28

The Integrated Archaeological Database

S. D. Stead
Scottish Urban Archaeological Trust

28.1 Background

The Scottish Urban Archaeological Trust purchased a computer in August 1983, with a view to providing an integrated recording system. Immediate use was made of the word-processing facilities and work started on a suite of customised databases to handle most aspects of the Trust's data management. The first set of applications fell into four broad categories: Accession Finds, Bulk Finds, Contexts and Gravestones. By June 1985 it had become apparent that these customised databases, written in Basic, lacked many of the facilities necessary to manipulate data and still contained many 'bugs'.

A Database Management System (DBMS), dBase II, was purchased and rushed into service. Potential users were polled for their requirements and database structures were developed to cope with immediate projects. Data was entered, edited and printed out in the required formats from the dBase command line. Scant attention was paid to the problems of consistency within data fields during data entry and extensive corrective editing had to be performed. This, coupled with the large number of data fields needed for some pro formas, (*e.g.* 140 for the bulk finds sheet), led to the idea of separating the user from the data structures by a front-end processor, which performed data verification and then split the data into a data bank of different structures.

At about this time some thought was given to how archaeologists interrogate the level II archive in order to interpret the site. The amalgamation of these thoughts and the experience that resulted in the development of front-end processors, led to the development of the conceptual model of an Integrated Archaeological Database (IADB). In the light of this model the existing system of *pro forma* recording was re-appraised. It was found that many of the data fields on the *pro formas* were ill-defined and that some of the data fields were applicable only in certain circumstances. For instance data fields concerning the nature of walls are applicable only to walls and not to cuts or fills. After informal discussion had failed to resolve these problems, a working party of archaeologists and computing staff was set up with the brief of 'producing a workable on-site recording system which was computer compatible'.

The working party decided that the problems inherent in 'all-in-one' *pro formas* were insoluble and a new approach should be adopted. The work of Clark (Clark in prep) on the 'toolbox' flexible approach to context recording was agreed to fulfil the requirements for both a pragmatic on-site recording system and computer compatibility. In this approach each context is recorded on a basic Context Recording Sheet (CRS), according to its type. There are four CRS's for Layers, Cuts, Walls and Interfaces. Further data about pedology, finds distribution, skeletons etc is recorded on Specialist Recording Sheets (SRS) according to an explicit manual.

Choice of which recording tools to use is defined by the site research strategy and is obviously influenced by the resources available.

In parallel to this work a series of trials was conducted with data structures for finds data. These showed that some of the data considered essential by people involved with a particular part of the recording process were not in fact part of the basic data set that they recorded. They were derived from other sections of the site record and was appended to the manual system at a later stage. This meant that the data-structures could be pared down leaving only a skeleton of information: those details derived from other parts of the data bank could then be accessed through paths laid out in the overall model.

These lessons and solutions suggested that reliance on a user's own interpretation of their needs and processes would not lead to a data structure which followed the rules of data structure design as laid out in Castro *et al.* 1985, in that data duplication would be rife and structures would rely on multiple field keys. This produced a need for users to prepare detailed statements of their current procedure and also of any tasks that they wished to perform in the future. Usually these tasks were forms of data presentation which would consume too much time if performed manually. Armed with these statements systems analysis could proceed. Consumer resistance was at first high but soon faltered as users realised the benefits of closely defining their tasks.

Two other major policy decisions were made. First that the system must be useable, at least at the input and routine retrieval levels, by inexperienced and archaeologically illiterate MSC staff. Secondly that the system must provide a communication link between specialists and archaeologists. The major result of this second decision is that the data bank must first allow the archaeologist to browse through the multiple levels of specialist data, and second, update the archaeological data that the specialist needs.

28.2 The conceptual model

At the centre of the model (Fig. 28.1) is one of four CRS's. This' forms the starting point for an archaeologist's interrogation of the data about a single context and is linked to other CRSs by a representation of the Harris matrix. During interrogation, SRS information from a context can be displayed. The archaeologist also has direct access to, for example, details of sections, photographs, specialist reports and finds illustrations. Many individual sections are linked to other sections via independent routes. For example the sherd-by-sherd pottery analysis, which details things like fabric type and internal and external colour, may be called direct from the CRS or from the tabular bulk-finds summary. In addition to this network of data exploration links, modules dealing with other tasks will be added. One of these is the finds management system, which monitors the movement of finds during conservation and the preparation of specialist reports.

