Dingwall, L., S. Exon, V. Gaffney, S. Laflin and M. van Leusen (eds.) 1999. Archaeology in the Age of the Internet. CAA97. Computer Applications and Quantitative Methods in Archaeology. Proceedings of the 25th Anniversary Conference, University of Birmingham, April 1997 (BAR International Series 750, CD-ROM). Archaeopress, Oxford.

Databases - Research Tools and Communication Aids

Hans-Dieter Bader

Abstract

Any research model which claims to be testable by fellow researchers depends on explicit categories, explicit methods and interpretations based on these categories and methods. Some archaeological research still fails in this respect resulting in unproductive controversial interpretations. Database management systems usually force the researcher to apply systematic categories and criteria in a consistent way to the research subject. By expressing these categories in an explicit way, they are open to communication and discussion. The transformation of the criteria in each category in the progression from data capture to data analysis shows clearly how research methods are applied to the data. As a result of this process the research method is made explicit and open for discussion. Archaeological research looks at material objects in three dimensional space. The material objects are experienced and communicated via categories fitting the chosen research method. Any database application working with a two or three dimensional visualisation tool like a GIS should enable any archaeological research to be transparent and therefore open for communication. Database management systems can be shown to play a vital role in a systematic research approach and the following discussion of the research.

1 Introduction

The following remarks on relational databases are aimed at the day to day use of databases to enhance systematic research approaches and to ease the difficulties of comparing research results based on similar archaeological find categories. This paper intends to explore some basic handling techniques of databases rather than illuminate the complex theory behind relational databases [Note 1] . Regional or national data collections aim for archaeological information preservation and archaeological site management and are usually supported by a Management Information System team. But most archaeological data gathering is still undertaking by individual researchers without access to the support of MIS programmers. It is in this arena where databases can enhance significantly the systematic approach to research questions and exchange of results [Note 2] . Most modern relational database management systems are sophisticated and easy to use. Programs like DBase for Windows, Paradox, Access, Superbase and similar products can be used in a very similar way and data tables can be moved from one system to another either via DBase files or ASCII delimited files (with few restrictions).

This article is based on experience gained during a Ph.D. project (Bader 1993) and by assisting students at the University Marburg to build data tables for a variety of projects.

2 Categories and criteria

Categories hold the data which is used to resolve a given research question. They can describe one or several artefact types or archaeological feature types. Defining research categories means defining the research methods and limiting the range of possible research questions. Thus this step in the research design channels and limits all following analysis. It may seem to be tedious to mention this basic fact but it is surprising how often researchers neglect to define their categories accurately, before they start gathering data. A database design forces the researcher to define all fields (-> categories) she will be using for data gathering. Either a field exists to put data into or it does not. It is not possible to invent new fields or drop old ones during data acquisition, as the database information would be inconsistent at the end of the data gathering. As it is not always easy to decide on the best solution to the structure of categories before the data gathering, a pre-run of data gathering to test the data structure is an important step before data acquisition. Usually only a small number of records is enough to indicate problems in the data model [Note 3] . The pre-run is also an indispensable step before data acquisition to build up the back bone of any reference collection (see below).

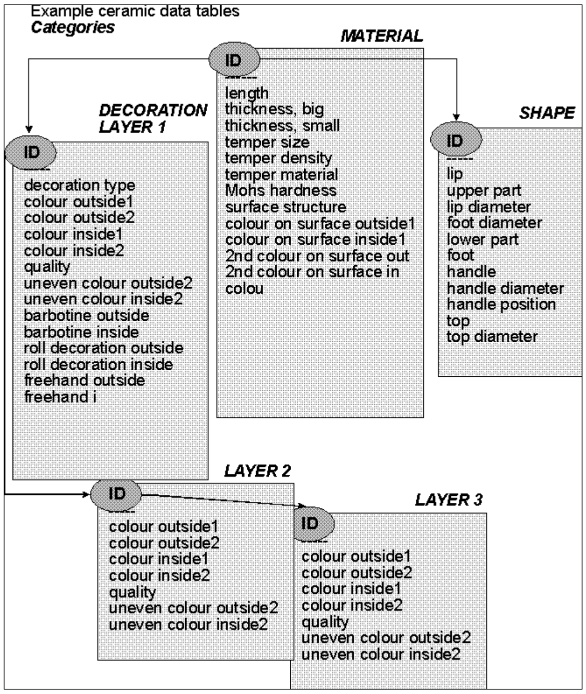


Figure 1. Categories (example ceramic data tables) .

