INFORMATION RETRIEVAL

AND THE INTRODUCTION OF 'SEABASE'

BY

T. J. MARRIOTT, B.A.
(*Librarian, Sea Systems Controllerate*)

One of the most fundamental features which distinguishes *homo sapiens* from all other animals is the ability to learn from cultural experience; that is to say that an individual can learn not only from his own experience but also from the accumulated experiences of other men and women in his own and other societies. In the most primitive societies all such experiences and knowledge were passed on by word of mouth, but for the last 10 000 years or so knowledge has also been recorded in some form of written document. In whatever way knowledge is recorded, it poses the problem of identifying and (at a later time) accessing information relevant to a particular situation; the processes and techniques involved in this activity are those of information retrieval.

Information retrieval therefore embraces everything from the workings of human memory to the intricacies of database management systems. The scope of this article is limited to a brief sketch of the development of information retrieval techniques as applied to the identification of documents, and the exploitation of such techniques in the management of technical documents within the Sea Systems Controllerate.

Purposes and Limitations

The objectives of bibliographic information retrieval are to allow and facilitate the twofold exploitation of information embodied in documents such as technical reports, journal and conference papers, textbooks, etc. Firstly, it should be possible to identify, at a later date, a document whose details are only imperfectly known to the enquirer: for example, it may be known that a test report was written on a particular item of equipment, but neither the report number nor title are given; or one document may be referenced in another by its report number only, and the enquirer may need clarification of the nature and scope of the referenced document.

Secondly, information retrieval techniques should facilitate the exploitation of information for purposes other than those for which that information was originally gathered. This is a little more difficult to explain, but taking again the example of a test report, at the time of writing this may have been envisaged as simply a routine evaluation of a piece of equipment: at some later date, it might be beneficial for someone else to utilize part of the information contained in that test report for a comparative study of one particular aspect of equipment performance.

It has been estimated by some researchers that such secondary use of information contained in technical reports may in some instances be worth tens of thousands of pounds; an information retrieval system that is limited to simple primary retrieval of already half-known documents is therefore achieving only a fraction of the true potential of the information that it encompasses.

The techniques of information retrieval were for a long time those of the card index and printed index. These are excellent tools for simple requirements but suffer from a number of major limitations. The first is that unless some machine assistance is used, the number of index entries that can be made for each document is limited by the clerical effort available for filing index cards or collating a printed index: there is therefore a need to reach an economic trade-off between the costs involved in compiling the index and the benefits from extra index entries.

The second major limitation is that of flexibility in the use of such indexes. If the enquiry is of a simple, single-concept nature that can be satisfied by a single index term, then a card or printed index may quite adequately give access to a list of relevant documents, though the list may be daunting in its length. However, if the enquiry is more complex, involving two or more concepts, the card or printed index may be cumbersome to use, involving much comparison of entries back and forth, and scribbling of lists whose meaning soon becomes obscure to the enquirers themselves.

Computerization

There were many techniques developed earlier this century which sought to overcome the second of the limitations mentioned above: devices such as optical coincidence cards, edge notched cards and punched card systems allowed the enquirer to more easily combine lists of index entries for single terms into a compound enquiry. But such techniques were always awkward to handle, resulting all too often in a pile of cards on the floor and a frustrated enquirer out of the door.

Information retrieval is therefore one area where the development of cheap computing power has been of great benefit. It was not one of the early growth areas in computing, as it typically requires relatively simple computing operations on large stores of data; but, as the costs of storage devices decreased in relative terms, the benefits from computerized information retrieval soon came to outweigh the costs.

What then are the major benefits from the computerization of information retrieval, as compared to the tried and trusted manual method?

The first is that the index is compiled by the computer, without clerical filing effort: therefore a larger number of index entries per document can be allowed. There are overheads in terms of the size of index files and access times for large files, but in practical terms the computer does not place a limit on the number of useful index terms that can be stored.

