Comp. Appl. Arch. 13, 1985.

THE ARCHAEOLOGICAL DATABASE APPLIED? NORTH YORKSHIRE COUNTY COUNCIL SITES AND MONUMENTS RECORD AT THE UNIVERSITY OF BRADFORD

P.N. Cheetham

Undergraduate School of Archaeological Sciences,
University of Bradford, Richmond Road, Bradford, West Yorkshire BD7 1DP

Introduction

This paper is intended to outline briefly the implementation of a sites and monument record (SMR) database at the University of Bradford. Some potential strengths and weaknesses are pointed out. An example is used to illustrate how such a large body of data, provided the time and access are available, may produce results directly applicable to field and academic Archaeology.

The NYCC SMR at Bradford

The University of Bradford has been able to obtain copies of the North Yorkshire County Council (NYCC) SMR data files as the first step in creating a Yorkshire-wide database at Bradford. It is hoped that this will allow multi-county synthesis and intercommunication of data, while also providing a powerful research tool within the University. This would be available to undergraduates undertaking various course work projects, postgraduate students and staff. The obvious problems of the compatibility and integration of the other county SMRs will not be dealt with here. Brief details of the Bradford implementation of the NYCC data are outlined in the relevant section of this paper. The NYCC system being one of the most comprehensive available, provides a credible example on which to consider some broader issues of which the author has become aware during his work. The setting up of the database formed the basis of his BSc honours dissertation. The system will not be available for general use until late June 1985.

SMRs and general databases

Although some of the following points are covered in more detail by Richards and Ryan (1985), they are considered to be of fundamental importance to the future development and ultimate usefulness of more general archaeological databases. The difference between an SMR and a general database may seem slight, but the concept is quite different. SMRs are compiled to aid Heritage management while their use as research tools is often secondary. This difference is accentuated during computerisation. Should SMRs be all-embracing general comuterised databases or be left simply as indexing systems with manual backup for detailed information?

While the arguments for the restriction of the range of a database due to storage limitations may be about to disappear with hardware advances, it is true to say that this range of entities and attributes could be extended ad infinitum and possibly should be. The contemporary range is defined by past experience and present ideas. No account can be taken for future or incorrect present ideas. Too strict an application of a policy of including only relevant data certainly

runs the risk of fossilisation and stifling new approaches. Although a problem general to Archaeology, it is particularly acute when considering computerised systems where some limits of size must be imposed. By having data in large, easy and repetitively accessible databases, novel options and possibilities can be explored with an ease not feasible under a manual system and, therefore, never attempted.

The more basic SMRs may be considered spring boards to more detailed analysis which if thoroughly produced can dramatically reduce the time of collation of distinct groupings of material. The more sophisticated the data set the more such a database moves from this function towards becoming a self-contained analytical tool. This is a trade-off between the benefits of ease to access and the problems of producing classificatory structures of general application given the logistics of their implementation.

Problems encountered during the transfer of data to the Bradford computer highlight the responsibility of all centres producing SMRs to ensure that the facilities and expertise exist to allow data to be disseminated in a form usable by the recipients. This should obviously be a flexible give and take situation, although the provision of some standards would be helpful.

The general level of accessibility of archaeological databases and their level of use must be a central concern. The NYCC SMR database seems to have been little used beyond its primary planning control functions. This is apparently also true for the greater Manchester SMR (Walker personal communication). It should be understood why this situation has arisen and how it can be improved. The reasons why SMRs seem to be little used appear to be that: access to the system is restricted; its existence is either ignored or unappreciated; the time, money or interest does not exist to exploit such databases; they are not regarded as reliable or credible for hard archaeological analysis. The usefulness and need for market surveys and monitoring studies of the needs of potential users are covered by other papers in this volume. The author fully agrees with such an approach, for although opinionated or theoretical concepts should be allowed as expansions to a system, its basic relevance to its users' needs must be established and maintained. A measure of how useful a database is should not be based on its size, number and types of fields, speed of access, conceptual structure, etc., but on the number of archaeologically meaningful questions that it can answer. Feedback between the users and the organising body is essential to assess this.

A major problem area is the mixing or segregation of interpreted data as opposed to raw or objective data. SMRs in particular are the compilation of information from many sources, some of which may be out of date or incorrect in the opinion of some archaeologists. The compilers, faced with a vast range of information with all of which they are unlikely to be fully conversant at the specialist level, may well be compelled to amend entries or impose their own interpretations on data they do not fully understand in order to fit it into an implemented scheme of recording. This could jeopardise the credibility of database information and their usefulness to specialist users, who are likely to have their own criteria for data validation based on their own applied knowledge.

