Archeologia e Calcolatori
16, 2005, 83-91

THIRD-PARTY DATA FOR FIRST CLASS RESEARCH

1. Introduction: data re-use

The re-use of digital data has become an important theme in archaeological research. Once upon a time, primary data collection was at the heart of "doing" archaeology, and undertaking fieldwork was a defining part of an archaeologist's professional identity. Those who undertook their data collection in a museum or library were seen as slightly dilettante and the epithet "armchair archaeologist" implied they were slightly "soft" as well. The last decade, however, has seen the growth of a new role of desk-based archaeologist whose favourite data collection tool is more likely to be *Google* than a trowel. Whilst the popular stereotype of the archaeologist is still *Indiana Jones,* the reality is more likely to involve databases and desk-based assessments. A number of underlying factors can be identified behind this trend. These include:

1. The recognition that the archaeological record is finite, and an appreciation that publication, even if completed, does not constitute "preservation by record". Fieldwork is now often regarded more as a limited exercise to answer specific research questions than the starting point of any archaeological project.
2. The increased costs of doing fieldwork, and the reduced availability of funding, combined with reduced opportunities for undertaking excavation. This has gone alongside, at least in some countries, pressures for a "quick win" in research terms, which work against long-term field research projects.
3. Similarly, doctoral researchers no longer have the luxury of spending several decades undertaking data collection. In order to finish on time, as well as complete the range of training activity now required, they are encouraged to identify and use existing data sets, rather than create their own.
4. The development of techniques and theoretical frameworks for data analysis, as articulated by David Clarke in his seminal work, *Analytical Archaeology* (Clarke 1968), combined with the subsequent computer and Internet revolutions and the availability of large scale on-line research resources.

Research funding councils, recognising the costs of primary data collection and the fragility of digital data, now require researchers to deposit their data with recognised digital archives (Richards 2002). Whole funding streams are devoted to resource enhancement and digitisation, whereby the primary aim is to create a re-usable digital resource which will be of use to a range of researchers. If they want to collect new data, researchers now have to justify the need.

However, whilst there are plenty of books and manuals on how to conduct primary data collection by fieldwork and excavation (e.g. Haselgrove *et al*. 1985; Roskams 2001), there is very little literature regarding the re-use of second-hand and third-party data. This is worrying, as data sets do not exist in isolation and third-party data comes with second-hand assumptions and biases (Richards forthcoming). New generations of archaeologists require training in dealing with the re-use of data sets. These are often aggregated from several original sources and possibly divorced from their initial context of observation. They frequently include ill-defined terminologies, and regularly come with inadequate metadata on how and why they were collected. This is not the place to attempt such a manual, although this paper will attempt to draw some conclusions from a single case study of re-use.

2. The Viking and Anglo-Saxon Landscape and Economy project, and third-party data

In Autumn 2004 the UK Arts and Humanities Research Council awarded us a major grant to undertake research into the "Viking and Anglo-Saxon Landscape and Economy" of England (VASLE). The first two aims of the project are:

– To map national distributions of metalwork types c. AD 700-1000 and to compare these with distributions of early medieval coinage, and landscape factors, in order to understand the visibility, recovery and distribution of metal-detected sites of the early medieval period.
– To characterise the finds assemblages of individual known sites, graphing percentages of coins and other object types in order to examine change through time and to derive artefact assemblage "fingerprints" which will help define a hierarchy of settlement types.

However, the project is unusual not for its aims but for its data sources. Rather than undertaking fieldwork to collect new data, the project set out to use data which already existed in two large computer databases, the vast majority of which had actually been collected not by archaeologists, but by metal-detectorists. For those who know the early medieval archaeology of England this is perhaps not so surprising. Archaeologists have been largely unsuccessful in locating Viking Age settlements and it is the so-called "treasure hunters" who know far more about the location and density of archaeological sites of the period (Richards 2004). They have even given rise to a new term "productive site" to describe a category of site which appears to be unusually rich in coins and metalwork (Pestell, Ulmschneider 2003).

