Title page

Authors' names: Anwen Cooper and Chris Green

Title: Embracing the complexities of 'big data' in archaeology: the case of the English Landscape and Identities project

Affiliation and address: Institute of Archaeology, University of Oxford, 36 Beaumont Street, Oxford OX1 2PG

Contact details: email: anwen.cooper@arch.ox.ac.uk
Telephone: 01865 278256
Mobile: 07812196116

Embracing the complexities of 'big data' in archaeology: the case of the English Landscape and Identities project

Anwen Cooper and Chris Green

Abstract

This paper considers recent attempts within archaeology to create, integrate and interpret digital data on an unprecedented scale - a movement that resonates with the much wider so-called 'big data' phenomenon. Using the example of our work with a particularly large and complex dataset collated for the purpose of the English Landscape and Identities project (EngLaID), Oxford, UK, and drawing on insights from social scientists' studies of information infrastructures much more broadly, we make the following key points. Firstly, alongside scrutinising and homogenising digital records for research purposes, it is vital that we continue to appreciate the broader interpretative value of 'characterful' archaeological data (those that have histories and flaws of various kinds). Secondly, given the intricate and pliable nature of archaeological data and the substantial challenges faced by researchers seeking to create a cyber-infrastructure for archaeology, it is essential that we develop interim measures that allow us to explore the parameters and potentials of working with archaeological evidence on an unprecedented scale. We also consider some of the practical and ethical consequences of working in this vein.

Key words: archaeological 'big data', GIS, landscape archaeology, grey literature

Introduction

In recent years, in line with researchers more broadly in the humanities, archaeologists across the world have taken important steps towards building what has been described as a cyber-infrastructure for the discipline (Kintigh 2006; Snow et al. 2006). Taking a lead primarily from scientific researchers, archaeologists have begun to appreciate both the fragility of their (increasingly wholly digital) datasets and also to imagine the incredible interpretative possibilities of assembling these datasets in their various guises on an unprecedented scale (see also Onsrud and Campbell 2007). Generously funded projects such as the Digital Archaeological Record (tDAR) in the US (http://www.tdar.org/), Archaeological Records of Europe - Networked Access (ARENA) in Europe (http://archaeologydataservice.ac.uk/research/arena2) and Transatlantic Archaeological Gateway (TAG) that spans the two continents (http://archaeologydataservice.ac.uk/research/tag) have invested considerable effort in testing out the practicalities of realising this vision. Meanwhile the practices of sharing and providing improved access to digital archaeological data are increasingly encouraged by funding bodies and researchers alike (see for instance Amorosi et al. 1996; Atici et al. 2013, 674; articles in a special issue of *The SAA Archaeological Record*, 11(1) on Digital Communication and Collaboration: Perspectives from Zooarchaeology; and the recent prioritisation of the 'Digital Transformations' theme and, in particular, the big data aspect of this theme by the Arts and

Humanities Research Council in the UK http://www.ahrc.ac.uk/News-and-Events/News/Pages/Digital-Transformations-in-the-Arts-and-Humanities---Big-Data-Projects-Call.aspx). Indeed it has been argued that in the longer term, it should be possible to link archaeological datasets to research data much more broadly in the humanities and sciences (Kintigh 2006, 567). The National Inventory of Natural Heritage in France provides an important example of work in this mode (Callou et al. 2011). While the interpretative benefits of such initiatives have yet to be explored fully (e.g. Spielmann and Kintigh 2011, 24), the strong drive towards a situation in which digital data are easily accessible and combinable undoubtedly has the potential to transform significantly archaeological practice (see Boyd and Crawford 2012; Latour et al. 2012 for broader discussions of the capacity of technological movements to change knowledge production practices much more widely).

In relation to this movement, archaeology's digital data have themselves been subject to scrutiny. This scrutiny has, of course, been enabled at a broad level by what has been described as a 21st century computing 'revolution' in archaeology (Levy 2014). Without this sea change we would arguably not be in a position to conduct such review. Analyses of this kind have also been prompted explicitly by the aforementioned burgeoning urge to collate digital datasets on a massive scale (Atici 2013). In short, such developments have brought about an appreciation that if we are to make the most of emerging technologies, we need to understand our data better. At a general level, key proponents of data sharing initiatives have noted the complexities of digital archaeological data, in particular the variable ways in which archaeologists record similar archaeological entities, the diverse formats and structures of datasets, and differing attitudes towards access to these data (Kintigh 2006, 570; Snow 2006, 958; Dam et al. 2010). On this basis it has been argued that attempts to build a cyber-infrastructure for archaeology should draw on technologies developed in scientific domains in which, similarly, 'primary data are highly contextual and inconsistently collected, many inferential steps separate scientific understandings of major phenomena from observational data, and competing ontologies need to be maintained and used' (Kintigh 2006, 571). Accordingly, efforts have focused on developing methods and technologies (involving metadata and ontologies, see Spielmann and Kintigh 2011 for a concise explanation) that make the integration of disparate datasets feasible.

More specifically, researchers in Europe and the US have raised concerns about the 'quality' of the primary datasets that many archaeologists aspire to merge. Such anxieties have been expressed mainly in the realm of zooarchaeology - a disciplinary area that is widely accepted to be leading the way in data integration initiatives given the broad consensus amongst its practitioners regarding recording practices. Mikkelsen (2012) highlights a range of omissions, irregularities and possible interpretative errors within key national archaeological datasets in Denmark. Dam and Hansen (2005) see the achievement of a 'certain quality' of information that allows for interoperability as a key challenge for archaeologists in coming years. Noting the multitude of issues involved in harmonising disparate digital archaeological data, Atici et al. suggest that data quality and integrity are key issues to address, and that detailed documentation of datasets (the creation of metadata) could provide a means of verifying such attributes (2013, 677). Based on similar observations, Gobalet made recommendations as to

how data consistency and quality might be improved, for instance via peer review of primary datasets (2001, 385).

In the UK - a geographical area that is recognised to be spearheading the support and development of data sharing networks both in archaeology and more broadly (Kintigh 2006, 572; Nature Editors 2009) - the critique of digital archaeological data has been particularly intense and has focused mainly on secondary rather than on primary archaeological datasets. Complexities, gaps, inconsistencies, uncertainties and inadequacies in these records have duly been exposed. Hierarchies of usefulness and unifying interpretative schema have been put forward (Roskams and Whyman 2007). In contrast to the situation more widely, the potential for considering diverse data stores as artefacts in their own right has also been advanced (Newman 2011). One specific and very important element of the situation in the UK is that many of the digital datasets that have developed differ not only in format, structure, detail, and in the terminologies employed, as has been observed elsewhere. They also overlap considerably in content - a phenomenon that is unlikely to be unique to the UK but that has not received significant (if any) coverage in wider discussions about data integration.

In this context, the work of researchers on the English Landscape and Identities project (EngLaID) based at the University of Oxford (http://www.arch.ox.ac.uk/englishlandscapes-introduction.html) can make an important contribution. In parallel with wider recent UK initiatives aimed at developing technologies to integrate primary digital datasets (e.g. STELLAR http://archaeologydataservice.ac.uk/research/stellar/), EngLaID has embarked on a project that requires the integration of highly diverse secondary (and to a lesser extent primary) digital datasets from across English archaeology and, most importantly, the interpretative use of these merged datasets. This places us in a unique position both to comment on the peculiarities of accessible digital archaeological datasets in one particular region of the world, and also to share our experiences of using integrated digital data on a sizeable scale (we have amassed close to 1,000,000 digital records). In this way, we are able to respond to Atici et al.'s important observation that 'despite wide acknowledgement that approaches to data collection, recording, analysis, presentation and interpretation vary among researchers, few studies have explored challenges researchers may face in the analysis of datasets produced by others' (2013, 664). Our work might also be viewed as an example of the kind of effort to address the particular character of regional dataset specifics (definitions and uses of monument and finds terminologies, temporal periods and so on) advocated by Kintigh (2006, 574).

Additionally, and more unusually within the context of recent discussions about archaeology's digital data, we have also revisited one of the main claims of reflexive archaeologies of the late 20th century; that all archaeological data (and outputs more broadly) are socially and historically produced (e.g. Clarke 1968; Hodder 1984, 1986; Patrik 1985; Shanks and Tilley 1987; Wylie 1985). Our approach draws at a broad level on insights both from this archaeological movement and from wider discussions in the social sciences regarding the important role played by non-human (as well as human) participants in knowledge production practices (materials, technologies and so on, e.g. Latour 2005). More

specifically, we have been inspired by science and technology studies that have examined information infrastructures well beyond archaeology (e.g. Bowker 2005; Bowker and Star 1999; Musen 1992; and papers in Gitelman 2013; Lampland and Star 2009). In doing so we are able to offer an alternative slant on the aforementioned data concerns that recent analyses of digital archaeological datasets at an international level have raised. We argue that as well as investigating the complexities of digital data and developing ways to make these work better, it is vital that researchers explore and appreciate the value of the histories and relationships - involving an array of human and non-human entities - that archaeological data embody. By doing so, it becomes possible to balance archaeology's widely endorsed quest for data that are 'complete', 'certain' and 'coherent' with the practicalities of working with data that often make these ideals seem impossible to achieve. It becomes feasible to foreground the exciting potential of archaeological data as bearers of disciplinary histories. It may even bring us closer to accepting the likelihood acknowledged much more broadly in information studies that, for very good reasons, ideal information schema are beyond our grasp (see papers in Gitelman 2013).

We begin with a summary of recent discussions about digital records in English archaeology. Much like the data they refer to, these critiques have not previously been assembled in one place. Together, they provide an important overview of the topography of English archaeological datasets, aspects of which will be familiar to researchers much more broadly. We augment and extend these important arguments, as well as reframing them slightly, through a consideration of the relationship between archaeological entities and the records they become. Once the insights of archaeological practitioners are set alongside those of researchers who have examined information infrastructures much more broadly, the particular difficulties experienced in archaeology gain important context. Moving on with the aim of understanding better and working practically with these complex and often intriguing records, we provide a detailed account of relationships between key datasets in English archaeology - making links wherever possible, and highlighting the mutual interdependency of the schema concerned. We finish by outlining methods that researchers on the EngLaID project have devised for transforming intricate and overlapping datasets such as these into a working research platform. The immediate details of this paper are perhaps of most direct relevance to researchers in the UK. However, as this preamble hopefully makes clear, we raise fundamental questions relating to the broad character of archaeological data and our capacities, as practitioners, to work with them at different levels that are pertinent much more widely.