The researcher who has to define field by field on the computer is forced to develop a comprehensive but lean data model which ensures that all envisaged analysis methods are covered by the data gathering. Relational database concepts are essential to develop a lean data model to speed up data gathering. If we take the example of ceramic sherds out of a settlement context; there are mainly undecorated sherds, some decorated sherds and many of them (decorated and undecorated) show part of the rim or the base. There will be categories which will apply to all sherds, independent of being decorated or not, showing shape features or not. Some categories will only apply to decorated sherds, others only to shape features. In a relational database it is easy to separate these categories into separate data tables and use them only when needed. All separate data tables are linked with one common field, usually the find number/identifier (see Fig. 1). Separate data tables not only speed up data gathering, they define the various subsets of data which can be used for various analytical steps. In the example of the sherds, all records can be used to look at temper, clay, clay colour, etc. to answer questions about raw material sources etc. The subset of sherds with decoration can be used to interpret stylistic variations and the subset of rim and base sherds can be used to look at the various functions of the pots (see Fig. 2). Thus it is obvious that a relational database concept encourages a more systematic analytical model.

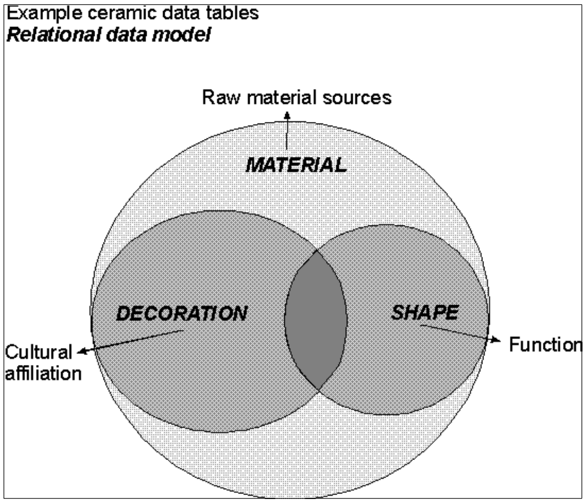


Figure 2. Relational data model (example ceramic data tables).

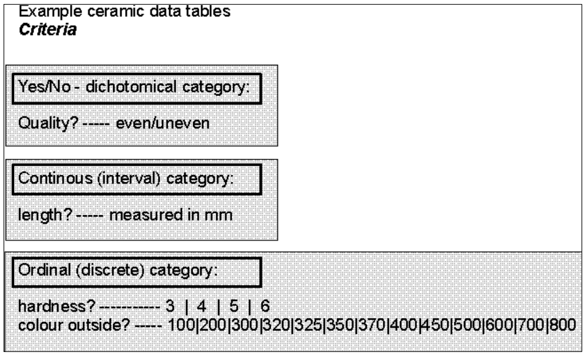

Example ceramic data tables
Criteria
Yes/No - dichotomical category:
Quality? ----- even/uneven
Continous (interval) category:
length? ----- measured in mm
Ordinal (discrete) category:
hardness? ---------- 3 | 4 | 5 | 6
colour outside? ----- 100|200|300|320|325|350|370|400|450|500|600|700|800

Figure 3. Criteria (example ceramic data tables).

