# 26

# A review of approaches to controlling archaeological vocabulary for data retrieval

Amanda Chadburn*

## 26.1 Introduction

This paper is *not* concerned with a specific computer application in archaeology, and therefore might at first appear to be out of place in a book devoted specifically to quantitative methods and computer applications in archaeology. It will concentrate instead on an important area, which is sometimes neglected due to lack of interest, but which the more sophisticated development of computerisation within the discipline is forcing into prominence. This area is the classification (used in the widest sense) of archaeological data and the vocabulary used for this purpose. Such vocabulary is required for retrieval purposes in textually-based archaeological records.

As I hope to show, if the oft-quoted 'Garbage-in, Garbage-out' is not to apply to our results, it is as essential that we think about the *conceptual* problems of organising, structuring and naming our data, as that we get the *technical* aspects of a computer application right. Indeed, without solutions to the conceptual problems, there is little point in making progress on the technical side.

It is true to say that there is a current trend in archaeology towards standardisation and precision, partly as a result of the theoretical rigour advocated by the 'New Archaeology' of the early seventies. For example, recording procedures on-site have been tightened up and standardised enormously over the last 20 years, as have many aspects of post-excavation work. Yet there are still many instances of 'sloppy thinking' regarding the organisation, input and retrieval of archaeological data, which could be improved to the benefit of all. To quote Lavell (Lavell 1985): 'Professionalism in archaeology is hardly in doubt these days; what I am asking for is professionalism in information retrieval, if we are to realise the immense potential of our data'.

There are obviously many areas within archaeology where vocabulary control will be of benefit, both on-site and off-site. This paper concentrates on site-type vocabulary, which is used in a variety of records including county-based sites and monuments records (SMRs) and national databases such as the National Archaeological Record of the Royal Commission, and the record of Scheduled Ancient Monuments of English Heritage.

* English Heritage
Fortress House
23 Savile Row
London W1X 2HE

## 26.2 Objectives of vocabulary control

Why is vocabulary control necessary at all? What objectives are in mind when we decide to control vocabulary?

Archaeology—be it excavation, post excavation, or survey work—generally produces many hundreds and often thousands of pieces of information. It is the task of the archaeologist to make sense of that data, and to find common themes and patterns within it.

When attempting to interpret these huge databases, the archaeologist needs to be able to access the data and use it efficiently. Without some form of manual or mechanical aid, such as an optical co-incidence system or computer, the only method of retrieval is either to rely on personal knowledge of the contents of the record, or to go through the entire record systematically. This latter method is the only one which ensures complete retrieval, and this applies equally to both computerised and non-computerised databases.

The drawbacks are obvious; to rely on personal knowledge is, to say the least, somewhat unsafe, and to go through an entire database extremely time-consuming. Such problems become more and more acute with the increasing size of a database.

The types of queries which are put to national archaeological databases and SMRs, often involve combining a number of pieces of information. For example, for management purposes, an archaeologist might need to know all the earthwork sites under the plough, and all cropmark and soilmark sites under the plough, and in which local authorities these occur. Conversely, some enquiries are site-specific, such as 'is there anything of archaeological significance this particular area?' Such a search is a simple matter to perform in a well-structured computerised database, and is also relatively easy to deal with in a properly mapped manual SMR. However, enquiries needing a global search involve a great deal of work in manual databases where there are no retrieval facilities. Even apparently simple requests for, say, all medieval moated sites within a county, or all archaeology within a parish, can prove incredibly time-consuming. We must be able to get at our data, and we need methods to do so.

The following non-archaeological example demonstrates how sometimes the effects of *not* being able to ensure complete retrieval can have dramatic consequences. Elizabeth Orna (Orna 1983) cites an example from the United States army, where some years ago, a weapons component was found to become unstable over time. A letter was sent out throughout the army, ordering all stocks of the component in question—fuse cap junctions—to be destroyed immediately. In one unit the components had been stored under two different names; 'fuse cap junctions', and 'junctions, fuse cap', but only *one* of the names was recorded. All 'fuse cap junctions' were taken out and destroyed according to the instructions, but in due course the remaining 'junctions, fuse caps' blew up and caused casualties. Non-retrieval of archaeological data is unlikely to be such a matter of life or death, but the point is surely clear!

