Promise and Paradox: Accessing Open Data in Archaeology

by Jeremy Huggett

Citation


Abstract

Increasing access to open data raises challenges, amongst the most important of which is the need to understand the context of the data that are delivered to the screen. Data are situated, contingent, and incomplete: they have histories which relate to their origins, their purpose, and their modification. These histories have implications for the subsequent appropriate use of those data but are rarely available to the data consumer. This paper argues that just as data need metadata to make them discoverable, so they also need provenance metadata as a means of seeking to capture their underlying theory-laden, purpose-laden and process-laden character.
Promise and Paradox: Accessing Open Data in Archaeology

by Jeremy Huggett

1. Introduction

Archaeology is currently well-served with free access to archaeological data via organisations such as the Archaeology Data Service (ADS)\(^1\) in the UK, tDAR\(^2\) and Open Context\(^3\) (USA), DANS\(^4\) (Netherlands), as well as national heritage organisations (for example, RCAHMS,\(^5\) English Heritage,\(^6\) etc.) and regional Historic Environment Records. There is little doubt that this accessibility has transformed the practice of archaeology.

This is despite the fact that much of this data is not truly 'open', with limits frequently placed on redistribution and reuse, for example. In part this relates to standard arguments about open data: the risk of reducing confidence and authority as a consequence of revealing discrepancies and errors in the data, for instance. One issue regularly raised in relation to archaeological data is that they frequently include spatial information which may risk facilitating looting (Bevan 2012a, 7-8). Hence it may be argued that spatial data should be degraded and full resolution data made available only to 'approved' users (as is the case with the Portable Antiquity Scheme database,\(^7\) for example). This paper, however, is less concerned with the content of the data or limits placed upon the use of the data and more with the consequences of access to the increasingly large amounts of data arriving on our computer screens.

2. Accessing Data

As an example of the kinds of problems which might arise, nationally available datasets are frequently in disagreement with each other. For instance, a search for medieval castles in Scotland using the ADS catalogue\(^8\) currently returns 39 records. This seems a very low figure, an impression which is confirmed by a search for castles in Canmore,\(^9\) the Scottish National Monuments Record held by the RCAHMS, which generates nearly 1500 records. This apparent discrepancy arises despite the ADS catalogue
including a copy of the RCAHMS NMR. What is a user to make of this? In part this is a consequence of the Scottish NMR generally not using dating terms because their meaning varies between Scotland and the rest of the UK ('iron age' or 'viking' refers to quite different periods of time in Scotland than in England, for instance). A decision was therefore taken years ago to not use such terms in order to avoid confusion or error, especially in those border areas which changed hands between the English and Scottish kingdoms. However, this fact is nowhere visible to a user accessing either dataset, which underlines the importance of understanding the context and background of the data themselves. Furthermore, removing 'medieval' from the search criteria used in the ADS catalogue and limiting the search to the Scottish NMR records results in the return of just over 1200 sites, leaving some 300 'missing' castles. Whatever the explanation for this, issues such as these raise concerns about accuracy and authority but may go un-noticed by the user.

Data comparability may be an issue in other respects. For example, accessing publicly available downloadable datasets relating to legally protected ancient monuments in Scotland\textsuperscript{10} and England\textsuperscript{11} is not difficult from their respective public websites; however, there is no direct data download available for Wales via the CADW website.\textsuperscript{12} The resulting datasets themselves are not comparable without further work — the attribute data accompanying the site locations differ in terms of level of detail, despite the fact that the datasets have been collected for the same purpose by organisations with the same legal responsibilities regarding such sites. Of course, Historic Scotland, English Heritage, and CADW make their data available on request, and also offer access via Web Map Services (WMS) through data.gov.uk, although the latter is not for the faint-hearted, requiring rather more expertise than downloading a spreadsheet or similar, and is less flexible in terms of what can subsequently be done with the data.

These examples illustrate just some of the problems that a user can encounter in accessing open data and these issues become all the more significant in a world of linked data and the semantic web. As data are brought together from different sources within semantic ontologies, how much information about the origins of that data is lost in the mapping process? When those linked data are drawn down from different resources, how much information about the transformations that have been applied between the different points of data collection and their final delivery to the screen is accessible to the end user? Currently no such information is available — or at least, not without the expenditure of considerable effort.
As Bevan (2012b, 493) has recently pointed out, the availability of these large-scale datasets should shift our goalposts and enlarge our interpretative ambitions, at least for medium- to coarse-grained analysis although, as he also points out, they may bring with them issues associated with recovery and recording biases. The problem is how we might recognise these issues when the emphasis is perhaps inevitably focussed on facilitating the availability and ease of delivery of the data.