Setting up the criteria for each category is the next step (see Fig. 3). Continuos (interval) categories reflecting the taxonomy (like the height, weight, thickness, etc.) are straightforward to set up. Dichotomical categories (present/not present) are even easier to set up. Developing criteria for ordinal (discrete) categories can be more complicated and ordinal categories are often the most important ones for archaeological questions. Some ordinal categories have a complete and comprehensive set of criteria and are as easy to set up as continuos interval categories. Returning to the example of ceramic analysis, Mohs' hardness is an ordinal category with a comprehensive set of criteria. There exist only 9 levels of hardness and in ceramic analysis 5 are the most to be used, whereby most ceramic assemblages can do with only 3 or 4 different levels of hardness (see Fig. 4).

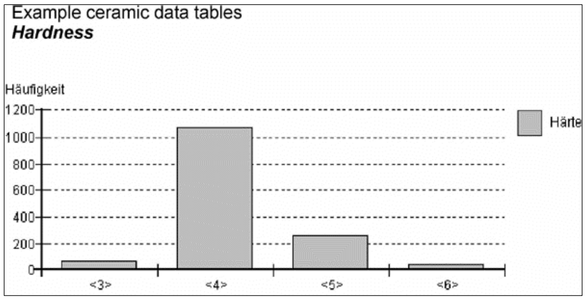


Figure 4. Hardness (example ceramic data table) .

For other categories there exist comparative tables which are neither complete nor comprehensive. Colour is one of these categories. A number of different tables exist to code colour and in our example of ceramic analysis the 'Munsell Soil Colour Chart' is the coding system most widely used. There are several disadvantages using a chart like Munsell. It is very expansive and very cumbersome to use in the field. Another more important problem is comparing the colour of a sherd to the colour of a print. This involves a certain degree of guesswork. A much easier and more accurate approach is to build up a reference collection (see Fig. 5 for a part of a reference collection) of small pieces of ceramic showing all colours existing in a specific ceramic collection. As the reference collection can be only considered complete after the end of the whole data gathering, the coding system needs to be an open system. In an open system it is possible to add new codes at any stage in the system. From a practical viewpoint the pre-run of the database system (see above) is used to gather most reference pieces and develop a coding system as the back bone for the open system [Note 4] . During the real data gathering a new colour requiring a new code and reference piece will only pop up occasionally.

There is another theoretical advantage of an open system developed on the basis of the real material compared to a closed system coming from the outside. It reflects the composition of the real material more accurately. E.g. the Munsell Soil Colour Chart may contain much more brown colours than there are clay colours of a specific assemblage but the yellowish clay colours could be more finely separated than on the Colour Chart [Note 5] .

The four steps of setting up a database management system - defining categories, separating tables, defining criteria for each category and pre-run, involve and ensure systematic thinking about the aim of the research, the methods to be used and the approach to data gathering chosen for the specific research aim. Databases do not prevent sloppy work but they just make it so much harder to dive into a research question without systematic thinking before the data acquisition step.

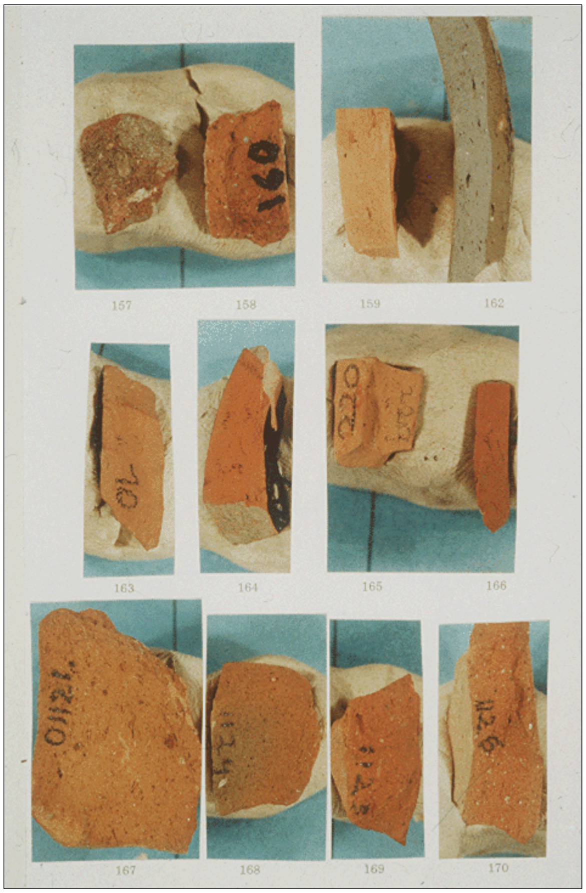
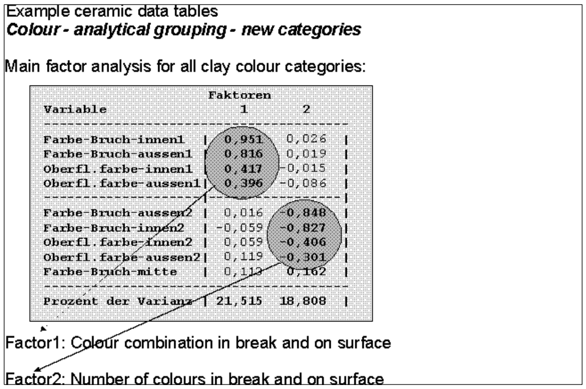


