COMPUTER RECORDING OF EXCAVATION DATA FROM HUNTER-GATHERER SITES

Ian Johnson

Introduction

The purpose of this paper is not to describe any specific system of excavation data recording but rather to make some general comments about the role of computers in this field and on the way in which computerised systems of excavation data recording should be implemented. My main contention is that, in hunter-gatherer archaeology, the recording of data from excavations has been largely neglected in favour of analysis of selected subsets of 'exciting' data (such as tool attributes) and that the potential of computers as a means of data management has not been realised. Besides potential benefits in terms of management and storage of excavation data, computer recording forces reassessment and explicit design of data collection techniques, resulting in better defined and better balanced strategies and a consequent increase in data quality. The potential for inter-site comparison arising from the reaccessibility of computerised excavation data records should provide a stimulus for systematisation of excavation and data collection techniques. It is to be hoped that this in turn will result in more thought being given to the storage and cataloguing of both excavated material and data.

In order to promote the use of computer recording techniques and to encourage comparability between different workers, what we require is a flexible data base system which is convenient to use, provides immediate short-term benefits in terms of ease of recording and of analysis but which does not stultify research by restricting the data recorded to a fixed list of variables, fixed classificatory schemes and fixed data structures. By limiting the scope of such a system to the recording of data from hunter-gatherer sites (which generally have few complex structures) it becomes a practical proposition without the need for state-of-the-art programming techniques or machine-specific features. Furthermore, it should be possible to develop such a system for low-priced microcomputers, allowing its use away from a University or Administrative environment and perhaps the direct exchange of data between researchers via floppy disks or cassette tapes. A valuable side-product of any effective database system would be the systematic cataloguing of excavated material. This would be invaluable when it comes to final storage, as well as being the first step towards the establishment of centralised data repositories.

Getting Away From Cards

Although a number of quite sophisticated computer applications have been published in the field of hunter-gatherer archaeology, the vast majority belong to the "cards to crunch" approach, in other words the use of the computer as a glorified number cruncher (and occasional plotter of graphs or plans). Whilst this is partly a function of equipment availability, a similar approach is often followed when interactive terminals are used. The problem lies in the fact that most archaeologists are not sufficiently interested in computing to exploit, or even to be aware of, the potential of interactive programs or database systems.

Typically, if some sophisticated numerical analysis is required, the data to be analysed are filled in on coding forms, keyed in in fixed columns and run through SPSS, BMD, or some other program that someone has kindly provided. The data may then be abandoned or possibly stored but all too frequently separate from the coding information which serves to interpret the data. Re-use of the data by someone else is generally tedious, if not impossible. If efficient data entry and storage programs were available, the same amount of effort might have served to create a useable database onto which future analyses or re-examination of the material might then build. I would stress that such a database should incorporate, as an integral part, the definition of the data, i.e., the methods used in its collection and any other comments that may be relevant to its interpretation.

One of the main drawbacks of the "cards to crunch" approach lies in its inflexibility. Corrections of data already entered are tediously carried out by repunching cards or the use of a text editor and the lines of data are often difficult to read and interpret, particularly if they are packed without spaces and/or coded numerically throughout. The effort of producing special-purpose editing programs is not generally justified unless the particular file format is going to be re-used many times (in which case one is talking about setting up a database system).

More serious, however, is the fact that one is generally restricted to a cases-by-variables approach and one must decide in advance what data is to be entered and in what columns. These restrictions deny the ongoing nature of research; often one will want to add new variables after an initial analysis of the data, or someone carrying out a specialist study will generate data at a later stage which should be added to that already recorded. Although such data additions are possible with fairly simple special-purpose programs or routines of packages such as SPSS, the need for them encourages one to wait until everything has been recorded on coding forms before the data is keyed into the computer. Many projects grind to a halt (or into suspended animation) before all the data has been collected and, all too often, we are left with an uninformative preliminary report and a mass of unanalysed data from which it is impossible to assess the potential of the site. On the other hand, if we had a database system capable of accepting data in dribs and drabs, just as easily as in one block, then such projects might produce quite detailed preliminary reports on the basis of which other workers might be interested in following up the project if it is abandoned. Such a system would also avoid the duplication of effort arising from data being lost or mislaid, as well as providing a detailed systematic catalogue of the material for storage purposes.