## 26.3 Problems in vocabulary control

Putting all this into practice may sound simple, but of course there are enormous problems, some of which are detailed below.

| | | |
|---|---|---|
| BAKERY | *use for* | BAKEHOUSE |
| BREWERY | *use for* | BREWHOUSE |
| HOTEL | *use* | INN |
| IRON WORKS | *use* | BLOOMERY (for all medieval sites) |

Table 26.1: Detail from (fictitious) thesaurus

### 26.3.1 The lack of an established site-type vocabulary in archaeology

Firstly, as yet there is no established vocabulary for describing monuments within British archaeology. This is a serious drawback, because as Gross (Gross 1975) points out, an 'automated information retrieval system calls for the use of an information retrieval language'. The relative youth of archaeology as a serious discipline, has meant that, unlike other disciplines such as biology, we have no major eighteenth or nineteenth century classifications for the subject on which to draw, apart perhaps from Christian Jurgensen Thomsen's three prehistoric ages of stone, bronze and iron! (Daniel 1971, p. 32). Thus, it is still possible to call a round barrow by a wide range of names such as 'tumulus', 'burial mound', 'barrow', 'mound' or 'tump' or by a more specific name such as 'twin bell barrow', 'disc barrow' or 'saucer barrow' depending on the database under interrogation. This has obvious implications for those attempting a regional or national search for a particular monument type in a number of SMRs.

On the other hand, there are disadvantages to setting up a very rigid classification, as the 'Use for' and 'Use' index instructions in Table 26.1, (which could cause a loss of meaning if followed) show. If an index, or thesaurus contained such instructions as these, then some monuments could be incorrectly classified i.e:

1. Bakeries are generally defined as commercial establishments as opposed to bakehouses which were attached to farms and country houses.

2. Breweries, again, are defined as commercial establishments as opposed to brewhouses which were found in domestic situations.

3. Not all medieval iron working sites were bloomeries; some were charcoal blast furnaces which were significantly different.

4. Although clearly there is an overlap between hotels, public houses, inns, and alehouses, strictly speaking hotels only started to exist around 1830, and became more and more distinct from coaching inns after that period.

'Hospital' is another example of a problematic term, being applied to a wide range of differing medieval and post medieval buildings.

### 26.3.2 Adoption of unsuitable vocabulary

A second major problem is that many archaeological databases have been set up without adequate thought having been given to *end-uses* of the record. This means that the vocabulary which has been adopted may not be suitable for the questions which will be asked of the database, rather than it being an *integral* part of the system.

Thus, archaeological vocabulary may well need to be different for retrieval and input purposes, or data might have been classified at the 'wrong' level. For example a

specific enquiry for the site-type 'Hermitage' may also pull out everything from a 'Monastic cell', 'Abbey', 'Convent', 'Friary' and 'Pilgrim's rest-house' if these have all been classified under the broader term of 'Religious house'. It is therefore essential that user requirements are fully understood before attempting to control vocabulary. If, however, sites have been classified under very precise and mandatory terms, we may wish to re-classify and rename these sites with academic advances in the discipline, and increasing archaeological knowledge. Terminology does have a habit of going out of vogue; for example, today a 'Causewayed camp' may well be classified as an 'Interrupted ditched enclosure', or a 'Causewayed enclosure'.

The nature of the subject itself sometimes creates problems: some feel that any vocabulary should strictly define the factual evidence, and not be interpretative. Hence a 'Round barrow' might become a 'Circular mound', with perhaps a separate qualifier of 'Funerary', if funerary material had been recovered. And there will always be the problem of whether to classify a few post holes and some pottery as, for example, a 'Settlement' or 'Farmstead' or simply as 'Finds'.

Lastly, we may need to be able to search under all possible names which can be given to a particular site-type. Homonyms such as 'Lock' which can mean both a river/canal structure, and a fastening device, and 'Dyke' which can mean a wall or an earthwork depending on which part of Britain the monument is situated, can cause particular problems here. What, too, about the problems of total retrieval of site-types such as 'Dovecote' or 'Fishpond' if these have classified as *components* or elements of a 'Manor'?

## 26.4 Approaches in use to meet these objectives

There are several approaches currently in use which aim to provide access to archaeological records, and which attempt to solve some of the problems discussed so far. Most involve indexing the data in some way in order to facilitate retrievability. These are: the thesaurus approach, the indexing approach, the word-list approach and the non-controlled approach, each of which will be briefly discussed. In addition, there have also been recent moves towards a classificatory approach to site-type vocabulary.