The English Landscape and Identities project

Before beginning our account, it is worth noting briefly the specifics of the EngLaID project, and the key datasets that it employs (see also Gosden et al. 2009). The EngLaID project is a five-year European Research Council-funded research project based at the Institute of Archaeology, Oxford. It aims to produce a history of the English landscape from 1500BC to AD1086 using digital and published data from across English archaeology. An analysis of the

evidence available at a nationwide level will be interwoven with more specific studies of the material from fourteen widely distributed regional case studies.

Full title	*Current acronym*	*Date of dataset inception*	*Date of earliest record held*	*Types of data held*	*No. records assembled by EngLaID*
Historic Environment Records[a]	HERs[b]	Varies, but the earliest formal Sites and Monuments Record (SMR), for Oxfordshire, developed in the late 1960s (Benson 1972)	Not known	Monuments (with findspots as a subcategory) and Events (archaeological investigations)	439,609
National Record of the Historic Environment	NRHE[c]	1947 (within the Ordnance Survey). NAR itself created in 1983 (Newman 2011)	mid-18th century (at least)	Monuments (with findspots as a subcategory)	132,587
Portable Antiquities Scheme	PAS	2003	1850	Finds made by the public (in particular metal detecting finds)	277,734
Corpus of Early Medieval Coin Finds	EMC	1987 (1999 as an online resource, Rory Naismith pers. comm.)	18th century	Finds (early medieval single coin finds only)	10, 131
Archaeological Investigations Project	AIP	1995	1982	Events	24,398
National Mapping Programme	NMP	Pilot projects began c.1985	n/a	Mapped aerial photographic data	n/a
Excavation Index	EI	1978	1201[d]	Events	123,326
Online Access to the Index of Archaeological Investigations	OASIS	1996	1984	Events	n/a

Table 1: Summary of key datasets from English archaeology assembled for the EngLaID project

As well as investigating a series of closely-related interpretative themes including identity, the forcefulness of nature, continuity and change and making place, one key aim of the project is to explore the character and interpretative capacity of the various publically-accessible digital datasets available in English archaeology. These datasets include catalogued and mapped digital data from Historic Environment Records (HERs), the National Record of the Historic Environment (NRHE, implemented in the Archives Monuments Information England database (AMIE)), the Archaeological Investigations Project (AIP), the Portable Antiquities Scheme (PAS), the National Mapping Programme (NMP), the Excavation Index (EI), the

[a] 84 in total spanning the whole of England, and including Urban Archaeological Databases (UADs).
[b] Referred to previously as Sites and Monuments Records (SMRs).
[c] Referred to previously as the National Monuments Record (NMR). In turn, the NMR was derived from the earlier National Buildings Record (NBR) and National Archaeological Record (NAR) (http://ads.ahds.ac.uk/catalogue/collections/blurbs/398.cfm).
[d] A record of stonework found at Corbridge when 'digging for treasure'.

Corpus of Early Medieval Coin Finds (EMC), and smaller datasets provided by individual researchers (for instance a database of prehistoric and early Roman salt working sites collated by Kinory for the purposes of her doctoral research, 2012). To summarise, the project aims to undertake analysis at a unique scale in terms of the variety of datasets, the time-scale and geographical area, and the array of archaeological evidence under consideration. Almost 1,000,000 text-based digital records have been amassed formally in a database. Alongside these, we are working with a large volume of mapped digital data and will draw on a range of supplementary information (e.g. records from the Oxford Radiocarbon Accelerator Unit (ORAU) database https://c14.arch.ox.ac.uk/login/login.php?Location=/database/db.php) as required.

Key English datasets employed by the project and referred to in this paper are summarised in Table 1. Further information about each of these datasets is available online (see Appendix 1). Specific details about the datasets made available to the EngLaID project are provided in Appendix 2. Important points to highlight are the variable periodicities, purposes, and broad contents (e.g. finds, monuments, events[1]) of these datasets. As has been observed to be the case with digital datasets much more broadly in the humanities (Prescott 2013), each of these English archaeological datasets has evolved in different social and economic contexts; they are diverse in their content, structure and in the technologies and terminologies they employ. Additionally, it is worth mentioning that the overlap in the substance of these datasets (noted above), occurs mainly between records held in HERs and those held in other datasets - a trait that is examined in more detail below.

Recent discussions about archaeological data in England

As mentioned above, within England, archaeological datasets have recently been a topic of considerable critical interest. In addition to the technological rationale already noted, this focus can be viewed, in part, as relating to a broader 'empirical turn' within archaeology (see for example Alberti et al. 2013; Lucas 2012). It may also be linked to the fact that practitioners curating English datasets are currently weathering substantial government funding cuts. The future survival of several key datasets is under threat and there may understandably be a perceived need to increase their prominence academically.

Critiques have been produced both by the practitioners responsible for curating and managing records (e.g. Boldrini 2006; Newman 2011; Robinson 2000), and by university-based researchers either focusing specifically on them in their research (Evans 2013; Robbins 2013) or attempting to use them interpretatively (Holbrook and Morton 2008; Roskams and Whyman 2007; papers in Worrell 2010). While most of these studies examine specifically one dataset or group of datasets (e.g. the PAS, or the NRHE, or HER records) a few seek to compare or integrate different datasets within English archaeology (Evans 2013; Roskams and Whyman 2007). One interesting broad attribute of these accounts is that they tend to be

[1] The term 'events' is used throughout the text to mean archaeological fieldwork investigations.

published in forums targeted specifically at information managers (*The Historic Environment: Policy and Practice*) or at least those with a more technical leaning (e.g. *Internet Archaeology*, within a special issue about computing) rather than in contexts aimed at archaeological researchers more widely. This almost certainly relates to a general reticence within archaeology about discussing data. Understandably emphasis is placed instead on building interesting stories with these data. Indeed it has been suggested that this is true of researchers in the humanities much more broadly (Ell 2010, 164; Gitelman 2013, 3). As a consequence, however, these very important studies - potentially of relevance to archaeologists working in an array of professional contexts - have not received the prominence and thus the widespread attention they deserve.

It is not necessary to describe in detail here the insights of each of these analyses. It is worth, however, summarising briefly the main issues they raise, and also how archaeologists in England have broached these complexities. Additionally we outline historical and practical factors that researchers have identified as a means of explaining how such issues have arisen. In doing so, it is argued, an appreciation has been gained of the wider interpretative value of archaeological datasets that is not evident in discussions about digital data beyond the UK.

Biases are identified at a wide level in the recovery and recording of archaeological evidence (e.g. Newman 2012; Robbins 2013; Roskams and Whyman 2007). Variability is highlighted both within and between datasets in term of data curators' use of common vocabularies (for instance the English Heritage Monuments Thesaurus) and the numerous standards and guidance documents that have been issued to aid their work. Further disparities are noted in data curators' technical expertise (and thus their capacity to handle the data), and in their use of software (e.g. Lee 2012; Robinson 2000). Major gaps, errors and discrepancies are highlighted in three key archaeological events databases (the AIP, the EI and OASIS) (Evans 2013). Roskams and Whyman (2007) emphasise the incompatibility of datasets in English archaeology together with the scarcity of inter-referencing between them. Indeed the 'inadequacy' of HERs as indices of the remains of past practice was noted fairly early on in their history (Wainwright 1989). Evans sums up the current situation for archaeological events databases as follows: 'a researcher working in the early twenty-first century is confronted with a myriad of recording systems, each subtly different and recording events that are not recorded elsewhere' (2013, 31). This argument could easily be extended to cover archaeological datasets much more broadly in England (not just those for events).

In relation to this situation, most critiques mention that data managers have called repeatedly for improved data collation, characterisation, interoperability, consistency of practice, and so on. In line with such requests, the researchers who have conducted these studies convey a broad sense (or perhaps hope) that their detailed analyses will put data curators in a better position to 'improve' the datasets concerned, and, as one researcher puts it, to bring them closer to the 'accuracy, reliability and completeness' that are usually seen as fundamental to the academic enterprise (Evans 2013, 20). Roskams and Whyman (2007) go one step further and develop a method for integrating diverse datasets. Interestingly, several of these analyses also mention that many university-based archaeologists have failed to engage with the

datasets under consideration (Robinson 2000; Evans 2013). As the work of the EngLaID project hopefully confirms, this situation is now changing. However it is certainly possible that a reluctance to use certain secondary datasets in English archaeology for research purposes has arisen due to concerns regarding their accuracy and usability.

Moving on to consider some historical and practical factors that these studies raise, Robinson (2000) sees social dynamics as key to the complexities he observes in the makeup of HER data. For instance, he highlights a shift that occurred in the primary agenda driving archaeological indexing in HERs between the late 1960s and early 1970s. While in the late 1960s the research potential of the data being indexed was a key motivating force, by the early 1970s the potential use of these for satisfying Local Authority (planning-related) agendas became more significant. He also notes that an ongoing tension exists between these two impetuses (ibid., 94-5). Robinson observes that changes in local political and economic circumstances affect substantially the resources made available for record making, and thus the makeup of records (ibid., 95). He notes the changeability (and multiplicity) of the common terminologies, standards and guidelines that HER professionals are required to work with, and the difficulties that archaeologists have had in negotiating these, despite their common commitment to consistency and coherence (ibid., 97-8; see also Cooper 2013, Chapter 8). He identifies the challenges involved in classifying archaeological entities systematically in the context of a rapidly developing discipline in which shared vocabularies and categories continue to evolve (ibid., 96). He also recognises that varying interpretations of such schema relate both to levels of expertise and also to personal tendencies towards the 'lumping' and 'splitting' of data (ibid., 97, see Atici et al. 2013, 667 for a wider archaeological discussion of this issue). Indeed, several authors highlight the importance of what we might term categorisation practices, stressing the difficulties involved in pigeonholing 'fuzzy' archaeological entities. For instance Roskams and Whyman (2007) contrast the fairly consistent recording of easily-recognised archaeological entities such as Bronze Age funerary monuments, with the much more variable recording of less easily defined (and also less well understood) entities such as Mesolithic occupation sites.

Additionally, two studies highlight the important role played by the materials and technologies employed in recording archaeological data. Newman charts (with a tinge of nostalgia) the transformation of records held in the NRHE from hand-written notes on maps and index cards to digital records on a 'dumb' computer terminal (2011, 2-4). Both he (ibid.) and Roskams and Whyman (2007) comment on how the digitisation of archaeological records (beginning in the 1970s, but primarily enacted in the 1990s and 2000s) had a reifying and dehumanising effect on these data, increasing ambiguity about their origin (see Bawden and Robinson 2009; Boyd and Crawford 2012 for broader discussions of this issue). Indeed Newman argues that the physical manifestation of data has a profound effect on their functionality and also on how they are perceived. Drawing on recent studies examining the concept of materiality in archaeology (e.g. Jones 2009) he goes so far as to suggest that data repositories such as the NRHE can be considered as material culture in their own right, whatever their mode of physical presence in the world (ibid., 8).

