[*Translating and the Computer 8*. Proceedings of a conference… 13-14 November 1986, ed. Catriona Picken (London: Aslib, 1987)]

Principles of information retrieval for personal information systems

Paul F. Burton

University of Strathclyde, UK

At some time or another, practically every researcher, student, lecturer or whatever has been faced with the same problem: knowing that an article, a book, a report or some document relevant to a problem exists and is available, but unable to find it in the collection of material which we all accumulate during our professional lives. Information scientists, it seems, are no exceptions to this, but, whatever the occupation, it can be a time-consuming business to search through files (often very large) in the hope that something will strike a chord. The reverse situation, when the researcher is trying to build up a sub-set of his or her documents in order to deal with one particular project, also holds good: there may be uncertainty as to whether everything relevant has been retrieved, or too much may be found, with the consequent task of reducing the information to manageable proportions. Information overload and underload are facts of life, and any methods which can help to optimise information retrieval are of value.

In this presentation, I shall attempt to outline the main techniques for personal information retrieval systems, both manual and computer-based. Whilst I shall be talking primarily about retrieval of bibliographic records, the principles and the techniques are applicable to a wide variety of information, including facts, comments, notes, etc. The term ‘personal information systems’ can cover a wide range of possibilities, and such systems can vary considerably in size, from a file of one hundred or so records to several hundreds: however, most of the techniques I shall describe are capable of handling this range with more or less success. It remains true, however, that the computer-based systems offer the greatest flexibility and retrieval power.

FIRST PRINCIPLES

In order to design an efficient and effective system (manual or computer-based), a careful analysis of the requirements of the system should be carried out first. This may seem an obvious point, but it is astonishing how often researchers, etc., simply start writing out index cards or setting up machine-readable files without first considering precisely what information is to be recorded, how much there is, and (perhaps most important) how it will be used. The result is often a system which consumes a great deal of time and effort but which does not produce the right results or which cannot be relied on, because there is too much uncertainty about what is retrieved.

This analysis need not take a great deal of time, provided that it is done properly. In essence, it involves considering what kinds of information are to be stored in the system and in what form. Bibliographic records obviously have a pre-determined structure (author, title, source, date, keywords, etc.) which makes it a little easier to design a suitable format, but other types of information are less structured and may require more thought in order that a consistent and easily read format can be developed. Consistency and suitability for the specific application are probably more important than conformity with established codes and rules such as those used in library and information systems, although you should be aware that such codes exist and should examine them to see if they are suitable. It is also advisable, for bibliographic records at least, to take into account the established practice of relevant journals, particularly if you plan to contribute work to those journals: if your records are in the same format, it will make life considerably easier.

The quantity of information (i.e. how many records) will have a bearing on the ultimate choice of system, if only because some of the systems described below are better suited, for example, for large numbers of records. It may only be possible to estimate the total size of any file, but in doing so some consideration should be given to the growth rate of the file and thus its likely size in two or three years' time. What works now may no longer be feasible when the number of records has doubled or tripled, though growth rate should obviously take account of discarding from, as well as of adding to, the file. Indeed, depending on the particular file, this may provide the opportunity to develop an archiving or discarding policy which eliminates outdated or obsolete material.

A major design factor will be the way in which the system is used to retrieve information, i.e. how will you approach it? For bibliographic files, this will probably be via authors, titles or subjects, so each will have to be provided, if it is economically feasible. Titles may only be needed when there are no associated authors, and the subject approach will obviate the need for title entries for every document, particularly if titles are nonrepresentative of the subject content. By providing at least author and subject entries, the system will satisfy searches for material by a given author, and will indicate whether a document by a specific author is available and what is available on a given topic or subject. There are, however, other relevant features of a bibliographic record which can be considered as access points, such as date of publication, language used, and possibly format. If these features are required, then the relevant information will have to be provided in the record.