Integration of Excavation and Data Recording

The use of computers as glorified calculators is also partly attributable to the compartmentalisation of data collection into discrete steps, typically 'excavation', 'measurement', 'analysis', and 'storage'. Frequently each step is completed before the next one is planned and this applies particularly to the excavation/measurement and analysis/storage junctions. The problem is particularly aggravated by the fact that many first-time Directors have quite extensive field experience but have never been faced with the pile of excavated material and its attendant problems once back in the laboratory; the result is generally excessive amounts of excavation without the backup of a well-oiled processing and recording system to get them through the laboratory work. I speak here from experience!

The moral is, of course, to plan the whole process right through as an integrated procedure, if possible building on other people's experience as well as ones own. It is therefore important that people should publish details not only of excavation techniques but also of laboratory procedures. Such publication is generally neglected as being insufficiently 'academic' but it is the only way of avoiding duplication of effort and badly organised procedures. A procedure thought out right through from excavation to the eventual storage of the material is likely to be far more effective and less prone to either time wastage or errors than one which is planned in steps.

Although a well documented and flexible database system will aid in the design of effective laboratory procedures, these observations apply whether or not a computer is used for data recording and analysis. The use of a computer does, however, force one to a greater awareness of the shortcomings of the data collected through the need to quantify systematically and to ask oneself questions about all that 'missing data' which it is easy to ignore.

A final reason for the adoption of computers in excavation data recording, lies in the nature of the changes in excavation techniques. Ever since the beginning of hunter-gatherer archaeological studies there has been a steady trend towards tighter spatial control and the collection of more and more material from the same excavated volume. This trend may be slowing as one gets into a situation of diminishing returns but it is certainly not reversing. Any excavation satisfying currently accepted 'minimum standards' criteria will generate huge amounts of numerical data with very low individual information content. Even the smallest excavation will probably generate 10,000 or so items of data, making computer analysis virtually essential if we are going to present the information effectively to other people without spending enormous amounts of time on repetitive and mechanical manipulation.

Storage and Retrieval

Systematic recording of data on a re-accessible medium is obviously the first step towards being able to constitute an archaeological data bank. I do not, however, see the provision of a sophisticated on-line data bank system for excavation data as being a pressing necessity; such a system would only become worthwhile if data were being accessed very frequently and were required at very short notice; this may be the case in many business or administrative situations but is most unlikely to apply to excavation data. On the other hand, a standardised data-base system could form the basis of a centralised data repository which could be accessed conveniently simply by loading the appropriate physical medium (tape or disk). The capital cost and running expenses of such a repository would be minimal in relation to the amount of data stored when compared with an on-line database, yet access speed could be quite adequate for the sort of situation in which excavation data is required.

It should be noted, however, that whilst access to any particular block of data (i.e. site) need not be particularly rapid, access to data within that site must be rapid once the site has been located, in order to allow additions and analysis to proceed quickly. This implies on-line disk storage of selected blocks of data and programs run interactively from a terminal. The restricted size of the database (i.e. a few sites rather than all the sites in the repository) means that a sophisticated retrieval system is not of prime importance.

As I conceive it, a central repository of excavation data of the type described would consist of separate files for each site stored on magnetic tape, each file being accompanied by a file giving information relating to the way in which variables for that site were recorded and the way in which the data file is organised. Whilst the structure of each data file would be similar, the actual variables recorded would vary from site to site, although it would be hoped that the establishment of such a system for a given geographical (and potentially chronological) zone would slowly lead to a concensus on the recording of major variables, with individual research orientations being expressed in extra variables specific to particular projects. As an alternative to magnetic tape storage the data might be recorded on interchangeable disk packs if money was not in short supply, or it could be recorded on floppy disks allowing easy on-the-spot copying of data with a dedicated low-cost microcomputer and easy distribution of duplicated data (by post). Whatever the form of storage, security arrangements (i.e. access to particular data) could be maintained manually, as in a library, rather than through passwords and the like, as would be the case in an on-line system and this considerably simplifies the task of programming.

