Information Technology: Research and Development (1983), *2* (169–186)

INFORMATION RETRIEVAL AT THE SEDGWICK MUSEUM

M. F. PORTER

Department of Earth Sciences, University of Cambridge, Downing Street,
Cambridge CB2 3EQ, UK

(*Received 16 May 1983*)

ABSTRACT

The Sedgwick Museum at the University of Cambridge now has a high quality and comprehensive online IR system covering its collection of 450 000 catalogued fossil objects. The indexing process, and the retrieval capabilities are described in detail, and an example is given of how the IR system is used with real museum enquiries. It is also shown how the IR system is used as an aid in many different aspects of data management, such as catalogue updating and editing, and dealing with loans of specimens and movements of specimens between drawers.

1. INTRODUCTION

The computerization of the catalogue of the Sedgwick Museum, Cambridge (Porter *et al.*, 1977), hereafter simply SM, has had a long and chequered history. The early feasibility studies go back to the late 1960s, although almost all the work of data analysis, program writing, and the monumental data preparation of the catalogue of over 400 000 specimens was carried out throughout the seventies. This work was all done under the direction of Dr Jonathan Cutbill, who is now at the National Maritime Museum, Greenwich, and it bore fruit in a number of ways, with the development of useful software, the analysis of different kinds of museum data, and the formation of the Museum Documentation Association, but by the beginning of 1977 when funds effectively ran out, the system had still not been made to work and there was no expectation at that time that it would ever do so. But happily this expectation has proved false, and at the beginning of 1981 the project was revived, this time with a successful termination. Currently the entire catalogue exists in a computerized form, and the computer is used for all aspects of cataloguing and data management. The catalogue is also accessible through a powerful IR (information retrieval) system.

The IR system now in use at the SM was written at the beginning of 1982, and in some ways the decision to write it went counter to previous thinking. Certainly by the end of the earlier phase of the project in 1977 it was supposed that information needs would be satisfied by consulting a range of hard copy indexes. These indexes would be large, and new editions would be produced at certain regular but

0144-817X/83/03 0169-18 $03.00 © 1983 Butterworth & Co (Publishers) Ltd

infrequent intervals. In fact the taxonomic index to the fossils is still treated in this way, and the capability is there for generating the rest, but we have found that the IR system as it stands, supplemented by the taxonomic index, provides a fully satisfactory mechanism for dealing with museum enquiries. Furthermore the IR system can be used as an aid in such activities as data correction, lending specimens to research workers, moving specimens between drawers, and indeed the generation of hard copy catalogues and indexes. Precisely how will be seen later.

2. THE DATA OF THE SM CATALOGUE

The SM catalogue is represented as a single set of records in GOS. GOS record structure will be familiar to users of the services of the Museum Documentation Association, and is briefly described in Porter (1982). Each GOS record in the catalogue contains a set of *fields*. Each field contains data, and also has a name, which is written preceded by an asterisk. Thus *pers is a person field, *date is a date field and so on. The complete set of field names in use in the SM catalogue may be organized as a hierarchical structure, in which certain field names effectively contain a list of other field names. (Thus *oh 'contains' *f, *pers, *date, *rn and *note.) The broad outline of this structure may be represented as follows:

*key	specimen identity number	e.g. A. 10397
*bcat	broad category of the specimen	e.g. Trilobite
*part	information about the specimen's constituent parts	
*store	storage location	e.g. ii.t.2
*oh	ownership history group, containing	
*f	method of acquisition	e.g. purchased
*pers	donor	e.g. Smith, R. W.
*date	date of acquisition	e.g. 1907
*rn	identity number in donor's collection	
*note	general note field	
*cs	collection statement group, containing	
*pers	collector	
*date	date of collection	
*rn	field number of specimen	
*note	general note field	
*ss	stratigraphy group, containing	
*loc	locality name	e.g. Scarborough
*ll	latitude/longitude	e.g. 12d 15m N 27d 6m E
*gr	grid reference	e.g. SH 33323241
*gd	geographical detail	e.g. 3m below embankment
*rk	lithostratigraphic information	e.g. Crugan Mudstones
*age	chronostratigraphic information	e.g. Jurassic
*zone	biostratigraphic information	e.g. Lower Cordatum Zone
*sd	stratigraphic detail	
*note	general note field	
*re	research event group, containing	
*f	nature of research	e.g. identified

*pers	person who conducted the research	
*date	date of the research	
*tax	taxonomic name	
*td	detail of taxonomic name	
*keyw	morphological information	e.g. upper right molar pygidium
*pres	preservational information	e.g. internal mould
*lith	lithological information	e.g. black shale
*ref	documentary reference, containing	
*a	author	
*d	date	
*t	title	
*r	name of journal etc.	
*v	volume	
*pp	page and figure information	e.g. pp. 34–35 figs. 4a, b

In fact this is a simplification of the structure which is actually in use, which has provision for over 100 fields, but is adequate for purposes of discussion. It must be remembered that in an actual record of data, only a certain subset of these fields need occur, but that it is possible for individual fields, or groupings of fields, to repeat as often as necessary. This repetition can be put to a variety of uses. Thus, for example, a list of authors of a paper can be represented by repeating the *a field:

*a MacKenzie, W. S. *a Donaldson, C. H. *a Guilford, C.