Relations between archaeological entities and the records they become

While it would certainly have been possible - given the EngLaID project's experiences of assembling and manipulating digital data from across England pertaining to a c. 2600 year time-period - to add detail to these critiques (highlighting further gaps, variations in recording practices, incompatibilities and so on) this is not our main intention here. Project researchers have recently undertaken such an exercise as part of a national English Heritage-led initiative aimed at improving the interoperability of English archaeological datasets (Kamash et al. 2014) and aspects of these observations will appear in future project publications. Rather our immediate interest is to extend and also to contextualise certain observations made in the studies outlined above, particularly those regarding the social and historical character of archaeological data. By doing so, we argue, it is possible to shift attention from focusing mainly on the 'inadequacy' of these complex datasets and on seeking primarily to perfect them, to appreciating the importance of delving further into their social and historical qualities, and also to making these data (with all their flaws) work for us. Arguably one benefit of the considerable body of data being handled by the EngLaID is that our capacity to improve these records is extremely limited. Unlike researchers undertaking smaller regional studies or investigating a fairly narrow timeframe, it is simply not possible within the context of our project to edit records to a point at which they are broadly consistent across datasets, or to unite, or even to interconnect records from the various datasets being employed into a single impeccable interface. In this context, it is vital that alongside being aware of the intricacies of our data, we also devise new ways of dealing with them pragmatically, in many cases *as they are*.

One helpful way of reframing the arguments made in existing critiques of English archaeological data might be to see them as concerning, at a broad level, the relationship between archaeological entities - a Bronze Age spearhead recovered from a ploughed field, an Iron Age pit replete with broken pottery excavated during a watching brief, a geophysical survey of early medieval settlement remains - and the digital records they become - maybe a 'Bronze Age findspot', a 'later prehistoric pit', an 'archaeological evaluation'. While this link is apparently simple, what the studies outlined above elucidate (more or less explicitly) is that in fact it condenses a whole series of other relationships and practices involving people, archaeological materials, organisations, technologies of various kinds (digital and otherwise), and so on, that have sometimes unfolded over a very long time period, and that are always evolving. Sometimes the relationship between an archaeological entity and the digital record it becomes seems quite close; if questioned, there would be widespread consensus (amongst archaeologists and datasets) that the digital record concerned provided a reasonably accurate and sufficiently full depiction of the past entity it referred to. This might be the case, for example, for a PAS record representing a 2nd century AD Roman brooch of a widely-recognised type, recovered during a metal detecting rally in 2010.

In many other cases however, and for a variety of reasons, this connection feels very distant. As Newman (2011) observes is the case when paper records are digitised, a degree of ambivalence develops about the precise nature of the archaeological entity the record

represents. In some cases this sense of distance might arise because a substantial disparity exists between the complexity of the archaeological entity concerned (for instance a large multi-phased excavation which took place over a considerable time period) and the simplicity of the record that represents it. In such instances, the reification process that Newman (2011) and Roskams and Whyman (2007) highlight seems particularly extreme. In other cases, a sense of remoteness might be engendered by distance in time. One example might be a record marking reports of a 19th century encounter with human remains and 'handmade pottery' during the construction of a house, first plotted and indexed by the Ordnance Survey, then transformed through various stages into a digital format. Again, albeit in a different way, the relationships and practices involved in the creation of the record have been condensed. Another, slightly different sense of detachment might be brought into focus by ambiguities concerning the archaeological entity in itself. In particular, a recent and extremely productive wave of aerial photographic and lidar surveys in the UK has generated a raft of records relating to archaeological features about which, beyond their basic form, we know very little - for instance an uncertainly dated cropmark ditch. Understandably, this uncertainty can be, and is, handled by archaeological practitioners in various ways.

It is also worth noting that because of all the other connections and practices they embrace, relationships between archaeological entities and the records they become have elastic properties. The research interests or theoretical leanings of one particular curator, changing disciplinary values, the capacities (and limitations) of a particular software package, the provision of a new scientific date, broader interpretive trends, funding cuts, the issuing of a new set of 'standard' terminologies, and so on might draw a record either closer to the archaeological entity it depicts, or push it further away as time passes. As an example, a small scatter of burnt stones surveyed in a field in the low-lying fenlands of East Anglia in the 1980s may initially have been recorded by hand on a map as a 'small, later prehistoric pot boiler scatter, probably a burnt stone pit'. When, several years later, this note was translated into a database format, the resulting record could have read 'later prehistoric flint scatter' (since 'pot boiler scatter' and 'burnt stone pit' do not appear as a monument types in the EH thesaurus). The findings of a study suggesting that many burnt stone scatters in this region actually represent much more substantial features, may well have led to a reclassification of this feature as a 'later prehistoric burnt mound'. Further changes in the makeup of the record - perhaps to 'Bronze Age burnt mound', and ultimately to 'Neolithic pit' - could have been prompted by the discovery of Beaker pottery sherds in the same location by an amateur archaeologist, and an ensuing small-scale excavation by a local archaeological group. Newman claims that the digital records curated today in the NRHE are still 'in essence the same entity as the original although added to, deleted, amended and enhanced' (2011, 8). We would suggest instead that in some cases the entity has been significantly transformed, while still retaining vestiges of its previous form(s).

Overall, extending a point touched on initially by Robinson (2000) in relation specifically to HER data, we argue that once the emphasis is placed on the temporal and relational qualities of digital archaeological data rather than focusing mainly on the data themselves and their often seemingly inadequate depiction of past remains, it seems less surprising that

archaeological datasets are complicated. This argument is strengthened if we extend two points raised to an extent in previous critiques of English archaeological datasets. Firstly, many archaeological data are not just 'fuzzy', they are also fluid - they hold the potential to morph. Secondly, many of the decisions that archaeological curators make in handling their data are not, to use Bowker and Star's terminology, black and white in an 'Aristotelian' sense, but might better be classed 'prototype' (1999, 62-3). These decisions are based on metaphorical understandings of a given entity and are undertaken in relation to the specific context in which they are made. Viewed in this way it seems entirely reasonable that very similar or even the same archaeological entities are represented in diverse ways in different datasets in which quite different historical circumstances and sets of relationships are embedded.

What *is* perhaps surprising is that although the processes through which *excavated* archaeological data come into being and are 'fixed' into a textual grid that enables them to operate as an iterative record have been scrutinised academically (see for example Edgeworth 2003; Lucas 2001; Yarrow 2003, 2006), those through which other forms of archaeological data are created and subsequently shaped by research communities have evaded detailed analysis. Additionally, given the likely extent of the data idiosyncrasies unearthed in these studies, it begins to seem extraordinary (and quite admirable) that the campaign to make archaeological data 'better' (fuller, more coherent, and more consistent) forges onwards. As Roskams and Whyman (2007) suggest, one might expect alternatively that such findings would urge archaeologists to abandon the entire enterprise of creating trustworthy data for dissemination beyond their initial contexts of production.

In order to provide some context for these arguments, we will draw on insights from studies undertaken more broadly in the social sciences, starting with the work of Bowker and Star (1999). As part of a much broader examination of categorisation practices in a range of social contexts, Bowker and Star examined the parameters and workings of the International Classification of Diseases (ICD) - an index, designed to aid at a worldwide level the categorisation of diseases for the purposes of death certificates, insurance claims etc.. In doing so, they also explored the broader 'information infrastructures' in which the ICD participates. Clearly this system does not constitute a dataset directly comparable to the archaeological examples considered here. However given that many of the themes they raise seem pertinent, this study is worth considering in some detail.

One noteworthy point to draw from Bowker and Star's study is that several of the properties they raise concerning the ICD and its related information systems resonate with those observed by authors commenting on archaeological datasets. This includes the ICD's tendency to obscure the relationships, practices and histories through which it is formed (ibid., 9), ICD users' variable employment of terminologies and standards (ibid., 11), the problems which arise due to practitioners' personal tendencies towards 'lumping' or 'splitting' data (ibid., 45), the problems caused by entities (in their case viruses) that evade neat categorisation (ibid., 98), the variable application (often for financial reasons) of technologies employed in handling data (ibid., 129), the mobility of research environments

and thus the pressure for information systems to respond to them (ibid., 69), and the politics - negotiations, conflicts, organisational pressures - involved in defining and employing categories (ibid., 45). Consequently Bowker and Star describe how, when they first came to investigate the ICD, they were somewhat startled by the 'panoply of tangled and crisscrossing classification schemes' they encountered, held together by 'increasingly harassed and sprawling international health bureaucracy' (1999, 21). Drawing on a study by Musen (1992) into attempts to share and combine medical knowledge in the US, they also discuss the incompatibility of disease-related data produced within different research environments at this time (Bowker and Star 1999, 68-9).

As well as echoing (in a very different context) many of the observations archaeological researchers have made in relation to their own information systems, Bowker and Star make further observations of the ICD that are relevant here. They contend that information systems that attempt to describe past entities are particularly susceptible to revision and that through this process varying opinions can be both embraced and silenced (ibid., 41). They note how the tension between conforming to standards and the contingencies of practice leads to interpretative movement (ibid., 15), and how data categories become compromised through attempts to make them 'accurate' (useful) for multiple users (ibid., 146). Interestingly, they suggest that attempts to standardise practices amongst ICD users typically provoke a florescence of local modifications and variations rather than leading to greater conformity (see Millerand and Bowker 2009 for a detailed discussion of the effects of standardisation practices). Rather than viewing such activities in a negative light, however, they argue that modifications and variations such as these are important. Through such practices, data are made relevant at a local level for their primary users (ibid., 151-2). Moreover, they contend, resistance to standardisation is necessary, since it is through acts of resistance that information systems are able to evolve (ibid., 49). In relation to this last point, it is worth noting that resistance to conformity has been viewed as a defining feature of the archaeological community, at least in the UK (Cooper 2013, Chapter 8). In fact, recent discussions about archaeological data sharing initiatives suggest that the importance of building capacity within information infrastructures for data to retain their specific relevance at a local level, and to allow for changes in practice through time, is increasingly acknowledged within the archaeological community, difficult though it might be to handle these attributes (Spielmann and Kintigh 2011).