Non-bibliographic information is, as has already been said, less structured and often without an obvious format. A format will, therefore, have to be devised which includes all that is needed. A subject approach is common with such files, so primary access will usually be by subject keyword (though individuals' names may be regarded as subjects in this context). A useful feature of non-bibliographic records is the reference to a source of further information or to the source from which the recorded data were taken, and space should be allowed for this. The material recorded may be simply a free text note or abstract, or a series of figures and facts. Whatever is recorded, consistency in presentation will aid the retrieval and use of such records at a later stage.

At this point it may be worth considering whether non-bibliographic and bibliographic files should be maintained separately. Again, the answer will depend on the use made of the files, but a single file has the advantage of indicating all the available information on a subject in one place, providing a 'one-stop' point of reference.

Some thought should also be given to the physical arrangement of the documents to which the records refer. There are many ways of arranging the documents themselves, depending on their use, format, size and shape, etc., but ideally the reference in the retrieval system should lead you straight to the document, without the need to look up another record in order to determine location. If documents are stored alphabetically by author, the file record will be sufficient by itself, but if the arrangement is, say, by report number (or something similar), then that number will have to be clearly indicated. Similarly, if the reference is located in a library (and not within the office or at home), it is useful to provide the shelf mark or classification number used by that library. You may also like to consider whether to include a facility to show that the document is (temporarily) not in its accustomed place, because it has been lent to someone else or because you are using it for some other purpose, for example.

Consistency was mentioned above as important in designing the records which the system will contain. It is particularly important when choosing the terms through which the records will be retrieved and this applies to individual names as much as to subject headings or keywords. Although it is possible to use free indexing of terms chosen directly from the documents, it is too easy to be inconsistent, and inconsistency will result in failure to retrieve all the relevant records or information, and may also result in the wrong information being selected. From your knowledge of the subject-matter and your use of the material, you should select suitable terms and use them consistently, thus ensuring a controlled vocabulary. The documents themselves will, of course, suggest relevant terms to use: these can be underlined in the text, if necessary, but you should ensure that terms are 'converted' to the form you have decided on.

For personal indexes, single terms, rather than pre-coordinated composite terms, are preferred, since it is less likely that records will be lost in the retrieval process, and it is, in fact, easier to be consistent about single terms. This will not always be true: there are compound terms (such as 'machine translation'), which should not be split up. Some of the problems relating to consistency will include the treatment of synonyms, where, assuming that two terms are exactly synonymous, it will be necessary to select one of them and to refer to it from the other term. (Related to this is the use of trade names for chemical compounds.) You will also have to make decisions on the treatment of homographs, though in a personal collection which is limited in scope, this may not be a major problem: terms will tend not to be ambiguous in such a context. Record the decisions you make about terms used in preference to others, and refer from the latter to the former. You may wish to add terms which are related in some way, to act as reminders about other, possibly useful approaches, and so on. As this develops, it will serve as a guide and ensure that the correct terms are used on each occasion. Such methods of vocabulary control can be developed to suit a particular application, and this is perfectly reasonable, provided that it is consistent. On the other hand, thesauri and subject headings are available for a wide variety of subjects, and it may be convenient to use one of these instead of developing your own. You will have to balance the advantages of a tailor-made system with the time it takes to develop, and the ready availability of a thesaurus with the possibility that it is too sophisticated or wide ranging for your application. Aslib can advise on suitable thesauri for many subjects, and they can be examined in the Aslib Information Resources Centre. Figure 1 shows a typical thesaurus entry, in which the term 'natural gas' has been preferred to 'condensate gas' or 'formation gas', and so these are listed with the symbol 'U[sed] F[or]'. More specific terms are indicated by 'N[arrower] T[erm]', while 'R[elated] T[erms]' points to headings which could be relevant, depending on the document being indexed.

