our interpretative ambitions, an observation that can be widened to incorporate open data in general. However, as he also points out, access brings with it issues associated with recovery and recording biases – and, as is argued here, potentially a lot more besides. The challenge is to recognise these issues when the emphasis surrounding openness is instead, perhaps inevitably, focussed on facilitating the availability, interoperability, and ease of delivery of the data.

Acknowledgements

I should like to thank Kath Baker for her comments on drafts of this chapter, and also Katie Green and Michael Charno for their assistance with the Archaeology Data Service access figures. As ever, all errors and misconceptions remain my own.

Bibliography

Anderson, C. (2008), 'The end of theory', *Wired magazine* **16**(7), 16–07.

Anichini, F. and Gattiglia, G. (2012), #mappaopendata. from web to society. archaeological open data testing, *in* 'Opening the Past: Archaeological Open Data', Vol. 3, Metodologie Applicate alla Predittvita del Potenziale Archeologico, pp. 54–56.

Arbesman, S. (2013), 'Stop hyping big data and start paying attention to 'long data'', *Wired magazine* **1**(21).

Archaeological Data Service (n.d.), 'The terms of use and access to ads resources'.
URL: http://archaeologydataservice.ac.uk/advice/termsOfUseAndAccess

Atici, L., Kansa, S. W., Lev-Tov, J. and Kansa, E. (2013), 'Other people's data: A demonstration of the imperative of publishing primary data', *Journal of Archaeological Method and Theory* **20**(4), 663–681.

Baker, D. (2012), Defining paradata in heritage visualisation, *in* 'Paradata and Transparency in Virtual Heritage', Ashgate Publishing, Ltd., pp. 163–175.

Barrett, J. C. (1988), 'Fields of discourse reconstituting a social archaeology', *Critique of Anthropology* **7**(3), 5–16.

Beagrie, N. and Houghton, J. (2013), 'The value and impact of the archaeology data service: A study and methods for enhancing sustainability'.
URL: http://archaeologydataservice.ac.uk/research/impact

Beale, N. (2012), 'How community archaeology can make use of open data to achieve further its objectives', *World Archaeology* **44**(4), 612–633.

Beck, A. and Neylon, C. (2012), 'A vision for open archaeology', *World Archaeology* **44**(4), 479–497.

Belisario, E., Cogo, G., Epifani, S. and Forghieri, C. (2011), *Come si fa Open Data? Istruzioni per luso per Entie Amministrazioni Pubbliche Version 2*, Associazione Italiana per l'Open Government.

Bevan, A. (2012*a*), 'Spatial methods for analysing large-scale artefact inventories', *Antiquity: a quarterly review of archaeology* **86**(332), 492–506.

Bevan, A. (2012*b*), Value, authority and the open society. some implications for digital and online archaeology, *in* 'Archaeology and Digital Communication: Towards Strategies of Public Engagement', Archetype, pp. 1–14.

Binford, L. R. (1987), 'Data, relativism and archaeological science', *Man* pp. 391–404.

Boellstorff, T. (2013), 'Making big data, in theory', *First Monday* **18**(10).

Borgman, C. L. (2007), *Scholarship in the Digital Age: Information, Infrastructure, and the Internet*, MIT press.

Bowker, G. C. (2005), *Memory practices in the sciences*, MIT Press Cambridge, MA.

Bowker, G. and Star, S. (1999), 'Sorting things out: classification and its consequences', *Inside technology* .

Bowler, G. (n.d.), Data flakes: an afterword to "raw data" is an oxymoron, *in* "'Raw" data is an oxymoron', Cambridge, MA: MIT Press, pp. 167–171.

Boyd, D. and Crawford, K. (2012), 'Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon', *Information, Communication & Society* **15**(5), 662–679.

Bradley, R. (2006), 'Bridging the two cultures–commercial archaeology and the study of prehistoric britain', *The Antiquaries Journal* **86**, 1–13.

Byrne, K. and Klein, E. (2010), Automatic extraction of archaeological events from text, *in* 'Making History Interactive: Computer Applications and Quantitative Methods in Archaeology 2009', Oxford: BAR International Series, pp. 48–56.

Canadian Archaeological Association (n.d.), 'Canadian archaeological association principles of ethical conduct'.
URL: http://canadianarchaeology.com/caa/about/ethics/principles-ethical-conduct

Carr, N. (2010), *The shallows: How the internet is changing the way we think, read and remember*, Atlantic Books Ltd.

