# Information Retrieval and Graphics at Danebury and York

Myra L. Shackley (Institute of Archaeology, Oxford)

A. G. MacGregor (York Archaeological Trust)

J. M. Duncan (Research Centre for Computer Archaeology, North Staffordshire Polytechnic)

## Introduction

This paper describes the information retrieval facilities of the PLUTARCH system (Wilcock 1974; Shackley and Wilcock, 1974) and their use in two archaeological applications: the excavation of the Iron Age hillfort at Danebury; and the cataloguing of small finds from excavations carried out by the York Archaeological Trust. The paper is in three sections:

The first section (Duncan) describes the computer-based information retrieval system used in both of these applications, giving details of the methods of recording, storage, retrieval and presentation of archaeological data available in the system.

The second section (Shackley) describes the nature of the Danebury site, the way in which data from the site were recorded and stored on computer files, and the results of the excavations to date, with reference to the manner in which computer-produced lists and diagrams have been used in the examination of the site.

The third section (MacGregor) describes the nature of the information on small finds handled by the York Archaeological Trust, and the way in which computer-produced catalogues of this information will be used by the Trust both for its own purposes and to supply the needs of other archaeological researchers.

## I THE COMPUTER SYSTEM

The information retrieval facilities of the PLUTARCH system were outlined at a previous conference (Shackley and Wilcock, 1974). The system described therein has been substantially altered, and this section describes the current facilities of the system.

This system is designed to be a general purpose one for the storage and retrieval of archaeological data. It was implemented on the ICL 4130 machine at Keele University, and the programs are written in ALGOL 60.

The system was designed to handle collections of archaeological data which have the following two basic characteristics: (i) a large number of archaeological entities about which common categories of information are to be recorded, e.g. dimensions, provenance, etc. (ii) the need to store and retrieve this data in various forms, e.g. lists, diagrams, catalogues, etc. The following paragraphs give details of the components of the system.

1.1) Data Definition and Description

This is largely the responsibility of the user, who must decide what data to record about each entity, and how to record it. Upon this decision will rest the effectiveness of retrieval. There are three possible approaches:

i) Data from an existing record may be transferred directly to the computer

ii) A system such as the IRGMA record card may be adopted.

iii) An individually designed recording system may be devised, and all data recorded in this format.

The first approach has the advantage of involving minimal reorganisation of data. The second brings standardisation on a national scale, and expert guidance on data recording methods. The third approach applies to cases where no adequate data format is in existence, such as a format for Roman Inscriptions. It is possible to combine (ii) and (iii).

When the user has decided upon the recording method, a table is generated which defines each item of information in a record, and this table is stored and used to interpret input data. New fields may be defined and added if the need arises.

1.2) Manual Recording of Data

The system of data recording is based on 80 column punched cards. Each card has three major fields:

i) Record number: this is punched in columns 72 to 80. A sequence of cards with the same record number constitutes a record for the computer. Record numbers may be derived from an already existing identity number associated with each entity, or may be arbitrarily assigned.

ii) Card type number: this is punched in columns 2 to 3 of each card in a record and denotes what information is to be found on that card. The user may designate certain types as common to a series of consecutive records, so that certain information may be associated with a number of records until superceded by different information on a card of the same type. This allows hierarchical arrangement of information, e.g. pit data common to consecutive pit layers (see Section 2). It also avoids duplication of information in the manual preparation stage.

iii) The data field: data is punched in columns 11 to 70 of the card, each card having one or more sub-fields, the size, position and contents of each being entirely at the discretion of the user. Four kinds of data are allowed: alphanumeric strings, integers, real numbers (i.e. with decimal point), and boolean (1 or 0 for presence or absence). Where alternative entries occur, e.g. two possible dates, a repetition figure is placed in column 1 of a subsequent card of the same type, and the alternative entry placed in the appropriate position. There is also a facility to allow entries longer than the allotted field to carry over onto subsequent cards, thus allowing long descriptive sentences to be recorded.

1.3) Automatic Checking of Data

The punched cards are automatically checked for such errors as mis-alignment of data, mispunching of numbers, etc. Manual checking for such errors as spelling mistakes is also necessary. Cards with errors are listed, and correct card images are stored on magnetic tape. Invalid cards are corrected manually and re-input via an amendment program.