Figure 5. Reference collection, temper type (example ceramic data tables), open system number still visible on some pieces, analytical code number shown.

3 Data analysis

The first step in data analysis is checking for data errors. Descriptive statistics and visualisation such as bar charts are some of the means to check for errors. The same methods can then be used to transform the raw data into analytical data. There exists a hierarchy of the various data types (Fletcher and Lock 1991, 2-7). Ordinal categories (like the colour codes) can be changed to interval categories by various methods such as normalisation, or to simple attributes (red? yes/no). Interval data can be reduced to fewer intervals or to dichotomical categories. It is not possible to change categories to reverse their hierarchy. Some categories can be summarised and accumulated and then presented in a new 'dummy' category. In the ceramic analysis example the colours on inside and outside and the colours in the break can be channelled into one or two categories describing colour classes rather than real colours (see Fig. 6). Main factor analysis and cluster analysis can assist in this mainly subjective, interpretative step. One has to be aware though, that the use of these multivariate statistics implies an active, archaeological interpretation which determines the possible outcome of any subsequent analytical step.



Example ceramic data tables

Colour - analytical grouping - new categories

Main factor analysis for all clay colour categories:

Variable	Faktoren 1	Faktoren 2
Farbe-Bruch-innen1	0,951	0,026
Farbe-Bruch-aussen1	0,816	0,019
Oberfl.farbe-innen1	0,417	-0,015
Oberfl.farbe-aussen1	0,396	-0,086
Farbe-Bruch-aussen2	0,016	-0,848
Farbe-Bruch-innen2	-0,059	-0,827
Oberfl.farbe-innen2	0,059	-0,406
Oberfl.farbe-aussen2	0,119	-0,301
Farbe-Bruch-mitte	0,113	0,162
Prozent der Varianz	21,515	18,808

Factor1: Colour combination in break and on surface

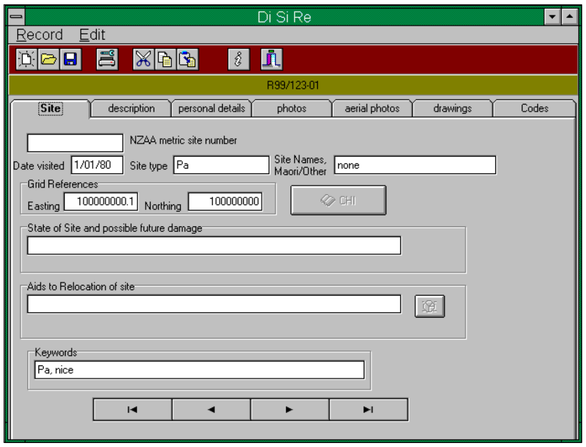
Factor2: Number of colours in break and on surface

Figure 6. Colour, analytical grouping, new categories (example ceramic data tables) .