A geographical location can be represented by a list of place names:

*loc Crugan Farm *loc Llanbedrog *loc Gwynedd *loc Wales

Successive transfers of ownership of a specimen can be represented by repeating the *oh field, and most importantly, successive identifications can be represented by repeating the *re field. It should also be pointed out that many of the fields in the above list are really groups of two fields, containing a keyword part (the important bit) and a detail part, which is of lesser importance. The detail part is written prefixed by the character '<', e.g.,

*pers Arber, E. A. N. < Trinity College

Records are input to the system by prefixing each category of data by its appropriate field name, e.g.,

```
*key B.20230
*re*f fig'd *ref*a Duncan *d 1868 *t Brit. Foss. Corals
*r Mon. Pal. Soc. *v Ser.2 part iv *pp p. 53 pl. xii fig. 15
*tax +Lepidophyllia +stricklandi Duncan *f holotype
*re catalogued *a Woods *d 1891 *t Cat. Type Foss. Woodw. Mus.
*pp p. 20
*ss*age Lower Lias *zone \UA. bucklandi\N Zone *loc Chadbury
*loc Worcestershire
*oh*pers Strickland Coll.
*bcat Anthozoan *store xx.t.c.
#
```

This can be compared with the first entry in Figure 1, which is how this record is printed out by the retrieval system.

The total size of the catalogue, including both the textual data and the machine representation of the field indicators is about 100 megabytes—i.e., 100 million characters of information. The original manual catalogue contains this information in about 200 sturdy binders which require their own catalogue room within the museum. The computer printout of the data and its taxonomic index fill over 36 000 pages. Even so this quantity of information can be made to fit on just over 200 microfiche, which sit comfortably in a small card file. This compaction of the catalogue on microfiche is important in using the IR system, since it means that specimen identity numbers can be retrieved at a VDU, and the corresponding records inspected on a microfiche reader beside it. As we will see, this approach has certain advantages over a system in which the records themselves are made to appear on the VDU screen.

3. THE INDEXING OF THE SM CATALOGUE

In order to make the SM catalogue available through an IR system, it is necessary to extract index terms from the GOS records which constitute the machine-readable version of the catalogue. This is done automatically; in other words a program goes through the catalogue file, and extracts a set of terms from each record which will index the specimen or specimens which that record represents. There are of course a large number of different ways in which the indexing might be done. The rationale behind the way we have chosen to do it should emerge more clearly with the examples of the use of the IR system.

Each *term* in the system consists of an upper case letter followed by a series of characters each of which may be a lower case letter, digit, full stop or comma. Terms are written between quotes, e.g.,

'Abarker,r.w.1927'
'Roxfordclay'

The initial letter gives a broad categorization for the term. Thus 'A' indicates authorship information, and 'R' indicates rock information. The first of these terms indexes specimens which have been described in R. W. Barker's paper of 1927, while the second indexes specimens which derive from the Oxford Clay. As an example of the indexing for a complete record, B.20230 of Fig. 1 is indexed by the following terms:

'Aduncan1868'
'Awoods1891'
'Dstricklandcoll.'
'Ffigd'
'Fholotype'
'Ganthozoan'
'I'
'Lchadbury'
'Lworcestershire'
'Qlias'
'Sxx.'

B.20230
(Anthozoan) store: xx.t.c
Strickland Coll.
A. bucklandi Zone, Lower Lias; Chadbury, Worcestershire.
Fig'd, holotype, Duncan, 1868, Brit. Foss. Corals, Mon. Pal. Soc., Ser. 2 part iv, p. 53 pl. xii fig. 15, as *Lepidophyllia stricklandi* Duncan;
Catalogued, Woods, 1891, Cat. Type Foss. Woodw. Mus., p. 20.

B.20231
(Anthozoan) store: xx.t.c
Strickland Coll.
A. bucklandi Zone, Lower Lias; (This specimen, with B.20,231, was associated with Strickland's ms. label "Chadbury, Worcestershire", but this locality is not among those quoted by Duncan.).
Fig'd, Duncan, 1868, Brit. Foss. Corals, Mon. Pal. Soc., Ser. 2 part iv, p. 14 pl. xxi fig. 8, as *Thecosmilia martini* de Fromental.

J.11576
(Anthozoan) store: xx.t.61
a, *b* slides
Presented, Part, G.M. Coll., 1945.
Lusa Coral Bed, Broadford Beds (lower part), Lower Lias; Ob Lusa, Broadford (4 miles E.N.E.), Skye.
Topotype, as *Isastraea murchisoni* Wright.