Secondly, the software used for accessing the index files allows for flexible combination of index terms. In most information retrieval software, at least the following operations are permitted:

- A search for alternative forms of the same term by entering a word stem terminated by a special symbol; thus SHIP* may be used to search for any term beginning with SHIP and so would include SHIPS, SHIPPING, SHIPBUILDING etc
- Combination of search terms using Boolean logic; thus alternative terms for the same concept can be entered using the 'or' operator (,) so that for example

  SHIPS, BOATS, VESSELS

  might search for entries having any of those words.

  The Boolean 'and' operator (+) can be used to specify a list of terms which must all be present for the enquiry to be satisfied; thus for example

  TURBINE + BLADE + CORROSION

  requires that all three index terms are present.

  (Note—the special symbols used above are from one particular computer system and are not universally applicable.)

Thirdly, the computer allows the enquiry to be conducted interactively, that is the enquiry can be modified according to the results achieved. In a manual index enquiry, it is normally necessary to complete the enquiry before finding that the results are inadequate and that a modified version of the enquiry has to be tried. In a computerized information retrieval system if, for example, too many references are retrieved on the first attempt, then additional index terms can be added immediately; alternatively if too few references are retrieved, the scope of the enquiry can be broadened. A sample of the references retrieved at each stage can be examined to assist in adjusting the enquiry to achieve the optimum result.

Fourthly, what might seem a minor benefit, but one which is appreciated by anyone making frequent use of manual indexes, the end result of a computer search can be a printed list of entries; in manual index searches it is not infrequent that the most time-consuming part of the operation is the copying of entries retrieved so that the documents can be followed up at a later time.

Computerized information retrieval is obviously not without considerable costs. The processes of data collection and data entry may involve a large amount of staff effort; information retrieval applications typically require large amounts of disc storage and, although the software does not have the same level of complexity as sophisticated scientific and technical applications, it does normally impose a large loading on the computer system.

The development of such retrieval systems therefore shows a pattern of growth from relatively few centralized applications, to a more recent blossoming of specialized and tailored local systems. Most of the early systems were dedicated mainframe services allowing access to indexes of reports, papers, etc., mainly in the aerospace industries: these mainframe services were (and still are) available to subscribers equipped with suitable terminals and modems; by spreading the costs over a large number of users, large files of information could be made available at economic rates.

Such services have multiplied and diversified over the last ten years or so, and many files of information of relevance to marine engineering are now accessible. These files index only documents in the public domain but those available include BMT Abstracts (until recently the *British Ship Research Association Abstracts*), Compendex (based on the printed *Engineering Index*), Metadex (based on *Metals Abstracts*), Weldasearch (compiled by the Welding Institute) and Fluidex (compiled by BHRA and covering hydraulics, pumps, turbines, etc.).

In the defence field, such services are obviously not publicly accessible. DRIC (the Defence Research Information Centre), which acts as a central repository and source for defence related technical reports, has for some years now used computer techniques to produce Defence Research Abstracts, and as part of its operations maintains its own computerized information retrieval system. This is not directly available on computer terminals outside DRIC, but searches can be performed for suitable applicants.

Information Retrieval in Sea Systems Controllerate

The Sea Systems Controllerate does not have a long history of use of computerized information retrieval techniques, but is now rapidly building up services to assist in the exploitation of the masses of docments stored in the Controllerate. The most obvious developments are the establishment of databases on the Bureau West mainframe computer services using the STATUS software package, one of the standard commercially available information retrieval packages. In this article only the SEABASE database will be described in any detail, though other databases are also now well established (for example REMUS in the field of specifications).

SEABASE is the Sea Systems Controllerate database of bibliographic information. It is a computerized information retrieval system indexing documents such as technical reports, books, journal articles, conference papers, etc. It must be emphasized, to eliminate any false hopes, that SEABASE is an index only; it does not contain the full text of any of the documents.

Following from what was stated above about the objectives of information retrieval techniques in general, the aims of SEABASE are:

- To enable enquirers to identify a document from incomplete details.
- To identify documents relevant to a specific topic, and thus maximize the benefit from information held in documents and minimize the risk of repeating work already reported elsewhere.