Retrieval and Analysis

No rules can be laid down about analysis; the number of possible analyses is only limited by the number of people involved and any attempt at standardisation would lead to stultification of research. Retrieval programs should therefore be aimed at interfacing with the data rather than attempting to carry out specific analyses (cf. Caelli, in press). I do not, therefore, want to go into any detail on data retrieval and analysis, although I am aware of the danger that what goes into a database system may never come out again if satisfactory retrieval programs are not available. However, I believe that the best way an excavation database system can sell itself to archaeologists is by providing convenience not only in data entry but also at the analysis stage. If this is the case it will represent an attractive alternative to analysis by hand or the straight coding-forms-to-cards-to-crunch approach. It also implies that the problem of reaccessing the data has been solved for at least the intra-site level; access to data for different sites need then only be a matter of physical organisation of the repository and access facilities.

For sites excavated using a given recording method and common classificatory and measurement systems, it may well be possible to set up standard programs which will access the database for a site and prepare commonly calculated statistics or graphical information, e.g. cumulative typological diagrams, distribution of artefacts by excavation area and stratigraphic level or proportions of different components in the sediment matrix. In this way the excavator can get a range of commonly needed information for a very small input of time, leaving more time to concentrate on points of specific interest in the data. One way of providing such a feature, which I used for a very simple database designed for use on Australian sites (see below), is to write retrieval programs which output data in a form suitable for input to a package such as SPSS. The 'standard' analysis can then be written in the language of the package, onto which any individual worker can add further analyses as required. Such an approach means that the user can expand the analysis without having to learn anything beyond the package instruction set, which is generally simple and well documented.

A second approach to providing 'standard' analyses is to write special-purpose retrieval and analysis programs. Whilst these should potentially be more convenient in use, more efficient in computer resources and give more easily interpreted output, they will not have the easy expandability of the first approach. The best solution is probably a combination, in other words the provision of special purpose analysis programs as well as general purpose retrieval programs which can be used to prepare datasets for input to a package or other programs.

Towards an Excavation Data Recording System

As part of my PhD research in Australia (Johnson 1979) I put forward some suggestions for improved excavation techniques and developed a very simple database system aimed at providing a means of recording the large amount of numerical data generated by the excavation of typical hunter-gatherer sites. The system uses fixed format records for data storage in a single data file containing several types of data record; each record type has a different format and is identified by a code in Column 1. The file is initially created by keying in coding forms but, thereafter, data may be added or amended through a series of conversational programs which combine a freefield file of the data to be added with the main excavation data file. Other programs allow specialised editing directly from the terminal and preparation of files in a simple cases-by-variables structure for input to SPSS. A number of SPSS instruction sets serve to interface with these files and carry out a standard set of analyses.

As the system stands it can accommodate only numerically coded data for individual objects or for each excavation unit within the site. There are restrictions on the ways in which objects and excavation units must be identified (e.g. they must be numbered, whilst excavation areas must start with a letter). No allowance is made for free format text such as commentary, for data on the site as a whole or for each stratigraphic level, or for the incorporation of variable definitions within the data file. Although the system could be expanded using the existing file structure it was only intended as a stop-gap measure to cope with the flood of numerical data being generated by Australian excavations and I feel that it should be replaced by a fresh start rather than attempts at improvement. Data recorded on the present system should, however, be easy to transfer to any more sophisticated database when this becomes available.

To conclude this paper I wish to give a brief overview of the excavation database system which I am now developing. First, the database will consist of two types of file; a DEFINITIONS file containing information required for the interpretation of the data and one or more DATA files containing the actual data for a particular site. Different types of data record within the data file or different types of definition within the definitions file will be identified by a record-type code in each record, allowing subsequent additions of completely new types of data; the use of data files which may contain any sort of information rather than strictly separate files for different types of information, reduces the need for a priori restrictions on the nature of the data to be recorded. Such a context-independant structure is only possible in the present situation because we are dealing with fairly small amounts of data to which access need not be very rapid, allowing us to do away with rigid structuring and indexing of the data file as would be required in, say, a site-registry or business application. Indexing (to avoid searching the whole of the data to read or insert a particular piece of information) can be effectively performed simply by manipulation of files or physical media (tapes or floppy disks) and run-time indexing.