Similar principles should be applied to the form of name used in the system, though this need not be as complicated as subject terms, unless there are a significant number of foreign names which may be transliterated in different ways. The simplest answer is to establish a set of rules about the way in which forenames are entered (initials only or spelled out in full), the presentation of prefixes to names such as 'von', 'de', 'de la', and so on,

Fast reactors (nuclear)
Power reactors (nuclear)
Production reactors (nuclear)
— Thermal reactors
Natural gas 2104
UF Condensate gas
Formation gas
NT Liquefied natural gas
Sour gas
Sweet gas
RT – Crude oil
Enriched gas
— Fossil fuels
Gas caps
— Gases
Gas pipelines
Gas production
Gas reservoirs
Gas storage
Heating fuels
— Helium
Liquefied petroleum gases
Manufactured gas
Natural gas liquids
Oil fields
Oil reservoirs
Reserves
Underground storage
Natural gas liquids 2104
UF Gas condensates

Figure 1. Example of a thesaurus entry

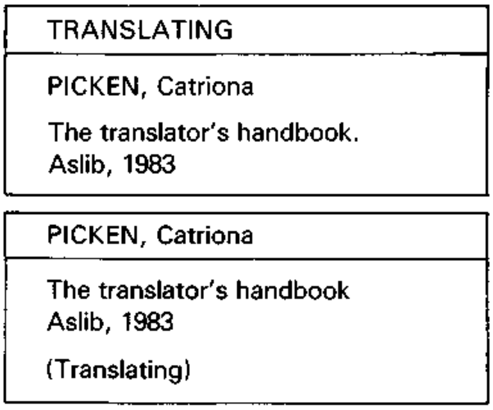
and, if it is relevant, the preferred form of the names of corporate bodies who do, of course, act as authors on occasion! Your choice of format will be determined by your own needs and the documents you handle, but the important thing is to be consistent and to stick to the ‘rules’ you develop. Librarians have had to contend with these problems for centuries, and have developed the *Anglo-American Cataloguing Rules,* 2nd ed. (AACR2) for just this reason. You could look at AACR2 to see if it would be suitable for your purposes, though, like the published thesauri, it may be too sophisticated for all but very large collections.

MANUAL INFORMATION RETRIEVAL SYSTEMS

Manual information retrieval systems have the advantage that they are relatively easy to prepare and use and do not require particularly sophisticated or expensive equipment. However, they can be inflexible in the approaches to the file which are possible (though there are exceptions) and they may be cumbersome to use with very large files. I think it also has to be said that manual files are less 'trendy' and can be regarded as having a poor 'image', facts which may not be altogether irrelevant. However, if the analysis described above is properly carried out, manual systems are perfectly capable of providing information retrieval — indeed, for a very long time, they were the only means of information retrieval!

The simplest manual system is to use ordinary index cards (5 " x 3" or larger), with the entry term written at the top and details of the document entered on the body of the card (see Figure 2). Cards are filed alphabetically by the entry term, and both sides of the card can be used, of course, Alternatively, the non-bibliographic details can be written on the card, though if there is a great deal of information, it may have to continue onto a second card which can be fastened to the first. Whilst this is a simple system to set up and operate, it has the disadvantage that normally only one or perhaps two cards per document will be created, making it difficult to search for complex subjects. As many cards as are required to express the subject matter of the document can be created (one for each topic), but it remains difficult to combine terms and quickly ascertain which documents refer to a complex subject. This method is probably best suited to small collections of documents with a very specific vocabulary.

Optical coincidence cards (Figure 3) are an improvement on the edge-notched card technique, though suitable equipment is needed to use optical coincidence cards effectively. In this case, the system works indirectly, by referring to a document number. The document then has to be retrieved using this number. The entry term is written on the top right-hand corner,



TRANSLATING

PICKEN, Catriona

The translator's handbook.
Aslib, 1983

PICKEN, Catriona

The translator's handbook
Aslib, 1983

(Translating)

Figure 2. 5"x3" index cards (not to scale)