Different statistical methods need different data characteristics. The transformation of raw data to analytical data has to accommodate this. E.g. cluster analysis requires normalised numbers for all criteria to ensure a sensible calculation of the Euclidean distances, and neural networks need a specific data preparation to weigh the data correctly.

At this step of the data analysis the question of 'what sort? and where?' should become transparent. This question is still at the centre of most archaeological interpretation. So database management systems are obviously not only a tool in the hands of archaeologists aiming for multivariate statistics but also provide a powerful tool for the 'old fashioned' archaeologist. The interpretative step of transforming the data can be combined with sorting the data (by site, layer, etc.) and can be an end in itself.

Nonetheless the data has been prepared for further analysis and use of multivariate statistics, such as main factor, cluster, neural networks or Bayesian classification.



R99/123-01

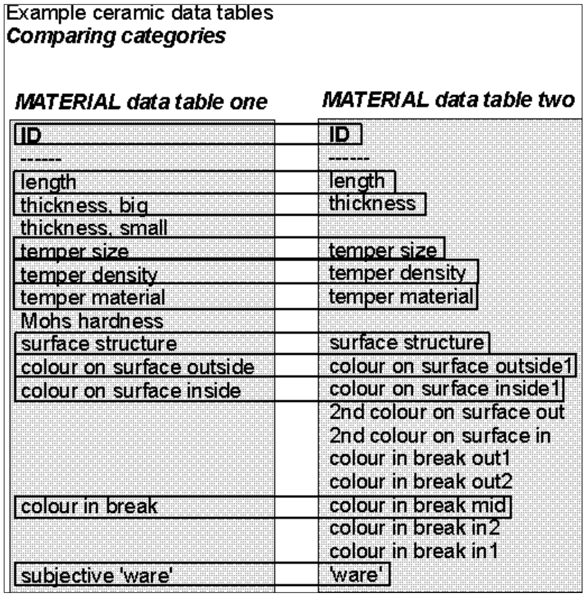
NZAA metric site number	
Date visited	1/01/80
Site type	Pa
Site Names, Maori/Other	none
Grid References Easting	100000000.1
Northing	100000000
State of Site and possible future damage	
Aids to Relocation of site	
Keywords	Pa, nice

Figure 7. Delphi front end (example site record data tables).

The data transformation can be done in any of the commercial database packages as well as in some of the large statistical packages. The choice of any of these packages depends rather on availability than on specific features they offer. I grew up with DBase and thus still use various versions of it to transform data. I prefer WinStat to other statistical packages because of its ease of use. But any other combination of software packages generally provides the same results. For the future there seems to be an opportunity to work very differently with software. A software package like Delphi opens professional database programming to any reasonably computer-literate archaeologist (see Fig. 7) for an example of a Delphi front end). Design and use of any data table compatible with Delphis data base engine (ODBC) makes the knowledge of a variety of database packages unnecessary. Any analytical tool developed as a module (.DLL), like the neural network software developed at Otago University at present, can be incorporated under the common platform of Delpi. This would mean that every researcher could start with a data model similar to the one she envisage, change it till it meets her needs and add any analytical tool necessary for the data analysis thus building up a personalised computerised research tool. At the Centre for Archaeological Research at the Department for Anthropology in Auckland, NZ three different Delphi applications are being developed at present, two of them are individualised research tools and one is a common site record database.

4 Data comparison

Researchers comparing their results may have to compare their grouping of the archaeological material. So the main question is: 'is entity A the same or different to entity B?'. Entity A is a group of finds, features, etc. defined in the work of one researcher and entity B a group of similar finds, features, etc. defined in the work of another researcher. Very often this crucial question is impossible to decide. The result is a battle of opinions and interpretations without consideration of the different data models or different use of analytical tools. Databases open the possibility of comparing research not at the end of the circle - the interpretation, but at the start - the data model and the raw data. Comparing two different relational database management systems starts with comparing their field structure and table structure, thus comparing the categories which are used. By transferring all similar categories into a new database model, we start really comparing apples with apples instead apples with peaches (see Fig. 8). The next step compares the criteria lists for each category, transforming them into a common criteria list with the least possible loss of raw data (see Fig. 9). Differences in data acquisition between researchers become apparent and thereby opened for discussion. Another step forward towards data standards is made by talking about the way that finds, features, etc. are described.