As far as possible data will be split up into blocks corresponding with different types of data so that, for example, data relating to stratigraphic levels as a whole will be in one contiguous block as will be data relating to a particular excavation unit or set of excavation units from one particular excavation area. This in turn means that the data can be split up into a number of separate files, each dependant on the definitions file and relating to a particular type of data and/or physical subdivision of the site; these separate files may then be stored on a series of floppy disks which may be mounted according to the data which is to be accessed. For example, one might have one floppy disk for each excavation square or group of adjacent excavation squares, one for variables relating to stratigraphic levels or sediment samples, one for commentary referring to the site as a whole and one containing the definitions file; however, there is no a priori division into data types so that sediment sample data and site commentary could appear on the same disk.

The subdivision of data in this way brings several advantages. Firstly, it circumvents the need to store indexes as part of the data since the index for a particular floppy disk can simply be built up at run time, if required, without excessive CPU usage or delay. Secondly, one can produce backup copies of parts of the data which have been modified without having to backup the whole database. Thirdly, it of course allows one to handle a large database with only a small computer. In addition, a single definitions file may serve data files for several different sites if they have been recorded using an identical system (commentary relating to each individual site will appear on one of the data files not on the definitions file).

Individual items of data will be accessed by mnemonic names rather than by their position on a record. Data storage will not be in fixed format records but by means of a series of variable code/variable value pairs which will be encoded and decoded by a subroutine in the programs. Whilst this system is less economical in storage and CPU time than a fixed format for a fixed list of variables with few missing values, it should prove more economical where variables are not present consistently and provides much greater flexibility. If the system encounters an unknown mnemonic name it will query the user and, if required, add the new variable and its definition to the definitions file. It will also 'learn' new input or output formats in a similar way, allowing them to be called by name rather than being typed in when required or stored within the programs. Field widths and variable types will be determined automatically. This means that the system for a particular site can be built up bit by bit from the terminal as processing progresses and new data becomes available, without the user requiring any knowledge of programming language or having to make a priori decisions on the list of variables to be recorded.

In the first instance data retrieval will be limited to the preparation of fixed-format files for subsets of the data, which can then be analysed using a package or other programs. As a pattern of common variables emerges for a given geographical zone special-purpose programs could be added to carry out particular commonly required manipulations.

As far as possible the entire database system will be designed to run on any 32 - 48K microcomputer with two or more floppy disks. For final storage and analysis the data would probably be transferred to a mainframe computer and, at this stage, the separate data files would probably be consolidated into a single indexed data file for ease of handling. On the basis of the storage requirements for my excavations at Capertee 3 (New South Wales, Australia) I would expect most sites to occupy from one half to ten megabytes of storage; the latter figure would be for an excavation of 50 cubic metres or more. Typical storage capabilities of a minifloppy disk lie in the range 100 to 300 kilobytes, so we are talking in terms of extremes of 2 to 100 disks per site. Larger projects might go for larger floppy or hard disks, so the maximum figure is unlikely to be reached.

Conclusion

Up till now archaeologists have been largely concerned with computers as number-crunchers for carrying out sophisticated or tedious analyses. This concern is reflected in an inflexible and unimaginative 'cards to number-crunch' approach. In this paper I have argued that computers have an equally important role to play as a means of data management. Given an effective database management system they may be able to alleviate the perennial problems of poorly published sites and excavated material stored without a systematic catalogue.

It now appears possible to design a flexible database management system for excavation data which is suitable for the so-called 'personal' microcomputers. This should encourage much greater interaction between the system and the user, owing to the possibility of dedicated machines in laboratory or even field situations. Not only will this lead to systematic cataloguing of excavated material in a form which can be duplicated for storage with the material, but the increasing availability of such machines should also encourage greater interchange of data, perhaps leading to the establishment of centralised data repositories.

Bibliography

Caelli, W. in press Bits from pits: information analysis and archaeology. *in* Johnson, I. (ed) *Holier Than Thou: Proceedings of the 1978 Kioloa Conference on Australian Prehistory*. Publications of the Department of Prehistory, Australian National University, Canberra.

Johnson, I. 1979 The Getting of Data: a case study from the recent industries of Australia. Unpubl. Ph.D. thesis, Australian National University.