The suggestion of a retreat to an attribute-only approach may not be the appropriate response, since this effectively removes any traditional access pathways to the data which are of proven usefulness. We should endeavour to avoid the loss of interpreted data based on sound archaeological reasoning. Relegating archaeological data to equally nebulous classes of ill-defined

attributes such as shape may be of little help. To have a truly attribute-only approach to archaeological data in which complex definitions of class or period are used to generate real time interpretation, termed a Virtual Database, would require a vast bank of expert systems whose knowledge bases would make the source database comparatively minute. However, this is one future direction to keep in mind.

There seems little harm in including interpreted data provided it is treated with the reservations of which archaeologists should already be aware. To this end, any database without bibliographic information is at best suspect and more likely to be useless or even detrimental because of the potential immortality of out-dated and simply incorrect data.

Above all, archaeologists must be made aware of the greater flexibility of analysis offered by computerised systems and develop their analyses to utilise them effectively.

Thus for an archaeological database to be applied it must be:

flexible
accessible
credible
appreciated

Without these qualities, while devouring large amounts of money and expertise, its contribution to mainstream Archaeology will sadly be minimal.

Assessment of NYCC SMR and the Bradford derivative

It is important to appreciate that the Bradford implementation of the NYCC SMR is not simply a direct copy. By examining the four desirable qualities above, it can be shown that although neither fulfills all of them, this dual system could form the basis of a productive unity.

The NYCC system is flexible in that its main thesaurus is not static but it is unlikely that the 75 relevant data fields will be modified or expanded at present. The large range of fields does however make it comprehensive enough to meet many varied needs and until these are exhausted or shown to be inadequate, flexibility is not a major problem. The stricture of the fixed field format used has imposed certain restrictions on the number of repeating fields. For example, only two age periods are permitted per SMR primary record. At Bradford, using a relational database model, one-to-many relationships may be set up for repeating fields. New relations are also possible for data derived from other sources such as geophysical survey details. Thus a more flexible system has been produced. Strict third normal form has not been adhered to, but then the problems of data integrity are reduced since updates occur across all the relevant fields simultaneously during the loading process. The structure has been formulated to minimise storage and maximise ease of access, including additional relational fields to index areal units and give broader site and find class groupings.

Access to the NYCC SMR is restricted more by how, than by whom, it is interpreted. Even with the greatest goodwill the NYCC unit can only answer direct requests or issue listings. Analysis using the data on-line seems unlikely in most cases and the range of software and peripherals is limited compared to the facilities available at Bradford. There, although access will be restricted, the mode of access will allow totally individual exploitation with no restriction

on money or, comparatively speaking, time. This type of access by exploration or by direct access and manipulation of particular fields is much broader than that allowed to external users of the NYCC system. Access to the Bradford system by other organisations through the universities' network, may also be a possibility.

While credibility as outlined in the previous section, will always remain a problem, Mike Griffiths of the NYCC is keen to increase the quality of data content and to be made aware of any potential improvements. Thus, while the NYCC is as credible as any other SMR, the opportunities for additional academic input and critical assessment of the content and bias within the data set through its use within the university can only improve the database, provided that constructive feedback is established.

Appreciation or lack of it for the potential of the NYCC database may be due in part to its incomplete nature or, as with new facilities in general, its need to gain acceptance through proven use. Undergraduate courses at Bradford in Archaeological Sciences and Scientific Methods cover the application of databases as tools of archaeological analysis. The opportunity to utilise this database for many coursework projects or postgraduate and staff research may produce results that will improve both the credibility and appreciation of the database by a wider range of users.

The Bradford system

The system is only partially operational due to problems in the transfer of data from the main SMR basic site data file. These should be resolved shortly. The fields containing the aerial photographic, bibliographic and archive data are now on the Bradford system. Both computer centres (NYCC and Bradford) have co-operated, but have been unable to arrive at a suitable transfer format. This has led to expensive wasted effort.

A relational DBMS, RAPPORT-4, has been set up on the Bradford CDC Cyber 170/720 mainframe. An interactive query language is not at present available at Bradford for this version of RAPPORT. Access is, therefore, by means of user-written programs. Storage is provided by a 236Mbyte disk pack used on the one interchangeable disk drive available. The heavy daytime demand on the mainframe has affected the availability of this drive unit and the compilation of programs requiring the large RAPPORT library.

Procedures to attach the datafiles and run standard applications programs have been integrated into a user-helpful front end. The standard applications will include a search suite to enable a range of regional searches by parish, 1km or 10km grid squares, radially or polygonally followed by a period and class search as required. This will be an exploratory tool, providing a limited summary of each SMR. Additional programs will allow more comprehensive details of a particular SMR record and distributional plots to be obtained. However, it is envisaged that normal access will be via specific user-written programs, although a good knowledge of the interface languages, FORTRAN and Pascal, as well as of the detail of the database structure and indexing, will be required to use the database effectively.

