PITFALLS IN THE PREPARATION OF A COMPUTER-BASED ARCHIVE FOR PALAEO-ENTOMOLOGICAL RECORDS

John A.W. Lock

Department of Biology, University of York, York, England.

Introduction

This work derives from a research project funded by the Scottish Inspectorate of Ancient Monuments to prepare an account of the environmental conditions prevailing in the medieval city of Perth, based on a study of the insect remains (and beetles in particular) preserved in the waterlogged archaeological deposits found at a large site in the High Street.

The production of such reports is time-consuming and labour-intensive. This paper describes an attempt to use computers to speed up the process, concentrating on problems that had to be overcome and the way in which this was achieved, starting from a consideration of the problem, through selection of hardware and software, to the development of a database system which is still in use.

Why use computers? To understand this a brief description of the production of a report is given below. Essentially it comprises three stages, extraction, identification and interpretation.

1. Extraction

Samples of soil are taken from the site, and washed down using a paraffin/water separation technique to extract the insects (Coope & Osborne, 1968). The insect remains are then sorted and mounted on card slides to facilitate subsequent examination and identification. This stage takes several days for each sample.

2. Identification

The insect remains are not complete individuals, readily identified to species using conventional keys. Instead they are disarticulated fragments, only certain of which (the head, the wing-cases (elytra) and the pronotum) can usually be assigned to a particular species. Even for these, identification to the level of species is not always possible.

Identification requires access to a reference collection of identified insects and sufficiently comprehensive collections are few and far between. Thus definitive determination of the insect fragments cannot take place immediately on extraction but has to await a visit to one of these collections.

Until then fragments of closely-related species are mounted together on the same card slide.

Each sample processed produces between 40 and 80 slides each of which has one or more species represented on it. In the case of the Perth High Street site, there were about 7000 slides. A smaller site would produce of the order of 500 slides.

3. Interpretation

The first step in interpretation is to bring all the information stored on the slides together in a 'species-distribution' table. This correlates species with samples.

Once this has been achieved, report-writing can be quite straight-forward. However it can be an extremely lengthy procedure if one is working from handwritten notes. It is in the preparation of this table that the computer can be of most use.

Selection of Hardware and Software

In 1980, when the magnitude of the problem of storing and manipulating the insect data was becoming apparent, computer-archiving seemed to be the answer. There were two possible options, either to use a micro- or a mainframe- computer as both were available.

1. Micro-computer option

There were two Apple II microcomputers available through an SRC-funded research team (Hardy, 1982). These computers had the advantages of being constantly available, and able to transfer data stored on their floppy disks to the university's mainframe computer (a Decsystem-10) using a purpose-built file-transfer system.

Disadvantages were that the size of data files was potentially limited (although this could be avoided by frequent transfer of stored data to the mainframe computer) the small core size meant that sorting programs ran extremely slowly, no machine-language sort module being implemented. The greatest disadvantage was the ease with which data stored on floppy disks could be lost irretrievably, a considerable problem for the SRC team (Hardy, 1982).

2. Mainframe computer option

The Decsystem-10 had the advantage of effectively unlimited data-storage capacity, with a low risk of disastrous data loss. It also had various sorting packages implemented.

Its limitations were that it had to be shared and was therefore not always available for interactive data input. Apart from the problem of physically limited access was that response time tended to be unacceptably show when it was used during the day.

A compromise solution was arrived at, similar to that chosen by the SRC researchers. An interactive BASIC program was written to input data which, with minor modifications, could be run on either the mainframe or the micro. It was thought that this would save time by removing restrictions on the availability of the mainframe computer.

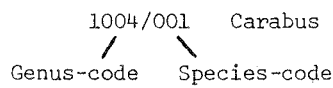
At the same time the entire checklist of British beetles (Kloet & Hincks, 1977) was encoded. This was done because there are almost 4000 species of beetles in Britain, each with a unique name. These names vary in length from 8 to 36 characters. For example, typing in

'Pterostichus adstrictus'

(an average-length name of 23 characters) gives at least 23 opportunities for a typing error to occur. Trapping such mistakes at the input stage requires cross-checking the input name against a file of some 4000 names adding considerably to the execution time of the program and also to the time taken to input data. Failing to cross-check can be disastrous if records are sorted by species-name.

Using codes facilitates rapid ordering of species-lists into the checklist sequence. In addition it gets around the taxonomic problem of synonymy. A glance at the British Checklist should confirm the importance of doing this.

Each British species was assigned a code, comprising a genus-code and a species-code. This was preferred to a unique numerical code because it was considerably more flexible, permitting the interpolation of new species into the list. Also two numbers seem to be easier for people to remember than one on the 'name-surname' principle.



For identifications to levels other than that of species two further codes were added, 'Group-status' and 'indet-species'.

```
 1004/0002   Carabus 2 spp. indet.
T1004/0001   Carabinae spp. indet.
F1004/0002   Carabidae 2 spp. indet.
      /          \
Group-status  indet-species
```

Implementation

However in use this system proved very unsatisfactory because it was extremely inefficient at the date-entry stage. It took much too long to input and confirm each record because of the error-checking subroutines and disabling these to save time allowed many typing mistakes to pass unnoticed.

Other ways to reducing the input time for each record were either not to enter any data at a VDU but to use some other means, or to reduce the number of data items in each record. In the event both steps were taken.