3. The Reality of Open Data

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

Although such a characterisation may seem fairly uncontroversial, the concept of the completeness and primacy of the data is problematic from an archaeological perspective since it loses sight of what these data actually are. This is one aspect of the paradox alluded to in the title of this paper: that, while the concept of open data is undoubtedly seen as 'good', the definition of open data is bound up to an extent in technical definitions which are problematic when the nature of the data themselves are considered. Some of the problems arise because much of the data are removed from their original context of use — management inventories turned into online public databases, for instance — but more fundamentally what is referred to as the 'archaeological record' is not itself complete or objective: it is "produced by disciplinary practices and identity processes and discourses" (Hamilakis 107). Data are not 'out there', waiting to be discovered; if anything, data are waiting to be created. As Bowker has commented, "Raw data is both an oxymoron and a bad idea; to the contrary, data should be cooked with care"(184). Information about the past is situated, contingent, and incomplete; data are theory-laden, and relationships are constantly changing depending on context. To put it simply, the archaeological record created from the destructive excavation of the
past is a poor reflection of what was found in the ground, let alone what used to be there; the consequences of pre-depositional and taphonomic issues are further constrained by our ability to recognise, recover, and record what we see. Furthermore, these data are created by specific people, under specific conditions, for specific purposes, all of which inevitably leads to data diversity (e.g. Kansa and Bissell 43). In some respects these are not unique features of archaeological data (e.g. Van House 271), but archaeologists are perhaps especially sensitive to them given the peculiarly destructive nature of a primary data collection methodology. Such characteristics also have significant implications for standards associated with data content, data documentation, and ontologies (Huggett 2012).

Consequently, an understanding of the context of the data is critical: for example, not only are the records and observations that archaeologists collect theory-laden, they may be purpose-laden as well, collected not so much with research in mind but resource-management, for example (Huggett 2004). Additionally, they may be process-laden too, with aspects of their creation and subsequent modification embedded, often invisibly, within them. The operationalisation of data within a computer environment strips out the context of recording — or at the very least, increases the distance from it (e.g. Huggett 2004).

For example, in excavation recording, the more or less standardised recording sheets (paper or digital) are not infrequently supplemented with daybooks which record events, thoughts, interpretations as the work progresses. Although recording sheets contain some elements of process (for example, whether an excavated layer was trowelled, shovel-scraped, or mattocked), much of the process information remains locked within the daybook which, while it may be incorporated within the physical archive, will not be included within the digital archive. Worse still, if the recording is entirely reliant on digital or paper proformas the process information may remain within the mind of the excavator themselves. Similar issues arise with born-digital data such as that associated with the increasingly widespread use of 3D laser-scanning of archaeological sites. The physical size of the basic raw data can be considerable — several billion points constituting terabytes of data in extreme cases — and behind each three-dimensional point is attribute information including colour, reflectance etc. which may or may not be used. The original dataset is subsequently processed and reprocessed, using various algorithms to decimate the data, to smooth out the surfaces, to fill any holes in the data, remove errors, and so on. Finally, the dataset may be used as the basis for adding polygon surfaces to construct three-dimensional reconstructions, analytical views of
elevations etc. However, the processes lying behind such data are rarely specified alongside the datasets themselves, and any subsequent user of such derived datasets may be entirely unaware of any historical sequence of alterations undertaken before the data were made available. Consequently the theory-laden, purpose-laden, and process-laden nature of the data remains largely hidden.

4. Dealing with the Knowledge Deficit

In many respects, therefore, Neil Postman's 1993 prediction has come true:

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

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

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

Subsequent debates about the nature of Big Data recognise that the reality is rather different: that Big Data is driven more by storage capabilities than by improved methods of gaining useful knowledge (Bollier 14). However, delivering data in increasingly large amounts but without accompanying information about the theories, purposes and processes which lie behind those data means that the data arrive at the end user contextless and consequently open to misunderstanding, misconception, misapplication, and misinterpretation.