### 26.4.1 The thesaurus approach

A thesaurus is 'a list of words organised by the ideas they express'. (RCHM (E) ). It provides a method of bringing terminology under control, gives consistent terms for indexing, and through its structure, gives guidance in searching through a database by pointing the user to related areas, and by giving specific alternative terms or synonyms.

Going into these concepts in more detail, we have seen that consistency in terminology is necessary to make sure that nothing relevant is missed because it has been indexed under a variety of different names. A list of keywords is produced to achieve this. For example, 'Rock carving' might be the keyword for 'Petroglyph'.

Secondly, guidance to the standard term is given for those users searching for information under a non-keyword term e.g. users searching under 'Campanile' may be told to use 'Bell tower', whose hierarchical or generic term might be 'Religion', and which might be related to terms such as 'Church' and 'Cathedral'.

Thirdly, information may be classified under as precise a term as possible, so that data which is not required is not retrieved, although as we have seen (section 26.3.1)

| Key: | SN | Scope note |
|---|---|---|
| | Hr | Hierarchy(ies) under which term falls |
| | UF | Use for |
| | BT | Broad term |
| | RT | Related term |
| | NT | Narrow term |
| | Use | Mandatory term |
| Alphabetical list of terms | | |
| ROUND BARROW | | |
| | SN | Circular earthen burial mound |
| | Hr | Funerary |
| | UF | Tumulus |
| | BT | Barrow |
| | RT | Ring ditch |
| | NT | Pond barrow |
| | | Saucer barrow |
| | | Disc barrow |
| ROUND CAIRN | | |
| | Hr | Funerary |
| | BT | Cairn |
| | RT | Cairn cemetery |
| | | Clearance cairn |
| TUMULUS | | |
| | Use | Round barrow |

Table 26.2: Detail from a (fictitious) thesaurus

there are also arguments against this as being too inflexible.

Lastly, those seeking to retrieve on a range of similar monument types should be able to find them through directions and guidelines given in the thesaurus. For example, the Broad Terms, Narrow Terms, Related Terms and so on. The very structure of the thesaurus itself should also provide this facility. For example, the Royal Commission on the Historical Monuments of England's Architectural Thesaurus (RCHM (E) ) organises terms into hierarchies and sub-hierarchies by type or common idea. Terms can relate to other terms both within a single hierarchy or across hierarchies. That particular classification system includes hierarchies of *form, function* and *building complex*.

There are several essential items of any thesaurus, as illustrated by Table 26.2:

1. Firstly, a list of **Keywords**, which are used for indexing information. Sometimes these are known as standard, index, lead or entry terms. These are organised alphabetically within hierarchies or within a structure which reflects the meanings of the terms.

2. Second, a list of **Alternative** terms, sometimes, confusingly, also known as entry terms. The purpose of these is to aid complete retrieval, and they often consist of synonyms, or words having the same meaning as the keyword.

3. Third, clear **Instructions And Rules** to aid users, by which alternatives are

differentiated from the keyword, and by which broad and narrow terms, and related terms may sometimes be shown.

4. Lastly, **Guidance** on the use of terms—or scope notes as they are sometimes called, which indicate the meaning of the term and when it should be used.

A good thesaurus is an invaluable aid to retrieval, but it does require a great deal of hard intellectual effort to produce. The overall categories themselves are difficult to decide, and any thesaurus also requires rigourous maintenance if it is to be updated. It also has to be very well cross-referenced in order to cope with the problems of searching under regional, obsolete and foreign terms—in fact any term which is not a keyword.

### 26.4.2 The indexing approach

Indexes, unlike thesauri, are *not* ordered by the ideas which the words express, but consist of an alphabetical list of keywords and alternatives, fully cross-referenced to each other. Links to other topics are *not* provided by the structure of the index itself, as with thesauri, but only by the cross-references throughout the list.

Indexes have the advantage of flexibility as they are not tied to a structure or classification of any kind; these latter could prove to be a constraint with changing interpretations and theoretical advances in archaeology. Some indexes if they are well-structured and cross-referenced, do not even recommend preferred terms, making it possible to use *any* term for describing or retrieving a monument. This also has the advantage of not 'fossilising' the index into the terminology of the 1980s, as any term can be used for retrieval, and the index can be updated as new concepts and terms come into use. Of course, like thesauri, indexes are also time-consuming to construct, and need to be correctly maintained.

### 26.4.3 The wordlist approach

Many SMRs currently use this approach. An alphabetical list of mandatory keywords, usually without any alternatives terms or cross-referencing, is used to classify types of monument. The advantage is that it is easy to set up and maintain, but has the drawback that users are not shown related terms of interest under which they might also like to search. An example of a wordlist can be seen in Table 26.3.

