A NATIONAL ARCHAEOLOGICAL ARCHIVE - COMPUTER DATABASE APPLICATIONS

David M. Evans London Institute of Archaeology

INTRODUCTION

This paper concerns itself with making recommendations for the computer implementation of an archaeological information structure. Essentially, we hope to develop an integrated archive standard for any form of archaeologically related data, in the course of which we shall define what we feel that the range of our sources should be, and what are our aims for data quality and integrity.

The value of computerisation in archiving is, arguably, limited. With modern abilities of high speed, high memory, bulk on-line data storage and reasonable cost, we can perform weel in terms of large-scale information modelling and also in storing and searching data as required. Data security, expression of error and confidence and data validation may all be optimised. However, the major drawback to such a system lies in its 'fossiliation' of both the data and, more importantly, its structure. If rigorous limitations are placed on the types and relations between data (as we find in individual projects) then the scope for research and development may be stunted. We would argue that a logical 'Least Common Denominator' structure on computer would be best, which may suit any project definition, and the fields of data stored therein.

Many individual bodies have produced record structures based around archaeological sites, contexts, finds, samples or archives. These include, in England, Sites and Monuments records, the Royal Commission on Historical Monuments National Monuments Record (NMR), the Museums Documentation Association (MDA), the Inspectorate of Ancient Monuments (IAMHB, now HBMC) and the full range of other Heritage-related bodies. Up to now, only the MDA have produced a data standard which could be considered as generally applicable to all the above data (the Museums Documentation System or MDS). The NMR (England) is in the course of developing such a system specifically for archaeological and monuments data, in conjunction with the London Institute of Archaeology and SERC.

In the following we will consider the range of data structures and their software application (where possible) and make recommendations for data management in the future. Hardware considerations for these recommendations assume a 16-bit microprocessor is to be used, preferrably a 68000-based for improved efficiency. A more detailed hardware assessment will be carried out at a later stage of this project.

DATA STRUCTURES

In order to examine the data needs of any national database structure, we have broken down the record forms of several national and local bodies. These include those mentioned above, as well as the ex-Ordnance Survey (now RCHM) Archaeological Index and the English County Sites and Monuments records. We have isolated several of the major data-analysis problems that have been thrown up by these forms, as outlined below.

1. The Primary Key of each index is based on Topographic Locational data, with a secondary breakdown by period.

2. There is no consistent definition of a 'site' which would differentiate a structure or building from an isolated findspot, or an archaeological site form an occupation site, for example. We would propose that a 'site' be defined as a single structure. An archaeological site would combine several of these meaningful 'sites' into an 'Associative Site'. A farmyard or terrace are likewise examples of associative sites, although this time as structural complexes. The recording of such 'Associative' types would be the same as for a 'Structural Site', except that an 'Associative Site' owns 'Structural Sites', as 'Structural Sites' own 'Contexts'.

^S The definition of a 'site' as a structure allows us to use the Ordnance Survey Site number to differentiate easily between different buildings, parts of archaeological sites, and structural complexes in a logical and topographically based way.

3. There is a variable emphasis on the analysis of within-site components. That is, some records detail the finds, contexts etc. in a consistent way for each site, others refer to excavation or other varied archives for further research.

4. There is no approach to the digitisation of topographic boundaries. This results in secondary indexing to survey plans, location maps and aerial photographic plots and hence a loss in assessment of data error.

5. There is no consistent approach to the recording of position within site, in terms of 'From', 'To', 'Distance', Perhaps automated text analysis of this data is preferrable to a structured breakdown, in terms of the effort required. It must be remembered that the ex-Ordnance Survey Archaeological Index assumed the prescence of detailed site plans, as did the RCHM inventories. However, as these plans are not themselves annotated in a way which would divide a site into its parts, for individual recording, they do not supplant the positional detail which we would suggest.

6. Bibliographic referneces are limited to whole text. There is the potential for linking specific parts of a text to the structured fields of databases.