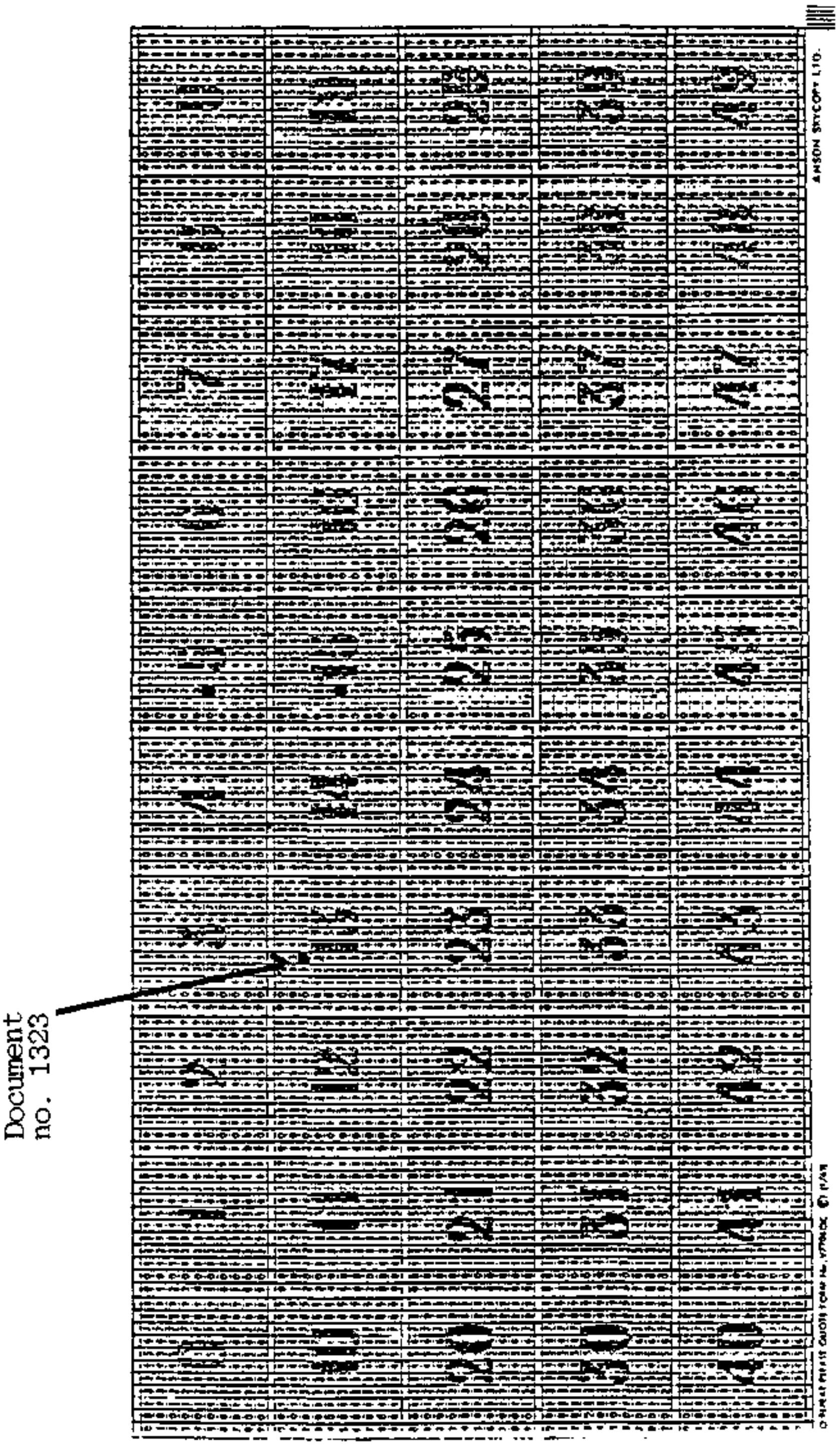


Figure 3. Optical coincidence card

and the appropriate document number is punched through on the body of the card. Obviously, cards large enough for the collection must be used, and they are available with up to 10,000 positions.