In England and Wales all finders of gold and silver objects, and groups of coins from the same finds, over 300 years old, have a legal obligation to

report such items under the 1996 Treasure Act. Prehistoric base-metal assemblages found after 1st January 2003 also qualify as Treasure. The Government also recognised that there was an urgent need to improve arrangements for recording all "portable antiquities" which fell outside the scope of the Treasure Act, and as a result the Portable Antiquities Scheme (PAS) was established. The PAS is a voluntary scheme for the recording of archaeological objects found by members of the public.

The heart of the Scheme is a network of Finds Liaison Officers based in local museums, who identify and record these finds. In 1997 the Department of Culture, Media and Sport provided funding to institute pilot schemes in six regions. Another five pilot schemes were established five years later, funded by the Heritage Lottery Fund (HLF). In 2003 the Scheme was extended to the remaining areas of England and Wales. At the time of writing, details of 86,596 finds were available from an online database (http://www.finds.org.uk/), of which 4,900 were dated to the early medieval period (AD 410-1066).

For the period AD700-1000 a second database is also of key importance. The PAS database contains few coins but the Fitzwilliam Museum in Cambridge hosts the *Early Medieval Corpus of Single finds of coins in the British Isles, 410-1180* (EMC: http://www.fitzmuseum.cam.ac.uk/coins/emc/). The *Corpus* collects data relating to all single finds of coinage from across the UK regardless of recovery method, with 7,074 entries currently available online. The database is managed by Dr Mark Blackburn, and is funded by the Leverhulme Trust.

Each of these data sets raises questions which are specific to those interested in the early medieval period. Nonetheless, the decision as to whether to undertake primary data collection, or to use third-party data sets, is a choice which faces every researcher, and for the reasons stated above, there are increasing pressures to adopt the latter solution. This rest of this paper will therefore discuss some of the issues we have faced in the VASLE project, in our re-use of third-party data. Furthermore, because VASLE requires the use and amalgamation of data held in two different databases, is raises specific questions of quality control and aggregation of data.

The wide geographical scope of VASLE highlights the potential importance of third-party data to academic research. The PAS and EMC both provide data for a geographical coverage well beyond current published archaeological survey, and their ready-made data sets have either never been analysed, or only used as smaller subsets, usually relating to defined regions or artefact groups. In addition, such re-use of data has allowed VASLE to begin its work immediately, subject to its cleaning and enhancement without a time-consuming data collection period. Indeed, in the case of this project data collection itself would create a range of potential problems. The relations between metal-detectorists and archaeologists have traditionally been difficult, and in many

cases data is only forthcoming once levels of trust have been built up over time, often via a local contact, such as one of the PAS's Finds Liaison Officer. In this respect, both the PAS and EMC have managed to become trusted by their local detecting communities, and gain large amounts of data, which may otherwise have gone unrecorded and would be lost in time. As a result, not only are their data sets useful in vastly decreasing likely data collection times for academic projects, but they also provide data that the researchers may otherwise not be able to obtain in any case.

However, as the two organisations collect data from the same groups of enthusiasts, there could be concern that the same finds would be recorded more than once at separate locations. In order to minimise such occurrences, the PAS Finds Advisor for post-Roman coinage also reports for the EMC, meaning that there should be consistency of recording between the two data sets, and any repetition will be cited in each case, i.e. the EMC number will appear in the PAS record and vice versa. As a result, for VASLE's requirements, we could confidently use the coinage only from the EMC data set.

3. The data sets

The format of the EMC and PAS data sets was obviously of prime importance for VASLE in order that they could be easily edited and interrogated, and could be linked to other data sets without problems. The PAS and EMC are both available online as searchable databases but were not suitable for our needs in this format. In part, this was due to the limits of the publicly available data. This includes the publication of only crude grid references even where accurate findspots are known in order that the risks of illegal detecting were minimised. Both organisations, however, agreed to supply the project with databases in either MicroSoft Excel or Access format, including details not available to the public.