Commenting on the temporal qualities of information systems, Bowker and Star suggest that although such schemes are required to participate in, and react to, changes in the world around them, they are also somewhat inert. This is due, at least in part, to the considerable costs involved in revising them (ibid., 76). Achieving a balance between the stability and evolution of information systems like the ICD is important, since shifting the parameters of a particular record can make it irretrievable to future researchers (ibid., 69; see also Latour 2000). Thus, they argue, data managers in general are required to act as 'future forecasters' - predicting what and how data might be used beyond their own involvement with it. They make the important point that, once defined, the classifications employed in information infrastructures become 'realities' that in turn define communities of use (ibid., 96). Within the

UK, the example of Iron Age hillforts is apt here. However ill defined this category is in practice, its existence as a class has created a 'reality' that has nurtured a strong community of researchers (the Hillfort Studies Group http://www.arch.ox.ac.uk/hfsg.html). Indeed, Bowker and Star argue importantly that due to the tremendous work involved in fashioning information systems, these artefacts embody 'moral and aesthetic choices that in turn craft peoples' identities, aspirations and dignity' (ibid., 4).

Turning to the practicalities of dealing with these difficulties, Bowker and Star argue that a shift in attitude has occurred within the public health sector. Whereas formerly information managers had faith in creating a unified system for representing the world, they now focus their attention on developing strategies that allow systems to handle better the diversity and the practicalities of the world (ibid., 129). One way in which they enact this is by creating what they describe as 'boundary objects' - similar to what Latour (1988) described previously as 'immutable mobiles'. These are data entities that have common identities over several communities of practice and yet are able, via customisation, to satisfy the informational requirements of each of them (ibid., 16; see also Millerand and Bowker 2009). Again, this chimes with the recent focus within archaeology (see above) on developing methods and technologies that enable the integration of disparate datasets (Spielmann and Kintigh 2011). More importantly, Bowker and Star observe that for pragmatic reasons, those who use such information structures are required, at times, to operate 'as if' the data are accurate, whatever the ambiguities concerned (ibid., 115-6). Bowker and Star provide guidance along these lines for anyone seeking to 'collect precise, uniform and complete information from a large domain over a long time - and at the same time invoke the necessity of ambiguity, fuzziness and plastic meanings for their real use' (1999, 158). Alongside this, however, they argue that a new kind of research is required - 'a plate tectonics rather than a static geology', exploring the (often untold but still fascinating) nuanced topographies and histories of information infrastructures (ibid., 31 and 133).

In considering these arguments we certainly do not want to suggest that all information structures share exactly the same issues or that archaeologists should lose faith entirely in improving the integrity of their data. Indeed, we argue, it is vital that continued attempts are made both to critique and enhance archaeological data, and to develop the interoperability of the various schemas in which it is held. Rather we have provided this context since it reveals that many of the complexities of digital archaeological data that practitioners have worried over in recent years, do have much wider parallels (see also Boyd and Crawford 2012; Levi 2013; Prescott 2013). This suggests that these intricacies are not straightforwardly symptomatic of what many have described as a chronically 'fragmented discipline' (e.g. Evans 2013; Tilley 1998). Rather, the partial, disparate and mutable attributes of digital archaeological datasets that researchers in the UK have brought to the fore, can be viewed as broad properties of many information systems that attempt to describe the intricacies of our 'world of becoming' (Connelly 2011). Indeed it has been argued that certain 'gaps' perceived to exist within archaeological data can play a vital role in disciplinary practices and can even be seen as advantageous. Such gaps require that archaeologists work creatively and industriously in their interpretative endeavour in contrast, for example, to the situation in

other disciplines such as anthropology where researchers can sometimes be hampered interpretatively by the fullness of their data (Yarrow 2012). In this context, we argue, it is very important firstly, that we invest effort in exploring further the 'topographies and histories' of our datasets so that we can better understand the circumstances of their creation and generate an appreciation of their almost 'magical' qualities (Bowker and Star 1999, 9) - their capacity to perform amazing interpretative feats. Secondly, we contend that in some circumstances (and in particular where very large datasets are concerned) it is vital that we treat archaeological data 'as if' they are accurate and develop pragmatic strategies for doing so in a way that is also sensitive to their complexities. As Boyd and Crawford note, it is vital that researchers engaging with large datasets retain a level of critical awareness (2012, 666). Indeed as Levi importantly points out, humanities researchers are particularly adept at approaching their data sensitively and with humility; an approach that she feels scientific researchers could learn from (2013, 36).

The remainder of this paper makes a start in this vein, drawing on data collected for the purposes of the EngLaID project. There is not scope in this context to undertake detailed historical work of the kind advocated by Bowker and Star (1999) and Newman (2011). Such an investigation, building on the approach developed by Cooper (2013, Chapter 6) in examining social aspects of research practices in British prehistory over a 35-year period, will form another output of the EngLaID project. Below, we focus primarily on exploring a different set of relationships that are also vital to understanding the makeup of our dataset - those between key datasets in English archaeology. Additionally, we offer some pragmatic strategies that we have developed for dealing with these data 'as if' they are accurate, that are potentially useful much more broadly. As Spielmann and Kintigh (2011) and Atici et al. (2013) point out, advocates of major data integration initiatives in archaeology face immense social and technological challenges. It is perhaps for this reason that despite over 15 years of work in this area, such initiatives have yet to make a marked impact on broader research practices. In this context, it seems particularly important that archaeologists develop interim measures for handling incongruent amalgamated data in order that the interpretative potential (and limitations) of such assemblages can at least be explored.

Exploring the topography of datasets in English archaeology

One way in which EngLaID project researchers have begun to explore what we might call the topography of information infrastructures in English archaeology - their qualities, substance, limits, dynamics etc. - is by investigating in detail the character of connections and disconnections that exist between key datasets. Given the recent interest in analysing archaeological data (see above), it is perhaps surprising that, with two exceptions, previous studies have focused largely on single datasets. As noted above, Roskams and Whyman (2007) do develop a method for integrating an array of records pertaining to archaeology in Yorkshire, and highlight various challenges that arose during their attempts to do so. Their emphasis, however, is necessarily upon building a strategy for dataset integration - on creating a single, more coherent user interface, and on defining an additional tier of broad

interpretative categories in order to unite the specific descriptive categories used in separate datasets. In this context, the subtleties of the links between the datasets they employ are understandably somewhat overlooked. Since the relationship between events databases (primarily the AIP and the EI) is the focus of Evans' recent study (2013), the analysis below investigates the nature of connections between other key datasets in English archaeology - HER datasets, the NRHE, the AIP and the PAS/EMC. It is only by establishing these relationships that it is possible to move on and develop methods for using these data pragmatically.

In order to explore this issue, thirty-five 100 square km test areas (10km by 10km squares where possible) were selected; one for each HER represented within the EngLaID project's fourteen case study areas (Figure 1). For each test area, project researchers looked manually for connections/overlaps between HER records and records in each of the other three datasets in turn (the PAS/EMC, AIP, NRHE, Figure 2). Where records in different datasets clearly represented the same archaeological entity (either partially or completely), the reference numbers were noted (see Appendix 3 for further methodological details). No links were identified exclusively between datasets beyond HERs (e.g. between PAS/EMC and NRHE records). Very occasionally, however, the same archaeological entity was represented in three separate datasets (an HER, an AIP, and an NRHE record). Some overlaps were also identified within single datasets (e.g. between NRHE polygon and point-based records). During this exercise, further specificities of the datasets concerned also, unsurprisingly, came to light.

Figure 1: Map of EngLaID case study areas (outlined in black) showing 100 sq. km test areas (grids within case study areas).

Figure 2: Example 100 sq.km test area, showing all records from the AIP, HER, PAS / EMC and NRHE.

Based on the findings of this experiment, the following points are noteworthy. Firstly, and most importantly, it is possible for the first time to generate proxy numerical distributions for the degree of connection or overlap between the foremost datasets in English archaeology (Figure 3). Data provided for the EngLaID project in 2012 for these thirty-five test areas suggest that a typical interconnectivity of 50-80% exists between HER and NHRE records, of 0-20% between HER and PAS/EMC records, and of 10-40% between HER and AIP records. It is also worth noting that the interconnectivity of these datasets varies widely in different parts of the country. For the test square on the Isle of Wight the respective figures are 87% (NRHE), 1% (PAS/EMC) and 19% (AIP), whereas for the Northamptonshire test square, they are 76% (NRHE), 90% (PAS/EMC) and 19% (AIP). Some of the variations observed in different test areas can be very easily explained. PAS data, for example, is either almost entirely included within the HER database for a given area or is otherwise included only very selectively (this probably represents primarily the inclusion of more substantial sites, such as hoards). Variations in the % overlap between HER and NRHE datasets across England may relate to differences in the geographical distribution of antiquarian activity (the NRHE certainly appears to provide the fullest record of 19th and early 20th century discoveries) or to varying relationships between HER and English Heritage professionals (in this case the

strength of personal relationships may have either encouraged or inhibited the exchange of information). Other differences undoubtedly relate to a spectrum of specific circumstances that require further detailed investigation beyond the scope of the exercise outlined here. Such details may well come to light during our broader proposed investigation into the histories of English archaeological datasets (see above).

Figure 3: Box and whisker plots (a form of statistical representation which shows differences between populations without making assumptions about their underlying statistical distribution) showing the numerical distribution of percentage overlaps between the 35 different individual sets of HER records and: NRHE records, AIP records, and PAS / EMC records[2]. The circles show actual data values for each test area.

While not wishing to dwell on this topic (since this would simply add to the slightly disabling aura of disjointedness and inadequacy developed in previous datasets critiques), it is also worth mentioning at this point some of the details of the relationships that were established between records in these four datasets. For example, a connection might be made between a PAS record of an 'Iron Age coin' (for which details including the date range of issue, the mint, the ruler, the materials employed etc. are also provided) and an HER record of an 'Iron Age findspot' (assigned to the entire Iron Age period, and with its identity as a coin revealed only within a descriptive field). Another link might be established between a NRHE record of a 'later prehistoric earthwork' and 'Iron Age findspot' and an HER record of an 'Iron Age hillfort'. In each case, although the same archaeological entity is represented, the records it has become in disparate datasets differ in structure and substance. Alternatively, a connection might be made between an HER record of a 'Bronze Age round barrow' and an NRHE record of a 'later prehistoric cairnfield'. In this case, the same archaeological entity is almost certainly represented in both records. In one dataset, however, the archaeological entity has an entire record to represent it, whereas in the other, the relevant record denotes a broader group of similar phenomena. Accordingly, the records themselves are not directly equivalent. Significantly in the context of recent discussions about creating a cyber-infrastructure for archaeology, in each of these cases it would have been very difficult to link the records in different datasets via automated digital processes. The connections between these records could only be made by undertaking painstaking manual explorations.

