MASS DATA COLLECTION:
THE VIEWPOINT OF A UNIT

J.S.F. Walker

Greater Manchester Archaeological Unit
Department of Archaeology
Manchester

In 1973 the term "naive user" was coined (Damodaran, Stewart and Eason) to describe computer users from a wide range of non-computing backgrounds forced to work with computers. This paper is the product of such a naive user.

It deals with three main topics. Firstly, a brief description of the Greater Manchester Archaeological Unit computer system, secondly, a review of some of the practical difficulties that the Unit has come up against; thirdly a short look at some of the slightly more subtle problems associated with Sites and Monuments Records.

The Greater Manchester Unit was founded in September 1980 as a full-time field archaeology unit to deal with the archaeology of Greater Manchester. For the last 18 months it has actively been involved in developing a computer system. At present the Unit has access to a Joint System ICL 1906A/CDC 7066, CDC Cyber 170-720, a Graphics Unit PDP 11/45, all of which belong to the University of Manchester Regional Computing Centre (UMRCC), as well as two micros, a CBM 8032/8050 and an APPLE II with full colour graphics.

The main responsibility of the computing division of the Unit is the recording and manipulation of large amounts of data from excavation and survey work, such as context sheets, sites and monuments records. The main data base system that is used to cater for this is FAMULUS, which runs on the Joint System 1906A/7600.

FAMULUS itself is available in various forms on most University computers in the UK. At Manchester it has been up-dated by B. Mandl, UMRCC, and Mr. C. Pettit, Computer Cataloguing Unit, Manchester Museum. Further revisions of FAMULUS are in the pipeline and it is currently under study by a national University Committee. The basic record within FAMULUS can be of up to 60 fields or 4000 characters in length. The Unit uses these records in a subject orientated way, so that what we have are a series of sheets dealing with such diverse objects as documents, contexts pottery, worked stone, bones, etc. The inputting of these sheets is done interactively, either on the Commodore or on the Cyber and the resultant files transferred through to the Joint System where they can be sorted, catalogued, or multi-logic searched. Output can be obtained either on paper, tape, disc or microfiche.

In addition to this primary data base system, we maintain systems for the production of graphical output which will be described by Mr. G. Briggs (see below). In addition to being used for graphics the Cyber, the CBM and the APPLE are also used to run various training programs, and occasional use is also made of vast numbers of packages available at UMRCC.

Since the inception of the Unit we have managed to use six different machines in an integrated way and, so far, have succeeded in capturing a significant proportion of the data with which we have to deal. In that sense the Archaeological Unit serves as an interesting test case because it is one of the relatively few full-time field organisations in this country that has adopted computers as standard tools.

What have been the successes and failures of this approach? And how has it been possible?

The most significant factor in the first place is that we obtained a strong policy decision that the high cost of implementing a computer system would be valid in the long run, as the system should prove timesaving and lead to improved efficiency across the board. Secondly, within the 41-man organisation direct responsibility for the handling of the system was placed under the second-in-command. It is a commonplace of data management text books that the Data Manager should be of a Senior Executive position for two main reasons.

(1) The implementation and design of a system cannot succeed without the officer in charge being able to control multi-role users and their disparate desires.

(2) The control of the whole data base should allow the manager an opportunity to review developments and spot inefficient operations within the organisation, but outside of the direct purview of computation.

These two factors were of considerable importance. It meant that the labour involved in running up to the implementation of the system was guaranteed and that one could avoid a vast number of minor political and workscheduling programmes by placing it in the hands of somebody capable of adjusting the work of sub-departments. Against this placing of control in the hands of a "senior person" are the problems that it is unlikely that he/she will have the time to devote himself fully to the project and that there is a danger that personal "pet" projects will receive undue consideration (see Chandor, 1976, pp 19-26).

Having obtained these two important decisions, we were faced by a task that really divided into two main parts:

(1) Mass data handling.

(2) Data Conversion Packages or Programs for handling such things as aerial photographs and resistivity surveys.

When it came to the problem of finding a mass data handling system one was confronted with the choice of developing one oneself or taking and amending an existing system. It was obvious that it would need considerable development time to develop a system and, even if one took a pre-existing system that was around on another site, it would still take considerable time to sort out the machinery at Manchester. At this stage it was decided that rather than try to design a system which was perfect in all dimensions, we would take an existing working system because it could be implemented quickly. We took the FAMULUS system because of the extensive development and refinement of it that has taken place in Manchester, and also because the advantage of the system was that the structured field approach to a record meant that data was manageable in a variety of ways. It is a very extensive system in that it is capable of making searches containing up to 60 logical operators. It is also possible to recapture data in a flexible way ready for conversion to format suitable for any other machine, should there be the need for a transferral of data.