At this point it is important to consider the target audiences for both data sets. The remit of the PAS makes it a highly public-oriented organisation. It is aimed at all potentially interested parties from teachers and school children, through local historians, metal-detectorists and amateur archaeologists to students and academic researchers, and as such the database – albeit extremely important – is only a part of its work with outreach and public events also a high priority. The EMC, whilst working with members of the public and making its data available online, has a more academic focus, from both its funding to the nature of its search facility.

Although the data obtained by VASLE comes from very similar, if not the same original sources, it has been collected, managed and presented by each organisation very differently, and as a result the context of the data becomes an important issue. As a single-material, single period data set, the

EMC does not contain the vast number of finds nor the variety of material seen with the PAS. Consequently, the organisation of each is very different. The EMC is produced by a very small team of numismatic specialists within a highly standardised data set and with rigorous use of terminology. The PAS, in contrast, has a two-tiered structure with its network of frontline Finds Liaison Officers providing overall national coverage from a local base and Finds Advisors contributing specialist advice and support.

In this way, all of the material obtained by the EMC is identified and catalogued by a specialist whereas the PAS cannot hope to achieve this, purely through the range of objects it comes into contact with. Finds Liaison Officers cannot be expected to be material culture specialists across all archaeological periods and all artefact categories within the UK (and on occasions Europe). Therefore it is unlikely that initial identification will be undertaken at a specialist level, and the resultant data sets can become problematic. In this, we mean that although the basic artefact identification undertaken by Finds Liaison Officers is highly competent, there is a lack of standardised approach to their description, dating, or classification.

The problem is compounded by the inconsistent availability of images on the PAS website, where records can be checked, and amended if necessary. Records are, however, checked by specialist Finds Advisors over time to ensure correct identifications have been made, and to make any amendments as required, although this will rarely occur prior to the record going online. Therefore, the structures of the two organisations lead to the production of very different data sets. The EMC requires little in the way of cleaning or enhancement, whereas the PAS must be carefully examined before analysis can begin, simply owing to the unavoidable complexity of its organisation.

4. Cleaning and enhancement

Within VASLE, data validation is a fundamental prerequisite of our first aim of mapping national distributions of artefact types. This can be divided into two procedures, cleaning and enhancement. In this case, cleaning can be taken as the deletion of all non-appropriate data, which in general relates to all of those records outside of the period of study. Enhancement ensures that all of the data is as accurate as possible, it is recorded in a standardised manner to allow for easy, confident interrogation, and that it allows for the greatest subsequent use to be made of the data, both by ourselves and researchers who may use these enhanced data sets in future. In many respects these procedures are less important for the EMC, which is already highly standardised, and closely dated requiring little work other than decisions regarding which coinage types would be included within the study. The PAS, however, provides a different prospect, mainly owing to its particular remit and organisation as outlined above.

Within the PAS database, the records are classified in two ways: date and object type. Dating is divided first by major archaeological period, e.g. Early Medieval (AD410-1066), and then into narrower categories where applicable, for example, Early Medieval divides into "early" (AD410-720), "middle" (AD720-850) and "late" (AD850-1066) subgroups. Additionally, if an object type is dated outside of these parameters, specific dates can be assigned, e.g. AD700-900. Object types are divided into two fields. Firstly, "Type", e.g. pins, represents broad groups of artefacts irrespective of period, and secondly, "Class" for more detailed information regarding the type of artefact, e.g. for pins, the class may be listed as "spherical-headed", or a decorative element may be given such as "ring-and-dot".