7. There is no useful method for logging administrative actions involving sites, parts of sites, people, meetings etc.. Topics such as Visits, Scheduling, Access, Data Accession are usually given the fields of 'Person', 'Date' and 'Text'. These comprise 'Action taken and to take', 'Contact' and reports of meetings etc.. An information modelling structure is needed to fully express the background, action and results of these procedures.

8. The current status of thesaurii or keywords is such that they recommend 'Preferred Words' and use loose definitions of terms. A more flexible, and yet more detailed, structure is required to model the meanings and relations of certain words WITHIN SPECIFIC CONTEXTS. Such an information structure is at present being developed by the NMR (England) for its archaeological thesaurus.

9. The results of Interpretative analysis of sites, eg. phasing, typology, value judgements, projections from sampling etc., are not made explicitly distinct from more securely controlled information, such as dates, locations, measurements, materials etc.. To this end certain fields must be linked to structures which outline their assignement process, or be given a reliability status. This is essential if we are to compare the work of, say, two people, or to compare two sites.

10. In general, the possible errors in data are not expressed in a readily understandable form. That is a '?' may mean that there is an uncertainty in one's own judgement, or, alternatively, there is an uncertainty in the reliability of the source of information.

An alternative approach to such data has been that of the MDA. The MDS has based its primary key of index upon any OBJECT. This has allowed not only archaeological materials (as above) to be described, but also archival materials, eg. photos, prints, textual media etc.. Each object is traced through its physical and administrative history, detailing its value, reason for accession and other information links for cross-referencing. To these ends it is very flexible, allowing a user-defineable level of data detail, in a structured way for any aspect of the data. It is designed for museum documentation of materials, not field description of localities and sites in their setting. Also, the archaeologist is little concerned with archival media, but with archival content, ie. the description of what is on a photo, and not the state of the photograph itself.

For archaeologists, the value of the MDS must lie primarily with the recording of finds and samples, as it leads to a detailed and consistent analysis, and with the public display of museum materials, using keys which have been developed for that purpose.

In our examination of the national records forms we have noted the different general pathways by which the data may be logically linked together and subsequently searched. Using these pathways any data structure, including MDS, may model archaeological information. These paths are as follows.

PATHWAY	DESCRIPTION
Topographical	Relative position in 3-D space
Interpretative	Theoretical or Abstract links between observations, eg. typologies, phases similarities, terminologies etc..
Temporal	Phased structure through time, Phases, periods and dates
Partiple	Heirarchical link between a site archive and the archives of its component parts
Descriptive	Assigns values to descriptive fields (equivalent to MDS STATEMENT structure) eg Responsible, Executor, Status, Date, Function or Reason, Environment etc..
Non-Archaeological	eg. Administrative, Biographical, Natural Sciences, Legal, Access etc..

It is through these categories that we may model the information from any archaeological source, by linking together structured data.

SYNTHESISED DATA STRUCTURE

Of the various computerised implementations of the above types of data structure, the heirarchical (eg Gos) and network approaches are the most difficult to visualise in a consistent way. The former only allows cross-linking between heirarchical branches as an add-on and has an unpredictable structure, whilst the latter comprises totally of cross-linking, containing a confused and variable structure for each record. The Relational Model, although just another representation of the same data structure, comprises a series of logically related tables, each having a consistent format (with correct design).

Therefore, we have selected the relational model to represent our information. However, alternative views of the data may be had by the use of other models. We have allowed for these by considering a broad range of data-linking pathways (above), which may hence allow for data reconfiguration on another software structure.

Our primary recommended structures, which describe purely ARCHAEOLOGICAL MATERIAL, are outlined below.

1. PARTIPLE Firstly, we wish to be able to define a site as a unity, and then to subsequently break it into its component parts for more detailed description, depending on the detail and interests of the particular recording project.

We have already discussed our ideal 'site' definition, above, and how the Ordnance Survey Index number may be used to assign a Primary Record number to each site (being structures or negative structures of any type). Each site may be broken into contexts (elements) which may further be divided into finds (materials). From these we may take samples, and samples of samples. At each level the entity being described, being site,context,find or sample, also contains its archive. This allows us to integrate archives into the material breakdown of site information.