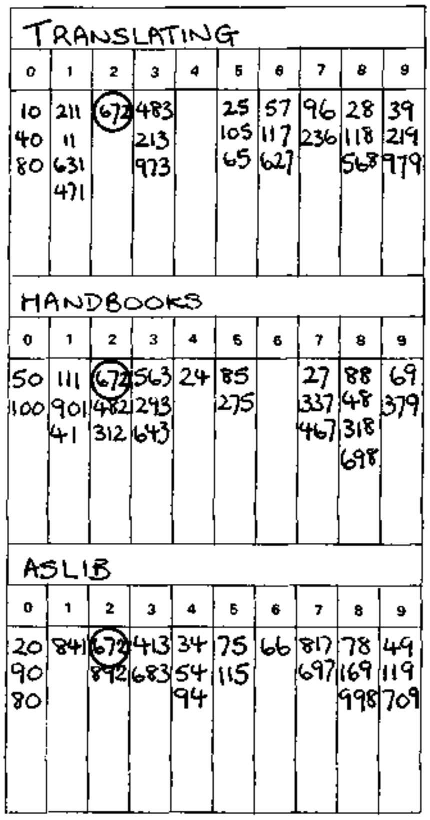
To retrieve documents (or document numbers), the cards with the relevant terms are selected from the total file, carefully aligned and either held up to the light or placed on a light box. The light will then shine through the holes representing the numbers which refer to the complete combination of terms. If there are no documents in the collection which refer to the specific term combination, those which refer to some of the chosen terms may be evident, since the light will shine more faintly through the hole for that number. This document may be close enough to the subject enquiry to be acceptable.

The disadvantages of optical coincidence cards may be obvious: hole punching must be very accurate, and this may not be easy with the larger cards where the holes are quite small. There is little or no room for annotations or comments, and some users may find the reference to a document number inconvenient, as the document itself must be found in a file arranged by document numbers. The possible alternative, of a separate listing of numbers and document records, only compounds the problem. It is also difficult to withdraw documents from the collection, since this involves blanking out the hole in some cards, and it may not be entirely successful.

A simpler variation on this technique is the use of Uniterm cards, which, like optical coincidence cards, refer to document numbers, but do so by the simpler expedient of writing the numbers in columns under the term heading used on the card. Document numbers can be traced by removing all the cards relevant to the subject enquiry and examining them for the numbers which appear on all the cards: the number or numbers appearing on all cards represent documents which cover the specific subject. One way to ease the scanning of the columns of numbers is to divide the card into 10 columns and put all numbers ending in the same digit in the same column. It is, however, a tedious operation and prone to mistakes, since it is very easy to miss a number, particularly if the card is full or almost full. For small files, however, it may be a less expensive and easier-to-operate alternative to optical coincidence cards. Figure 4 shows three Uniterm cards on which document number 672 is the only common document, and so the only one which covers all three topics.

COMPUTER-BASED INFORMATION SYSTEMS

Manual systems such as those described above have been used for many years, but there is little doubt that computer-based systems are set to replace them in the future. Computer-based systems (and I shall be describing microcomputer systems in the main) offer speed, flexibility and



TRANSLATING

0	1	2	3	4	5	6	7	8	9
10 40 80	211 11 631 471	672	483 213 973		25 105 65	57 117 627	96 236	28 118 568	39 219 979

HANDBOOKS

0	1	2	3	4	5	6	7	8	9
50 100	111 901 41	672 482 312	563 293 643	24	85 275		27 337 467	88 48 318 698	69 379

ASLIB

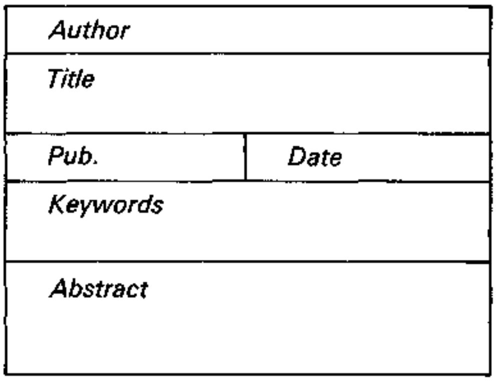
0	1	2	3	4	5	6	7	8	9
20 90 80	841	672 892	413 683	34 54 94	75 115	66	817 697	78 169 998	49 119 709

Figure 4. Uniterm cards (showing document number 672 **common to all and thus representing a document covering the complex subject)**

sophistication in searching and retrieval, whilst they are, in many ways, easier to establish and operate (once the initial training period has been worked through!). The main disadvantage of microcomputer-based systems at present is one of cost: despite the rapidly falling costs of hardware, and the possible fall in software costs prompted by inexpensive but powerful microcomputers such as the Amstrad PC1512, a basic software package capable of handling personal information retrieval will cost in the region of £100 to £200. It also has to be said that few microcomputer systems are as portable as a card index! However, a microcomputer will give a much greater return on the capital invested in it, since software is available for many other relevant applications, such as word processing, spreadsheets, graphics and statistics.

