Present and future trends of information retrieval (*)

HERBERT COBLANS
Formerly Research Department, Association of Special Libraries and Information Bureaux (Aslib), London, Great Britain

My subject is the communication of scientific knowledge, a problem which dates back to the sixteenth century, the beginnings of modern science in which Italy played so significant a part. What differentiates «modern» science from its antecedents in medieval times is essentially that it represents a consensus of accepted knowledge (*). To be accepted it must be widely available, published and documented so that it can be critically evaluated. Thus «Because Leonardo did not live to publish his discoveries, his influence on the history of scientific thought was considerably less than the evidence of his manuscripts suggests it could have been. He was dropping weights off towers a hundred years before Galileo is supposed to have performed the same experiment and long before the doctrine of geocentrism was abandoned he concluded that the sun did not move» (*).

Similarly Mendel’s famous laws of heredity which, although published in 1866 in an obscure Moravian periodical, did not become part of the scientific consensus until they were re-discovered around 1900 by de Vries and others. If this concept of science as a consensus is accepted, research and its documentation must be seen rather like mirror images. When scientific research was largely a «hobby» of gentlemen scientists, from the seventeenth century until the first part of the nineteenth century, the problems of communication, what we now call documentation, were not serious as the scale was limited. Increasingly with the growth of applied science and technology at the beginning of this century, effective documentation has become one of the most important pre-requisites of successful research and development.

Around 1950 when the full impact of the scientific and technical literature of World War II was being felt, an American, Calvin Mooers (*) invented a new term, «information retrieval». The phrase is ambiguous and

(*) Conferenza tenuta presso il Consiglio Nazionale delle Ricerche, Roma, il 13 novembre 1969.
misleading but it has stuck, in fact it has gone into the language and is even used in other languages. It has become very closely associated with the so-called «information explosion» and with a hoped-for revolution in documentation involved in the use of computers. Both these implications have been exaggerated and are, in fact, partly erroneous. Firstly, there is no explosion in any meaningful sense of that much abused word — that is if one is not indulging in rather crude journalism. Secondly, the computer is not necessarily the whole answer to the exponential growth of the literature of science and technology.

Exponential growth is one of the laws of organic nature, but it is also a statistical law of community life. This phenomenon has been well explained and illustrated by de Solla Price in his book «Little science, big science». He shows that «If any sufficiently large segment of science is measured ... the normal mode of growth is exponential ... multiplying by some fixed amount in equal periods of time, ... an empirical law (which) holds true with high accuracy over long periods of time». Some typical doubling times in years are given below.

Doubling times in years

<table>
<thead>
<tr>
<th>50 yrs.</th>
<th>Population</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>number of universities</td>
</tr>
<tr>
<td>20</td>
<td>Number of chemical elements known</td>
</tr>
<tr>
<td></td>
<td>accuracy of instruments</td>
</tr>
<tr>
<td>15</td>
<td>Number of scientific periodicals</td>
</tr>
<tr>
<td></td>
<td>number of scientific abstracts published</td>
</tr>
<tr>
<td>10</td>
<td>Literature of non-Euclidean geometry</td>
</tr>
<tr>
<td></td>
<td>literature of experimental psychology</td>
</tr>
<tr>
<td></td>
<td>number of telephones in the U.S.A.</td>
</tr>
</tbody>
</table>

Admittedly there has been a great increase in publication, but we must keep our perspective. The position varies with subject and what we are measuring. Thus a few years ago K. Mellanby, Director of the Monk's Wood Experimental Station estimated that in Britain «the number of pages used to communicate the new results of original research has only doubled between 1936 and 1966» — 30 years in contrast with the general doubling rate of 15 years. In some fields like classical taxonomy the rate of growth is negative. On the other hand the growth of the research literature of mathematics is linear — fairly steady at around 500 papers per annum.

This question of size and growth can also be approached from the point of view of quality. A recent study on British periodicals on science and technology uses the frequency of citation of papers in periodicals as a measure of the «goodness» of the periodical. Using the magnetic tape store of Science

Citation Index it was found that there were 68,800 citations in the periodicals of 1965 to a total of 29,000 papers which appeared in 1842 British scientific periodicals in 1963 and 1964. However, 95% of these citations referred to only 165 periodicals. In other words less than 10% of all periodicals (the best tenth) would give 95% coverage.