In addition SEABASE has the aim of assisting enquirers to locate physical copies of documents by recording their location.

SEABASE was set up on a co-operative basis between the Sea Systems Library and any technical sections at Foxhill with a requirement for assistance in information retrieval. The Library has a large collection of technical documents, but many reports are also held outside the Library. It was agreed that there were considerable benefits in having a single Sea Systems Controllerate bibliographic database at Bath. In particular, management of data entry, elimination of duplication, and consistency of indexing, database management and training, is all achieved more economically and effectively by having a focal point with a specific responsibility for information retrieval, rather than it being an 'add-on' task for a more or less enthusiastic member of each technical section.

At January 1987, SEABASE contained details of over 14 000 documents, and new entries are being added at between 600 and 1000 per month. Of those 14 000 entries, most were for technical reports (12 000) with the remainder split between books and journal articles and conference papers. Of the technical reports, coverage of different report sources varies greatly at present according to the convenience and priority of input. Thus for example there is complete coverage of the reports from some Admiralty Research Establishments over the last 10 years, but piecemeal coverage of reports from other AREs. A summary of SEABASE coverage at January 1987 is given in TABLE I.

TABLE I—*Summary of coverage at January 1987*

Category	*No. of entries*	*Notes*
Books	1 260	Mainly those held in Sea Systems Library
Journal and conference papers	1 200	Includes *Royal Institution of Naval Architects Transactions* back to 1970; *Institute of Marine Engineers Transactions* back to 1979. Articles from *Society of Naval Architects and Marine Engineers Transactions, Marine Technology, Naval Engineers Journal, Journal of Naval Engineering*, etc., will be added shortly.
Reports: total coverage	12 000	Coverage depends on: ease of availability of entries for data capture & subject interests of technical sections participating in SEABASE
examples of coverage		
ARE Haslar	1 200	
ARE Holton Heath	620	
ARE Dunfermline	1 600	
YARD Ltd.	660	

Of direct relevance to readers of the *Journal of Naval Engineering* is the coverage of journal articles. In selecting input for SEABASE it was necessary to avoid expensive duplication of services already available from the commercial databases mentioned above, such as BMT Abstracts. However, marine engineering has never been a field well endowed with comprehensive indexing services and so it was thought worthwhile including such core journals as the *Transactions of the Institute of Marine Engineers*. When it comes to MOD publications such as the *Journal of Naval Engineering* and the *Journal of Naval Science*, these are of course not included in the open commercial databases, and so it was planned from the outset that articles from such journals would be included in SEABASE.

Coincidentally the editor of the *Journal of Naval Engineering* was reconsidering the costs and benefits of continuing the former 8-yearly cumulative printed index to the *Journal*, and the develpment of SEABASE presented an alternative allowing potentially more flexible indexing of its accumulated wisdom. It is therefore intended that by the middle of 1987 SEABASE will include entries for all articles from the *Journal of Naval Engineering* back to the last printed cumulative index (1979), and that in due course coverage may extend further back if demand seems to justify it. The cumulative printed index will no longer be produced, although there will continue to be one at the end of each 'volume' of three issues. Sea Systems Library will be willing to answer enquiries relating to identification of articles from the

Journal of Naval Engineering for those without access to SEABASE (e.g. from ships, Dockyards and ARE establishments).

It may therefore be useful to specify precisely what sort of information is held in SEABASE and what sort of enquiries can be answered. The following information may be held for each document:

(*a*) Title, as given in the document.

(*b*) Reference number. For technical reports this will be the report number; for journal articles it is a code linked to the journal so that, for example, a search can be easily limited to articles from the *Journal of Naval Engineering* only.

(*c*) Date.

(*d*) Source. For a report this is the organization responsible for producing the report; for an article it is the issue, page number, etc.

(*e*) Authors.

(*f*) Security classification of the original document.