Example ceramic data tables
Comparing categories

MATERIAL data table one	*MATERIAL data table two*
ID	ID
------	------
length	length
thickness, big	thickness
thickness, small	
temper size	temper size
temper density	temper density
temper material	temper material
Mohs hardness	
surface structure	surface structure
colour on surface outside	colour on surface outside1
colour on surface inside	colour on surface inside1
	2nd colour on surface out
	2nd colour on surface in
	colour in break out1
	colour in break out2
colour in break	colour in break mid
	colour in break in2
	colour in break in1
subjective 'ware'	'ware'

Figure 8. Comparing categories (example ceramic data tables) .

The common data tables can then be transformed into analytical data using the solutions used by both researchers. This may result in two different analytical data sets. Once again, doing so opens the way towards a discussion of filtering and transforming data. The last step involves using the analytical methods being used by both researchers resulting, in a worst case scenario, in four different results (see Fig. 10). Even if it is impossible to synthesise the different results, both researchers will have learnt a great deal more about their own data and the data of the other researcher and the comparison of their results stands on a much broader basis than before.

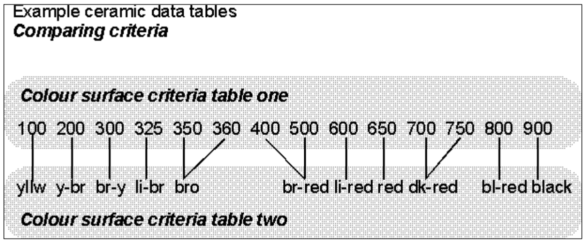

Example ceramic data tables
Comparing criteria
Colour surface criteria table one
100 200 300 325 350 360 400 500 600 650 700 750 800 900
yllw y-br br-y li-br bro br-red li-red red dk-red bl-red black
Colour surface criteria table two

Figure 9. Comparing criteria (example ceramic data tables) .

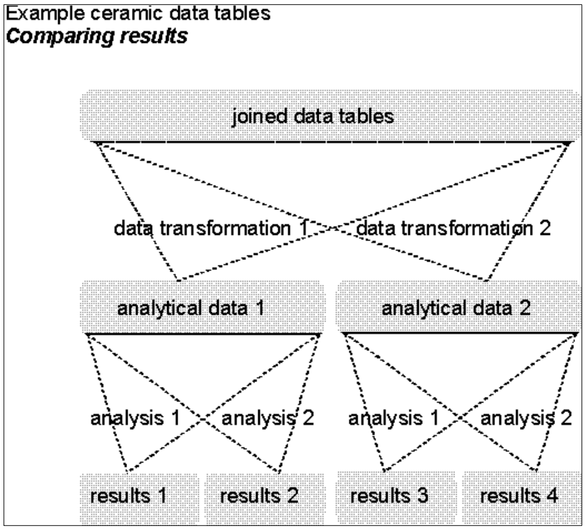


Figure 10. Comparing results (example ceramic data tables) .

In these days of competing database and software packages, the suggested approach of comparing data tables seems quite ambitious. But visual programming offers a way (see above) to agree on a common platform which could be used by a large part of the archaeological community. Integration of existing or future tools, specially developed for archaeology (like the Bonn Seriation Package), under this common platform, would make the suggested comparison of raw data and analytical ways a matter of days or even some hours. My suggestion is to form an international workgroup to develop such a common platform and to advertise the development of specialised archaeological tools under this common platform.

