Indexing of a computerized bibliography for London’s archaeology

Audrey Adams

Describes the work of producing the index for a computer-based bibliography of publications and archival records of the archaeology of London. Details are given of the determining factors: analysis of the data and of the users, the establishment of a keyword structure, with specific provision for computer applications. The importance of index structure and vocabulary control is emphasized.

Background information

The Museum of London was established in 1975 to collect, display and interpret items reflecting the history and archaeology of the nation’s capital, and to be a centre for education and research on all aspects of London’s past. This paper discusses the Museum’s vigorous rescue archaeology programme which is carried out by its Department of Urban Archaeology. This Department was set up in 1973 to record and preserve where possible London’s archaeological heritage. The importance of this heritage was argued by Victorian antiquaries who pleaded in vain with the Corporation of London for such a programme in the 1840s. World War II bomb damage revealed many Roman remains, and the Guildhall Museum (incorporated now in the Museum of London) did what rescue work it could, including recording of archaeological remains. The City of London is constantly undergoing structural changes and as these occur, the Museum negotiates with the developers to have access to the site to excavate, record and preserve any archaeological evidence it may contain. As a result, the Museum has produced a large number of publications and an even larger number of archival records. What has been missing is the vital link: an index to provide the needed access to all this archaeological evidence.

Therefore in 1982, following consultations with the Department of the Environment, Greater London Council (current project funder), and the Council for British Archaeology, the Museum’s Department of Urban Archaeology established a project to produce a computer-based indexed bibliography of London’s archaeology. Its aims are as follows:

1) To maximize the usefulness of the vast amount of information concerning London’s archaeological heritage, by making it readily available to the general public, museum staff, historians and other researchers.
2) To simplify and accelerate the writing of reports on new archaeological discoveries, by allowing rapid reference to previous relevant work.

It is intended that the bibliography will be stored on the Museum of London’s computer system, in the Museum Library, where it will be readily available to the general public, students, researchers, archaeologists, historians and town-planning authorities.

The need for the project

In order to interpret properly any archaeological discovery (whether an individual object or a complex series of features on an archaeological excavation), it is necessary to know about previous discoveries on the same site, and similar discoveries in the surrounding region. Archaeological evidence is like a jigsaw puzzle. Each of the pieces in itself may not seem significant, but each contributes to the pattern and is needed to complete it. Therefore all pieces should be recorded. In London, information of this kind has been collected and recorded since the 17th century and is now preserved in a vast variety of books, periodicals, newspapers and archives. There is not even a comprehensive list of publications and sources. When the researcher does manage to locate a relevant periodical it has either no index or one inadequate for his needs. Some, as an early Illustrated London News example shows, have most unhelpful indexes: a Roman lead coffin found near Old Bow is indexed under ‘C’ for ‘Curious’ lead coffin. A classic example of how not to index!

Thus ready access to archaeological information has been impossible, with the predictable result that research is time-consuming and so haphazard it is certain that important records are constantly being overlooked, to the detriment of our understanding of the past. Archaeologists faced with this problem have tended to fall back on an unsatisfactory solution: quoting only known sources instead of examining all the evidence. This has the advantage of saving time—and archaeologists never have adequate time for report writing—but the disadvantage of perpetuating error. Sins of omission rather than commission!

As this project has a vast amount of information to
handle and a limited budget, it is better to produce an index than to put the full printed text of sources onto the computer. It is the index that will provide the needed access to information.

**Methodology**

Where does one begin to produce a computerized index to a bibliography of archaeology? Several things had to happen at once or have equal priority—if such a concept is not a contradiction in terms. But as this was to be a computerized index, I began by analysing what pre-computer work needed to be done. Computers have come a long way from the early, clumsy, code-using models, but they still cannot think. They can only do exactly what they are asked to do. They cannot in fact provide anything a manual system cannot provide, except that they have the potential to provide faster retrieval of information and a more viable storage system for quantities of information. This in turn implies the possibilities of fully articulated indexing.

In other words, the principles of indexing information are the same for either a manual or a computer information system. The basic aim is to extend the availability of the information to as many users as possible. What makes a system a ‘good’ one is one which enables users to retrieve all the applicable information within a reasonable timespan; or, in the new jargon, is ‘user-friendly’.

As users’ requirements vary, to achieve the basic aim means providing extensive cross-references to give different routes to the same information. Theoretically, a manual system, either back-of-book index or library card catalogue could provide these cross-references, but there comes a point when the sheer volume of space required makes it a practical impossibility. Once the volume of data reaches a certain size, it is necessary to computerize.