Carr, N. (2013), 'All can be lost: The risk of putting our knowledge in the hands of machines', *The Atlantic (November 2013* **November 2013**.
URL: http://www.theatlantic.com/magazine/archive/2013/11/the-great-forgetting/309516

Carson, C. A. (1996), 'Laser bones: Copyright issues raised by the use of information technology in archaeology', *Harv. JL & Tech.* **10**, 281.

Charter, L. (2009), 'The london charter for the computer-based visualisation of cultural heritage (version 2.1)'.
URL: http://www.londoncharter.org

Chippindale, C. (2000), 'Capta and data: On the true nature of archaeological information', *American antiquity* **65**(4), 605–612.

Darvill, T. (2007), 'Research frameworks for world heritage sites and the conceptualization of archaeological knowledge', *World Archaeology* **39**(3), 436–457.

Drucker, J. (2011), 'Humanities approaches to graphical display', *Digital Humanities Quarterly* **5**(1).

Duguid, P. (2005), "the art of knowing': Social and tacit dimensions of knowledge and the limits of the community of practice', *The information society* **21**(2), 109–118.

European Association of Archaeologists (1997), 'European association of archaeologists code of practice'.
URL: http://e-a-a.org/codes.htm

Faniel, I., Kansa, E., Whitcher Kansa, S., Barrera-Gomez, J. and Yakel, E. (2013), The challenges of digging data: a study of context in archaeological data reuse, *in* 'Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries', ACM, pp. 295–304.

Fowler, P. (2001), 'Time for a last quick one?', *Antiquity* **75**(289), 606–608.

Gitelman, L. (2013), *Raw data is an oxymoron*, MIT Press.

Gramsch, A. (2011), Theory in central european archaeology: dead or alive?, *in* 'The Death of Archaeological Theory?', Oxbow Books, pp. 48–71.

Heritage Gateway (2007), 'Terms and conditions'.
URL: http://www.heritagegateway.org.uk/gateway/termsandcondition

Hodder, I. (1999), *The archaeological process: An introduction*, Blackwell Oxford.

Huggett, J. (2000), *Computers and archaeological culture change*, Oxford University Committee for Archaeology Monograph 51, pp. 5–22.

Huggett, J. (2004*a*), 'Archaeology and the new technological fetishism', *Archeologia e calcolatori* (15), 81–92.

Huggett, J. (2004*b*), 'The past in bits: towards an archaeology of information technology', *Internet Archaeology* **15**.

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

Huggett, J. (Forthcoming), 'Promise and paradox: accessing open data in archaeology'.

Institute for Archaeologists (2013), 'Ifa 2013 bylaws of the institute for archaeologists: Code of conduct, reading'.
URL: http://www.archaeologists.net/codes/ifa

Jeffrey, S., Richards, J., Ciravegna, F., Waller, S., Chapman, S. and Zhang, Z. (2009), 'The archaeotools project: faceted classification and natural language processing in an archaeological context', *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **367**(1897), 2507–2519.

JISC/Research Information Network (2011), 'Jisc 2011 data centres: their use, value and impact'.
URL: http://www.jisc.ac.uk/publications/generalpublications/2011/09/datacentres.aspx

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

Kansa, E., Kansa, S. W., Burton, M. M. and Stankowski, C. (2010), 'Googling the grey: Open data, web services, and semantics', *Archaeologies* **6**(2), 301–326.

Knox, K. T. (2007), 'The various and conflicting notions of information', *Issues in Informing Science and Information Technology* **4**(1), 675–89.

Lake, M. (2012), 'Open archaeology', *World Archaeology* **44**(4), 471–478.

Llobera, M. (2011), 'Archaeological visualization: Towards an archaeological information science (aisc)', *Journal of Archaeological Method and Theory* **18**(3), 193–223.

Lucas, G. (2001), *Critical approaches to fieldwork: contemporary and historical archaeological practice*, Routledge.

Lucas, G. (2012), *Understanding the archaeological record*, Cambridge University Press.

May, K., Binding, C. and Tudhope, D. (2010), Following a star? shedding more light on semantic technologies for archaeological resources, *in* 'Making History Interactive: Computer Applications and Quantitative Methods in Archaeology 2009', Oxford: BAR International Series, pp. 227–233.