It seems likely that the database will be used extensively within the School of Archaeological Science at Bradford, to judge from the interest already shown and the many varied requests for data generated by the project. It is intended to provide some means of monitoring the system both for in-house development

and as feedback to the NYCC Archaeological Unit. It is unfortunate that the use of the database by outside agencies will inevitably be limited.

The current storage capacity of the DBMS should allow it to accomodate the 70Mbyte of data expected to form the complete NYCC SMR. Inclusion of all the other Yorkshire SMRs in this database would be difficult. RAPPORT requires 30% redundancy in data filespace, this plus space for utilities and applications programs would exhaust the capacity of the DBMS. The system is at present set up to accomodate 25,000 complete records. At the latest update the datafile contained 12,000 individual SMR records. The following section outlines one example application developed and applied to data extracted from the Bradford database.

A method for the estimation of aerial photographic potential

The problem of assessing the potential or exhaustion of a particular area for aerial photography is a central question to the producers and users of photographic data. To be able to get an accurate estimate of the potential of an area based on its past performance, as a ratio of the expected new sites discovered to total sites photographed, is a step in this direction. The estimation of the absolute number of potential sites is not considered and could not be extrapolated with any confidence by the proposed technique.

To investigate the potential of the method, the Bradford/NYCC SMR database proved invaluable as a source of data to characterise the fall off in new aerial photographic sites discovered over time. Hence, it will provide a useful example application. The aerial photographic data have been indexed by 10km grid square to assist such an analysis and also to allow sampling scheme analysis and aid regional searches. Thus 10km square areas were analysed for crop, soil and parch marks on a yearly basis.

In a privately circulated paper in 1984 White and Lee, of the NYCC Archaeological Unit, pointed out the implications for the development of the Unit's database of new:repeat photography. This author's work, developed independently, illustrates how the availability of data, directly stimulates fresh approaches.

To account for the variation in overall response between years, the ratio of new sites to total recorded sites was calculated. Whimster (1983) suggested that this ratio is not affected by large variations in annual response and used this value to illustrate the fall-off in potential for sites in a 100km square around Cambridge. Although he discusses the proportion of new:repeat photography, it is the new:total photography proportion which is represented in his figure 76. When smaller areas were examined, although a trend is apparent its character is not clear (Fig. 1). By smoothing the trend using a plot of the cumulative sum of the ratios, a consistent and characteristic form of curve became apparent. Figure 2 illustrates one such plot with a simple modified exponential fitted using a least squares regression coefficient on the liner form. The fitted curve is of the form:

$$y = \delta + \alpha e^{\beta t}$$ where y=cumulative ratio
t=time in years

This curve does not have all the characteristics expected, that is a gradient where $\delta=1$ at t=1 y=1 but gives a reasonable fit especially towards the top end of the plots if the linear regression coefficient is maximised. From this fitted curve the gradient and hence the expected ratio of new:total sites for the next flying year can be calculated.