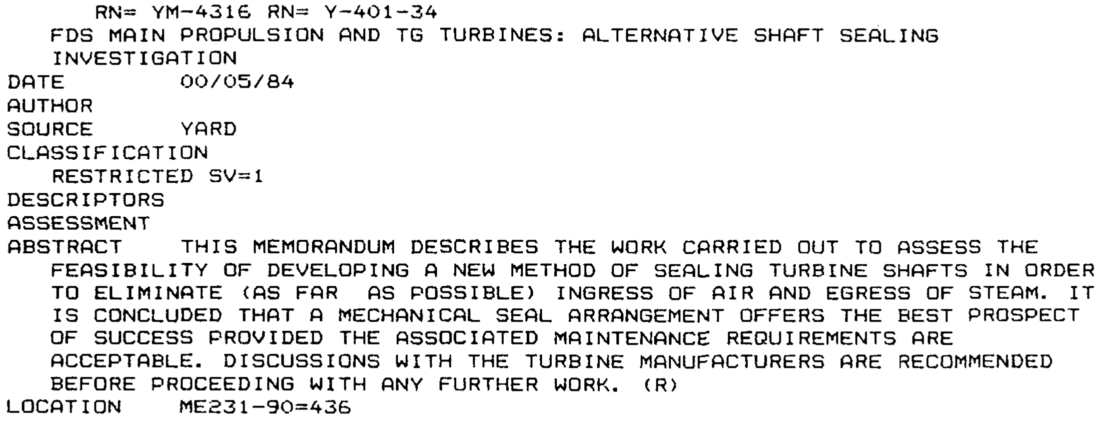
(*g*) Keywords or subject descriptors. These are the index terms describing the major concepts covered by the document.

(*h*) Assessment. Where appropriate, comments can be recorded of the Controllerate view of a document; for example to record that it must only be read in conjunction with another document, or that further work has been commissioned.

(*i*) Abstract or summary, where this is provided with the document.

(*j*) Physical location of copies of the document; normally limited to locations at Foxhill.

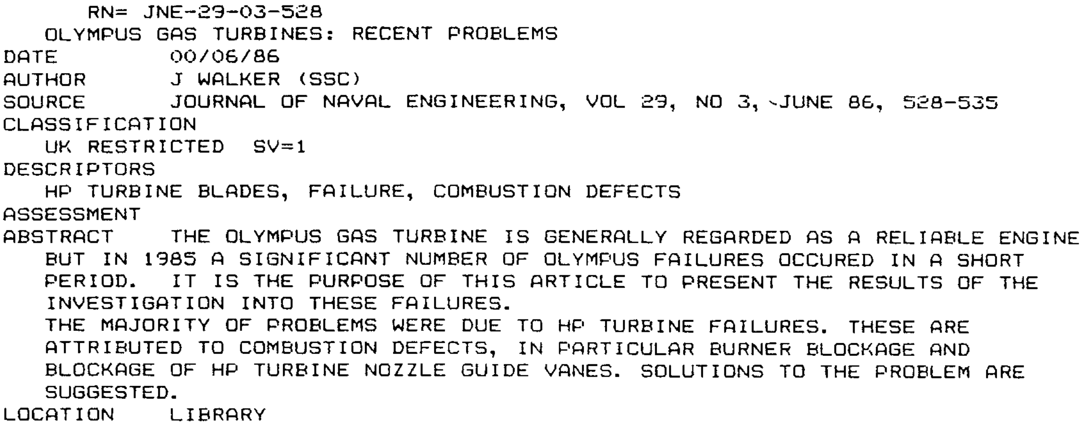
Samples of SEABASE entries are given below at FIG. 1 (an entry for a technical report) and FIG. 2 (an article from the *Journal of Naval Engineering*). When such entries are fed into SEABASE, each and every word of the entry is automatically indexed, together with its context in the entry. Thus it is possible to search for words occurring only in certain fields; for example for words taken from the title, for authors, for keywords. Conversely it is possible to search for a word wherever it occurs; so for an obscure topic the whole of the abstracts can be searched for any mention of a word. Because the context of a word is indexed, it is simple to search for phrases. In accordance with the information retrieval operations outlined above, it is possible to search for word stems, to use Boolean operators, and to modify the search interactively according to the results.



```
        RN= YM-4316 RN= Y-401-34
   FDS MAIN PROPULSION AND TG TURBINES: ALTERNATIVE SHAFT SEALING
   INVESTIGATION
DATE         00/05/84
AUTHOR
SOURCE         YARD
CLASSIFICATION
   RESTRICTED SV=1
DESCRIPTORS
ASSESSMENT
ABSTRACT      THIS MEMORANDUM DESCRIBES THE WORK CARRIED OUT TO ASSESS THE
   FEASIBILITY OF DEVELOPING A NEW METHOD OF SEALING TURBINE SHAFTS IN ORDER
   TO ELIMINATE (AS FAR  AS POSSIBLE) INGRESS OF AIR AND EGRESS OF STEAM. IT
   IS CONCLUDED THAT A MECHANICAL SEAL ARRANGEMENT OFFERS THE BEST PROSPECT
   OF SUCCESS PROVIDED THE ASSOCIATED MAINTENANCE REQUIREMENTS ARE
   ACCEPTABLE. DISCUSSIONS WITH THE TURBINE MANUFACTURERS ARE RECOMMENDED
   BEFORE PROCEEDING WITH ANY FURTHER WORK. (R)
LOCATION      ME231-90=436
```