J.34980
(Anthozoan) store: xx.t.c
Strickland Coll.
bucklandi Zone, Lower Lias; Chadbury, Evesham (near), Worcestershire.
Fig'd, [holotype], Duncan, 1867, Brit. Foss. Corals Second Ser., Mon. Pal. Soc., Part IV, p. 54 pl. xiii figs. 1–4, as *Isastraea stricklandi* Duncan;
Listed, Woods, 1891, Cat. Type Foss. Cambridge, p. 20, as *Isastraea stricklandi* Duncan.

J.34981
(Anthozoan) store: xx.t.c
Strickland Coll.
bucklandi Zone, Lower Lias; Chadbury, Evesham (near), Worcestershire.
Described, [paratype], Duncan, 1867, Foss. Corals Second Ser., Mon. Pal. Soc., Part IV, p. 54, as *Isastraea stricklandi* Duncan.

J.37177–37180
(Anthozoan) store: xx.t.61
Whidborne Coll.
Lower Lias; Bream Down, [Gloucestershire].
Montlivaltia haimei Chapuis and Dewalque.

J.37181
(Anthozoan) store: xx.t.61
Lower Lias; Lincolnshire.
Montlivaltia haimei Chapuis and Dewalque.

J.37182
(Anthozoan) store: xx.t.61
Walker, J.F. Coll.
Lower Lias; Fenny Compton, Warwickshire.
Montlivaltia haimei Chapuis and Dewalque.

J.37183
(Anthozoan) store: xx.t.61
Lower Lias; Fenny Compton, Warwickshire.
Montlivaltia polymorpha Terquem and Piette; xxviii 11 2.
Montlivaltia mucronata Davidson;
Oppelismilia mucronata (Duncan).

FIG. 1

'Sxx.t.'
'Sxx.t.c'
'Tlepidophyllia'
'Ustricklandi'

Note that the correspondence between the fields of the original GOS record and the terms generated is not entirely simple. Thus the term 'Aduncan1868' is derived from an *a and a *d field which come within the same *ref group, while the storage location, which is in a single GOS field, gives rise to three index terms: 'Sxx.' indexes all specimens in bay xx of the museum, 'Sxx.t.' indexes all specimens in cabinet xx.t, and 'Sxx.t.c' indexes all specimens in drawer xx.t.c. Thus it is possible for the IR system to search for specimens in particular bays or cabinets or drawers in the museum. Again in the *f field, the fact that the specimen is a holotype and has been figured is important, and so the terms 'Ffigd' and 'Fholotype' have been generated, while the less important *f field containing 'catalogued' has been ignored. Finally the *tax field has given rise to two terms in separate categories 'T' and 'U'. The 'T' term is the genus part of the taxonomic name, and the 'U' term the species part.

Note also the term 'I'. This indexes every specimen in the system.

4. THE PROBLEM OF VARIANT FORMS OF THE SAME WORD

To make the IR system truly effective it is essential to find some solution to the problem of words in the catalogue having slightly variant forms throughout. To some extent the 'normalization' process of stripping out all but the essential characters and forcing lower case during the generation of index terms helps here, but does not solve the problem. As an example, one of the donors to the SM is R. Wright Barker, whose name is liable to appear in the manual catalogue in various forms, e.g.,

R. Wright Barker
Wright Barker
R. W. Barker

This name was variously transcribed into the computer, depending upon whether the data preparation assistant judged 'Wright' to be a middle name or part of the surname:

*pers Barker, R. Wright
*pers Wright Barker, R.
etc.

Any of these forms is likely to be followed by 'Coll.' (short for 'Collection'), so that attempted retrieval by a single term, e.g.,

'Dbarker,r.w.'

will miss all those specimens indexed by the other related terms. This is got over by using a technique which is well known now, although not in wide practical use (Adamson and Boreham, 1974; *see* Freund and Willett, 1982, for a full discussion). The terms are divided into fragments of a fixed size. Three character fragments—

trigrams—are popular, but the SM fragment index uses tetragrams—four character fragments. Thus the tetragrams of the term 'Dbarker,r.w.' (forgetting the initial letter) are:

```
-bark-    -arke-    -rker-    -ker,-
-er,r-    -r,r.-    -,r.w-    -r.w.-
```

The IR system maintains a dictionary of these tetragrams. When a tetragram is looked up in the tetragram index it will give back a list of the terms in which it occurs. In order to find the terms that most resemble a given character sequence S, S is split up into tetragrams, and the corresponding lists of terms are looked up in the tetragram index. These lists are then read as input streams and merged together to extract the K terms that most resemble the string S. K is some number chosen by the user of the IR system. The merging process organizes the terms so that they can be presented to the user in decreasing order of similarity with the string S. This technique is fast, and the result can be presented to the user almost instantaneously. (An approach based upon a linear search through the terms is not really practicable, since the total number of terms is actually in excess of 60 000.)