Overall, the main findings of this experiment are that each of the datasets in English archaeology is substantially partial in its makeup. While this has long been suspected amongst archaeological practitioners, it has never previously been quantified to this extent. This underlines the mutual dependency of the datasets concerned (in other words, any study based on a single dataset inevitably would be partial) and the requirement that, in order to develop interpretations based on the widest possible array of information, archaeological researchers should create ways of handling this complexity (see below).

Additionally, it is clear from this investigation that difficult judgements must be made during any attempt to integrate records within different datasets that represent the same

[2] Percentages recorded where there were less than ten records of a particular dataset within a 100 square km test area (often the case with NHRE line data) are discounted as statistically irrelevant and so are not represented.

archaeological entity. At one level the process of integrating overlapping records gives researchers access to a broader range of information and makes this information simpler to use. At another level, however, this process adds further interpretative complexity to the data (see Yarrow 2006 for a discussion of how reconfigurations of archaeological data can both reveal new aspects of these data yet also elide others). By merging equivalent records from disparate datasets (for instance an HER record of an 'Iron Age hillfort' and an NRHE record of a 'later prehistoric earthwork' and 'Iron Age findspot'), particularities of the separate records must either be obscured (by choosing one interpretation and level of detail over another) or otherwise incorporated (by adding all potential interpretations and levels of detail), in which case superfluous information will be brought into consideration.

It can, therefore, be contended that the process of data integration raises important issues of trust - if one interpretation is chosen over another, which record is the most faithful representation of the archaeological entity at hand? If we combine interpretations, can we actually trust the hybrid outcomes? Our initial probing of this point suggests that the process of melding data from different datasets in fact creates a different form of interpretative doubt. Using the example provided above, the archaeological entity under consideration becomes a 'later prehistoric/Iron Age' 'earthwork/findspot/hillfort'. While, of course, it can be all of these things, such openness to interpretative disparity evidently brings with it compromises in terms of interpretative clarity that are not always helpful. Discussing a scientific research context (ecology) Millerand and Bowker contend that the recent blossoming of attempts to improve the interoperability of datasets produced in different milieux requires that researchers trust data produced by others (2009, 150). Whether this will also be the case for archaeologists, as they begin to work in a similar vein, remains to be seen. In this respect, however, it is notable that although the need to consider issues of confidentiality and trust was raised fairly early on by proponents of major data sharing initiatives in archaeology (Snow et al. 2006, 959), Atici et al. suggest that even very recently, and even among the pioneering zooarchaeological community (in terms of their involvement in data sharing ventures), data sharing was still relatively rare (2013, 664).

Using data 'as if' they are accurate

Inspired by the practices of information managers working in the medical arena (Bowker and Star 1999) together with researchers that have engaged much more broadly in working with very large datasets (see papers in Gitelman 2013), we will turn now to consider how strategies might be developed for employing intricate archaeological data 'as if' they are accurate. As noted above, such strategies are particularly pertinent in situations where, in relation to the sheer scale of analysis, researchers' capacities to 'improve' data are limited. Again work undertaken in the context of the EngLaID project provides a useful example for practitioners much more broadly. In order to use data from a range of English archaeological datasets 'as if' they were accurate, one important step was to develop ways of integrating records from the overlapping datasets to form a working research platform.

It should be mentioned at this point that Roskam and Whyman's (2007) integration of archaeological records from disparate datasets in Yorkshire could in some ways be viewed as a groundbreaking attempt to use English archaeological data 'as if' they are accurate. The approach outlined below, however, differs from theirs in two important respects. Firstly, the EngLaID project is operating at a substantially different scale to Roskams and Whyman: we are dealing with almost 1,000,000 records, compared to the 45,000 records that they collated (ibid.). For this reason, innovative automated processes play a more substantial role in the EngLaID project methods. Secondly, the approach outlined here tackles the tricky issue of how to handle multiple and varied representations of the same archaeological entity within different datasets. It is not entirely clear how Roskams and Whyman dealt with this matter in their own study. Indeed, as mentioned above, the issue of how to tackle dataset overlaps is certainly not prominent in wider discussions about data integration in archaeology.

In practice, we have applied two different data integration methods: one for smaller case studies (involving up to 8,000 records), where it is practical (in terms of time taken) to build links through the application of human interpretation between equivalent objects from different data sources, and one for our national-level survey (and also for some of our larger case studies involving up to 50,000 records), where building such links would be overly time-consuming. For the former case, the methodology briefly outlined above for 100 square km areas was expanded out to the whole case study area, with the relevant researcher going through the datasets manually in geographical information system (GIS) software and looking for equivalent records. Unique identifier codes were noted in a spreadsheet for each equivalence discovered. These were then imported into our database as a relation between two different records. As some HERs record relationships between records within their database internally, our database already featured the functionality required to represent these links. Whilst exploring data in the database, it is then simple to switch between a record and its equivalent(s) from other data sources. In order to avoid putting too many duplicate 'dots on the map' when exporting material for these particular case studies from the database into our project GIS, the export script written treats one record as the primary source and appends data from other records to it when creating the GIS-compatible computer files. As the dataset presumed (on the whole) to have greatest detail, any HER record is preferred. If no HER record exists, then other data sources are given a ranking of preference based primarily upon spatial precision (i.e. a PAS record is preferred over an equivalent EMC record as EMC records are only ever located to the nearest kilometre square).

For larger case studies and our national survey, going through every dataset and building links between equivalent records was not possible. It is a very involved and time-consuming task that cannot easily be automated due to several factors - inconsistencies in spatial coordinates/precision being the most obvious, but also different choices made in classification, as discussed above. As such, we had to come up with a more easily and robustly automated method. On this basis, and in agreement with Ell's (2010) recent assertion that GIS can play an important and productive role in managing large complex archaeological datasets, we decided to work with GIS 'spatial binning' techniques. Such techniques are fairly commonly used beyond archaeology, for instance in handling environmental ecology data.

They have also been employed to a limited extent within archaeology (see for instance Goodman and Piro 2013, Chapter 4; Boldrini 2006). Essentially, spatial binning involves applying a (relatively) coarse gridded mask across a series of data points and recording the presence/absence (of evidence) of objects of a particular period and type for each cell in the mask. This means that if a cell in the mask contains, for example, five records of Bronze Age barrows across three different datasets, this is simply recorded as the presence of the type 'barrow' for the period 'Bronze Age' within this cell (as we have no way of knowing without human consideration of the source data if these are truly five different barrows or whether some of the records refer to the same object, Figure 4). The form of the gridded mask can vary and we have experimented with several different types. The only requirement is that the cells are of a regular shape and form a complete tessellation across the area of interest, so they essentially have to be triangles, squares or hexagons. Squares and hexagons were considered to be most suitable for the purposes of this project, with both shapes having their own advantages. Squares are simple, but people tend to see false linear alignments when they are used. Hexagons are more complex to implement, but produce a visually more natural result. We have experimented with both types of tessellation at various spatial resolutions (e.g. 1km by 1km and 2km by 2km squares, hexagons of 3km and 5km span), with neither being given an absolute preference in the final analysis (the choice being dependent upon the spatial scale of any intended output, see below).

Figure 4: Spatial binning as a method of automated dataset integration.

The task of applying the mask chosen to a series of datasets is relatively simple to implement using GIS software (ArcGIS) and programming languages (Python). However, the major difficulty encountered was the large number of variant terms for sites of similar function used across our datasets, even taking into account the nigh-universal use of the EH monument type thesaurus (see Prescott 2013 for a discussion of how the operation of uniform categorisation schema - in this instance involving bibliographic records - can offer a guise of homogeneity that masks considerable heterogeneity in terms of how people use the schema). The same applies to terms used for periods, which are also often represented as start and end dates. As a result, we came up with a simplification thesaurus which grouped all the monument type terms encountered within our datasets into eight broad classes (e.g. 'agriculture and subsistence' or 'domestic and civil'). Each of these can then possess up to 26 sub-types (e.g. 'coaxial field system' or 'granary'). We also devised a simplification thesaurus for find types, with twenty simplified terms (e.g. 'coin' or 'tool'), some of which are allowed to overlap (e.g. 'axe' falls within both 'weapon' and 'tool'). Currently, dating evidence is simplified down to unspecified prehistoric, Bronze Age, Iron Age, Roman, early medieval, or uncertain, although we will improve where possible the temporal precision used over the lifetime of our project (see Green 2012 for more detail on this methodology).

The application of this spatial binning method then makes it quick to create distribution maps of any of our defined monument or find types for each of our six defined broad periods of interest (including undated). This can then be extended further analytically by testing different distributions against each other or against other geographic phenomena (such as average

elevation, etc.). The method involves a certain degree of loss of spatial resolution, but this is minimised by selecting a mask with an appropriate size of cell to the scale of analysis: if studying data at a spatial scale of 1:3,000,000 (at which England fits within the bounds of a sheet of A4 paper), differences of less than 1km are very hard to distinguish and, as such, 1km by 1km grid cells would be entirely appropriate. This means that the results mapped would be visibly almost indistinguishable from those achieved through the much more time-consuming task of human testing of object equivalence. Naturally, as the spatial scale examined becomes narrower, the resolution of mask needed becomes finer. Inevitably, there comes a point where equivalent objects become likely to fall in different cells and thus appear as duplicate objects. This is when we would switch to manual testing of equivalence, as discussed above, as this then becomes practical in terms of time spent. The exact point of this cut-off varies according to the spatial precision of coordinates recorded and the spatial extent of the features being studied and so can only be defined on a case-by-case basis. For English archaeology, this would be the nearest kilometre square at the coarsest (as the vast majority of coordinates possess at least this level of spatial precision and the vast majority of sites are considerably less than this area of extent) and, in most circumstances using finer resolutions would be entirely justifiable.