28.3 Linking and tokenisation

The links between structures vary in the level at which they operate depending on the type of interrogation expected. The majority of links are two-way and are based on the context number. Other reference terms are used; for instance the finds management system uses Accession Numbers and the Bibliographic system uses internally-generated reference codes.

The degree of tokenisation of data has been restricted to four-letter mnemonics for two reasons: firstly, users prefer them and indeed use them in the paper record on occasion, and

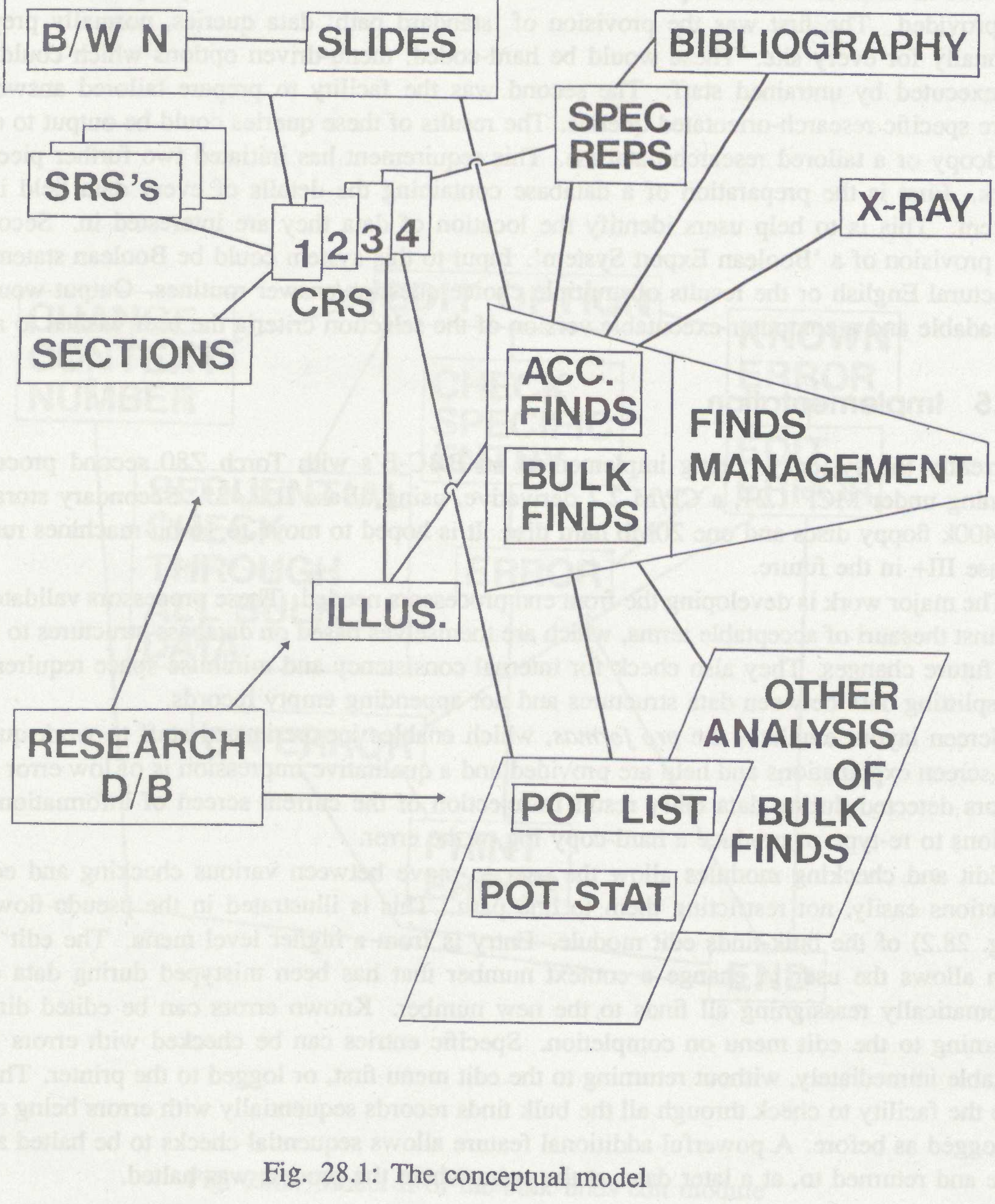

Fig. 28.1: The conceptual model

secondly dBase does not support true numeric fields, thus removing the incentive of saving space by coding.