It is possible to restrict the search to the category of donors by prefixing the string S with the letter D. Thus with

```
      S set to 'Dbarker,r.wright'
and   K set to 15
```

the term similarity test described above (TS test for short) gives the following

```
Dbarker,r.wright:
        2848 = Dbarker,r.wrightcoll.        (21)
        5513 = Dwrightbarker,r.             (284)
        5514 = Dwrightbarker,r.coll.        (38)
        2847 = Dbarker,r.w.                 (60)
        2850 = Dbarkerr.a.wright            (2)
        5512 = Dwrightbarker,h.coll.        (1)
        2843 = Dbarker,j.                   (4)
        2844 = Dbarker,j.m.                 (13)
        2845 = Dbarker,jessie               (1)
        2846 = Dbarker,missm.               (4)
        2849 = Dbarker,t.w.                 (1)
        3040 = Dbrycem,wright               (5)
        2826 = Dbaker,r.                    (49)
        2885 = Dbecker,r.b.                 (37)
        3097 = Dbutler,r.w.                 (155)
```

The numbers on the left are term numbers, and any term number can be used as a shorthand form of its corresponding term in an IR request. The bracketed numbers on the right give the number of specimens indexed by each of the terms. It can be seen after a moment's scrutiny that the variant forms are given by the first six terms in the list together with term 2849.

The TS test is useful in a variety of ways. For example if S is set to be 'Ftype' and

K is set to some suitably high value (500 is adequate) we can get a summary of all uses of type terms (holotype, paratype, neolectotype . . .) together with frequency counts. Thus we can discover at once that there are 2225 holotype specimens in the SM. TS tests can also be used to spot important misspellings as a prelude to data editing.

5. THE IR QUERY LANGUAGE

The IR query language enables terms to be combined together according to the rules of Boolean logic to form complex search expressions. The approach adopted here is in fact entirely traditional, and seems to suit the nature of the index vocabulary quite well. The areas where it seems inappropriate will be discussed later. An IR query can go through three successive phases. In the first, the number of specimens which satisfy the query is printed out at the computer terminal. Then the simple query

```
'Ginsect'
```

produces the response

```
306 specimens found
```

which tells the user that the SM houses 306 specimens of fossil insect. In the second phase a command causes the identity numbers of these specimens to be displayed at the computer terminal. The response is then:

```
C.6652
C.13755-13779
C.15181-15189
C.76801-76994
D.21639-21679
E.10995
E.16907-16909
E.16937-16939
F.11460-11472
J.4729
J.34428-34430
J.34435
J.34441-34444
J.58307-58311
J.58689
J.58691
```

The records for these specimens can be inspected on the microfiche version of the catalogue. In the third phase, a further command will cause a computer file to be created containing the records for the specimens. What is then done with this file will depend upon why the IR query was issued. Typically the file could be printed out (without storage locations—a simple precautionary measure) and mailed to a researcher who has written to ask what fossil insects there are in the museum. In fact the third phase of retrieval can be done independently of the second phase. The first phase of retrieval can be thought of as a process which takes an IR query as input and produces a file of identity numbers as output. The second phase simply displays this file.

The Boolean logic provided is quite general. So for example the term 'Ginsect' can be combined with the term 'Sxvi.', which indexes the 10022 specimens in bay xvi of the SM, in the following ways:

Expression	meaning	no. of specimens
'Ginsect'	all insects	306
'Sxvi.'	all specimens in bay xvi	10022
'Ginsect' & 'Sxvi.'	insects in bay xvi	13
'Ginsect' – 'Sxvi.'	insects not in bay xvi	293
'Sxvi.' – 'Ginsect'	non-insects in bay xvi	10009
'Ginsect' ¦ 'Sxvi.'	specimens which are either insects or are in bay xvi	10315

(The symbols '&', '¦' and ' – ' provide logical AND, OR and SUBTRACTION respectively.) And with the help of round brackets, term expressions can be constructed of arbitrary complexity, e.g.,

$$((t \mathbin{\&} u) - v) \mid (w - (x \mathbin{\&} y \mathbin{\&} z))$$

where t, u, v . . . stand for index terms.

The query language has a few special features which are of interest. Any term can be replaced in an expression by an explicit list of identity numbers written between square brackets. (It should be explained that all SM identity numbers have the form letter–number, e.g., B.8927.) This has proved useful for administrative purposes. Thus to retrieve the records A.1000 to A.1999 in a file one begins with the IR query

[A.1000–1999]

and then goes through phases one and three of the retrieval process. Similarly, the name of a file of identity numbers created in phase one of the retrieval process can be used in place of a term in a subsequent query to stand for an explicit range of identity numbers.