In theory it should have been possible to search for data relevant to VASLE using the "middle" and "late" subgroups, or the specified dates, but many of the objects have only been dated through the generic "Early Medieval" term. In addition, the classification of artefacts lacks standardisation between the different Finds Liaison Officers, with the same artefact type often classified in a variety of ways. For example, late Anglo-Saxon stirrup strap mounts were in some cases classified as stirrups under "Type" and then under Williams's (1997) typology under "Class", but in others they were defined as "horse trappings" under "Type" and then "stirrup strap mounts" under "Class". Such lack of standardised classification within the PAS database proved problematic for our initial data searches, as the records could not be searched with confidence that all related records would be found. As a result, the project obtained all "Early Medieval" data to ensure completeness.

The enhancement process relates mainly to the PAS, given its wide range of artefact types, and variations in descriptive terms used. Where possible, all artefact groups were standardised, with the artefact type under "Type" and the sub-classification recorded either under a commonly used scheme, or by the main decorative element. A useful example of the enhancement process are the finds of small dress pins, which appear to be ubiquitous on most Middle Anglo-Saxon sites. Over 350 Anglo-Saxon pins have been recorded by the PAS, and each was checked in conjunction with the individual images available on the website to ensure accurate classification and enhancement. Within the PAS database, the pins are generally classified as either "pin" or "dress pin" under "Type", and the shape of the pinhead (the main feature for closer classification) given under "Class", unless a particular classification scheme has been used in which case the typological class is given (e.g. Hinton 1996, 14-37).

This works well until the data is taken as a whole when the variations in terminology used (especially under "Class") hinder straightforward analysis. For example, spherical-headed pins were described under four other terms, all of which occur within general archaeological literature. In order to standardise this data the pins were re-classified within Hinton's (1996, 14-37) typological

scheme. In addition, the occurrence of multi-find records (i.e. records containing more than one find from the same site) also proves problematic for studies where analyses that involve quantification are important. In these cases, enhancement also had to include the separation of such records into standard, single entries.

Alongside these aspects of cleaning and enhancement, the use of two databases also presents its own challenges. Whilst both the PAS and EMC utilise similar sources, the integration of the data sets is not without its difficulties, and is important for the successful analysis of any third-party data set. We have to be confident that the link between the PAS and EMC is such that our interrogation will provide VASLE with all the data as required. As VASLE is a large-scale geographically based analysis, the linkage of the two data sets must be via some aspect of location but this relationship is less straightforward than may have been imagined. Both data sets provide grid references, parish name and county for each find reported at least, and in some cases any name attributed to the site is provided. Ideally, this latter field would be most useful, and was originally our choice to be used in VASLE but its inconsistent use across the two data sets precludes it. Obviously the county in which finds were made lacks the required level of accuracy and the location by grid reference can also prove difficult given that a low, but not inconsiderable, proportion of entries lack such information.

As a result, the most realistic method to link the data sets from the PAS and EMC is to utilise the parish name. For the VASLE's initial objective (the mapping of metalwork types across the UK) this presents little problem as such accuracy is more than adequate, and has been used with success elsewhere (e.g. Naylor 2004). The comparison of individual sites for the project's second aim may, however, require the division of data from parishes between more than one findspot. Although this can only be undertaken on a case-by-case basis it nevertheless will require additional utilisation of the data sets by hand rather than through a simpler sorting of the data.

Overall, the above discussion has usefully summarised a number of problems that have been encountered by VASLE in our use of third-party data, and which may indeed be factors throughout the life of the project. These are not specifically related to quality of the data collected, but rather to its presentation, especially so when considering the PAS. With its lack of a standardised approach to data entry, interrogating its database with confidence is difficult prior to a full check and appropriate enhancement. Conversely, research on a smaller scale, for example including a small area or small number of sites, may find such problems less visible, and it is only once material is analysed in a large scale project such as VASLE that they become apparent. The nature of cleaning and enhancement in our case has highlighted that this can be a major task, and any project may begin with data sets whose size is far in excess of either expectations or need simply owing to the presentation of the original data.

5. Conclusions

What more generic conclusions about third-party data usage can be drawn from our experience?