1.4) Creation of the Main File

The file of correct card images is processed to create the main data file used for retrieval. In order to reduce the volume of information stored, and thereby the processing time, the data is coded, each item of data being stored with information to indicate its nature. This enables direct referencing of information in each record instead of serial access in main store. In addition, every occurrence of a word in string fields is replaced by a number, the original word being stored in a dictionary held on a disc for decoding the information for output. The effect of this coding has been found to reduce the volume of data stored by a factor of about 5.

The main file can be built up in a series of steps, with previous versions of the dictionary and coded main file being updated by successive runs of the coding program.

1.5) Retrieval of information and presentation of results

When information is to be retrieved from the coded main file, the user must supply three program parameters. These are:
i) the boolean search expression: this consists of one or more relations of the form

< field > relational operator > value <

or < field > relational operator > field <

The field indicates the item of information in each record which is to be tested for the specified relationship.

The relational operator is a two character item indicating the relationship to be looked for between the field and value or other field. The six arithmetic operators, EQ, NE, LT, GT, GE, and LE are allowed, together with four operators for strings, SE (equal), SN (not equal), SI (contains), and SX (does not contain). The value or second field indicates the number or string with which the first field is to be compared.

Complex boolean search expressions can be built up using the connectives AND and OR (the NOT connective is not included because boolean fields are treated as arithmetic 1 or 0), and levels of logic can be denoted by the use of the brackets (and).

ii) Mode(s) of output: records conforming to the search criteria may be extracted in whole or in part, and in a number of different modes, all of which can be in operation during a single run of the retrieval program. These modes are:
PRINT. Specified fields are output from each valid record to a backing store file for subsequent listing in a desired format. The user may control the listing format himself or may specify output in a standard format. A facility also exists for sorting the output records on any number of keys, so that catalogues may be arranged in various orders. In addition the output file may be re-input to the retrieval program, thus allowing subsets of the main file to be isolated and interrogated repeatedly without the need to scan the whole main file each time.
HISTOGRAM, PIECHART One or more fields may be taken from each valid record and output in a format suitable for input to the relevant diagram-producing section of PLUTARCH.
SCATTERGRAM. One or more pairs of fields may be output in a form suitable for plotting in two axes.

MAP PLOT. Records containing grid references can be output in a form suitable for plotting, superimposed upon a site plan or map.
(iii) The fields for output are specified in a similar manner to the search fields. The order in which these fields are specified is preserved on the printout of results, thus allowing the user to arrange the results to suit his own requirements.

## 1.6) CONCLUSION AND PROPOSED DEVELOPMENTS

The system in its present state functions satisfactorily for the quantities of data upon which it has been used to date. However, as the volume of information increases, it is plain that the serial organisation and access of the files will prove inadequate because of the relative slowness of searching. It is therefore proposed to employ direct access file organisation methods, using the existing retrieval programs to 'invert' the files, so that only those fields directly involved in the search expression and results specification need be processed, thus achieving faster processing of the data.

It is also proposed to integrate the information retrieval facilities described here with the graphics-controlled PLUTARCH system, in order to achieve greater communication between user and machine. The user will be able to specify the search expression and results via the graphics console, and will be allowed greater flexibility in deciding upon modes of presentation of results. In addition, the search expression capabilities could be expanded to allow arithmetic functions as well as values. The dictionary of words from string fields will be able to contain alternative versions of words, e.g. abbreviations, so that retrieval on string fields will not be impaired by discrepancies in the manual recording of data.

## 2. DANEBURY: THE PILOT PROJECT

2:1 The general archaeology of the Iron Age hillfort of Danebury, Hampshire, has already been described in previous papers (Shackley 1976, Wilcock and Shackley 1974). The fort was occupied from the 5th century B.C. until the Roman conquest (Cunliffe 1971), the massive univallate defensive earthworks enclosing an area of 13 ha. With the exception of some early structures all phases of the occupation were associated with numerous chalk-dug pits, varied in size and shape. Over 800 pits have now been examined and it is estimated that the proposed total excavation of the fort will yield at least 5000 more, associated with a regular settlement pattern of rectangular timber buildings. There is no surface stratigraphy inside the fort, except in close proximity to the ramparts, and nearly all the archaeological finds come from the pits.

The primary function of the majority of pits was clearly the storage of grain, although some may have been used for other products and some, lined with skins, for the storage of water. After the pits were abandoned they were either left open to silt naturally or were filled with rubbish and occupation material. Standardised recording and careful interpretation of the pit sedimentation pattern is clearly necessary if the development and chronology of the site is ever to be fully understood.