Some manual systems handling large amounts of information are a practical compromise. They have based their indexes on a system of ‘implied greatest use’, i.e., to answer a restricted range of preconceived questions. If a question does not fall within this range, it can only be answered by a prolonged search through the text of the book. It is the old needle in the haystack situation. The needles are there, but does the researcher have the time to locate those that are not indexed?

It is essential to understand that this particular problem will not automatically disappear by transferring the information from current manual systems to a computer. The information still needs to be analysed for its potential use and the data base constructed accordingly. The problem of structure not only remains, it is the key to the computerized data base. Structure cannot be stressed too strongly. A simple error made by the Patents Office in computerizing its patents illustrates this point.

The manual system, i.e., the card catalogue of the patents, was transferred to the computer without any alteration. The result was unfortunate. Patents held by The Ford Motor Company were filed under ‘T’; some were filed correctly under ‘F’ because the card read ‘Ford Motor Company’. As the computer cannot think, it files under the first letter of the first word. The human filer, of course, obeys the filing conventions and ignores the definite and indefinite articles at the beginning of a title and so interfiles ‘The Ford Motor Company’ with ‘Ford Motor Company’. Correct structuring of the computer program would have saved this massive error.

Computers operate on logic; therefore the structure of the index must be logical. Indexers know that a consistently ordered index is the most efficient one for all users, but for a computerized index it is a mandatory requirement. Without consistency, the index will, in computerese, be an exercise of ‘garbage in, garbage out’.

Two technical problems remain: writing a requirement specification for computer analysts so they can understand the programming needs, and then selecting the software/hardware that will best meet these requirements.

To sum up, if a computerized data base is to meet the needs of as many users as possible, it is essential first to undertake the appropriate analysis of the data and users’ needs before examining computer soft- and hardware.

How did I apply this abstract theory to this indexing project? As noted earlier, two things now had to happen at once. First, I circulated the curatorial staff of the Museum and the field archaeologists, i.e., the current users, to find out their needs and also to ask them for any thoughts on additional potential users. Second, I began analysing the nature of archaeological data in order to determine the implications for indexing.

**Analysis of archaeological data**

What is the nature of archaeology?

Archaeology is a study of residuals, the remaining material manifestations of previous cultures. Archaeology is primarily site-oriented. It seeks to answer these questions about our ancestors: who was where, when, doing what, with what? In terms of being able accurately to identify archaeological evidence, where and when can most often be assigned; doing what, with what are much more elusive. Thus for much of archaeological evidence, no function can be assigned. What part a structural feature or an artefact played in a former culture is unknown at the time of recording and, disconcertingly, may remain so, but there is a ‘something’ made by man to record and in time, to index.

Archaeological records go back to the 17th century and these contain ‘historic’ words not in current use. And regionally, there are variations in words to express the same concept.

The Indexer Vol. 14 No. 4 October 1985
The implications of the nature of archaeological data for indexing are:

1. No single reference will bring out all the evidence, therefore coded or classification systems based on hierarchical principles will not work, e.g., a medieval castle is both a fortification and a domestic dwelling. Also, as coded classifications are based on known function or type, they are not applicable to many of the artefacts and structures of prehistoric to industrial age archaeology, i.e., the greater part of recorded archaeological evidence.

Prehistory refers to that period without documentary evidence. This is followed by proto-history where written evidence is limited to inscriptions. History is the period for which we have documentary evidence. Documentary evidence is also a patchy source. Obviously it is limited to the documents that remain and to the contents of those documents. The Liberatic Rolls note in great detail even to the colour of paint the dwellings ordered by Edward III² for himself, but the dwellings of peasants were of no interest. Archaeological evidence alone can provide this undocumented information. Indeed, for most of the material remains, the remains themselves are the primary evidence, documents are secondary sources and not always reliable, as a recent excavation in London disclosed: 'Though a detailed plan exists of the buildings on the site in 1610, none of the 17-century and later brick cellars uncovered corresponded precisely with those depicted—a reminder that some quite late changes in development can only be elucidated by excavation.'²

2. Hierarchies shift according to the evidence of differing sites: e.g., a Roman hypocaust is sometimes part of a villa, sometimes part of a public bath-house; or the hypocaust alone remains with no structural evidence to show what it was once part of.

3. The evidence is dynamic with new data to be added, sometimes resulting in new interpretations of previous evidence. All interpretations have to be recorded, because a final 'right' one cannot be assumed.

