RELATIONAL PROBLEMS IN ARCHAEOLOGICAL DATA HANDLING.

Brendan J. Grimley

School of Archaeological Sciences, University of Bradford, Bradford, West Yorkshire BD7 1DP

Introduction

The paper gives a brief introduction to the history of Relational Data Management Systems, and illustrates a problem of implementing such a system on a microcomputer. A data storage and retrieval system developed at Bradford is then summarized, offering one solution to the problem.

Relational Systems

Hierarchic and network database systems have received extensive use since the early 1960's, though the hardware dependence inherent in these systems had caused some concern.

By the mid-1960's, it was felt thet a better method of representing conceptual information could be designed, i.e. an approach to so-called data independence, where the details of any individual data set are divorced from the details of a specific hardware installation. A major advance at this time was the use of Entity Set structuring methods, using tables to represent data, and, importantly, using entity identifiers to represent associations, rather than physical pointers.

With a foundation of theory from relational mathematics, E.F. Codd introduced the Relational Model in the late 1960's (Codd 1970). He "noted that an entity set could be viewed as a mathematical relation on a set of domains D1, D2,.....,Dn, where each domain corresponds to a different property of the entity set." (McGee 1981). Further, Codd defined a relation as "a time-varying subset of the Cartesian product D1xD2x....xDn" (McGee 1981,508), i.e. a set of tuples comprising single elements from each domain.

From this definition it is evident that the domain on which a relation can be constructed can be elements of any type,- even other relations. Significantly, Codd pointed out that such complexity was of no advantage, and proposed thst relations be constructed from domains of elementary values. This developed into normalized relations.

Emphasis was on the conceptual form of the recorded

information, (the users view), not on the hardware implemented structure, (the programmers view). Each record can be thought of as comprising a collection of fields and associated entities. Fundamentally, from Codd's work, each record has a fixed number of distinct fields (atomic fields), i.e. no repeating groups are allowed. Using tables as the basis of record storage, each row comprises the individual records, and each column a field (attribute). No two rows in the table are identical.

Problem of Normalization

The analogy with standard pre-printed sheets commonly used in archaeological recording is quite clear; data entries appear in the appropriate fields, and each completed form "...is characterized by conforming to the pattern and provisions of the appropriate blank form." (Cohen and Nagel 1934).

Codd introduced normalization of the data, the so-called normal forms, intended to avoid problems of insertion or deletion of items in a relation. In effect, the normal forms govern the amount of data redundancy and duplication. Codd recommended that all information should be stored in third normal form (3NF), with the values from non-key domains dependent only on the key; (where attribute A is functionally dependent on attribute B if the value of B determines the value of A. In practice, this is a one:one relationship).

This is where the problems arise in archaeological recording. Figure 1 is an example of a card in use by the Northampton Development Corporation for recording lithic finds information. It displays a common format of fields and associated entries, and it would appear reasonably easy to construct a normalized relation for this data.

The major problem is with the repeating groups (annotated fig. 1) which can logically occur where one record - in this case one find - can contain several observations for any one section. In order to maintain a normalized form, and certainly for 3NF, it is neccesary to create new relations separating the sections, and thereby increasing the conceptual complexity. Large mainframe computers running commercially available software usually have sufficient memory space to maintain several relations, though often programmers are required to create the relations using information provided by the archaeologist.

However, with microcomputer-based implementations, the limited internal space is soon consumed, and processing

NDC ARCHAEOLOGY UNIT — LITHIC FINDS

1	SF No. 4855				
2	Area: D6	Layer: 16	Feature: 165D	Phase (infill): 5	Site Phase:
3	Co-ordinates: 252·07/607·93		Level: 78·28		
4	Material: FLINT - GRANULAR INCLUSIONS				
5	Condition				
6	Cortex				
7	Primary Analysis: FLAKE.				
8					
9					
10	Additional Comment				
11	Classification: 1. UTILISED. (RIGHT SIDE)		2 PROBABLY UTILISED (LEFT SID		
12	LONGITUDINAL ACTION.		? TRANSVERSE ACTION		
13	SOFT MATERIAL		MEDIUM MATERIAL		
14					
15	Additional Comment: WEAR ON LEFT SIDE MAY BE THE RESULT OF HAFTING				
16	PRESSURE.				
17	Use wear 1 MICROFLAKING	GLOSS	2. MICROFLAKING		
18	FLAT	DIFFUSE	FLAT / (RIDGE)		
19	VERY SMALL SCARS		LARGER SCARS		
20	MODERATELY HEAVY		MODERATELY HEAVY		
21	LIGHT	TRACE	MODERATE		
22	RIGHT SIDE	RIGHT SIDE	LEFT SIDE		
23	BOTH FACES.	BULBAR FACE.	DORSAL FACE.		
24	Breadth: 23·5 MM		Length: 42·5 MM	Thickness: 7MM	
25	Breadth: Length ratio: 2·8 : 5				
26	Angle of edge				
27					
28	Weight				