FIG. 1—SEABASE ENTRY FOR A TECHNICAL REPORT

The establishment of SEABASE on a co-operative basis, whilst offering benefits suggested above, also creates problems when it comes to the retrieval stage. For the majority of the documents indexed, data entry forms are completed by the technical sections holding the reports, and so practice in the selection of keywords varies widely. Although the Library imposes a certain amount of editorial standardization on keywords, the basic philosophy is that the technical sections are best equipped to decide on indexing; unfortunately this assumption does not always match with the staff resources and priorities given to preparation of input forms. However, all is not lost: with a little imagination and ingenuity the enquirer can normally use the power of the retrieval system to overcome inconsistencies of indexing and variations in practice.



```
      RN= JNE-29-03-528
   OLYMPUS GAS TURBINES: RECENT PROBLEMS
DATE         00/06/86
AUTHOR       J WALKER (SSC)
SOURCE       JOURNAL OF NAVAL ENGINEERING, VOL 29, NO 3, JUNE 86, 528-535
CLASSIFICATION
   UK RESTRICTED  SV=1
DESCRIPTORS
   HP TURBINE BLADES, FAILURE, COMBUSTION DEFECTS
ASSESSMENT
ABSTRACT     THE OLYMPUS GAS TURBINE IS GENERALLY REGARDED AS A RELIABLE ENGINE
   BUT IN 1985 A SIGNIFICANT NUMBER OF OLYMPUS FAILURES OCCURED IN A SHORT
   PERIOD.  IT IS THE PURPOSE OF THIS ARTICLE TO PRESENT THE RESULTS OF THE
   INVESTIGATION INTO THESE FAILURES.
   THE MAJORITY OF PROBLEMS WERE DUE TO HP TURBINE FAILURES. THESE ARE
   ATTRIBUTED TO COMBUSTION DEFECTS, IN PARTICULAR BURNER BLOCKAGE AND
   BLOCKAGE OF HP TURBINE NOZZLE GUIDE VANES. SOLUTIONS TO THE PROBLEM ARE
   SUGGESTED.
LOCATION     LIBRARY
```

FIG. 2—SEABASE ENTRY FOR A JOURNAL ARTICLE

It is therefore hoped that SEABASE can be used to increase the use to which information embodied in technical documents is put: the system should help avoid repetition of earlier work and allow the application of information to new problems as they arise. Such benefits are, though, only realized if the system is used, so any comments on the sort of questions that enquirers would like to be able to ask, the sort of documents they think should be included and the level of indexing that it is justified to provide would be welcomed by the author.