1. The nature of indexing and the nature of computing.

Essentially the preparation of an index falls into two sections—the intellectual work of deciding what a document is about, and the non-intellectual work of collating several references to the same thing.

Essentially a computer can perform two functions—analysis of information, normally by the use of some form of statistics, and manipulation of information.

The costs of human intellectual resources are rising; the costs of computing are falling. It was inevitable therefore that sooner or later someone would try to mechanise the indexing process; the odd thing (in retrospect) is that instead of separating out the two parts of indexing, the earlier efforts tried to treat simultaneously both the intellectual and the 'chore' parts of the job. But before we consider the actual processes involved we must take a look at the background against which this work was carried out.

2. Indexing and Information Retrieval.

An index is an information retrieval tool—one of the earliest and still one of the most important information retrieval tools. What is an information retrieval system but a means of locating information by using a special sub-set of language to identify the concepts and a series of unique references to identify the place in which these concepts are set out or discussed? For the lifetime of, I guess, every reader, it has been taken for granted that this special sub-set of the language can be used in indexing in two ways: either by the use of a subject heading list of some kind, or by the use of a classification code which is often (though it need not be) numerically expressed. What has, I think, been much less generally appreciated is the essential difference between the two.

It is a characteristic of a classification scheme that any one code number represents the co-ordination of a number of concepts; it is a characteristic of a subject heading list that the concepts are clearly distinguished one from another. So while we can write '372.215.6=use of action songs as a teaching method in primary schools', we have to use 'action songs', 'teaching methods' and 'primary schools' simultaneously to identify the same concept from a subject heading list.

So our information retrieval tool can use either 372.215.6 or the combination of three phrases to define the concept, and by adding a reference to either can identify the location of the required information. If the retrieval tool is a pack of 5-in. x 3-in. cards, or a printed list, we need only the number to cover all three aspects of the concept; if subject headings are being used, provision has to be made for the possibility that a searcher might use as his own point of entry the term 'action songs' or the term 'teaching methods' or the term 'primary schools' and there must be three entry terms in the system to cater for him.

So a classification scheme pre-co-ordinates the elements of the indexed concept, whilst a subject heading list requires the user to exercise some form of co-ordination after the indexing has been done.

Now, pre-co-ordination imposes a point of view: the main field is subdivided into mutually exclusive fields, which are subdivided
again and again and the sequencing of the subdivisions is important. Because teaching methods are subordinate to type of school, if we want action songs as a teaching method in secondary schools we must use a different class number, and if we just want ‘action songs’ we must look in two different places. In fact, what matters is the context in which we are thinking about action songs, and in mechnised information retrieval systems (even down to the simplest of edge-notched card systems) it was early recognised that what was needed was a system which said: find me the references which are indexed by ‘action songs’ and ‘teaching methods’ and ‘primary schools’ or ‘kindergartens’ (i.e. but not ‘secondary schools’).

Now, manually prepared indexes have always recognised this fact throughout their history: an entry under ‘action songs’ will distinguish by means of sub-headings, between ‘in primary schools’ and ‘in kindergartens’, and have a further sub-heading ‘as teaching methods’. If the index is required in great detail, ‘teaching methods’ may appear as a further sub-heading both of ‘in primary schools’ and of ‘in kindergartens’. In any event, there will almost certainly be a corresponding entry ‘Primary Schools, Action songs in’.


It was this recognition that an entry term, generally speaking, ought not to stand alone but be given in its context that led to the earliest form of computer-prepared index, the KWIC (=Key Word in Context). This was an attempt to simplify the production of printed indexes to massive collections—primarily periodicals and in-house technical reports—by treating each of the significant words of the title (as opposed to the insignificant words such as ‘the’, ‘of’, etc.) as indexing terms. The computer’s facility of manipulation was brought into play to bring each significant word in turn into the centre of the page, with the rest of the title spread around it (Fig. 1). No intellectual effort was involved except the preparation of the list of non-significant words, and large indexes could be prepared in a very short space of time. Of course, the success of the system depended on two unproved assumptions—that the words in the title fully represented the concepts in the document and that there would be few synonyms, so that a single word could serve efficiently as an entry point. In the event, neither of these has proved to be the case. There is the famous example of the paper on binary arithmetic entitled ‘Fingers and Thumbs’ which, though properly (according to KWIC logic) indexed under both Fingers and under Thumbs, never served to lead the user to the concept of binary arithmetic; and KWICs have never been as useful as they might because of the lack of any form of vocabulary control, so that the user had to try to think of all the words for his sought concept which an author might have used in his title. However, KWICs have had quite an influence on the accuracy of titling of papers, and to solve the other points, programs have been written which perform the KWIC-type rotation (of the sentence round the indexing term) not on the title but on a descriptive phrase written by the indexer. Devotees of the original KWICs object that this defeats the main object of a KWIC, which is to produce an index with the minimum of human intervention, and KWICs still flourish; in highly specialised subjects, in which the terminology is controlled by the finiteness of the field, their speed of production and ease of use outweigh the disadvantages of an uncontrolled vocabulary and the rather ugly appearance on the printed page. KWOC (= Key Word Out of Context)—also illustrated in Fig. 1—does something to improve on the latter.