4. There is a need to cross-refer variant terms used for the same item over three hundred years of archaeological recording and for the extensive regional variations. For example, 19-century antiquaries used 'vault' where current archaeologists would use 'cellars' or 'basements'. And in Lincoln, 'gouts' are still used for 'drainage ditches', a regional specialty that was noted in 1849.³ 'Barge' is most commonly used to describe a flat-bottomed boat, but in the Exmoor dialect, it means a 'dirty great hog'.

Analysis of the users

1. Who is currently using the data?
2. What shortfalls do current users report?
3. Projected new users.

The current users reported back with the following requirements:

- 1 address of the excavation/discovery
- 2 O.S. grid reference
- 3 Parish
- 4 excavator's name
- 5 date of the excavation
- 6 where the finds are stored
- 7 where the archives are kept
- 8 site features and finds
- 9 dating of site features and finds
- 10 illustrations, maps, plans
- 11 correct bibliographic references automatically produced
- 12 location of publication holdings.

Potential users were given as historians, other researchers, planning authorities and the general public. For these users, subject categories by generic and function would be necessary. For example, they might be looking for evidence of 'Roman Religion' but not know the constituent terms to allow them access to the archaeological information they need. Current users also asked for priority to be given to periodical indexing. Archives and books are to be indexed later.

What keyword structure would meet data and users' requirements?

The basic assumption must be that the user knows nothing. He/she may be a newcomer to archaeology or, though skilled in archaeology, may be unfamiliar with retrieval methods or with the idea of a structured index. As the aim is to save the researcher time, the structure of the index cannot make any assumptions of prior knowledge.

Keyword structure

Any index requires a list of keywords. So first of all I looked around to see if there was a ready-made keyword list for archaeological information. Unfortunately, there was not an authoritative complete list for British archaeology. I had therefore to establish a keyword list. Again, because of the vast scope of the project, it was necessary to divide this task into workable stages. The initial keyword list for site features (such as pits, revetments, walls, crypts, etc.) was compiled first. The vast majority of these words were selected in the course of analysis of over 2000 articles in thirty periodicals of all periods from the 18th century onwards.⁴ Historic and specialist dictionaries provided the correct spelling. The initial keyword list for finds or artefacts (such as keys, bowls, dress-hooks, earrings, etc.) was then compiled from the same sources, plus specialists' lists from the various departments of the Museum. This is currently being combined with the features' keywords to provide the master list. It is important to user-test these lists before adoption, and accordingly they are circulated to

*There are hundreds of volumes of periodicals, including newspapers, remaining to be indexed.
users for comments, corrections and additions. New keywords are added as required.

The next thought was, could the computer provide a fully articulated index from data input, i.e., could software companies provide us with a thesaurus or index package? Unfortunately, not. Users’ needs vary with differing data bases so idiosyncratically that no such package can be produced. So, hard intellectual effort must be put into keyword structure. It is well worth the effort to produce a structured keyword list. Research is just beginning to analyse the results of the past twenty years’ experience of free access of online data bases. Some of these have no in-built word structure, leaving each user to be his own indexer. The research results show very low retrieval of data because the user can retrieve only those words he knows. If he asks for ‘cellars’ he will miss all the references to ‘basements’, ‘vaults’, ‘crypts’ and ‘undercrofts’. All this needed information is lost because there is no built-in synonym retrieval.

Hoping at least to avoid these known errors, my archaeological keyword list is structured thus:

1. The listing is alphabetical by noun (feature or find) followed alphabetically by distinguishing adjectives of type, shape, size, material, condition, etc., in any applicable combination,

   e.g., Drains, circular, lined

   The keyword entry is inflexible for the index; i.e., the noun followed by adjectives; the flexibility is built into the retrieval part of the computer program so that any required combinations can be retrieved. For example, the retriever could ask the computer for ‘lined circular drains’ or ‘circular lined drains’ and retrieve all relevant data.

   This reverse entry has been deliberately chosen because I cannot presume any prior knowledge of the users. For example, there are more than one hundred types of the burial mound called ‘barrows’. If I listed by adjectives, a person using this list manually would have to know every type of barrow to find them.

2. Keyword listings are plural following indexing conventions.

3. Keywords are cued by upper-case initial.

4. Words not used as keywords are given a ‘see’ reference, which refers to the correct keyword,

   e.g., ring cairns see Cairns, ring

   long barrows see Barrows, long

   This again is for recording and index purposes. If the user asks the computer for ‘ring cairns’ he will retrieve all relevant references from the data base.

5. ‘see’ words are cued by lower-case initial.