2. TOPOGRAPHICAL Secondly, the physical position of any sites or parts of sites must be defined by relation to their National Grid Reference and Ordnance Datum, or to each other. This will, of course, vary with time for any particular objects, for structures upon their destruction etc..

3. TEMPORAL Thirdly, the 'history' of materials and their investigation may be described by a 'Phased' analysis of a site or site-part. That is, we may note how its characteristics change through time, based upon stages such as construction, occupation, post-depositional disturbance and 'investigation' as examples.

4. DESCRIPTIVE Lastly, for our primary structures, the site or site-part may be described as it existed in time and space. The same descriptive structure may be applied to any object, action or concept, being

Responsible — Authority, Auspices, Archive Holder

Executive — Principal Actor(s)

Status & Value — Pure Description, eg Condition, form, dimensions, colour, legal status, Educational/Environmental or Aesthetic value.

Environment — Description of data in milieu of site, eg. land-use, topography, environment of decisions etc..

Function / Reason

Date — Absolute (more precise than 'phase')

The above structure may be represented as a series of relational table formats.

Beyond this primary data structure lie other data links and types. The first of these is the INTERPRETATIVE, or Theoretical, linkage between information. This is made when data is heavily subjective, eg. in the cases of value judgements and cultural affinity. This link structure would form a separate relational table, linking site-based elements.

The second broad type of linking pathway is NON-ARCHAEOLOGICAL, in that it does not itself deal with the contemporary processes effecting the materials of the past. Instead, it describes such links as ADMINISTRATIVE (eg. in site visits, archive maintainance, meetings etc.), BIOGRAPHICAL (as addresses, status etc.) and those areas concerned with the NATURAL SCIENCES (ie. geology, pedology, flora, fauna and climate). It is an administrative link which joins Bibliographic references to their source.

Within this theoretical data structure, we require assessments of the quality and integrity of the stored data. A controlled terminology leads towards this. By the use of a thesaurus, giving a term, its related terms, a preferred term (where required), a definition and the contexts in which the term should be used, we can usefully interrogate a structured database or keywords list. Table look-up on input for validation of terms and prompts for correct usage are envisaged.

To qualify the assignment of these terms, we may indicate

1. '?' Possible error in our assignment

2. '£' Possible error in source

3. '-' Not Applicable

4. 'NA' Not Available

Security control of restricted information may be built into the operating system, but record and field level locking should be built into DBMS software.

ARCHIVE MANAGEMENT

Having outlined a consistent data structure and the control of its content, we must consider the accession of data and its retrieval and indexing. The depth of detail that it will prove possible to store on-line (eg on 40-400 Mb Winchester disc, or video disc) will depend upon the implemented virtual memory size, and remains to be assessed by this project.

The potential sources of the data lie with the current range of archaeological archive bodies, including the NMR, SMR's, Museums and excavation bodies. There is also a role for data from local and national societies' projects and especially for the fruits of research. However, there is a need for informed monitoring of data quality.

Providing that the means and quality of data are controlled, any structured archaeohistorical data may be fully accessioned into the ideal database. This assumes NO CONTROL by the archive body over the conditions and fields of data aquisition, leaving the totla design of a project and its recording to the responsable body. We have outlined above a structure by which any such data may be placed in A consistent national structure. This does, however, require more detialed documentation of the meanings of terms and the definitions of fields used in the project, in order that their logical position in the overall database may be made more clear.

The archival duties described coincide with those of the National Monuments Records, and here, therefore, lies the most suitable repository for such a computerised archive.

Once a corpus of data has been produced, it will prove necessary to select fields for B-tree indexing. The essential fields for this are the Primary Keys of each table (in the Relational Model) and cross-references for joins on these keys (where found). Beyond these keys, essential to operation of the structure, all fields are eligible for indexing in priority order of applications interest, eg topographic, legal, administrative etc..

In the coming years future developments of such a system may include adavanced graphics, Image processing, telecommunications networks and links and optimal memory management. However, if a degree of national and international hardware and software compatability and standards could be reached, this would be the most encouraging development of all.