2:2 The Recording System. During 1973 a pilot computerised recording system was put into operation, recording both general characteristics of each pit and detailed observations concerning the nature and contents of the various layers. Data was recorded directly onto computer coding sheets, specimen layouts of which can be seen in Shackley (1976). These records incorporated codings for pit shape and fill, together with dimensional measurements and general descriptions. The base angle of the pit walls was measured and human and animal bone records were added, as well as other archaeological information. Particular attention was paid to the sedimentation pattern of each pit, in order to establish silting models and examine silting patterns in relation to topography, pit dimension and pit date. Some 20 cards were used to describe each pit and a further 16 cards, all in standard format, for each individual layer. This made a total of over 8000 cards for the 1973 pits.

The objects of the pilot scheme were as follows:

1) to explore the possibilities presented by the use of a computerised data storage system (PLUTARCH) for recording information from pits;

2) to attempt the total recording of sedimentological information from pits, since each pit forms a single discrete unit, and to use this information in the construction of models for pit silting;

3) to answer problems posed by the recorded sedimentation patterns and to compare these patterns with pit shapes and archaeological information;

4) to list all the compiled data for quick reference;

5) to pose a series of specific questions which could be answered from the stored data, in order to assist future excavation planning and the development of a final archaeology-orientated storage format.

2:3 Results. A simple print-out of the pit and layer records produced an excellent reference manual which proved helpful in archaeological interpretation. This was used to establish a pit shape classification (Shackley 1976), based on the occurrence of two main types of pit, the 'beehive' and 'cylindrical' forms, present both as eroded and uneroded examples and occurring in the ratios of approximately 3:1. Models were made of the sedimentation patterns occurring in both types, and it proved possible to provide some estimates on the rate of fill of the naturally silting pits. This data will eventually be used in conjunction with a dating factor to estimate the total number of pits which could have been open on the site at any one time.

Graphical expression of some of the data proved particularly helpful, and included distribution maps of various features (pits of particular type or containing particular features) superimposed on a computer-drawn outline of the hillfort. Histograms were made of the dimensional measurements and a series of 2 variable scatter diagrams constructed. Unfortunately the phrase (dating) characters were not available for this pilot scheme although they have now been completed and will be used in all future work. This has limited

the amount of meaningful archaeological (as opposed to palaeoenvironmental) data which could be obtained, since it would have involved the erroneous assumption that all the pits studied were open at the same time.

The map plots are proving helpful in plotting pits of different type and size, and in the future will be used to plot out plots of different phase in relation to structures of particular type. This may be useful in predicting what types and variations of pits are to be expected in areas soon to be excavated, and testing this hypothesis after the excavation has taken place. A scatter diagram which shows the quantity of animal bone (coded in fragment numbers)plotted against the type of deposit (fill pattern) of the pit shows that the largest quantity of animal bone occurs in pits which are wholly or partly filled with occupation material, and that those pits which are naturally silted contain relatively little bone. However a plot of the pit volume against the quantity of bone shows that the most bone does not necessarily come from the largest pits, and it seems on the whole to be unrelated to volume but closely related to the type of sediment.

The above remarks provide an indication of the type of results which have been obtained from the pilot survey, and these have been sufficiently encouraging to warrant the adoption of a computer based recording system for the whole excavation. This system, which began in the 1974 season, will be more archaeologically orientated and will aim to shed further light on the topics mentioned above.

## 3. YORK: THE ARCHAEOLOGICAL BACKGROUND

Since the formation of York Archaeological Trust in 1972, excavations have been mounted continuously within the city boundaries. An important and ever-increasing legacy of these excavations is the large body of finds which, by December 1975, already totalled over 9,000 individual items (excluding pottery). An invaluable feature of this new collection, already outweighing in sheer bulk many museum collections, is that nearly all of it comes from archaeologically-stratified contexts. Most of these new finds will eventually be published in the fascicule series instituted by the Trust, but in the meantime they form potentially an important source of information for researchers in the fields of Roman and later archaeology. In order to maintain maximum accessibility of this information, an efficiently cross-referenced catalogue is obviously of prime importance. After some initial experimentation with various manual schemes, it became increasingly clear that a computer-based system was going to be essential, and following a meeting at Cambridge in 1974 of museum staff and field archaeologists, held under the auspices of the Information Retrieval Group of the Museums Association (IRGMA), it was decided to adopt the proposed IRGMA archaeological record card as the basis of the York system. Since the associated IRGMA handling programs are still in the developmental stage, however, an interim system of machine-processing the recorded information was sought and was eventually provided by the Research Centre for Computer Archaeology, Stafford (see above ).