A full outline of the findings of our work in this vein is beyond the scope of this paper. However our initial results suggest that in many cases there is little noticeable difference at the broader spatial scales between analyses based upon human-cleaned and computer-processed data, albeit that analysis based on the latter can only be practically undertaken using relatively coarse periods/types and with some loss of spatial finesse. It is also worth highlighting that relying upon the categorisations produced by other bodies (i.e. treating data 'as if' it is accurate) does work better in some cases than in others. To give an example of a distribution map that appears somewhat unrealistic, in Cornwall - where the local authority team is particularly research active and confident in its interpretative approach - HER professionals have recently assigned all 'ridge and furrow' field systems to the early medieval period. This is not the case for most of the rest of the country. The result is that maps created of early medieval 'agriculture and subsistence' using our spatial binning method show a great density of instances in Cornwall, with the modern county boundary appearing as a very obvious step-change in the distribution (Figure 5). On the whole, however, such differences should occur at a county level and, as such, should be obvious once mapped, due to being constrained by county boundaries.

For a less contentious monument category - perhaps Bronze Age round barrows, Iron Age hillforts, or Roman villas - the distributions look much more realistic and compare well with plots produced by researchers working with a narrower range of data who have 'improved' their data to a substantial degree, for instance those presented in Taylor 2007 (Figure 6). Indeed for case study areas where we have undertaken lengthy manual data cleaning for the purposes of undertaking fine-grained analysis, it is possible to test the reliability of these computer-processed data further. When manually 'cleaned' individual records relating to Bronze Age funerary monuments are plotted against computer-processed records relating to the same category 'binned' within 1km x 1km squares, the results compare very well (Figure

7). Excluding those squares that overlap partially the county boundary (where the record may thus have been in the neighbouring county), only four uncleaned cells (out of 221) have no overlap with cleaned records and only three new records have been added through the cleaning process (all of which are on the edges of uncleaned cells in any case). It is also worth emphasising that differences of this degree would not be evident at all if this data category was plotted at a national scale.

Figure 5: Map showing Kernel Density Estimate surface of early medieval monuments relating to 'agriculture and subsistence' (10km radius, based upon 1x1km grid centroids), showing the effects of differences in categorisation practices: specifically the obvious step change at the border between Cornwall and Devon.

Figure 6: Taylor's map of Roman villas (displayed as black open circles) overlaid upon a shaded density map of Roman villa records from the EngLaID database (10km radius Kernel Density Estimate plot, collated by and based upon 1x1km grid centroids). There are differences in specific detail, but the overall pattern is very similar, with the greatest densities of circles being coincident with the highest (darkest) values on our density map (adapted from Taylor 2007, Fig. 4.9).

Figure 7: Map showing manually-cleaned barrow records for Cambridgeshire within our East of England case study (white circles) plotted over presence / absence 1x1km spatial bins for uncleaned / computer processed barrow data (grey squares).

On the whole, therefore, the method seems relatively robust with problematic results being very obvious since the patterning produced conforms to HER boundaries (as previously stated). It is worth noting, however, that one interesting outcome of our work in this area so far is that treating data 'as if' they are accurate can sometimes be as revealing about archaeological categorisation practices as it is about the spatial patterning of past practices. One productive consequence of this finding is that where we have reached interpretative limits in trying to map certain data categories, this has prompted us to consider ways in which we might move beyond (geographically-fixed) traditional forms of representation for archaeological data, such as distribution maps. As an example of this (albeit on a small scale), we have undertaken work in which we analysed the various prehistoric / Roman field systems on Salisbury Plain, Wiltshire, in which re-plotting a geographically distributed series of field system boundaries from a single fixed origin enabled us to create graphs demonstrating the degree of 'coaxiality' of each sector of field system (Figure 8). In this way, precise spatial location becomes less important than spatial character more broadly as a fruitful route into archaeological interpretation. Importantly, although attempts to map overall distributions of field systems of any date within our study period at a national level have proved to be perplexing, by approaching and visualising these data differently, it is still possible to generate interpretatively interesting results.

Figure 8: Graphic visualisation of field system bank and ditch alignments at Figheldean on Salisbury Plain, Wiltshire. The thick black line shows the orientation of the coaxial field system as defined by previous researchers (McOmish et al. 2002, Fig 3.4). The thin black lines show the length and bearing of each boundary feature. The shading shows the density of these boundaries in 15˚ / 100 metre slices. The strong clustering of boundaries around the previously defined orientation and perpendicular to it verifies clearly the coaxial character of this field system.

Discussion

Recent studies of digital archaeological datasets have revealed that these datasets are 'characterful' - they have diverse histories, contents and structures and are riddled with gaps, inconsistencies and uncertainties. Due to these qualities, data curators have worked hard over many years - particularly in recent years as the extent of these peculiarities has become clearer - to improve their datasets via a range of mechanisms (standards, policy documents, software packages, professional forums etc.). Ultimately, it is hoped, these strategies will help to make the datasets more accurate in their representations of past entities, more coherent, more interoperable and, importantly, more useful for the various functions they fulfil (including employment in archaeological research). In England, some recent analyses have also appreciated that the characterfulness of archaeological datasets can be of value in itself. They can be considered as objects (or materials) of study in their own right, yielding insights into shifts in archaeological approaches and working practices, sometimes over very long time periods. This appreciation importantly reconnects recent examinations of digital archaeological data with insights made by late 20th century theorists regarding the contingent character of archaeological (and much broader) knowledge production practices.

In this paper, we have taken these arguments further and also added some important context drawing on broader studies of information infrastructures and categorisation practices. We argue that once the emphasis is placed more strongly on the history and relationships these datasets entail - the many people, organisations, technologies and materials involved in their creation and maintenance, sometimes over very long time periods - it is hardly surprising that they are complex (both individually and collectively) and that at times the relationships between the data and the past entities they represent seem ambiguous. Importantly examinations of information infrastructures elsewhere also show very clearly that archaeologists are not alone in their experiences of what Edwards (2010) has termed 'data friction' - anxieties and arguments over the quality, consistency and effectiveness of data - and what Ribe and Jackson (2013) have described as 'ontological choreography' - the manoeuvering that is required to retain data consistency in constantly shifting research milieux. One recent science and technology study recounts the struggles that a small team of ecologists in Baltimore had in maintaining consistency in their records of stream chemistry over a mere 16-year period (Ribe and Jackson 2013). Medical practitioners have puzzled over the incompatibility of their datasets at a national level and beyond, and the incredible capacities of data managers to interpret in diverse ways strategies aimed at greater coherence (Bowker and Star 1999). Meanwhile humanities researchers have exposed the unruly character of even seemingly consistent bibliographic records, shaped as they are by layers of 'intervention and disruption' over extended time periods (Prescott 2013, 57). Consequently it seems very important that we look beyond disciplinary and national confines, not only in order to find models and technologies that might assist with attempts to create a cyber-infrastructure for archaeology (Kintigh 2006, 575; Snow et al. 2006, 959) but also as a means of contextualising the data issues we are experiencing.

At least three important outcomes arise from our analysis. Firstly, it is important to remember that whatever their idiosyncrasies, archaeological datasets can perform amazing work in terms of helping us to build an understanding of the past practices we are interested in - there is little point in simply giving up because the evidence that we use in our research is 'subjective'. Recent attempts to encourage the creation of detailed metadata (in order to record and get a handle on such subjectivities), and to develop ontologies and software that enable the linking of diverse datasets, represent positive steps in this vein (Kintigh 2006, Snow et al. 2006, Dam et al. 2010, although see Levi 2013 for a discussion of the potentially unwanted effects of 'taming' digital data via the creation of metadata).

Secondly, it is vital to continue with our quest to understand our datasets better in their own right - probing not only their flaws, but also the social and historical aspects of which they are formed, and the extent to which they also form us as research communities (cf. Ribes and Jackson 2013). In this respect, it is worth commenting further on the contention noted above that the complex qualities of English archaeological datasets are symptomatic of a socially fragmented discipline (e.g. Evans 2013). Cooper's study of research practices in British prehistory suggests that while on the one hand, British archaeology has for a long time been rendered as being chronically fragmented, alongside this view, it is described by practitioners as being incredibly close-knit and even incestuous (2013, Chapter 9). One explanation for this seemingly paradoxical understanding is that, in practice, archaeologists are able to move freely between institutional contexts over the course of their lifetimes and even to identify themselves in relation to different archaeological bodies at one time (for example a local archaeology group and a university department). In this way, she argues, archaeologists are able to bridge perceived disciplinary fractures and to curb the potentially disabling effects of such disjunctures. Given this situation, it is perhaps surprising that English archaeological datasets do not show greater levels of affinity. Indeed the modest dataset correspondence demonstrated above could suggest that, although we might think of digital archaeological data as being much more mobile and tractable than the people who create, tend, and work with them, in some ways these data are more firmly fixed to their institutional settings than we imagine. This may be a further dimension of the inertia of information infrastructures that Bowker and Star describe (1999, 76). More broadly, it is worth bearing in mind that we need to be careful about making straightforward assumptions about how archaeological datasets relate to the wider research communities of which they form part (see also Boyd and Crawford 2012).

Thirdly, given the acknowledged characterfulness of many digital archaeological datasets, it is essential that researchers develop methods for making data work for them at a range of different analytical levels. At one level this means sustaining the existing (and admirable) efforts of archaeological data curators and software specialists to enhance the faithfulness, unity and interoperability of digital datasets both individually and collectively. For researchers working at a local or regional scale or examining particular sets of data from a specific time period (a single archaeological period or period transition), the process of creating an effective dataset can involve many hours of meticulous work, editing the data they are employing to a level at which they are satisfied with them. As Boyd and Crawford (2012,

670) and Levi (2013, 35) point out, fine-grained studies are a vital counterpart to work that operates at much larger scales. Alternatively, as several recent larger national-scale research projects in England have done very successfully, this process might sometimes require working primarily with the 'best' available data - those derived from recent fieldwork projects and held in 'grey literature', which have the greatest chronological resolution, offer the highest level of detail, and are most easily cross-comparable (see for example the impressive work currently being undertaken by the Rural Settlement of Roman Britain project based at the University of Reading http://www.reading.ac.uk/archaeology/research/roman-rural-settlement/). By taking a selective approach, the size of the dataset under consideration is reduced to a level at which it is still feasible to make the improvements that are seen to be required. In future, once some of the many challenges that face major data sharing and integration initiatives are resolved (Spielmann and Kintigh 2011, 24), it may even be possible to transform incongruent datasets of varying qualities much more effortlessly into coherent research platforms.