The main type of program used for information retrieval on microcomputers is known variously as a file handler or, more commonly, a database management system (DBMS). DBMS are available for practically every make of microcomputer, varying in power and sophistication. Prices are equally varied, ranging from £25 to several hundreds of pounds: it is possible to spend £2,000 on information retrieval software, though for that price you will have a very sophisticated system capable of handling very large files of records which contain both structured and unstructured (free text) information. As a rough guide, however, you can expect to pay between £100 and £250 for a program capable of handling personal information systems on any scale, though small systems can be operated satisfactorily with less expensive software.

For effective use, DBMS still require the careful analysis described earlier. It also remains important that you are consistent in the use of terms used to describe documents for retrieval purposes, and the establishment of a controlled vocabulary, or the use of an existing thesaurus, is still necessary. Based on the analysis, it is possible to create, on screen, a record structure suitable to the application and which contains the necessary fields of information. A sample record structure, created using Cardbox Plus software, is shown in Figure 5. Once such a record structure is devised, the relevant information can be added for each document, until a database is created. Figure 6 shows the earlier record structure 'filled' out with appropriate detail. There is, however, a significant difference between DBMS and manual systems, a difference which gives the microcomputer-based system its advantage. A single record is sufficient for retrieval purposes (assuming, of course, that it contains all the information required for retrieval) since the software will search and retrieve using any field or combination of fields in the record: there is no need to make multiple



Author

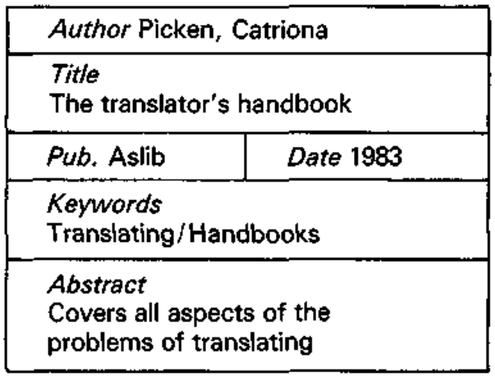
Title

Pub. *Date*

Keywords

Abstract

Figure 5. Suggested screen layout of record using DBMS



Author Picken, Catriona

Title
The translator's handbook

Pub. Aslib

Date 1983

Keywords
Translating/Handbooks

Abstract
Covers all aspects of the problems of translating

Figure 6. The completed screen record using DBMS

entries under various subject headings or authors. It is also the case that computer-based systems will retrieve more quickly than manual systems, at least with large files.

Computer-based retrieval systems can, in fact, provide access to records in many more ways than would be feasible with manual systems: effectively, any field in a record can be used for retrieval, though it may not always be necessary to do this. Much will depend on the software and its techniques for indexing records: it will be wasteful of disk space to have every word in the record indexed for searching, when only a few terms are required in normal use. Once again, the initial analysis will reveal just which fields are required for searching and therefore have to be indexed by the system. Some software, on the other hand, does not create inverted indexes to files (which require separate disk storage): these programs are less space-consuming, but tend to be slower in retrieval. As a point of interest, and to give an indication of the storage space required by computer-based systems, 1,000 bibliographic records will usually require at least 1 megabyte of disk space, and will usually require more, allowing for indexes, system overheads, etc. For this reason alone, hard disk systems are preferred, and they are faster in operation as well.