A useful feature of the IR system is that the files of identity numbers which it handles are stored in compact ranges. Thus the identities A.1, A.2 . . . A.1000 would be stored as the range A.1–1000 rather than as 1000 separate identity numbers. This means that a file of identity numbers representing the entire collection is small enough to be easily manageable within the IR system, and that one can therefore make good use of the term 'I', which it will be recalled indexes every specimen in the SM. Thus the IR query consisting simply of

'I'

produces the response

448123 specimens found

—the number of specimens currently catalogued in the museum (clearly a very useful piece of information). An IR query of the form

('I' – [A.1–1 000 000]) & expression

would restrict the search of the given expression to all specimens except those in the range A.1 to A.1 000 000—the so-called A-section of the catalogue.

6. AN EXAMPLE OF AN IR REQUEST

We will now give an example of how the IR system is used in practice. This particular enquiry was received late in 1981 and the IR system was used to answer it early in 1982. It was to ask whether the SM had any corals from British localities of Liassic, Callovian and Kimmeridgian age. To take these criteria in turn:

1. Every coral is indexed by the term 'Ganthozoan', and careful editing has guaranteed that there are no variant spellings of this term, so finding the corals is quite easy.
2. The term 'Britain' is virtually never used in the SM catalogue, British localities usually being defined in terms of a county or large town. But it so happens that the lettered sections into which the SM catalogue is divided correspond to broad stratigraphic and geographic divisions, and that the relevant specimens should be in three sections, namely B, F and J. Of these B and J contain British, and F foreign material. By excluding the F-section from the search we can trap the purely British material.
3. The age terms Lias, Callovian and Kimmeridgian are used extensively throughout the catalogue, and should cover, or nearly cover, specimens from those ages. The TS test is necessary however, to discover the variant spellings and usages of these terms.

If then the geological ages are represented by a series of terms q1, q2, . . . the IR query will have the following shape:

(q1 ¦ q2 ¦ . . .) & 'Ganthozoan' – [F-section]

So first we do the TS test on the strings 'lias', 'callovian' and 'kimmeridgian'. For the string 'lias' no variant forms are found. For 'callovian' we get:

```
callovian:
*24207 = Qcallovian          (482)
*24208 = Qcallovien          (467)
 48282 = Ucallovience        (7)
 48283 = Ucalloviense        (63)
 48284 = Ucalloviensis       (40)
 24483 = Qludlovian          (133)
 59830 = Uswallovi           (2)
   381 = Acallomon1955       (13)
   382 = Acallomon1960       (5)
  3113 = Dcalloman           (6)
```

Terms judged relevant for the subsequent retrieval are marked with an asterisk. Similarly the TS test for 'kimmeridgian' gives:

kimmeridgian:

*24429 = Qkimmeridgian	(541)
*24430 = Qkimmeridgien	(12)
*26996 = Rkimmeridgien	(1)
*24427 = Qkimeridgian	(1313)
*25094 = Rbasalkimmeridgephospha	(48)
*26994 = Rkimmeridgeclay	(909)
*26995 = Rkimmeridgegrits	(1)
*24428 = Qkimeridgien	(22)
*26992 = Rkimeridgien	(1)
40338 = Tetheridgia	(8)
14609 = Lkimeridge	(2)
14610 = Lkimeridgebay	(57)
*24426 = Qkimeridge	(9)
*26989 = Rkimeridgebeds	(3)
*26990 = Rkimeridgeclay	(2888)
*26991 = Rkimeridgelimestone	(6)
38266 = Tbembridgia	(54)
50777 = Uetheridgi	(1)
50778 = Uetheridgii	(23)
56016 = Upartridgiae	(17)

Note that we are using rock names involving the word ‘Kimmeridge’ as indicators of Kimmeridgian age. The IR request can now be formulated as follows:

```
                    (24207 ¦ 24208 ¦
                     24429 ¦ 24430 ¦ 26996 ¦ 24427 ¦ 25094 ¦ 26994 ¦
                     26995 ¦ 24428 ¦ 26992 ¦ 24426 ¦ 26989 ¦ 26990 ¦
                     26991 ¦
‘Qlias’) & ‘Ganthozoan’ – [F.1–1000000]
```

90 specimens are then retrieved, their identity numbers being given as follows:

B.20230–20231
J.11576
J.34980–34981
J.37177–37223
J.37231–37254
J.45982–45985
J.45994
J.46234–46235
J.48516
J.49741–49742
J.49977
J.58890–58891
J.68700

The printout for these specimens which was sent to the enquirer was about five pages long. The first page is shown in Figure 1.

7. RESOURCES NEEDED BY THE IR SYSTEM

The SM catalogue itself occupies about 100 megabytes. The indexes—the term index and the tetragram index—occupy about 10 megabytes. Currently on the IBM 3081 computer of the University of Cambridge Computing Service the indexes are held on the public filing system, while the catalogue, as a direct access file, is held on a 200 megabyte private disk pack. This means that submitting an IR query and getting back a file of identity numbers (phases one and two of the IR system) are effectively instantaneous, or at any rate no slower than the delays one normally encounters in online access to a large computer. Getting back a file containing the records themselves (phase three) is slower, since one has to wait for the private disk pack to be mounted by the operators. The delay time could be up to an hour. Accessing the indexes online while consulting the catalogue on microfiche is therefore an economic and convenient way of using the Cambridge resources for our IR work. In fact as far as the Computing Service is concerned our requirements are quite small. Equally the same system (but excluding the third phase) could be set up on an in-house mini or large micro, equipped with GOS and the IR software, a disk of suitable size (40 megabytes or so), a printer, and some means of loading the disk from an external source—a magnetic tape reader say.