Mayer-Schönberger, V. and Cukier, K. (2013), *Big data: A revolution that will transform how we live, work, and think*, Houghton Mifflin Harcourt.

McKeague, P., Corns, A. and Shaw, R. (2012), 'Developing a spatial data infrastructure for archaeological and built heritage', *International Journal of Spatial Data Infrastructure Research* **7**, 38–65.

Mudge, M. (2012), *Transparency for empirical data*, Farnham, Surrey: Ashgate, pp. 177–188.

Naylor, J. and Richards, J. (2005), 'Third-party data for first class research', *Archeologia e Calcolatori* (XVI), 83–91.

Negroponte, N. (1996), *Being digital*, Random House LLC.

Niccolucci, F. and Richards, J. (2013), 'Ariadne: Advanced research infrastructures for archaeological dataset networking in europe', *International Journal of Humanities and Arts Computing* **7**(1-2), 70–88.

of America, A. I. (2008), 'Archaeological institute of america code of professional standards'.
URL: http://www.archaeological.org/pdfs/AIA_Code_of_Professional_StandardsA5S.pdf

Open Definition (2014), 'Open definition version 1.1'.
URL: http://opendefinition.org/okd/

PastMap (2013), 'Terms and conditions'.
URL: http://pastmap.org.uk/

Patrik, L. E. (1985), 'Is there an archaeological record?', *Advances in archaeological method and theory* pp. 27–62.

Postman, N. (1993), 'Technopoly: the surrender of culture to technology'.

Räsänen, M. and Nyce, J. M. (2013), 'The raw is cooked data in intelligence practice', *Science, Technology & Human Values* **38**(5), 655–677.

Richards, J., Jeffrey, S., Waller, S., Ciravegna, G., Chapman, S. and Zhang, Z. (2011), *The Archaeology Data Service and the Archaeotools Project: Faceted Classification and Natural Language Processing*, Los Angeles: UCLA Cotsen Institute of Archaeology Press, pp. 31–56.
URL: http://escholarship.org/uc/item/1r6137tb

Richards, J., Naylor, J. and Holas-Clark, C. (2008), 'The viking and anglo-saxon landscape and economy (vasle) project'.
URL: http://archaeologydataservice.ac.uk/archives/view/vasle_ahrc_2008/index.cfm

Richards, J., Naylor, J. and Holas-Clark, C. (2009), 'Anglo-saxon landscape and economy: using portable antiquities to study anglo-saxon and viking age england', *Internet Archaeology* .

Royal Society (2012), 'Science as an open enterprise: open data for open science'.
URL: http://royalsociety.org/policy/projects/science-public-enterprise/report/

Society for American Archaeology (1996), 'Society for american archaeology principles of archaeological ethics'.
URL: http://www.saa.org/AbouttheSociety/PrinciplesofArchaeologicalEthics/tabid/203/Default.aspx

Trigger, B. G. (1998), 'Archaeology and epistemology: dialoguing across the darwinian chasm', *American Journal of Archaeology* pp. 1–34.

Tuomi, I. (1999), 'Data is more than knowledge: Implications of the reversed knowledge hierarchy for knowledge management and organizational memory', *Journal of Management Information Systems* **16**(3), 103–117.

Turkle, S. (1997), 'Life on the screen: Identity in the age of the internet', *Literature And History* **6**, 117–118.

Vlachidis, A., Binding, C., Tudhope, D. and May, K. (2010), Excavating grey literature: A case study on the rich indexing of archaeological documents via natural language-processing techniques and knowledge-based resources, *in* 'Aslib Proceedings', Vol. 62, Emerald Group Publishing Limited, pp. 466–475.

Weinberger, D. (2010), 'The problem with the data–information–knowledge–wisdom hierarchy', *Harvard Business Review Blog Network* **2**.

Weinberger, D. (2011), 'Too big to know'.

Yarrow, T. (2008), In context: meaning, materiality and agency in the process of archaeological recording, *in* 'Material Agency', Springer, pp. 121–137.

Zuiderwijk, A., Janssen, M., Choenni, S., Meijer, R. and Sheikh_Alibaks, R. (2012), 'Socio-technical impediments of open data', *Electronic Journal of e-Government* **10**(2), 156–172.