First, we have seen that data sets produced using similar information can become extremely different depending upon how the background organisation is structured and this must be considered when assessing timescales for data standardisation in project proposals. Secondly, the audience and background organisation for the data sets to be used should be considered carefully. The specific terms of reference for the original data collection are of prime importance here and can greatly affect how that data is processed and presented. Certainly the collection of data by large organisations and relatively high staffing numbers, such as the PAS, can affect how immediately usable it is for academic research simply owing to the way in which its own data storage is organised. Although far less than collecting comparable data from scratch, the potential labour costs of re-use of third-party data sets cannot be underestimated.

With regard to the two specific data sets considered in this paper, the PAS database is an ideal data set for those working on particular areas or sites and in local history, whilst the nature of the EMC makes it a more immediate and accessible tool for academic research. The PAS is extremely valuable but it requires more work prior to analysis.

Our research certainly demonstrates that third-party data sets are worth the bother. Although we have encountered some unforeseen problems with the data sets, especially from the PAS, the simple fact that they exist is important. The enhanced data resource has become an invaluable tool in answering our research objectives. With regard to this type of data, we could not collect such large amounts of data ourselves, and much of it would either go uncollected or would be lost.

In conclusion, the nature of archaeological research is changing. The days of the solitary fieldworker, or the research student spending three years compiling a card index or database, are over. Future researchers will be expected to make far greater usage of existing digital data sets, and to justify any primary data gathering. They will expect to use powerful web-based technologies to aggregate and analyse online databases and archives. However, the value of these data sets, and the quality of the research undertaken from them, can only be as good as the underlying data. Re-use of data requires a close understanding of the context of data collection and of the vocabulary used to describe the observations. The archaeologist of tomorrow needs training not so much in methods of data collection, but in data analysis and re-use.

John D. Naylor, Julian D. Richards
Department of Archaeology
University of York

REFERENCES

Clarke D.L. 1968, *Analytical Archaeology*, London, Methuen.

Haselgrove C., Millett M., Smith I. 1985, *Archaeology from the Ploughsoil: Studies in the Collection and Interpretation of Field Survey Data*, Sheffield, Department of Archaeology and Prehistory, University of Sheffield.

Hinton D.A. 1996, *Southampton Finds Volume 2: The gold, silver and other non-ferrous alloy objects from Hamwic, and the non-ferrous metal-working evidence,* Southampton, Monographs 6, Stroud, Sutton.

Naylor J.D. 2004, *An Archaeology of Trade in Middle Saxon England*, BAR British Series 376, Oxford, Archaeopress.

Pestell T., Ulmschneider K. 2003, *Markets in Early Medieval Europe. Trading and 'Productive' Sites, 650-850*, Macclesfield, Windgather Press.

Richards J.D. 2002, *Digital preservation and access*, «European Journal of Archaeology», 5, 3, 343-367.

Richards J.D. 2004, *Viking Age England,* Stroud, Tempus.

Richards J.D. (forthcoming), *Do we believe in the 'data' in the Archaeology Data Service*, in S. Roskams, M. Beck (eds.), *Interpreting Stratigraphy 2001 Conference Proceedings*.

Roskams S. 2001, *Excavation*, Cambridge, Cambridge University Press.

Williams D. 1997, *Late Saxon Stirrup-Strap Mounts: A Classification and Catalogue*, CBA Research Report 111, York, Council for British Archaeology.

ABSTRACT

The use of third-party data is becoming an increasingly important part of archaeological research but there has been little critical analysis of such data sets, or their use. This paper highlights both the challenges and benefits of third-party data through discussion of the experiences of the UK's Arts and Humanities Research Council-funded project "Viking and Anglo-Saxon Landscape and Economy". It shows that the background organisation and intended audience of third-party data set can greatly affect how the data is collated and presented, and the enhancement of such resources for particular research aims may be labour intensive and time consuming, and should not be underestimated. However, it is argued that the usefulness of third-party data sets outweighs any potential problems which may be encountered, but that there needs to be recognition of these challenges and appropriate training provided for future archaeologists.