Potential means of resolving this situation do exist, but they, like the
underlying problem, have not been given sufficient prominence or focus. For example, the London Charter concerning the documentation of computer-based visualisation of cultural heritage refers to the need for documentation of the interpretative decisions made in the course of a 3D visualisation process (London Charter 2006, 7). In version 2.1 this is explicitly defined as 'paradata':

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

There has subsequently been some debate about the utility of the term and its relationship with metadata (Baker, Mudge) although it is worthwhile separating the two concepts at least as a means of distinguishing and emphasising the need for each. Indeed, paradata as a term was originally coined in order specifically to distinguish between metadata which describes the data and paradata which describes process data (Couper 393). In archaeology, metadata is generally understood as describing the nature of the data, the location of the data, and the existence of similar data (Wise and Miller). Consequently the metadata employed by the Archaeology Data Service,¹³ for example, focuses on issues of authorship, rights, and sources, and carries only limited descriptive information and nothing relating to process or derivation. At least within the archaeological context, therefore, there is scope for including something akin to paradata but extending it to data beyond the immediate scope of the London Charter.

The term 'provenance' is perhaps a more relevant and understandable alternative to paradata (for example, Mudge 180-181) but here we have to distinguish between object provenance in the archaeological/art historical sense (its origin, ownership history etc.) and — what is of interest here — data provenance (its collection, recording, modification, etc.: what Mudge refers to as a 'process pipeline').

As an extension of this, the W3C Provenance Working Group, looking at the development of provenance information for Web resources, define three categories of provenance (Gill and Miles):
• Agent-centred provenance (the people or organisations involved in generating or manipulating the data).
• Object-centred provenance (the origins of the data and their derivation).
• Process-centred provenance (the actions and steps taken to generate the data).

In fact there are numerous approaches to defining provenance, including a variant on the familiar resource discovery what/where/when triple, in which where-provenance describes the origin of the data, why-provenance describes the reasons behind the production of the data, and how-provenance identifies the methodology behind the production of the data (for example, Cheney et al., Buneman et al.). However, as a broad generalisation such definitions have clear commonality and are essentially variations on a theme. Whatever terminology is used, it would be highly desirable for data to be accompanied by something akin to paradata/provenance; however, this applies to all data, not just the 3D visualisations which are the focus of the London Charter.

5. Capturing Provenance

Some provenance metadata may already be captured automatically by the software tools used. For example, ESRI’s ArcGIS automatically captures certain types of metadata, including derivable properties of the data (such as extent or number of features), and some (but not all) of its geoprocessing tools update the metadata with an account of the processing that has taken place. Similarly, EXIF metadata automatically captured by digital cameras includes information about shutter speed, focal length, exposure compensation, metering pattern, and date and time the photograph was taken, for instance. Technically-derived provenance metadata such as these are relatively simple to capture as long as the tools and technologies employed are configured to generate it. Currently, however, much information about data capture and processing typically remains tacit and unrecorded and hence represents a significant recording challenge.

Post-hoc automated metadata capture techniques typically rely on a battery of text processing methodologies, but their application to provenance metadata would presuppose that the information were present in the first place. That being the case, automated mapping of datasets using overarching ontologies such as CIDOC-CRM have the potential to capture aspects of the process, and archaeological variants such as the CIDOC-CRM
EH extension have been developed (e.g., May et al, Tudhope et al, Vlachidis et al). Indeed, the CIDOC-CRM and extensions such as CIDOC-CRM EH already contain aspects of provenance within their definitions, albeit deeply embedded within them and not currently fully developed. For example, the CIDOC-CRM EH model of the Centre for Archaeology's Information Domain (Cripps et al) includes a section relating to archaeological survey which contains a \textit{ProcessSurveyDataset Event} covering the processing of the survey data. In CIDOC-CRM terms, these are described as an \textit{E7: Activity} and an \textit{E65: Creation Event}. The Activity class covers actions that result in changes of state and includes properties such as 'had specific purpose' \textit{(P20)}, 'used general/specific technique' \textit{(P32/P33)} (Crofts et al 5). Such properties could be employed as aspects of provenance data: for instance \textit{P32} "identifies the technique employed in an act of modification"(Crofts et al 46) with the technique itself identified using a controlled vocabulary, although extensions describing aspects such as the parameters or conditions relating to the use of the technique would need to be added. However, the extensibility of an ontology such as CIDOC-CRM makes it feasible to model process data if desired, at the cost of greater complexity. It also underlines that while it helps to distinguish between management or discovery metadata and paradata or provenance metadata as a means of emphasising the significance of both, that is not to say they cannot be incorporated within the same schema.