6. To provide the user with maximum information available to him, related keywords are given ‘see also’ refer-

ences on the following bases:

   i) Synonyms

      e.g., Jugs, drinking see also Blackjacks; Bombards

   ii) Similar functions

      e.g., Cisterns see also Aqueducts; Reservoirs; Tanks, water; Wells

   iii) Alternative interpretations

      e.g., Dovecotes see also Kitchens, detached

   iv) root terms

      e.g., Aisles see also Barns, aisled; Chapels, aisled;

      Churches, aisled; Halls, aisled; Houses, aisled;

      Villas, aisled

   v) for collective nouns, parts are shown

      e.g., Battles see also Crenellations; Crenelles;

      Crenels; Embrasures; Machicolations; Merlons;

      Murderholes; Parapets

   vi) any combination of the above 5 divisions

      e.g., Tracks see also Cart-tracks; Causeways; Paths:

      Road, three-tracked; Trackways; Ways

   vii) distinguishing tags are added where necessary

      e.g., Hospitals (sch) see also Schools, charity

Because of the nature of archaeological data, it was considered essential to list keywords as specifically as possible first. The next task is to add applicable generic and function tags where possible, so that, for example, someone could ask for ‘Roman Religion’ or for ‘Medieval Water Supply’ and retrieve references to all the components.

Recording form

A recording form was designed with appropriate boxes for entering details (see user requirements) from each periodical article. Most important to archaeologists is the dating of evidence. Chronological tags are not shown on the keyword list; allowance is made for them on the recording form. Recording forms using controlled vocabularies provide the necessary consistent entries required by computers.

Computer applications and requirements specification

In order not to impose too technical a language on computer users, computer companies selected established words for computer functions wherever possible. Hence ‘keyword’ for computers refers to any sorting field, in this instance, any user requirement, e.g., author, Saxon knife, Bronze Age hut, O.S. grid reference. How well one is able to provide quick but complete retrievals of data from the computer for users depends upon how adequately the requirements specification is prepared for computer analysts. The computer cannot think; it will not do anything it is not specifically programmed to do. Therefore the requirements must be worked out not just in principle but in detail.

For this particular index, the cross-references have to be unpicked and sorted into needed categories:
1) True synonyms by chronology.
   e.g., (Roman) Caligae see also Sandals, military

2) Related words by chronology.
   e.g., Bombard as medieval drinking jug.
   Bombard as post-medieval piece of artillery.
   Bombard as medieval early form of bassoon

Another requirement is connecting every query to bibliographic reference, site location, O.S. grid reference and so on. And noting if feature or find is illustrated and/or mapped.

Next it is necessary to work out what prompts should be put into the computer to inform the user of all possible relevant information that is in the data base. For example, the specialist may want only mosaic floors having gods and goddesses. Automatic referrals to all mosaic floors will load him up with references he does not want. But such referrals may be needed by other users, so the automatic prompt will be offered as a choice, i.e., the user will push a given key if he does not want the referrals to be shown.

It is particularly important to have this prompt facility because of the great variations in recording detail over the three hundred years that have passed and the possibilities for the future. For example, Victorians were not always able to date their archaeological evidence closely. But this evidence should not be overlooked on this count as it may be the only evidence for a particular site. For example, the specialist may ask for 'Boudiccan burnt buildings', some of which will be listed under the close dating of 61 A.D. The user will be prompted to ask more broadly for 'Roman burnt buildings' in order to pick up the not so closely dated references. He can call up all these references on the screen but note only those which in his judgment might be useful.

As part of the specification requirement, two separate computerized keyword lists will probably meet all users' needs: 1) a 'dictionary' listing and 2) a retrieval listing. The dictionary listing will contain any necessary scope notes. It is an aid for researchers uncertain of the keywords necessary to access the data they require:

   e.g., User asks dictionary for 'medieval hospitals'.

Dictionary offers:
1 (medieval) Hospitals see also Almshouses; Bedehouses; Farmeries; Hostelries; Hostels; Houses, spital; Houses, spittle; Infirmaries; Maisons-dieu
2 (medieval) Hospitals (sch) see also Schools, charity

User may go to retrieval listing with his choice, i.e. (medieval) Hospitals (sch) and be given the bibliographic sources.

The dictionary facility eliminates the need for printed manuals of keywords.

What should the computerized index provide for archaeological information?