8. OTHER USES OF THE IR SYSTEM

Much of the strength of the IR system comes from the fact that it sets up the catalogue as a direct-access file. This means that the cost of extracting N records from the file is proportional to N and is small when $N=1$. With a sequential file held on magnetic tape the cost is approximately constant, since the entire file must be read through in order to extract the records. To read through the entire SM catalogue, for example, takes over one minute of CPU time, even if no other operations are done. To pull out a subfile using the IR system takes a fraction of a second. Most activities carried out on the catalogue involve therefore the use of the IR system to access the records, and GOS to operate on the retrieved subfile. In fact practically every cataloguing operation involves some use of the IR system.

To take the case of record editing. All editing is currently done using the GOS processor COMBINE. COMBINE takes two GOS files, a base file and an edit file, and, as its name suggests, combines them together to form a new one. The edit file is prepared in exactly the same way as an ordinary file of new records, except that the field indicators (*re, *tax *pers etc.) may be followed by markers which say whether they are to precede, follow, replace etc. the adjacent fields in the corresponding records of the base file. For many simple editing purposes these markers may be completely omitted. COMBINE is expensive to run on the whole SM catalogue however, since it needs the one minute CPU time to read the catalogue through. The process actually adopted is therefore as follows:

1. Using the IR system, a subfile of the catalogue is formed which contains just those records which the edit file is attempting to alter.
2. COMBINE is run on this subfile (as base file) and the edit file.
3. The result is printed out for checking.

Edits usually require correction, so this process will be repeated until the edits are perfect. The edits are not actually run to form a new version of the catalogue, but are in fact archived in anticipation of a major update. The frequency of major updates will depend upon cataloguing activity in the museum, but we may suppose that it is once every six months or so. Once every six months then, the corrected edit files are themselves combined together to form a single joint edit file which is then combined with the complete catalogue to form a new edition of the catalogue. (In fact combining the edit files together correctly is an operation which GOS is unable to do! We have a separate program for this activity.)

This process applies to all editing procedures, whether it is the correction of spelling errors or the adding in of new identifications of specimens in *re groups.

Again to take the case of loans. If Dr X wishes to borrow specimens A.14077–14080 and B.6481, the records are retrieved via the retrieval request

```
[A.14077-14080 B.6481]
```

and the resulting records are printed out in three ways:

1. As a piece of catalogue. This is kept in the museum for reference.
2. As a piece of catalogue but with storage locations removed. This is sent to Dr X with the specimens.
3. As an array of small records ready to be guillotined up as slips for the specimen trays from which the specimens will be removed. The slips contain the identity number and storage location of each specimen and the name of the borrower together with the date.

When the specimens come back to the museum, printout (1) may be referred to to obtain the storage locations to which the specimens are to be returned.

The case of drawer movements can be dealt with quite simply. If all specimens in drawer *x* are to be moved to drawer *y*, we use the IR system to extract the identity numbers of all specimens in drawer *x*, and from these identity numbers create an edit file which replaces the storage location by *y* in all records with the given identity numbers. The edit file can then be archived ready for the six-monthly update.

Generating the large taxonomic index also uses the IR system. The *bcat field of each record contains a broad categorization of the specimen, and care has been taken to ensure that this field is always present in a record, and contains a word from a short fixed vocabulary. If the TS test is used with the string 'G. . . .' (the four full stops effectively define a dummy tetragram) and K set to a high value, the complete vocabulary can be seen. It begins

```
G. . . . :
    5578 = Gacritarch          (39)
    5580 = Galga               (225)
    5581 = Gammonoid           (19123)
    5582 = Gamphibian          (59)
    5583 = Gamphineuran        (100)
    5584 = Ganthozoan          (13055)
    5585 = Garachnid           (295)
    5586 = Garchaeocyathid     (14)
    5587 = Gartefact           (812)
    5588 = Garthropod          (37)
    . . . .
```

Generating the taxonomic index is too large an operation to be done in a single job. We therefore retrieve each group of specimens in turn, Acritarchs, Algae, Ammonoids and so on, and generate an index for each group. Each group gives rise to a separate index file, which is printed on its own set of microfiche.

9. LEARNING FROM THE IR SYSTEM

Experience in using the IR system with the SM data has caused us to modify our ideas about what the data should be. From this point of view the whole exercise has been quite educational. In this section I would like to make some comments about the SM data in the light of this experience. It must of course be remembered that the original SM catalogue was formed without any expectation that it would ever be computerized, and subsequently computerized without any expectation that it would be accessed by an IR system like the one described above. Consequently it is hardly surprising that one notices shortcomings in the data.