When it comes to choosing between programs for information retrieval, one other factor may be significant. Those programs which set up separate inverted indexes to files do so by adding all the indexed terms for each record as it is added to the file from the keyboard. This can take some time: one program known to the writer takes over one minute for each record, and this can make data entry a time-consuming task. Such programs often have a batch entry facility, which will allow you to create the records using a word processor and then 'convert' them to the database in batches. While this can also take some time, it is automatic and means that you can get on with something else in the meantime. This would normally only be a problem when setting up the file in the first place: the wait will be acceptable when adding only a few records for updating. The other side of the coin is, as suggested above, that searching is much faster, since the program uses the index to find relevant records, just as we use a back-of-book index to find a specific page in a book. Programs which do not create inverted files are faster at the data entry stage, but slower in retrieval, since essentially they have to examine each record in the file to see if it contains the sought term (this is not always the case, since techniques have been developed to speed up this sequential searching, but they are still rarely as fast as programs with inverted indexes).

All but the least sophisticated software allows searching for combinations of terms, usually using the Boolean operators 'and', 'or', and 'not', so, as was suggested earlier, combinations of terms are possible in order to refine searches. Different fields can be combined, so that searches for, say, an author and a subject, or title keyword and date, are possible.

You should also be aware that a number of microcomputer systems can now be used with optical character recognition (OCR) devices in order to speed up data entry. If you already have a file of several hundred (thousand?!) records, it is a daunting task to consider typing them all again from the keyboard. OCR systems mean that existing index cards can provide the input (assuming, of course, that they contain all the relevant detail): the OCR reader is passed over the card and the image is converted to digital format and stored on disk. Files can subsequently be edited to correct mistakes, layout, etc. However, such OCR systems are often only suitable for typewritten entries (though there are some which will recognise handwriting), may be limited in the typefaces that they can recognise, and may be expensive. Their cost, of course, has to be balanced against the time you will spend typing everything in at the keyboard.

It was suggested earlier that microcomputer systems can be multipurpose, in that a variety of software is available for a variety of applications. One such type of program is used for online search assistance, that is, it converts the microcomputer into an intelligent terminal capable of being used with the very large online databases of bibliographic records and the databanks of non-bibliographic information which are now available throughout the world. One of the many features of these programs is the ability to download records from the online database, i.e. to copy them onto the microcomputer's disks. It is then possible, with some programs, to convert these to an existing format and to merge them with an existing file of records. This obviously saves a great deal of time and money, and is a way of ensuring that your files are kept up to date. However, it must be said that the copyright situation in this application is far from clear: not every online database allows downloading (though it does go on and database providers will eventually have to come to terms with it), and if you use an online database, you must check your agreement to see if you are allowed to download and under what conditions. Normally, you will not be allowed to sell a copy of the records to a third party, but can use them for your own work.

Other microcomputer software which can be used alongside information retrieval systems includes word processors, which offer major advantages when preparing documents, reports, letters, etc. (particularly if they are to be used more than once), spreadsheets, which can be used for financial records, statistics, and modelling, plus statistical packages which are of some sophistication and usually provide a graphics facility, so that histograms, bar and pie charts or line graphs can be produced. A relatively new breed of software is the integrated package which combines information retrieval system, word processing and spreadsheet in integrated modules, such that data from one can be incorporated in another with no re-typing. If your work involves, as it may well do, all these routines, such a package may be the best buy, since you can be assured there will be no problems of compatibility.

CONCLUSION

In the relatively short time available, it has not been possible to cover every single aspect of personal information retrieval. I hope that enough has been said to indicate the potential systems available to you and that you will be able to select a system relevant to your needs. While manual systems cannot be discounted entirely (on cost grounds alone), there is little doubt that microcomputer systems offer a wide range of applications, with software at a range of prices. The multiple uses to which microcomputers can be applied make them valuable ‘Jacks of all trades’: developing your own expertise and ensuring that you examine each application carefully will ensure that they are also ‘masters of all’!

REFERENCE (received after the conference)

Heeks, R. *Personal bibliographic indexes and their computerisation.* Taylor Graham: London, 1986.

AUTHOR

Paul F. Burton, Department of Information Science, Strathclyde Business School, University of Strathclyde, Glasgow G1 1XH, UK.