The English Landscape and Identities project is taking a rather different, more immediate, and arguably quite bold, approach to this issue. Acknowledging the importance of at least trying to undertake analyses at scales beyond those mentioned above, we are asking what we *can* say about past practices over a very long (c. 2600-year) time period when we bring together digital data from as many as possible complex English archaeological datasets to beyond a level at which it is feasible to 'improve' these data other than in a very limited way. This approach chimes with, or at least tends towards, the much broader movement, both in scientific disciplines, and more recently in the humanities (see for instance Levi 2013; Prescott 2013; and the English National archives Traces through Time project http://nationalarchives.gov.uk/about/traces-through-time.htm), where disparate datasets are increasingly being interconnected over wide areas in what are termed 'big data' projects. It is worth emphasising at this point that big data is a relative term, and that in relation to this, understandably, definitions of this term are hard to pin down. Scientific and even some recent humanities big data applications operate at a scale that is in an entirely different league to that being employed on the EngLaID project. The Sloan Digital Sky Survey (http://www.sdss.org/) had amassed over 140 terabytes (TB) of data even by 2010 (The Economist 2010). Family Search servers in the US handle between twenty and fifty TB of information a day (Tilbury 2013 https://indico.cern.ch/conferenceOtherViews.py?view=standard&confId=246453). Meanwhile at the time of writing, the main database constructed by the EngLaID project amounts to just over three gigabytes (GB) (alongside over 100GB of GIS data). However, in their critical consideration of the big data movement at a wide level, Boyd and Crawford suggest that big data 'is less about data that is [*sic*] big than it is about a capacity to search, aggregate and cross-reference large datasets' (2012, 663). Levi (2013, 33) adds that the development of novel data-processing methods is a further defining feature of this mode of practice.

Certainly, in line with big data projects more broadly, the volume and intricacies of the data assembled by EngLaID project researchers have at times required us to relinquish some hold

on our academic ideals of 'consistency, coherence and accuracy' (Evans 2013; see also Weinberger 2012). In this sense, we have had to move beyond the concerns over data quality voiced by other researchers who have engaged with the possibility of working with other peoples' data (e.g. Atici et al. 2013). Similarly, working in this way has necessitated that we test out and develop the methods and emerging technologies that can enable both us, and archaeologists more broadly to work at this scale. As Gitelman and Jackson point out 'the temporary era of big data has been enabled by the widespread availability of electronic storage media' (2013, 6). They also highlight the importance of nurturing our capacities to deal with the accumulating mass of data: 'statisticians are on track to be the next sexy profession in the digital economy' (ibid., 3). In this sense, it could be argued that we, like the data managers in the medical profession that Bowker and Star (1999) studied, are seeking to operate as 'future forecasters' - we are keeping in mind how archaeological data might be used beyond the lifespan of our project. It is also vital to stress that, in pursuing this investigative aim, we are necessarily raising fundamental questions regarding the control of archaeological data. If we seek, as an research community, to continue to work in this vein (and we feel that it is very important that we do), it will require that we revisit our approaches towards data accessibility: to consider in detail what the politics and etiquettes are of connecting, employing and making available archaeological data on an unprecedented scale.

During our initial work in this broad, inclusive and experimental manner, one key task, as outlined above, has been to try to understand better how the various characterful datasets we seek to interpret can work together. On this basis, we are developing approaches that allow us to employ them effectively in unison in advance of a situation aspired to by proponents of a cyber-infrastructure for archaeology in which new technologies will be able to do this work for us. Perhaps more innovatively, and more in the broad spirit of researchers working with huge and diverse datasets beyond archaeology, we are thinking through pragmatic ways of using our assembled mass of digital data, with all its foibles, 'as if' it is accurate. The results of our work thus far are preliminary. Based on this initial research, however, one important finding is that working with data 'as if' they are accurate can sometimes be as informative about archaeological categorisation practices as it is about the past practices we seek eventually to illuminate: an intriguing outcome in itself. Interestingly, an early attempt to interpret integrated zooarchaeological data from the tDAR database in the US reached similar conclusions (Spielmann and Kintigh 2011, 23). Indeed researchers in the humanities more broadly have suggested that analytics designed for interpreting scientific big data might be employed fruitfully as a means of exploring and exposing the complexities of humanities big data as well as for interpreting these data (Prescott 2013, 57). It has also become evident to us that working in this way lends itself most effectively to analysing certain (less ambiguous) data categories, and will undoubtedly lead us to ask a different (perhaps bigger and simpler) set of research questions to those accessible to researchers using smaller, more refined datasets (see Amorosi et al. 1996; Boyd and Crawford 2012 for broader discussions of this issue). Perhaps most productively, our exposure of limits in the capacities of traditional modes of data representation (e.g. distribution maps) when working in this way, has also encouraged us to explore fresh forms of digital visualisation that are often not statistically viable or relevant when working with smaller datasets.

We certainly intend that over the next two years we are able to use our dataset to produce engaging histories of the English landscape from 1500BC to AD1086. As we have hopefully shown, however, the EngLaID project also has significant potential to offer fresh insights into the character of digital archaeological data, and to develop new ways of working with these data. These findings have repercussions much more broadly and, while to a certain extent unforeseen, they are seemingly a very rewarding upshot of the project.

Acknowledgements

This study was carried out as part of a five-year European Research Council funded research project. It also draws on the findings of work undertaken separately as part of an English Heritage-commissioned investigation aimed at developing a new information access strategy for England. The data upon which the study is based were provided by 75 separate HER Officers, English Heritage, the Archaeological Investigations Project and the Portable Antiquities Scheme. It goes without saying that our work would not have been possible without the support and expertise of the professionals involved in curating and extracting these data for us. We are particularly grateful to Sally Croft (Cambridgeshire HER), Simon Crutchley (English Heritage), Rebecca Loader (Isle of Wight HER), Dan Pett (PAS) and Emma Trevarthen (Cornwall HER) who kindly gave us permission to publish images of their data. Simon Crutchley, Martin Newman and Roger Thomas facilitated access to the NRHE data and have offered thoughtful guidance during our endeavour to get to grips with our various datasets. Ehren Milner gave us advice about accessing AIP data, and Dan Pett provided the PAS data. Letty ten Harkel, Zena Kamash and Laura Morley undertook the 100 square km test exercise along with us. Miranda Creswell, Duncan Garrow, Chris Gosden, Letty ten Harkel, Zena Kamash and Dan Stansbie provided helpful comments on an earlier draft of the paper. The input of four anonymous reviewers improved substantially the version of this paper that was originally submitted for publication.

References

Alberti, B. Jones, A. M., & Pollard, J. (2013). *Archaeology after interpretation: Returning materials to archaeological theory*. Walnut Creek: Left Coast Press.

Amorosi, T., Woollett, J., Perdikaris, S. & McGovern, T. (1996). Regional zooarchaeology and global change: Problems and potentials. *World Archaeology,* 28(1), 126-157.

Atici, L., Witcher Kansa, S., Lev-Tov, J. & Kansa, E. C. (2013). Other peoples' data: A demonstration of the imperative of publishing primary data. *Journal of Archaeological Method and Theory,* 20, 663-681.

Bawden, D. & Robinson, L. (2009). The dark side of information: Overload, anxiety and other paradoxes and pathologies. *Journal of Information Science,* 35, 180-191.

Benson, D. (1974). A Sites and Monuments Record for the Oxford region. *Oxoniensia,* 37, 226-237.

Boldrini, N. (2006). Planning uncertainty: Creating an artefact density index for North Yorkshire, England. *Internet Archaeology,* 21 http://dx.doi.org/10.11141/ia.21.1. Accessed 15 October 2014.

Bowker, G. & Star, L. (1999). *Sorting things out: Classification and its consequences*. Cambridge, MA: MIT Press.

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

Boyd, D. & Crawford, K. 2012. Critical questions for big data. *Communication and Society* 15(5), 662-679.

Callou, C., Baly, I., Gargominy, O. and Reib, E. (2011). National Inventory of Natural Heritage website. Recent, historical and archaeological data. *The SAA Archaeological Record,* 11(1), 37-40.

Clarke, D. L. (1968). *Analytical archaeology*. London: Methuen.

Connelly, W. (2011). *A world of becoming*. Durham, NC: Duke University Press.

Cooper, A. (2013). *Prehistory in practice: A multi-stranded analysis of British prehistoric research, 1975-2010*. British Archaeological Report, British Series 577. Oxford: Archaeopress.

Dam, C. and Hansen, H.J. (2005). The European digital resource in archaeology: Sites and monuments data as a common European web resource. *Internet Archaeology*, 18 http://dx.doi.org/10.11141/ia.18.4. Accessed 15 October 2014.

Dam, C., Austin, T. and Kenny, J. (2010). Breaking down national barriers: ARENA - a portal to European heritage information. In F. Niccolucci and H. Sorin (Eds.) *Beyond the artefact. Digital interpretation of the past. Proceedings of CAA 2004. Prato 13-17 April 2004* (pp. 94-98). Budapest: Archaeolingua.

Edgeworth, M. (2003). *Acts of discovery: An ethnography of archaeological practice*. British Archaeological Report International Series 1131. Oxford: Archaeopress.

Edwards, P. (2010). *A vast machine: Computer models, climate data, and the politics of global warming*. Cambridge, MA: MIT Press.

Ell, P.S. (2010). GIS, e-Science and the humanities grid. In D.J. Bodenhamer, J. Corrigan and T. M. Harris (Eds.) *The spatial humanities: GIS and the future of humanities scholarship* (pp. 143-166). Bloomington: Indiana University Press.

Evans, T. (2013). Holes in the archaeological record? A comparison of national event databases for the historic environment in England. *The Historic Environment: Policy & Practice*, 4, 19-34.

Gitelman, L. (Ed.) (2013). *"Raw data" is an oxymoron*. Cambridge, MA: MIT Press.

Gitelman, L. & Jackson, V. (2013). Introduction. In L. Gitelman, (Ed.) *"Raw data" is an oxymoron* (pp. 1-14). Cambridge, MA: MIT Press.

Gobalet, K. (2001). A critique of faunal analysis: Inconsistency among experts in blind analysis. *Journal of Archaeological Science*, 28, 377-386.

Goodman, D. & Piro, S. 2013. *GPR remote sensing in archaeology*. London: Springer.

Gosden, C., Kamash, Z., Kirkham, R. & Pybus, J. (2009). Joining the dots: Exploring technical and social issues in e-Science approaches to linking landscape and artefactual data in British archaeology (pp. 171-174). *E-Science workshops, 2009 5th IEEE International Conference*. Oxford: Institute for Electrical and Electronic Engineers.