9.1 Locality data

In the manual catalogue localities are almost invariably defined not by latitude/longitude or the national grid reference, which are specified only rarely, but by a description involving place names, for example in the case of specimen A.95193:

> "Old quarry 250m N. of Overbury Farm, Woolhope Inlier, Herefordshire."

These descriptions are cumbersome, they lead to the same place being described in a variety of different ways, and they utilize place names which are liable to become obsolete, and which do not necessarily overlap the point on the earth's surface which they are trying to define. Furthermore unless one is familiar with local geography they are often ambiguous. (Is the old quarry and/or Overbury Farm in Woolhope Inlier?) As far as computer aided retrieval is concerned, the grid reference is much more usable:

NGR SO 61113669

This is accurate to 10 metres, and the specimen from this locality could be indexed automatically by terms which give the various grid squares containing it:

'Lso6111 3669'	(10 metre square)
'Lso611366'	(100 metre square)
'Lso6136'	(1 km square)
'Lso63'	(10 km square)
'Lso'	(100 km square)

(Retrieval of localities by geographic coordinates is used in the bibliographic database GeoRef, available on DIALOG. Two search strategies are available, both different from the one suggested here. *See* Farrar and Lerud, 1982.)

In this scheme retrieving on localities can be done by specifying the grid squares which define the areas of interest. A similar, though slightly more clumsy, approach could be used with latitude/longitudes. A realistic approach for the SM data would be to index British material by grid reference, and foreign by latitude longitude (the collection is predominantly of British material).

9.2 Stratigraphic data

Each specimen in the SM comes from a particular rock, and has an associated age. The names of rocks can be organized as a hierarchy. Just as Ely is in Cambridgeshire, so the Farley Member is in the Much Wenlock Limestone Formation. In practice however the hierarchical aspect is not too important, since retrieval is most often by formation name (Oxford Clay, Wenlock Limestone . . .), and a formation name will usually be included in the catalogue record, if known.

Age names give rather more problems, however. The age of a specimen cannot be given in terms of the number of million years BC, since the beginning and end points of the geological periods known as Jurassic, Cretaceous and so on are subject to continuous enquiry. It is therefore not possible to translate age names into segments on a one-dimensional grid in the same way that place names can be translated into boxes in a two-dimensional grid. On the other hand the common age terms can be placed in a fairly simple two-level hierarchy, in which the upper level is a period term, and the lower level an age or epoch term. So part of the structure would be:

```
Ordovician (period) . . .  Jurassic  (period)
        |                      |
   Ashgill  (epoch)        Oxfordian (age)
```

One would then insist that each record in the system should carry at least one age term out of this structure, and that if it is a term from the lower level, the corresponding term from the higher level should be made to index the record automatically. Other, more specific, age terms could of course be included in the cataloguing. It has been found in practice that it is these higher level age terms that are required for retrieval purposes, but that they are often omitted from records of the catalogue when the name of the rock from which the specimen came was felt to imply a particular age designation.

Simple age and rock qualifiers such as 'upper', 'lower', 'middle' (as in 'Upper Jurassic', 'Lower Devonian') are used throughout the catalogue, but very informally, and they have been ignored in indexing the IR system. Occasionally however they are significant, as in 'Lower Greensand'. Ideally they should be excluded from cataloguing except when they form an integral part of the age or rock name.

9.3 Taxonomic data

Almost all the specimens in the SM are identified by a taxonomic name. A taxonomic name, of course, defines a species, and it is possible to organize the set of species names into a hierarchy, with a suitable name at each level, for example:

Trilobita
|
Agnostida
|
Agnostina
|
Sphaeragnostidae
|
Sphaeragnostus similaris (Barrande)

In setting up an IR system, I believe that all but the simplest use of such a hierarchy is best avoided. In its fullest elaboration the hierarchy of such names is complex and has no generally agreed form. At any given time different researchers will draw it up in different ways, and as palaeontology progresses it undergoes change. In any case there are unavoidable complications with the taxonomic names at the lowest level in the hierarchy. Each taxonomic name contains (among other things) a genus and a species part, with the genus acting as the lowest level grouping of the species. So in the example, the species *similaris* belongs to the genus *Sphaeragnostus.* But as palaeontological knowledge grows, the way in which researchers choose to classify species into genera changes, and so a given species name is liable to be linked historically with a number of different genus names. Of course this would not matter if one could simply use the species name as a basis for retrieval, but unfortunately the species names are not all unique, and must often be preceded by a genus name in order to become unambiguous.

In the SM catalogue the terms in the *bcat field, which give a broad categorization for the specimens, effectively supply a level of hierarchical subdivision above the species level. The terms in use have a certain arbitrariness, and reflect differences of interest among researchers as much as natural divisions in the animal kingdom. But as terms for retrieval they are in fact very useful.