The argument, therefore, is not so much that the capability for representing provenance data does not exist; it is more that the importance of provenance metadata needs to be recognised and accepted in the same way as the value of resource discovery metadata is now taken for granted. Tools to capture some of the information may already be available in some software and partially embedded in complex ontologies, but the reasons to develop and use them need to be highlighted and the benefits of capturing and making provenance metadata available need to be emphasised. These include:

- The provision of a better understanding of the collection processes and circumstances that lie behind the data themselves.
- An understanding of the pipeline of data transformations that sits between the data as collected and the data as received by the end user.
- An improved appreciation of the authority and reliability of the data as well as avoiding inappropriate applications of the data.
- A better understanding of the quality trade-offs between different methods of data capture and data processing methodologies by virtue of maintaining a record of the processes involved.
This kind of metadata also opens the possibility of new research questions based on provenance information which is currently not available: the ability to extract data based on common process or methodology, for example.

6. Resolving the Open Data Paradox

Provenance metadata can therefore be seen as a means of addressing the lack of contextual information typically associated with data delivered via our cyberinfrastructures, the absence of which should present significant issues when those data are situated, contingent, and incomplete. Provenance metadata has the potential to capture aspects of the theory-laden, purpose-laden and process-laden nature of data. On the other hand, provenance metadata increases the data load associated with any given dataset, especially since it cannot necessarily be assumed to exist simply at the collection level. For example, individual records or sets of records within an excavation database will be created by different people and individual contexts will be excavated using different methods; likewise a single individual might be associated with the creation of a GIS dataset but that dataset itself consists of multiple layers which have been created using various data sources and algorithms. Provenance metadata may therefore be required at all levels of a given dataset.

However, the need is for provenance to accompany data explicitly, rather than being hidden away or even absent altogether. This is key to addressing the underlying paradox behind open data. Increasing access to increasing amounts of data has to be set against greater distance from that data and a growing disconnect between the data and knowledge about that data. In the process, the promise of open data may be better understood in terms of both reassurance about the proper use of that data and the subsequent realisation of its transformative potential.

7. References


Anichini, F. and Gattiglia, G. '#MappaOpenData. From web to society.'
Archaeological open data testing', in Opening the Past: Archaeological Open Data, MapPapers 3-II (Metodologie Applicate alla Predittvita del Potenziale Archeologico), 2012. 54-56.


Bevan, A. (b) 'Spatial methods for analysing large-scale artefact inventories', Antiquity 86, 2012. 492-506.

http://www.aspeninstitute.org/sites/default/files/content/docs/pubs/The_Promise_and_Peril_of_Big_Data.pdf


http://www.cidoc-crm.org/docs/cidoc_crm_version_5.0.4.pdf


Hamilakis, Y. 'Iraq, stewardship and 'the record' An ethical crisis for archaeology', *Public Archaeology* 3, 2003. 104-111.


Huggett, J. 'The Past in Bits: towards an archaeology of Information Technology?', *Internet Archaeology* 15. 2004:  
http://intarch.ac.uk/journal/issue15/huggett_index.html


http://www.londoncharter.org/fileadmin/templates/main/docs/london_charter_1_1_en.pdf


Open Definition 'Defining the Open in Open Data, Open Content and Open Services'. 2012: http://opendefinition.org/okd/


Footnotes

1 http://archaeologydataservice.ac.uk/
2 http://www.tdar.org/
3 http://opencontext.org
4 http://www.dans.knaw.nl/
5 http://www.rcahms.gov.uk/
6 http://www.pastscape.org.uk/
7 http://finds.org.uk/
8 http://archaeologydataservice.ac.uk/archsearch/
9 http://canmore.rcahms.gov.uk/
10 http://data.historic-scotland.gov.uk/
11 http://www.english-heritage.org.uk/professional/protection/process/spatial-data/
12 http://cadw.wales.gov.uk/
13 http://archaeologydataservice.ac.uk/