The adoption of the new card means that the Trust's records will be compatible with those of all other users of the IRGMA system in rescue units and in museums; in particular, the harmonisation of the Trust's recording system with that of the Yorkshire Museum, in which it is anticipated that most of the finds will eventually be deposited, means that there will be only minimal administration involved in accessioning new finds. The machine-processing facility, on the other hand, will render the information recorded on the cards very much more accessible than has hitherto been the case.

Current practice at York, as in most excavation units, is to make only a minimal record of each find as it leaves the ground; its accession number, context number and site coordinates are noted in a finds register, along with the name of the finder and a brief description of the object. At the present time, this often remains the only formal record of the average find until it reaches the research stage, some months (or even years) later. With the adoption of the IRGMA system a rather fuller record of each artifact will be made during regular viewing sessions on site and the information transferred quickly to the computer, so that the catalogue remains up-to-date and comprehensive. Information recorded in this way may then be recalled under various headings as follows:-

BY SITE. Each site is identified by a date-linked code, e.g. 1972.21. The identity numbers ascribed to the various sites are taken from a block of such numbers reserved each year for the Trust within the Yorkshire Museum's system.

BY ACCESSION NUMBER. Accession numbers on each site start from 1 and are always quoted along with the site code, e.g. 1972.21/843. Thus every find handed to the museum is already fully identifiable within the museum's system.

BY SIMPLE NAME. For example, bowl, pin, buckle. Keywords are generated locally as required, but in the event of some uniform national system being developed these will be adopted, if possible.

BY MATERIAL. Expressed as simply as possible, e.g. iron, leather, copper alloy. When an object is made up of two or more materials, or when it would be useful to recall under more than one heading (e.g. under 'stone' or 'shale') several options can be recorded.

BY DATE. By recording both 'period' (e.g. medieval) and/or earliest and latest possible dates for the particular item (e.g. 1200-1300), a certain amount of flexibility can be maintained in the early stages of decision-making.

BY CONTEXT NUMBER. That is, by the number assigned to the particular layer on site from which the object was recovered. These numbers are assigned in a single series starting at 1 on each site.

BY CONTEXT DATE. Recorded in the same way as the date of the object; the two may or may not be related, since the widespread disturbance common on urban sites results in much early material being redeposited in later layers.

BY CONTEXT TYPE. Potentially useful in retrieving well-sealed pit-groups of closely contemporary material.

A free text description of each find is included and information is also recorded on dimensions, conservation processes, drawings and photographs.

As recorded above and reproduced in machine-printed catalogues, the information can be ordered in a number of useful ways: arranged by site and number, the catalogues will form the basis of the Trust's site record systems; by simple name or by date, the system will give easy access to particular classes of finds or to items of particular chronological periods respectively; by recalling on material, groups of items with common needs in terms of analysis, conservation, specialist reports, etc., can be easily isolated; the ability to consult up-to-date and comprehensive lists of finds from each context will ease the site supervisors' task of writing interpretative reports on the features which he has excavated. Together, all these catalogues should form a very adequate 'Level two publication' as defined in the Department of the Environment's recent PRINCIPLES of PUBLICATION IN RESCUE ARCHAEOLOGY (1975). The needs of research students and other interested parties should be efficiently met: the specialist in bronze buckles of the fourteenth century, for example, can be provided with lists of buckles designated on typological ground as being fourteenth century or of all buckles from fourteenth century contexts.

For the moment the scheme will serve purely as a recording system, but it does have the capacity to be used analytically and it is anticipated that this will form an important part of its function in future years.

## Bibliography

Cunliffe, B.W., 1971. 'Danebury, Hampshire: first interim report on the excavations', Ant. J., 51, 240-53.

Shackley, M.L., 1976. 'The Danebury project: an experiment in site sediment recording'. Geoarchaeology, ed. D. A. Davidson and M. L. Shackley, London, Duckworths.

Wilcock, J.D., 1974. 'The PLUTARCH system', Computer Applications in Archaeology 75-82.

Wilcock, J.D. and Shackley, M.L., 1974. 'The recovery of information from Iron Age pits: the Danebury project and PLUTARCH', Computer Applications in Archaeology, 82-90.