1. A 'user-friendly' computerized index that can be accessed directly; i.e., does not require consultation of a handbook or manual or the attendance of a specialist. It will all be on the computer.
2. A structured terminology for the recording and indexing of archaeological information which provides easy access to all words in past and current use, allowing the searcher to find all the variant terms to a given concept in a reasonable time.
3. A terminology that can be expanded in a controlled manner to incorporate new keywords, i.e., a system that is active and responsive, not a limiting classification system.
4. When the dictionary list is printed out, it is available as a recording/indexing tool that is automatically updated as new words are entered and can be used for manual or computer records.
5. In effect, a history of the use and development of archaeological terminology.

Computer implications for indexers

1. Structured keyword lists, i.e., controlled vocabularies, are the minimum essential for maximum retrieval from any data base. They require subject specialist indexers. Structuring a keyword list is initially time-consuming, but an easily expandable controlled vocabulary allows fast indexing of new information.
2. The resulting indexed data base provides the user with a means of retrieving all the relevant information he requires. In other words, the indexers' initial input ensures maximum retrieval with minimum effort and time on the part of the user. As Dubois points out, a researcher asking an unindexed data base for 'instant coffee' missed all the more numerous references to the trade synonym 'soluble coffee'. The more complex the data, the more references will be lost to the user where vocabulary control does not exist. Despite all the claims put forward for computers possessing artificial intelligence, they cannot index.

References

An additional useful reference is:

Audrey Adams is Bibliographer of the Museum of London’s project to produce a bibliography of London’s archaeology.

Indexes for the rugger club

We have received from an unimpeachable source the index only to Limericks historical and hysterical, ‘plagiarized, arranged, annotated and some written’ by Ray Allen Billington, Senior Research Associate, the Huntington Library, PhD, (MA Oxon) LittD, LHD, LLD (W. W. Norton & Co, 1981). 101 pages.

The work may be of interest to anthropologists who specialize in post-match mores among practitioners of the more macho sports. In a family journal, little indeed may be quoted from this index. It would appear that the text makes reference to the technicalities of the lyricist’s craft:

- Harvard University, as inspiration for limerick writers, 15-16; limericks on, 62
- Opechancanough, difficulty of limerick on, 28
- and some examination of experimental rhyme may be indicated by:
  - Beckwith, Duke of, foreshortened perspective of, 50

There is some treatment of the military arts:

- Parma, exploits of warrior from, 57
- Ephriam [sic], crusades with Kay, 76; (we are troubled by the further entry: Kay, crusades with Ephriam, 96 [sic]). Exotic place-names, universities, and young ladies from, and in, exotic, and other, places, and in universities, also receive the editor’s attention.

The index may have some claim to the most succinct entry ever, with:

- Sexual activity, pleasures of, 20-101; variant activities of, 20-101

Nevertheless, we fear that we cannot recommend this index to those of a nervous disposition, to the sober, to Queen Victoria, nor a fortiori to any reader of The Indexer. We must however suppose that this pioneering creation of an index designed for the clubhouse bar may of itself have an elevating influence on the tastes of the unfortunate inmates.

J.L.B.

A glossary for indexers IV

Martyn Gilchrist glosses for us the calls from The BBC microcomputer user guide listed in The Indexer 14 (3), 176:

- Write to I/O area SHEILA: with best wishes to our favourite Australian indexer, memorably yours FRED and JIM (c/o &93 and &95)
- Read key within time limit: or go directly to BREAK, do not pass GOTO, do not collect PRINTout.
- Flush selected buffer: Definitely! Actually a half size seed tray is more sturdy than a shoebox. For the larger job two trays can be used.
- Remove character from buffer: It happened at Waterloo railway station. It was the compiler of the index to the definitive 5-vol. edition Stress and metal fatigue in fire hydrants, The. Finished to a near impossible deadline and at an almost derisory fee. Nobody told him about the repagination until it was too late.
- Cancel VDU queue: further in-house investment in computer terminals resulting in more published indexes. Dependent upon no VAT element input to works.
- Disable character entering input buffer event: spurious headings and page references which somehow get into the index. These headings are always out of alphabetical sequence and the page references apply to the previous book. Take a twenty-minute break to restore concentration.
- Acknowledge detection of ESCAPE condition: a very satisfactory and pleasing state of affairs. The indexer’s special skills recognized. Allowed to decide style, length of index etc. Complimentary copy of work always sent.
- DIM GAP: this little known mnemonic sometimes appears in the more esoteric of text books and what it stands for is Don’t Index Mapnames Give All Places. It’s a helpful way of remembering that placenames on maps in general works are not indexed, but placenames in the text are.
- RUN: your options are
  - A = ALA filing rules
  - B = BS 1749
  - O = Other
  - please specify A/B/O?