4. Permuted Indexes.

Programs for preparing indexes by computer went hand in hand with programs for information retrieval and by the early 1960s it was becoming widely recognised in the latter that IR systems had two functions to fulfil which were to some extent mutually exclusive—relevance and recall. In high relevance systems, the retrieved references were all required to be highly relevant to the
KWIC (KeyWord In Context) INDEXING

The five titles here to be KWICed are:

Can Management Information be Automated
Source Data Collection in Automated Business Systems
Automatic Control — an Evaluation
Control through Information
Computer Management Control Systems.

Permuting the titles to bring to the middle each significant word results in:

CAN MANAGEMENT INFORMATION BE AUTOMATED 1
SOURCE DATA COLLECTION IN AUTOMATED BUSINESS SYSTEMS 2
AUTOMATIC CONTROL — AN EVALUATION 3
CONTROL THROUGH INFORMATION 4
COMPUTER MANAGEMENT CONTROL SYSTEMS 5

The reference number at the end of each line leads the user to the correct document in the list, which gives full details of author, title and source.

There can also be provided an author index, giving merely author's name and the relevant document numbers.

KWOC INDEXES Follow the same principle, but do not permute the title — they extract the keyword and put it at the beginning of the line, followed by the full title:

INFORMATION CONTROL THROUGH INFORMATION 4
CAN MANAGEMENT INFORMATION BE AUTOMATED 1
MANAGEMENT COMPUTER MANAGEMENT CONTROL SYSTEMS 5
CAN MANAGEMENT INFORMATION BE AUTOMATED 1

Figure 1.
enquiry; in high recall systems the enquirer wanted an assurance that all the relevant material had been retrieved. At first it was thought that different retrieval languages (i.e. different sub-sets of the language, as in para. 1 of this paper) were needed for this, but experimental work began to show that it was not the languages, but how they were used in retrieval, that was critical, and this had its effect on the form of printed index programs. SLIC (= Selective Listing in Combination) made its appearance (Fig. 2), and though it has never had the popularity of KWIC it has had a considerable influence on index preparation programs. In the first place, it required the use of a controlled vocabulary with terms precisely defined; in the second, it was recognised that display—a good, readable, well laid-out page—was important. CELT (Classified Entries in Lateral Transposition) used faceted classification schemes, rotating the classification codes as did KWIC but (since the codes take up little space) printing the title alongside each rotation of the codes. ‘Permuterm’ took each pair of keywords from the string of keywords describing the document, and then sorted them, printing the reference numbers of documents to which each pair applied. Another form of permuted index printed across the top of the page a series of broad subject classes, and down the side of the keywords actually used, and then tabulated under each class/keyword combination the reference numbers of the documents to which it applied. Many other systems have been developed—too many to specify here; but the critical thing about all of them is the increasing recognition of the importance of the human intellectual effort in indexing.

But though the intellectual effort, excluded by KWIC, came back into prominence, there did not return the one thing KWIC did give—an indication of the relationships of the terms to one another; and it is these relationships that bring with them greater meaning and therefore higher relevance, in the retrieval process. This problem has been overcome only within the last two years.

SLIC INDEXES work on keywords chosen by the indexer, not on words from the title. The words are put into alphabetical order and permuted. Any permutation which repeats the first part of a previous permutation is omitted—thus, since we would find A B C under A B C D, there is no need to enter A B C separately. A SLIC on keywords A B C D E would be:

A B C D E
A B D E
A C D E
A C E
A D E
A E
B C D E
B C E
B D E
B E
C D E
C E
D E
E

Figure 2.