28.4 Query methods

As well as interactive data exploration it was felt that two further data query methods should be provided. The first was the provision of 'standard path' data queries, normally prepared manually for every site. These would be hard-coded, menu-driven options which could then be executed by untrained staff. The second was the facility to prepare tailored answers to more specific research-orientated queries. The results of these queries could be output to either hardcopy or a tailored research database. This requirement has initiated two further pieces of work. First is the preparation of a database containing the details of every data field in the system. This is to help users identify the location of data they are interested in. Second is the provision of a 'Boolean Expert System'. Input to this system could be Boolean statements, structural English or the results of multiple choice question/answer routines. Output would be a readable and a computer-executable version of the selection criteria the user wished to apply.

28.5 Implementation

Currently the system is being implemented on BBC-B's with Torch Z80 second processors running under MCP 1.21, a CP/M 2.2 derivative, using dBase II 2.43*. Secondary storage is on 400k floppy discs and one 20Mb hard disc. It is hoped to move to 16-bit machines running dBase III+ in the future.

The major work is developing the front end processors needed. These processors validate data against thesauri of acceptable terms, which are themselves based on database structures to allow for future changes. They also check for internal consistency and minimise space requirements by splitting data between data structures and not appending empty records.

Screen layout emulates the *pro formas*, which enables inexperienced staff to work quickly. On-screen explanations and help are provided and a qualitative impression is of low error rates. Errors detected during data entry result in rejection of the current screen of information with options to re-type or produce a hard-copy log of the error.

Edit and checking modules allow the user to move between various checking and editing functions easily, not restricting them to one path. This is illustrated in the pseudo-flowchart (Fig. 28.2) of the bulk-finds edit module. Entry is from a higher level menu. The edit menu then allows the user to change a context number that has been mistyped during data entry, automatically reassigning all finds to the new number. Known errors can be edited directly, returning to the edit menu on completion. Specific entries can be checked with errors being editable immediately, without returning to the edit menu first, or logged to the printer. There is also the facility to check through all the bulk finds records sequentially with errors being edited or logged as before. A powerful additional feature allows sequential checks to be halted at any time and returned to, at a later date, at the point where the process was halted.

28.6 Proposed applications

Recent papers (Todd *et al.* 1986, Harris (this volume)) have discussed the use of a graphics-data exploration loop. It is hoped that this system will provide a similar exploration facility in a text environment. The archaeologists will 'explore' their data by roaming through the various levels