The combination of database tools with images, GIS tools and 3D digitisers widens the application of databases. Areas like art history in classical archaeology are still based on textual description of artefacts. In cases like the research of Roman portraits a small number of researchers have developed a specific textual code which is difficult to learn and to interpret. The personal investigation of the original material is often necessary, but expensive because of travel and time consuming. Databases with set categories added to 3D visualisations and linked via GIS programmes will not always replace this personal investigation, but they will open these research areas for a wider archaeological community, which enhances the interpretation of these artefacts [Note 6]. Again a common visual programming platform with specialised archaeological modules would make this high performance computing accessible to a large part of the archaeological community.

5 Conclusion

Database management systems force the archaeologist to start a research question with a defined set of categories. Relational data tables ensure that the data model is adequate for different situations during the data acquisition. Various methods to define criteria for each category lead to consistent data acquisition procedures. All this helps to develop a systematic approach towards any research question. Data description and transformation lay bare the basic distribution of the research objects. They are an essential step towards multivariate statistics which have the potential to visualise the finer aspects of the data structure/grouping.

Comparing two or more data sets of similar archaeological objects instead of comparing end results opens the discussion about categories, criteria, data acquisition procedures and analytical tools. This process allows a systematic exchange of ideas concerning the whole process of an archaeological research project rather than the exchange of opinions towards the interpretative results.

Modern visual programming tools allow the development of a common platform to run data tables and analytical tool and may prove essential to compare different research projects rapidly.

Acknowledgements

I would like to thank all members of the excavation team in Kaunos from 1986 - 1990 for their fantastic support in dealing with some truckloads full of sherds.

Notes

1 The body of mathematical theory behind the use of relational databases (e.g. Yang 1986) is of no concern in this context.

2 In this context the article of Jan Rulf describing the pre-processing of archaeological data (Rulf 1993) is of importance and should be read in conjunction with this paper.

3 complex find categories like fibulae or complete Greek vases may need only 10 records as a pre-run, simpler categories like sherds would need around 100 records. Complex finds display a high variation in only a few pieces as simple find categories need more pieces to show the same high variation. It is important to try to incorporate simple examples and very complicated examples of each find category to ensure that the data model is comprehensive.

4 During the pre-run there may be collected 20 or 30 reference sherds. They can be numbered 100, 200, 300, ..., 3000. Between two reference sherds another 99 reference sherds could be allocated - more than enough for an archaeological reference collection.

5 This is what exactly happened during a real ceramic analysis (Bader 1993, p.72). An additional advantage of a reference collection is the speed it can be used compared to a Colour Chart. Small pieces of ceramic in a tool box (for screws and small bits) held into place with bluetack and the code number beneath them can be scanned very quickly to find the code for any other sherd. Swapping pages on the Colour Chart back and forth to find the right code takes much longer.

6 Some research in Tibetan Thangka painting show the possibilities of imaging tools in art history (Makkuni 1992). 3D digitisers and imaging tools added to some research into Classical Attic grave stelai (Dallas 1992) seems to be the next logical step in this research.

Bibliography

Bader, H D, 1993 Mengenanalyse *der hellenistischen Keramik der sog. Tempelterrasse in Kaunos, SW-Tuerkei*, Tectum Marburg

Dallas, C, 1992 Syntax and semantics of figurative art: a formal approach, in Reilly, P and Rahtz, S (eds.) *Archaeology and the Information Age*, London, 230-275

Fletcher, M and Lock, G R, 1991 *Digging Numbers*. Oxford University Committee for Archaeology Monograph, 33

Makkuni, R, 1992 The electronic capture and dissemination of the cultural practice of Tibetan Thangka painting, in Reilly, P and Rahtz, S (eds.) *Archaeology and the Information Age*, London, 323-350

Rulf, J, 1993 Pre-processing of archaeological data, in Andresen, J. and. Madsen, T and Scollar I (eds.) *Computing the Past CAA92*, Aarhus, 329-332

Yang, C, 1986 *Relational Databases*, New Jersey

Contact details

Hans-Dieter Bader
Centre for Archaeological Research
University of Auckland
Private Bag 92019
Auckland, NEW ZEALAND
e-mail: hdbader@clear.net.nz