Under this system, then, we have sheets dealing with a variety of topics. Each sheet or record consisted of the series of fields. It turned out that it was not the data base system itself that posed the greatest problems, rather it was designing the sheets because this was found to be a difficult and subtle process, and I would not say that, so far, we have completely succeeded in all ways. Some basic points have come out though from the data capture and inputting part.

Firstly, if you give someone a fairly narrow number of terms for one specified field, or if you have restricted entries such as species on the bone, on an average it can take one minute per field to view and input the attribute. This slow rate gets even worse if one exceeds an apparent psychological barrier that lies between 20 and 30 fields; beyond this the encodation rate is extremely slow. We found that free text was essential in any sheet it only to maintain the interest of the inputters and to cater for exceptional cases.

As regards inputting, it is quite clear that commercial style card punching is not the answer because of the high level of errors that occur. As much, or even more, time can be spent editing data bases as creating them in the first place. We have found by practice that direct interactive inputting, in our case onto either CBM or Cyber Files, is by far the most successful. The actual pressing of the carriage return button makes people check what they have entered and, equally important, random data input can take place as and when events occur. The design of inputting programs can be made simple, and these inputting programs can have built into them checks to ensure that only the correct restricted terms occur.

As I hinted before, the layout and contents of sheets is a problem. In some areas it is very simple - for instance when dealing with bones. The number of observations that can be made about them are, in practical terms, finite and well defined. In other areas, such as pottery, one suffers primarily from a lack of a coherent philosophy about what are significant attributes. Rather than commit oneself to researching that question, a more empirical approach was adopted in which it was a case of "Oh! This seems interesting, or is pertinent to our study, therefore we will use it", which is at slight variance to what we felt was our archival role in creating the data base. After a pilot study using four different operatives on over 250 sherds of pottery, each encoded separately, it was clear that certain common concepts of attributes in pottery were invalid because different encoders would arrive at different results and consistency was impossible to achieve. The major failures were in the realms of texture and the condition of the sherd.

Having thus eradicated inconsistencies or dubious terms, the problem is to contract encodation to ensure that only useful attributes are recorded. At this stage one is to say that perhaps we have made a basic mistake. Surely the only attributes where they have been recorded are those that are pertinent to our study. If one hypothesised that the length of sherd was a significant factor, then surely only recording their lengths was what was needed. As the approach this statement seems very pertinent; but misses 3 practical points.

(1) An underlying emphasis in our system on the creation of an archive. Such an approach often necessitates broad spectrum recording in an attempt to cover future demands. (For a general approach to the problem, C. Pettit, 1981).

(2) Much has been said of the hypothetico deductive approach and how the original hypothesis per se is not a significant event. In practice, however, where do such hypotheses generally arise from? A perusal of the data base.

(3) Damodaran, Stewart and Eason (1973) have suggested that there exists three main types of data base users; managers, specialists and clerks. Managers are typified by their desire for generalisation and clear output at the cost of little effort. Specialists, on the other hand, require detailed information and complex functions. The manaers and specialists form our main users. The data base must clearly cater for both, and in practice this often means adding attributes to a specialist list, and so hence a further archival emphasis.

The two factors above, archival and deductive, are obviously the same thing in different guises. I decided that what was important was to generate a data base that was both pertinent to individual studies, but capable of generalisation. A data base which could be used as a tool directly in the generation and the testing of various hypotheses. A simple empirical approach has been developed to consider what are useful attributes. Firstly, the specialist develops his list, to this is added attributes needed by management and other specialists. A broad spectrum study is then done on a pilot group of objects and attribute use, and consistency tested. Inconsistent and unused attributes are then unreferred to and eradicated. With this in mind, let us go on to consider the problem of Sites and Monuments Archives.

The present Sites and Monument Record systems available in this country, which are operated by over 60 organisations, are largely based upon Ordnance Survey Record Cards, which record sites in terms of type and have large quantities of free text. At county and unit level across the country there has been a movement towards consolidating and expanding this basic record system, and at present some 8 are computerised. I should imagine that somewhere in the order of 200,000 sites 78 megabytes are recorded in the UK on this broad basis (survey of surveys). A random survey of the record terms or attributes used in recording these sites shows a broad standard pattern. Most systems of recording can be broken into two parts:

Administrative)	-	county boundaries
Local)	-	political needs

Archaeological) - type

At the moment the D.O.E. are using an OH10 C3C processor to handle information on Scheduled Ancient Monuments in England and Wales. They have also offered "advice on the organisation and development of sites and monuments records and appropriate software and hardware for machine-based systems" (D.O.E. Advisory Note 32. Ancient Monuments Secretariat). Their 31 field records contain only 11 which rely heavily on defined 'type' entries. To input, maintain and develop this system there are two individuals, an OH10 C3C and a 26 Mb Winchester disc. What is not discussed in the Advisory Note is the crucial future use to which the system is to be put. The data really will not satisfy the criteria we are discussing today, each site will be "unique", and whilst the D.O.E. can advise on a national system they have failed to publish a rationale of the system in terms of its end use other than as an internal administrative tool.