Repeating Groups

Figure 1. A pre-printed record sheet.

times become considerably increased, even for a small database of a few thousand records.

Bradford System

In Bradford, a storage/retrieval system has been designed to resemble more closely the information structures in archaeological use, and has been implemented on a microcomputer.

Importantly, "the relational model is a framework or philosophy for finding compatible solutions..." to the problem of data independence (Astrahan et. al. 1975:139). The advantage of entity set structuring methods (of which the Relational Model is one), is in the conceptual simplicity. Codd's important contribution has been in demonstrating the advantages of simple record forms, i.e. single entries from specified domains.

In consequence, the Bradford system has adopted the single domain/field structure in order to compile the information, but significantly, makes use of the concatenation of logically related fields into groups, termed subschemas in this application, (e.g. in figure 1, lines 7-10 comprise one such group, lines 11-16 another). Each subschema comprises individual fields, but can, if necessary, be repeated any number of times, while still being treated as individual units.(E.g. in figure 1, lines 11-16 occur twice, lines 17-23 three times).

This enhances the processing capacity of the computer, and the complexity of record form available to the user, while not detracting from any of the information the user wishes to record. Conceptually, the machine is operating on data that exists in rectangular form, while physically it is clear that the data does not represent the desired rectangle (a necessary part of systems such as Rapport).

The applications programmes comprising the storage/retrieval suite have been collected together and operate interactively on an overlay basis, controlled by a driving menu which provides the user interface. The same suite can be used on several different recording requirements simply by establishing a data dictionary, which is essentially an empty record card for each application, created via a set of interactive routines (Grimley and Haigh 1982).

The retrieval operations allow up to eight levels of request parameter (e.g. select all cores used as scrapers greater than 20 mm. in length...), linked by the AND, or

NOT operator, thus providing the facility of creating exclusion sets (e.g. select scrapers that are not broken).

Selected information can be displayed at the computer screen or at a printer, or a smaller subsidiary data-base can be established on disc and used as a work-file, improving access time to that data-set. Output can consist of the primary identifier (e.g. the find number of figure 1), individual subschemas, or the complete record.

Data selected can then be made available to statistical or plotting routines (e.g. figure 2, a distribution map of a specific find category produced on a plotter connected to the microcomputer, and which can be drawn at any desired scale).

Conclusion

Codd's work developed from a working environment concentrating on the information storage, and not on specific hardware considerations. Significantly he demonstrated the advantages of using simple, single attribute formats, and introduced normalization in order to control data redundancy.

Commercial Relational packages implemented at large computer installations can handle archaeological information efficiently, but often require an applications programmer to establish and interrogate the data-base. Maintaining the necessary normalization via a microcomputer however is time, and space, consuming; and considerably inefficient.

The Bradford storage/retrieval system arose to cater for the need to utilise microcomputers, and has been designed adopting the simplicity of the entity set models, while more closely emulating the structure of archaeological record cards. Processing and control routines can be invoked and operated effectively by the archaeologist to create and analyse information in a data-base.

References

Astrahan, M.M., Chamberlin, D.D., King, W.F. and Traiger, I.L. 1975 — System R -A Relational Data Base Management System. DATA BASE SYSTEMS 1975.eds Hasselmeier, H. and Spruth, W.G. 139-148. Springer Verlag.

Codd, E.F. 1970 — A Relational Model for Large Shared Data Banks. COMM. OF ACM. 13, 1970. 377-387.

Cohen, M.R. and Nagel, E. 1934 — An Introduction to Logic and Scientific Method. RKP.