The first step had several important implications:

1. There would be no need to use the micro-computer

2. Hence the progrmas would not have to be written in BASIC

3. Data would be entered either as punched tape or cards

As computing activities were now going to be restricted to the mainframe computer, COBOL was chosen as the language for the new programs because on the Decsystem-10 it had the most efficient sort module, capable of sorting thousands of records by several different keys extremely rapidly. In addition it was extremely good at file-handling.

Punched cards were chosen as the input medium because errors are easier to detect and remove than on punched tape and the university provided a card-punching service.

As insects were identified the codes were written direct onto standard coding forms (see figure 1) which were then handed to the card-punch operators.

So from being a progressive micro/maingrame linkup the recording system reverted to being extremely traditional in nature!

In the second step the volume of data stored per record was considerable reduced. Initially it had been thought best to record as many details as possible.

Fragments were divided into the following groups:

Heads - Complete, Left and Right fragments
Pronota - Complete, Left and Right fragments
Left-elytra - Complete, proximal and Distal fragments
Right-elytra - Complete, proximal and Distal fragments
Elytral fragments
Other fragments

- a total of 14 categories.

In retrospect, many of these categories appeared to be redundant. Given that the objective was the rapid production of species-lists to publication standard and that space in journals is extremely limited, there seemed little purpose in storing data which was unlikely to be seen again.

So the fragment-data held in each record was reduced to 6 categories:

Heads
Pronota
Left elytra
Right elytra
Elytra fragments
Minimum number of individuals (MNI)

This was achieved by combining certain of the fragment-types. It was felt that having successfully identified the insect remains the user could be relied on to perform simple additions!

This had the bonus of reducing the total length of a record to less than 80 characters. Hence each record occupied a single card, and a single line on a VDU screen.

One needs to store other information, principally location data. This falls into two groups, site- and laboratory- data. Site-data shows which part of the site the beetles came from and the laboratory-data enables one to relocate the beetles once they have been mounted, identified and stored.

The layout of a single record is shown in figure 1.

J. Lock
Biology

RECORD FORMAT FOR ARCHIVE

PH575 2434 A 123 4 F1004/0002 0001 0000 0004 0005 0003 0006
| SITE-DETAILS | LAB-DETAILS | SPECIES-DETAILS | FRAGMENT-DETAILS | MNI

SITE-DATA
- SITE: PERTH High Street
- SUB-SITE: 75
- LAYER #: 2434
- SUB-SAMPLE: A

LAB-DATA
- SLIDE #: 123
- FLOAT #: 4

INSECT-DATA
- TAXONOMIC: 'Carabidae 2 spp. indet.'
- MIN. HEADS = 1
- MIN. PRONOTA = 0
- MIN. LEFT-ELYTRA = 4
- MIN. RIGHT-ELYTRA = 5
- MIN. MISC. FRAGMENTS = 3
- MNI (THIS SLIDE) = 6

NOTES:

1. ONLY UNDERLINED KEYS ARE COMPULSORY
2. IF NO SPECIES ARE RECORDED FROM A LAYER THE CODE 1555/0050 IS USED.

Figure 1. Record format

Archiving published data

Up to now only 'raw' (laboratory) data has been considered. In order to understand the implications of the data produced in the laboratory one has to compare these archaeological assemblages of beetles directly with others of known age, locality and type. Therefore the COBOL system was used to produce a database of published records of interglacial, archaeological, and modern beetle assemblages. This presented certain problems.

1. Synonymy.

This was described earlier; suffice it to say that it was far worse here.

2. Non-British Species and Checklist Errors

The beetle fauna of Britain is declining (Hammond, 1974). Many species have disappeared in the past 40,000 years, and hence are not included in the Kloet & Hincks checklist which only covers species currently living in Britain. So a method of incorporating 'new' species into the code-list was required. The solution chosen was to use a facility of COBOL, the ISAM file (Indexed Sequential Access Mode), and to modify the species-code:

British genus + species	Carabus nemoralis	1004/011
British genus + foreign sp.	Carabus fuziwuzzi	1004/211
British species, list error	Genus neglectus	1004/311
Foreign genus + species	Genus novus	1004/501

Summary

The development of a computer system to aid the production of insect reports has been described.

Initially it involved a mainframe-micro link, with interactive data entry. This proved disappointing in use.

It was replaced by a conventional data-processing system written in COBOL with data being entered as punched cards. This works well and has been used to prepare a considerable database.

The essential difference between the two processes lay in a redefinition of the 'man-machine interface'.

Cross-checking of input data was better done visually by the user rather than electronically by the computer, and the number of fragment-types to be input was reduced by a half by simple arithmetic performed at the time the insects were identified and coded.

Conclusions

The development of a computer-based archive for palaeo-entomological records has illustrated that use of computers does not in itself save time or increase efficiency. This depends on how they are used.

References

Coope, G.R. and Osborne, P.J. 1968 — Report on the Coleopterous fauna of the Roman well at Barnsley Park, Gloucestershire, Trans. Bristol Gloucester Archaeol. Soc. 86,84-87

Hammond, P.E. 1974 — Changes in the British coleopterous fauna, in Hawksworth, D.L. (ed.). The changing flora and fauna of Britain. Systematics Association Special Volume 6:323-369. London and New York.

Hardy, E. 1982 — A microcomputer-based system of recording bones from archaeological sites, Proceedings of the Micro-computer Jamboree, 11-18, University of Bradford.

Kloet, G.S. and Hincks W.D. 1977 — A Check List of British Insects, 2nd Edn, pt 3 Coleoptera and Streptisers, Roy. Entomol. Soc. Handbooks for the Identification of British Insects 11 London