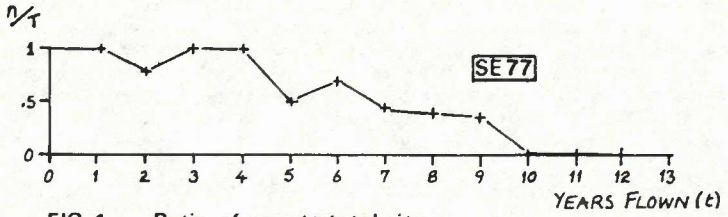

FIG. 1 Ratio of new to total sites

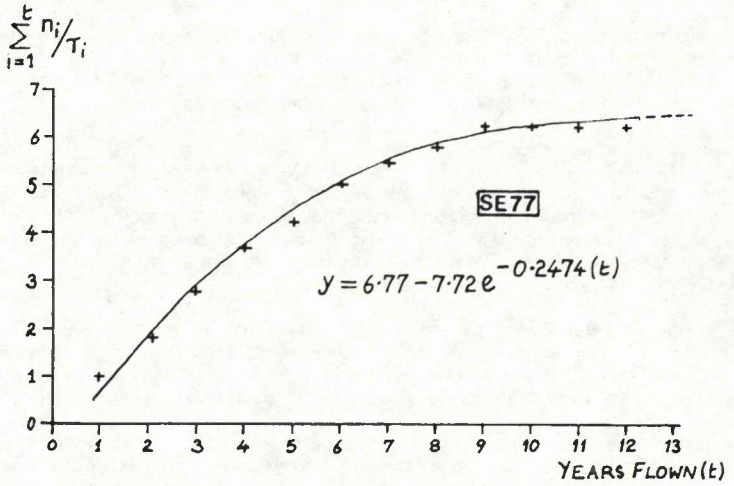

FIG. 2 Cumulative ratio and fitted curve

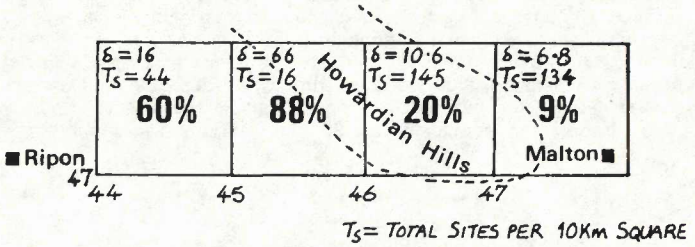

FIG. 3 Estimated potential. Ripon-Malton transect

In practice, a BASIC program on an Amstrad CP464 was used to obtain the maximum linear regression coefficient by iteration while varying the value of δ and the next year's potential was calculated, while a plot of the current configuration was produced to check the goodness of fit visually. The value of δ seems to relate to the intensity of survey in that the smaller δ is the less flying years have been considered and yet the potential is significantly reduced. Areas only clipped or passed over casually will maintain a high potential over a longer period even though fewer sites may be discovered.

Certain assumptions are implicit to the method. They are:

all observable sites are photographed and recorded on the database
the areas are flown in a random manner year after year
the coverage is representative of the whole of the unit analysed
all sites are discovered by aerial photographic survey

The last assumption excludes the use of earthwork or shadow sites since it is not possible to confirm whether they were discovered exclusively by photographic survey. During a smaller project looking at the aerial photographic response around York by using ground surveys to act as a control and hence attempt to relate significant variations to land use or subsoil conditions, the same problem was encountered. The lack of information on the mode of discovery complicated the problem. It is felt that any SMR should include this detail if it is to attempt any analysis of field survey bias, a task to which SMRs should be eminently suited.

The need for an assessment of aerial photographic potential is of interest in at least three areas of work:

in settlement studies intending to utilise such data it is important to know how good the existing data are and where additional flying would be appropriate
in general survey flying it is important to know which areas need most study and to see whether a uniform level of cover could be attained
it would be interesting to know the extent to which ideas on ground response corresponded with this analysis

It must be emphasised that absolute numbers of sites could not be estimated. A relatively barren area could give up one new site per year with little repetition and hence high potential, whereas areas having >100 sites may still show a 10% potential and, therefore, comparatively more sites. The following example based on NYCC data illustrates a possible interpretation of a small area to which the technique has been applied.

Figure 3 shows four 10km squares running from just east of Ripon to Malton. This area extends, therefore, from the Vale of York in the first square to the western edge of the Howardian Hills in the second, over those hills in the third and into the Vale of Pickering in the fourth. The percentages are estimates of the potential for new sites in the next flying year with the total number of sites Ts and the δ value. While the examination of accurate flight logs would confirm the situation, the results presented suggest that two flying regimes may have been operating: one, the more intense, centred on Malton and extending up the eastern flanks of the Howardian Hills; another in the Vale of York of a less intense nature. The interface between the two zones has, however, been largely excluded from both. The implications for any settlement synthesis using

aerial photographic evidence or alternatively for which areas to concentrate further survey work in are obvious. Although the technique is not regarded as giving precise values the characteristic trend of the fall-off observed suggests representative and quantifiable measures are being produced.

The next step will be to check for more suitable curves to model the fall-off and make a more valid assessment of the errors in the predictions. The implications for SMR databases are twofold:

> without good, full and easy access to data, such ideas could not be developed or put into routine use
>
> data that are apparently of simply administrative value, for example date and mode of discovery, may be of sound archaeological value. As such the exclusion of any data through opinionated judgement may be questionable.

Acknowledgements

This work was made possible by Mike Griffiths, County Archaeologist for the NYCC, who made the data available for use at Bradford. John Haigh of the Schools of Archaeological and Mathematical Sciences, University of Bradford, supervised the project and liaised with the NYCC. Mike Kelly made available two routines used in the system and helped sort out computing problems. This paper has benefitted from extensive discussions with all of them.

References

DAVIS, C.H. & RUSH, J.E. 1979 Guide to Information Science 3: 7-8. The Library Association, Wetsport.

GILCHRIST, W. 1976 Statistical forecasting. John Wiley & Sons, London.

RICHARDS, J.D. & RYAN, N.S. 1985 Data processing in Archaeology. Cambridge University Press.

WHIMSTER, R. 1983 Aerial reconnaissance from Cambridge: a retrospective view 1945-1980. in G.S. Maxwell (ed) The impact of aerial reconnaissance on Archaeology. 92-105. Council for British Archaeology Research Report 49. London.

SMAILES, M.H. & GRIFFITHS, M. 1982 Archaeological records system. North Yorkshire County Council Computing Service, Northallerton.