In the IR system, genus and species components of the taxonomic names give rise to different index terms. Thus there will be terms 'Tsphaeragnostus' and 'Usimilaris' corresponding to the name *Sphaeragnostus similaris* (Barrande). In retrieving a particular species with species name 'Ux', one can either use the request

'Ux'

by itself, or use

'Ux' & 'Gy'

where y is the broad category term under which species x is subsumed, or use

'Ux' & ('Tz1' ¦ 'Tz2' ¦ . . .)

where $z1$, $z2$ . . . are the genus names which have from time to time been used in conjunction with x. The approach adopted will depend upon the degree of ambiguity caused by use of the name x by itself. Of course purely taxonomic enquiries need not involve the IR system, since they can be referred to the taxonomic index to the specimens on microfiche.

9.4 Other descriptive data

In addition to the taxonomic information for each specimen (in the *bcat and *tax fields) the SM catalogue also contains a certain amount of other descriptive information about the specimens. When the catalogue was put into machine-readable form, these data tended to go into a keyword (*keyw) field, where the significant terms were extracted as keywords, and the rest called 'details', so for example 'upper right molar' might have been transcribed as

*keyw molar < upper right

'upper right' being a detail. The rules for extracting the keywords were however never formulated, and elsewhere the same data might be transcribed as

*keyw right molar < upper
or *keyw upper molar < right
or *keyw upper right molar

Unfortunately there were a number of other fields into which descriptive data might be placed, and the result is that data of this kind do exist in the records, but are not held in a sufficiently consistent form to make them really useful in retrieval.

It seems now that it would have been natural to break down the descriptive information as follows:

1. Information about the specimen that goes beyond what is contained in the taxonomic name, for example, whether it is male or female, young or old (the *td field).
2. Information about which part of the animal or plant has been preserved, e.g., 'mandible', 'upper right molar', 'leaf', 'thorax', 'pygidium' (the *keyw field).
3. Information about the mode of preservation of the specimen, e.g., 'cast', 'internal mould', 'external mould' (the *pres field).
4. Information about the surrounding matrix of the specimen, e.g., 'black shale' (the *lith field).

Further, it is now felt that keyword/detail analysis on these different categories of information should have been avoided. If, for example, teeth are described by a simple consistent vocabulary, then such terms as 'Kupper', 'Kright' (K = keyword) are as useful as 'Kmolar' in such combinations as

'Kupper' & 'Kmolar'
or 'Klower' & 'Kleft' & 'Kmolar'
or 'Kmolar' | 'Kincisor'

The strategy actually adopted for generating terms from the existing *keyw field is to generate a separate term for each word in the field, so that the descriptive phrase

*keyw upper left molar

gives rise to three terms:

'Kupper'
'Kleft'
'Kmolar'

This works reasonably well, but with a retrieval system based on Boolean logic it would work better if the phrases had been constructed in a more controlled manner. Fortunately use of these K-terms is not critical in retrieval. The situation would no doubt be different with other kinds of museum collection, say an archaeological, social history or decorative arts collection. Here a set of keywords derived from a phrasal description of each object could be very important, and possibly the most suitable retrieval system would be a probabilistic one. In this case the IR system could be made to operate at two levels: a 'Boolean' level, where the user presents the system with a Boolean expression of terms, the purpose of which is to restrict the view of the user to the subset of the catalogue comprising records which satisfy the Boolean expression, and then a 'probabilistic' level, in which the search and subsequent ranking of records take place on that subset.

10. CONCLUSIONS

Museum records are typically complex—the examples in this paper should have made that very clear—and in trying to set up an IR system for a museum catalogue a seemingly endless number of very general IR problems keep cropping up. Nevertheless it is possible to construct a very practical IR system for a museum catalogue which takes account of these inherent complexities without itself being complex or especially difficult to use. In fact teaching the IR system has proved to be quite easy, compared with teaching other aspects of the computerized cataloguing system.

Finally it is again worth emphasizing that despite the size of the catalogue, the resources required to run the IR system are surprisingly modest, and are hardly noticed by the Cambridge University Computing Service.

REFERENCES

Adamson, G. W. and Boreham, J. (1974) The use of an association measure based on character structure to identify semantically related pairs of words and document titles. *Information Storage and Retrieval 10*, 253–260.

Farrar, R. K. and Lerud, J. V. (1982) Online searching using geographic coordinates. *Second International Conference on Geological Information.* (C. M. Kidd, ed.) pp. 109–116. Oklahoma Geological Survey (Special Publication 82–4).

Freund, G. E. and Willett, P. (1982) Online identification of word variants and arbitrary truncation searching using a string similarity measure. *Information Technology: Research and Development 1*, 177–187.

Porter, M. F., Light, R. B. and Roberts, D. A. (1977) *A unified approach to the computerization of museum catalogues.* London: British Library (BLRD Report No. 5338HC).

Porter, M. F. (1982) GOS: A package for making catalogues. *Information Technology: Research and Development 1*, 113–129.