Green, C. (2012). Archaeology in broad strokes: collating data for England from 1500 BC to AD 1086. In A. Chrysanthi, D. Wheatley, I. Romanowska, C. Papadopoulos, P. Murrieta-Flores, T. Sly, & G. Earl (eds.) *Archaeology in the Digital Era: Papers from the 40th Annual Conference of Computer Applications and Quantitative Methods in Archaeology (CAA), Southampton, 26-29 March 2012* (pp. 307-312). Amsterdam: Amsterdam University Press.

Hodder, I. (1984). Archaeology in 1984. *Antiquity*, 58(222), 25-32.

Hodder, I. (1986). *Reading the past: Current approaches to interpretation in archaeology*. Cambridge: Cambridge University Press.

Holbrook, N. & Morton, R. (2011). Assessing the research potential of grey literature in the study of Roman England. Stage 1 report. Cotswold Archaeology. http://dx.doi.org/10.5284/1000368. Accessed 3 December 2013.

Kamash, Z., Cooper, A., Green, C., ten Harkel, L., Morley, L. 2013. *Transregional research using national datasets*. Unpublished report. Oxford: Institute of Archaeology.

Kinory, J. L. (2012). *Salt production, distribution and use in the Britsh Iron Age*. British Archaeological Report British Series 559. Oxford: Archaeopress.

Kintigh, K. (2006). The promise and challenge of archaeological data integration. *American Antiquity*, 71(3), 567-578.

Latour, B. (1988). *The pasturisation of France*. Cambridge, MA: Harvard University Press.

Latour, B. (2000). Did Ramses II die of Tuberculosis? On the partial existence of existing and nonexisting objects. In L. Daston (Ed.) *Biographies of scientific objects* (pp. 247-269). Chicago: Chicago University Press.

Latour, B. (2005). *Reassembling the social*: *An introduction to Actor-Network-Theory*. Oxford: Oxford University Press.

Latour, B., Jensen, P., Venturini, T., Grauwin, S., & Boullier, D. (2012). 'The whole is always smaller than its parts': A digital test of Gabriel Tardes' monads. *British Journal of Sociology*, 63(4), 590-615.

Lee, E. (2012). 'Everything we know informs everything we do': A vision for historic environment sector knowledge and information management. *The Historic Environment*, 3(1), 28-41.

Levi, A.S. (2013). Humanities 'big data': Myths, challenges, and lessons. In *Big Data, 2013 IEEE International Conference Proceedings* (pp. 33-36). http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6691667&isnumber=6690588. Accessed 15 October 2014.

Levy, T.E., (2014). Editorial. *Near Eastern Archaeology*, 77 (3, special issue on Cyber-Archaeology).

Lucas, G. (2001). Destruction and the rhetoric of excavation. *Norwegian Archaeological Review*, 34(1), 35-46.

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

McOmish, D., Field, D. & Brown, G. (2002). *The field archaeology of the Salisbury Plain Training Area.* Swindon: English Heritage.

Mikkelsen, M. (2012). Development-led archaeology in Denmark. In L. Webley, M. Vander Linden, C. Haselgrove & R. Bradley (Eds.) *Development-led archaeology in northwest Europe* (pp. 117-127). Oxford: Oxbow.

Millerand, F. & Bowker, G. (2009). Metadata standards: Trajectories and enactment in the life of an ontology. In M. Lampland & S. Star (Eds.) *Standards and their stories: How quantifying, classifying and formalizing practices shape everyday life* (pp. 149-166). New York: Cornell University Press.

Musen, M. (1992). Dimensions of knowledge sharing and reuse. *Computers and Biomedical Research,* 25, 435-67.

Nature Editors. (2009). Data's shameful neglect: Research cannot flourish if data are not preserved and made accessible. All concerned must act accordingly. *Nature*, 461(7261), 145.

Newman, M. (2011). The database as material culture. Internet Archaeology, 29. http://intarch.ac.uk/journal/issue29/tag_index.html. Accessed 3 December 2013.

Onsrud, H. & Campbell, J. (2007). Big opportunities in access to "small science" data. *Data Science Journal,* 6(Open Data Issue), 58-66.

Patrik, L. E. (1985). 'Is there an archaeological record?' In M. B. Schiffer (ed), *Advances in archaeological method and theory*, (pp. 27-62). New York: Academic Press.

Prescott, A. (2013). Bibliographic records as humanities in big data. In *Big Data, 2013 IEEE International Conference Proceedings* (pp. 55-58). http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6691670&isnumber=6690588. Accessed 15 October 2014.

Ribes, D. & Jackson, V. (2013). Data bite man: the work of sustaining a long-term study. In L. Gitelman (Ed.) *"Raw data" is an oxymoron* (pp. 147-66). Cambridge, MA: MIT Press.

Robinson, B. (2000). English Sites and Monuments Records - information, communication and technology. In G. Lock & K. Brown (Eds.) *On the theory and practice of archaeological*

computing (pp. 89-106). Oxford University Committee for Archaeology Monograph 51. Oxford: Oxbow.

Roskams, S. & Whyman, M. (2007). Categorising the past: Lessons from the Archaeological Resource Assessment for Yorkshire. *Internet Archaeology* 23, http://intarch.ac.uk/journal/issue23/2/index.html. Accessed 3 December 2013.

Shanks, M. & Tilley, C. (1987). *Re-constructing archaeology: Theory and practice.* Cambridge: Cambridge University Press.

Snow, D., Gahegan, M., Giles, L., Hirth, K. Milner, G., Prasenjit, M., & Wang, J. (2006). Cybertools and archaeology. *Science,* 311, 958-959.

Spielmann, K. & Kintigh, K. (2011). The digital archaeological record: The potentials of archaeozoological data integration through tDAR. *The SAA Archaeological Record,* 11(1), 22-25.

Taylor J. (2007). *An atlas of Roman rural settlement in England*. London: CBA Research Report 151.

Tilley, C. (1998). Archaeology: The loss of isolation. *Antiquity,* 72, 691-93.

Wainwright, G. (1989). Management of the English landscape. In H. Cleere (Ed.) *Archaeological heritage management in the modern world* (pp. 164-170). London: Council for British Archaeology.

Weinberger, D. (2012). *Too big to know: Rethinking knowledge now that the facts aren't the facts, experts are everywhere, and the smartest person in the room is the room*. New York: Basic Books.

Worrell, S., Egan, G., Naylor, J., Leahy, K. & Lewis, M. (Eds.) (2010). *A decade of discovery: Proceedings of the Portable Antiquities Scheme Conference 2007*. British Archaeological Reports, British Series 520. Oxford: Archaeopress.

Wylie, A. (1985). Putting shakertown back together: Critical theory in archaeology. *Journal of Anthropological Archaeology,* 4, 133-47.

Yarrow, T. (2003). Artefactual persons: The relational capacities of persons and things in the practice of excavation. *Norwegian Archaeological Review,* 36(1), 65-73.

Yarrow, T. (2006). Perspective matters: traversing scale through archaeological practice. In G. Lock & B. Molyneaux (Eds.) *Confronting scale in archaeology: Issues of theory and practice* (pp. 77-87). New York: Springer.

Yarrow, T. (2012). Not knowing as knowledge: asymmetry between archaeology and anthropology. In D. Garrow & T. Yarrow (Eds.), *Archaeology and anthropology: understanding similarities, exploring differences* (pp. 13-27). Oxford: Oxbow.

Appendix 1: URLs for cited datasets

NRHE: http://www.pastscape.org.uk/
NMP: http://www.english-heritage.org.uk/professional/research/landscapes-and-areas/national-mapping-programme/
HERs: http://www.heritagegateway.org.uk/gateway/
AIP: http://csweb.bournemouth.ac.uk/aip/aipintro.htm
OASIS: http://archaeologydataservice.ac.uk/archsearch/
PAS: http://finds.org.uk/
EMC: http://www.fitzmuseum.cam.ac.uk/dept/coins/emc/

Appendix 2: Specificities of the datasets provided to the EngLaID project

HER data (including several UADs)
75 of the 84 HERs and UADs in England provided data for the EngLaID project. Most of these datasets use a monument and event structure, with events associated with a monument linked to it and, presumably, with events not linked to an existing monument generating a new monument upon their taking place. Nottinghamshire is an exception in having an intervening third layer, which can be thought of as a "feature". Essentially, in that case, events are linked to features and then features are linked to monuments where a pre-existing site is known of or once they become important enough to merit consideration as a "monument". The majority, but not all, HERs also record sources/bibliography. Many record finds details, especially amongst HBSMR users. However, such details tend to be added on an ad hoc basis.

NRHE
NRHE data were supplied by English Heritage as shapefiles and associated PDF documents that contained the most important attributes of each record. These PDF files had to be scanned using a script to extract the relevant attribute data. This process is slightly imperfect due to some monument types running across multiple lines, which makes it impossible for any automated (digital) process to tell when a term finishes. As such, the resulting output results in a "stream of consciousness" list of monument types for each period, with one running into another, e.g. Roman: VILLA BATHHOUSE BARN ROUND HOUSE. This makes some queries hard to perform, but is largely functional.

AIP
AIP data was downloaded by EngLaID from the AIP website. This is not entirely ideal due to a restriction in the database software used by the AIP, which means that it is not possible to download more than a certain (undetermined) number of records at one time. Some categories of data are, thus, not extractable due to exceeding this limit even when filtered down to the greatest level.

PAS
PAS data were supplied directly by the PAS in the form of a single large CSV spreadsheet. PAS data maps reasonably well onto HER finds recording formats, but contains a lot more detail.

Appendix 3: Additional methodological details for the 100 square km exercise

Since, upon testing, a negligible degree of overlap between records held in the EMC and those within the PAS was identified, and since their structure and content is broadly compatible, these datasets are combined and treated as one.

100 square km test areas were selected so as to contain a reasonable, but not overly onerous number of records, so as to characterise as accurately as possible the nature of the dataset relationships. When tested specifically, almost all these squares included an above-average record density for the HER in question. If the HER area concerned included distinctly different landscape zones, which were likely to produce particular kinds of records (e.g. Norfolk, where the Fens produces very few PAS records, but the upland Breckland zone produces many) the 100 square km area was positioned, as far as possible to include both/all such zones. In two cases, West Berkshire and North East Lincolnshire, it was not possible to fit a full set of 100 1 km by 1 km cells within the case study area, so smaller areas were tested.

The findings of this exercise should be treated with some caution. All four EngLaID researchers involved in carrying out this exercise used the same broad methodology and at all times, we tried to maintain a consistent approach. However the challenges involved in identifying links between records in different datasets varied considerably from test area to test area. Consequently, researchers necessarily developed their own ways of dealing with the particular ambiguities they faced in conducting the test.