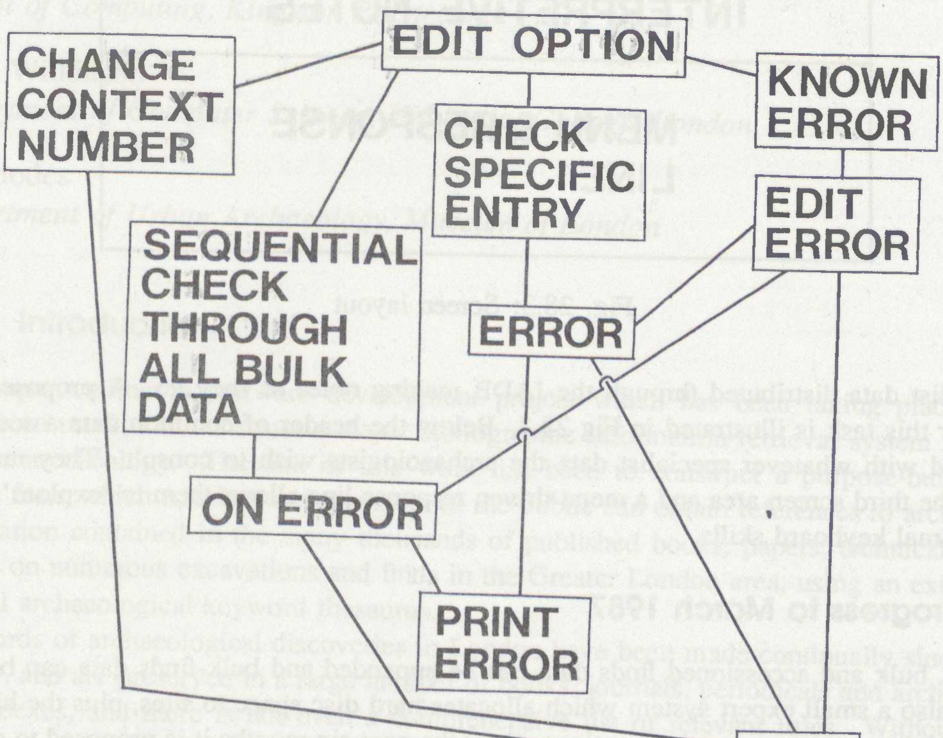

Fig. 28.2: Model II of the bulk-finds edit module

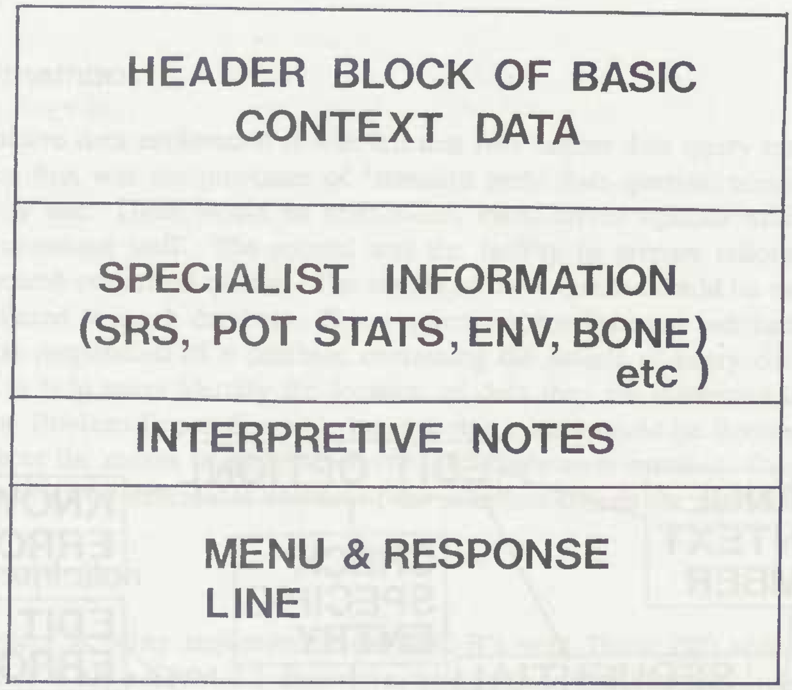

Fig. 28.3: Screen layout

of specialist data distributed through the IADB making notes as they go. A proposed screen layout for this task is illustrated in Fig 28.3. Below the header of common data a screen area is overlaid with whatever specialist data the archaeologists wish to consult. They may make notes in the third screen area and a menu-driven response line allows them to 'explore' the data with minimal keyboard skills.

28.7 Progress to March 1987

Currently, bulk and accessioned finds data can be appended and bulk-finds data can be edited. There is also a small expert system which allocates hard disc space to sites, plus the high-level menus and data structure creation routines. Over the next six months it is proposed to complete the accessioned finds edit system and produce the basic hard-coded finds query options.

References

CASTRO, L., J. HANSON, & T. RETTIG 1985. *Advanced programmers guide featuring dBase III and dBase II*, Ashton-Tate, Culver City.

CLARK, P. R. in prep. Recording complex stratigraphy: a flexible response. In preparation.

TODD, S., S. COLLEY, & N. CAMPLING 1986. '(Abstract) Relational databases and three dimensional graphics', *in* E. Webb, (ed.), *Computer Applications in Archaeology 1985*, Institute of Archaeology, London.