### 26.4.4 The non-controlled approach

Some sites and monuments records allow *any* terms which the compiler of the record feels appropriate, to be entered into the database. There are no preferred terms or keywords, and retrieval of data involves selecting all possible entries from a print-out of the words in use. The advantages of this approach are its total flexibility, but its drawbacks are the time it takes to search the database as there are no real methods for retrieval.

(It should be noted that this method may be the most suitable for a small database which is unlikely to expand greatly, as the time taken to construct an index or thesaurus should be weighed against its limited usefulness in such cases).

| |
|---|
| Abbey |
| Adit |
| Allotment |
| Almshouse |
| Altar |
| Amphitheatre |
| |
| Bailey |
| Bakehouse |
| Barn |
| Barrow |
| Bloomery |

Table 26.3: Detail from a (fictitious) wordlist

### 26.4.5 The classificatory approach

Some are now taking the view that the classification of monument type is the key to retrieval. This, of course, is a tricky area, as ultimately, in order for a classification to be successful, it would need to be broadly agreed throughout the profession. As a discipline, we would do well to learn from the mistakes of nineteenth-century natural science (recently, biological studies have been moving away from the taxonomic approach).

The Monuments Protection Programme, currently being undertaken by English Heritage (see also Booth, this volume), is concerned with the classification of monument type, partly through the production of monument descriptions. To satisfy the requirements of the programme, 'the preparation of monument descriptions for all the main classes of monument, all identifiable period-specific forms of relict landscape, and each period-specific form of urban area is an essential preliminary to evaluation.' (Darvill *et al.* 1987). Terminology may thus be created through the classification of monument type.

However, further vocabulary for retrieval may still need to be developed even if a site-type classification becomes accepted, depending on the objectives and end-uses of a database. There may still be a need to retrieve site-types at a different level, or indeed a variety of levels, which would not be possible using the classificatory terminology alone. For example, if a monument has been classified at a high level of detail e.g. 'Class I henge' or 'Chambered long barrow', someone wishing to search for all henges or megalithic tombs may have difficulty in searching under the classificatory terms without an index of words for retrieval purposes, linked to the classification terminology.

One thing is clear, classifications which *impose* a rigid hierarchy upon terminology, tend to be far less successful than those which have been formed from the 'bottom up' *i.e.* using the natural groupings of words themselves to create categories.

## 26.5 Conclusion

Over recent years, archaeology has seen an increasing number of computer applications, not least because computers cut out the boring tasks of searching and retrieving complicated sets of data, which would be tedious, time-consuming or impossible if performed manually.

The effectiveness of these searches is not only determined by variables such as the hardware and software, the quality, scope and depth of the information itself, and the data structure, but also by the degree of control exercised in the vocabulary used to input and retrieve the textual data. In turn, the *end-uses* to which the records are to be put, must be fully understood before a database can be structured, and vocabulary controlled in order to *meet* user requirements. This is true of all records, manual or computerised.

As archaeologists, we have spent much time over recent years improving the technical aspects of our computer programmes. But without taking all these factors into consideration, the full potential of computerisation will never be realised.

## Acknowledgements

I am most grateful to Nigel Clubb, Ben Booth, Robert Cruse, Helen McMurray, Rebecca Payne and Russell Man who have discussed and commented on aspects of this paper.

## References

DANIEL, G. 1971. *The Idea of Prehistory*. Pelican.

DARVILL, T., A. SAUNDERS, & D. W. A. STARTIN 1987. "A question of national importance: approaches to the evaluation of ancient monuments for the Monuments Protection Programme in England", *Antiquity*, 61: 393–408.

GROSS, R. 1975. "The application of thesauri in the state archives of the G.D.R.". *in* Bell, L. & Roper, M., (eds.), *Proceedings of an International Seminar on Automatic Data Processing in Archives*. HMSO, London.

LAVELL, C. 1985. "Information: Are we retrieving it?". *in* Burrow, I., (ed.), *County Archaeological Records, Progress and Potential*. Association of County Archaeological Officers.

ORNA, E. 1983. *Build yourself a thesaurus; a step by step guide*. Running Angel, Norwich.

ROYAL COMMISSION ON THE HISTORICAL MONUMENTS OF ENGLAND 1987. *Draft Thesaurus of Architectural Terms*. R.C.H.M., London.