It is the archaeological part of the sheets that is of most concern. The vast majority of sites are recorded by "type". These type entries consist of variably defined terms such as 'moat' 'deserted medieval village' 'farmstead' 'bell barrow'. Although some of these terms are closely defined, for instance 'bell barrow', and relate to a clearly defined attribute list, others such as 'deserted medieval village' are extremely open.

Let us consider what the use of these "type" terms may mean in practice. Any basic lack of definition would obviously mean that two dissimilar objects may be classified as one. A slightly more subtle problem, however, is that type classification in this form is rigid. Any re-definition of the vague term 'deserted village' for instance means massive data editing jobs, which will be bedevilled by the fact that the only entry of relevance will be that that says in the first place 'deserted village'.

At the moment the NMR is trying to tackle the type problem by compiling a complex thesaurus of definitive terms. Indeed, to quote Ziman (1979, p 160) "the problem of classification is fundamental, and cannot be settled by an arbitrary convention which has no roots in reality." It is easy to define formal categories but the crucial test remains as to whether they are sharp, significant categories: significance being whether they are "well defined, stable, consensible and meaningful as elements in a conceptual scheme" (Ziman 1979, p 162).

A practical answer might be to take the view that every site has a number (perhaps infinite) of attributes. These terms can be limited to those attributes that are readily seen and understood in the field. The basic units of the record should be the physical characteristics which when grouped may be considered to indicate a specific type of site. Such an attribute recording system is promising in that:-

(1) It is flexible in that a redefinition of a term merely means looking for different attribute groups, and not changing the data. In other words we may test hypotheses against it more easily.

(2) It is objective in the sense that a site may be recorded on

its positive characteristics and not on an instant interpretation.

The slow realisation of this obvious fact will mean large-scale restructuring of our own GMAU data base. It is already clear that it will be difficult in some cases to design a physical field attribute recording system that will clearly show one group of attributes as indicating a 'deserted village' for instance. Surely then, the system must fail. In reality not so; field archaeology is bedevilled by terms that are almost philosophical constructs nearly impossible of proof in this field because of their lack of sharpness. If a term is so loose as to need a great deal of unavailable information about the site before it is valid, then it is certainly redundant from the point of view of a field survey team, who generate the data in the first place. We must in this problem consider what our basal date is; are we dealing with an ever-expanding base of information and trying to apply to it growingly invalid terms that are divorced from a conceptual scheme?

At the moment the NMR are trying to obtain some £30-40,000. to operate a possibly XENIX organised data base, their primary objective being to record the data held by the Royal Commission at a basic level pending a further possible expansion.

I would estimate that on average each of the present SMR organisations holds some 1.75 megabytes of information. A lot of these organisations are at present moving towards computerisation. It is impossible at the moment to assess how much directly and indirectly such moves are costing the body archaeological, but to provide suitable micro material for them would cost between £150-250,000. From within my own limited orbit of experience with the GMAU system, it would appear that SMR systems are relatively rarely searched per se rather what is required are sorted catalogues and distribution maps. I would have thought that the most economical way to solve the SMR data base problem would have been regionalisation, because:

(1) Maintenance of hardware and software is simplified.

(2) Regional centres allow shorter travel to and communication distances.

(3) It will ensure limited duplication of effort in terms of inputting and programming.

(4) It should allow greater developmental and study time.

The cost of most of the present options open are prohibitive, but I would have thought that with care regional centres can be built around existing facilities and prove cheaper and more effective in the long run.

I have today, then, reported on aspects of GMAU's work. If this paper is to have had any value it will be to point out:-

(1) What the Greater Manchester Archaeological Unit is doing.

(2) To show some of the problems that the Unit have faced, and how they have been overcome.

(3) To hint that even simple tasks like recording sites are difficult.

BIBLIOGRAPHY

Chandor, A. 'Choosing and Keeping Computer Staff'. George Allen and Unwin, London, 1976.

Damodaran, L., Stewart, T.F.M., Eason, K.D. 'The Needs of the Naive User' in DATA FAIR 73 CONFERENCE PAPERS Vol. II pp 384-391 BRITISH COMPUTER SOCIETY, London 1973.

Pettit, C. 'The Manchester Museum Computer Cataloguing Unit: A Step in the Right Direction?" MUSEUM DOCUMENTATION ASSOCIATION, March 1981.

'Survey of Surveys 1978' ROYAL COMMISSION OF HISTORICAL MONUMENTS (ENGLAND), HMSO 1979.

Ziman, J. 'Reliable Knowledge' Cambridge 1979.