5. Articulated Indexes.

Analysis of a typical index entry shows that it is composed of a series of noun phrases articulated by prepositions:

Indexes
preparation of, by computer.

This can be ‘normalised’ into a sentence-like structure ‘Preparation of indexes by computer’. The first programs for articulated indexes used the prepositions as ‘articulation points’ and the words between them as entry terms, so that the sentence above could be converted into

Computer
preparation of indexes by
Indexes
by computer, preparation of
Preparation
of indexes by computer

Improvements came thick and fast. An early one was the recognition of the importance of ‘of’ in an index, so that the program was arranged to put terms preceded or followed by ‘of’ at the front of the entry. By this means, the second of the above examples became much clearer.
Indexes

preparation of, by computer.

Then came the facility to split a long noun-phrase into two or more entry points. In the original version, for instance, 'preparation of pressed steel plates' could not be indexed under 'steel' or 'plates', but the current version allows the indexer to choose any or all or none of the components of a noun-phrase for use as entry terms. Another refinement was the ability to distinguish between two kinds of 'and'—that which is a true conjunction, linking noun-phrases wanted in every entry, and that which links noun-phrases to which the other indexing terms apply independently. Thus 'Relations between King and Parliament' needs both 'King' and 'Parliament' in every entry:

King and Parliament, relations between

Parliament and King, relations between

On the other hand 'Preparation of Subject indexes and Author indexes' is satisfied with

Subject Indexes
Preparation of

and

Author Indexes
Preparation of

Figure 3.
Finally, it is worth noting that articulated indexes are as suitable for book as for periodical literature indexing—provided, of course, that the book requires the detail of such an index. The preparation of indexing sentences for articulated indexing by computer requires no skills not at present possessed by the competent indexer. But if I may be flippant for a moment, I must emphasise one important characteristic—he must not make spelling mistakes! In the index of which Fig. 3 is a sample, the author succeeded in spelling ‘emmission’ in three of its four possible permutations. Fortunately the program prints out a proof-sheet (in fact, Fig. 3 is a copy of the proof, not of the final output), so the computer’s inability to recognise that, for this indexer
emission = emmission = emision
was overcome in the final output.

(This particular program has the added advantage that the output may, if required, be a magnetic tape which can be fed to a photo-typesetting machine, thus eliminating all human activities (except proof-reading) between the writing of the individual indexing sentences and the production of the typeset page.)

6. Auto-indexing.

In my first sentence I distinguished between the human intellectual activity of recognising what a document is about and the non-intellectual work of generating, sorting and collating the several required entries. The intellectual part requires the expression of the meaning in a specialised sub-set of the total language.

I cannot close this article without making at least a passing reference to the experimental work now being done on auto-indexing. This is the use of the computer to identify this special sub-set and depends on statistical analysis of the frequency of the use of any given word. It can be assumed that if a word appears more frequently in a document than it does in its total usage in all communication, then it is significant to this document and is at least a candidate word for the special sub-set of indexing words.

At present, this work is wholly experimental and is designed to find words which can be used in computer information retrieval systems rather than in the less detailed printed indexing systems, but though results to date are not wholly satisfactory, it is beginning to appear that before long the selection of candidate words can be left to the computer. Other work on semantic analysis and further work on the way in which words are used together to form complex expressions is pointing the way to thesaurus generation and dictionary compilation and look-up programs which may eliminate much of the human part of the indexing process. But for myself, though my work lies wholly in the field of computerised documentation and though I have every confidence in its future, I do not see this ever taking the place of the human indexer of books.

I have little doubt that the time is approaching when technical papers now appearing in the periodical literature will be wholly computer-held, -indexed and -searched. A book, however, is something more than a report on a single piece of work, or even a state-of-the-art review: it is a synthesis by one man or a small group of men of the work of many men; one to which he has added something himself and his philosophy as well as his knowledge. As such, it is a thing one wishes to keep and to read again and again. And as such, it must have a printed index whose content must ultimately (whatever the computer processes which have subsequently manipulated it) be the work of a human.

I have not given a bibliography: this article covers a period of over ten years, and a multitude of techniques have been developed during this time on which I have no room to comment. If any reader would like to take the matter further, however, I shall be happy to suggest some further reading—though I should warn that little of it will be found in text-books and recourse will have to be had to the original papers.