Grimley, B.J. and Haigh, J.G.B. 1982 — A General Purpose Data Management System for Archaeologists. COMPUTER APPLICATIONS IN ARCHAEOLOGY 1982, 63-68. Univ. of Birmingham.

McGee, W.C. 1981 — Data Base Technology. IBM J. RES. DEVELOP. vol 25, no.5. Sept 198 505-519.

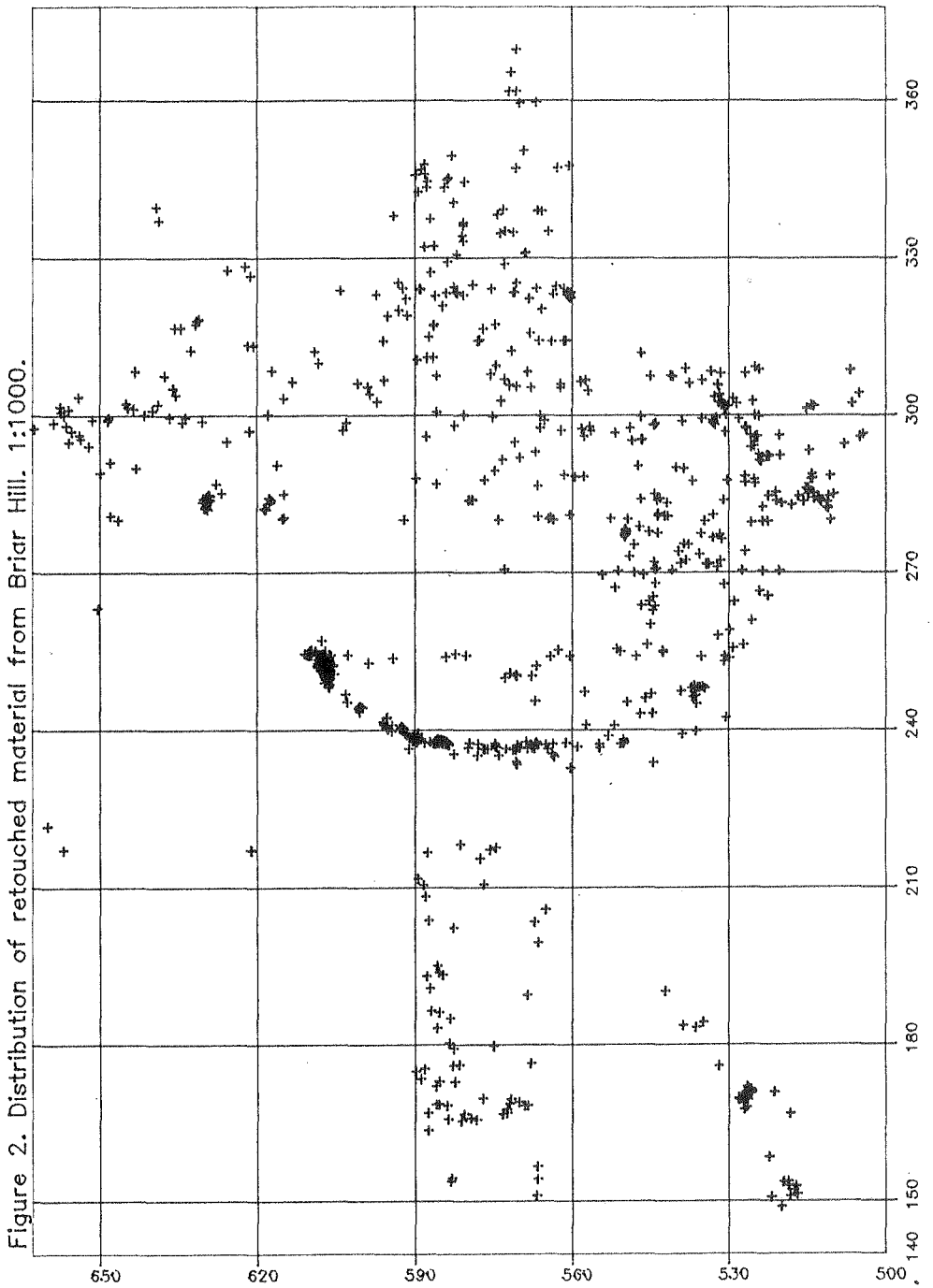

Figure 2. Distribution of retouched material from Briar Hill. 1:1000.