Another way of looking at this problem is from the point of view of coverage. A good example is that of «Chemical abstracts». To control the world literature of chemistry Chemical Abstracts Service (?) scans some 12,000 periodicals, but 90% of the abstracts that are published are based on articles in only 2000 periodical titles. To satisfy 90% of user requirements is comparatively easy and inexpensive; it is the extra 10% which requires a disproportionate increase in costs and effort.

Basically the size of the literature is a function of the amount of money spent, as major product of research and development is new information as recorded in documents. There is thus a direct correlation between research expenditure and publication. A recent Euratom investigation showed that in almost all countries involved in atomic energy, for every million dollars spent on research and development on the average 18 documents are produced. This is neither «explosive» nor unreasonable. The problem is how to extract, to retrieve the costly information embedded in these documents.

A glance at the statistics of national expenditures will help to give some quantitative basis to the emotive generalisations that are commonly bandied around about the information problem.

<table>
<thead>
<tr>
<th>1962</th>
<th>GERD in millions</th>
<th>GERD GNP</th>
</tr>
</thead>
<tbody>
<tr>
<td>U.S.A.</td>
<td>$18,000</td>
<td>3.1</td>
</tr>
<tr>
<td>U.K.</td>
<td>£634</td>
<td>2.2</td>
</tr>
<tr>
<td>Netherlands</td>
<td>Fl 860</td>
<td>1.8</td>
</tr>
<tr>
<td>France</td>
<td>P 4420</td>
<td>1.5</td>
</tr>
<tr>
<td>W. Germany</td>
<td>DM 6625</td>
<td>1.0</td>
</tr>
<tr>
<td>U.S.S.R.</td>
<td>2.6(?)</td>
<td></td>
</tr>
<tr>
<td>Japan</td>
<td>1.3(?)</td>
<td></td>
</tr>
</tbody>
</table>

GERD = Gross expenditure in research & development

GNP = Gross National Product (total monetary value of national production).

a) For the financial year 1965 the percentages for U.S.A. & U.K. were 3.5 and 2.8 respectively.
b) The figure for the U.K. in 1938 was 0.3%. Thus there has been a 9-fold increase over 27 years.

We, especially the scientists, are very concerned that more money should be made available for research and development, and therefore we must accept the implications of this. More documents will result and their handling and control from the point of view of their subject content—in a word, documentation—must be accepted as an important part of the whole cost of scientific research.

I have already referred to the lack of precision in the term information retrieval. What is involved in IR and what is the present state of the art? IR includes two different but related concepts, data-providing systems (fact retrieval) and reference-providing systems (document retrieval). Ideally the user wants information, data, but in practice he (especially if he is a scientist) will more commonly be given a set of documents which might give him some of the facts or ideas that will help him. Of course it depends on the nature of the original question. For example, a general question on the composition of a specific alloy, like Invar, or on «the number of suicides in a given country in a stated year», is in a sense trivial, since a proper reference service in a library can usually give an immediate answer. On the other hand a study of «the vacuum deposition of thin films» or «the epidemiology of suicide» would not have a simple precise answer, in the present state of our knowledge, and at best a number of possibly relevant documents would be the commonest response. To provide such a set of documents raises questions of very real complexity involving indexing, classification, semantics and linguistics.

The raw material which flows into a documentation service has many forms (book, periodical, report, trade literature etc.). But in practice the main effort is devoted to the scientific paper in a periodical as a typical bibliographical unit. Traditionally the paper, as presented, has elements for identification and should provide some clues as to its subject content. The following identifying features can usually be distinguished:

1. Authors (personal or corporate)
2. Title
3. Periodical issue
4. Date
5. Place of work of authors
6. Figure & table captions
7. Contents of tables & diagrams
8. References cited
9. Typographically differentiated words in text
10. Words, phrases, sentences providing clues to subject content
11. Abstract
12. Name of conference (if applicable)
The elements 1 to 5 and 12 are unequivocal tags of identification, though
1 and 5 can introduce issues of matching and ordering because of variant
spelling and the lack of standardisation in any fact expressed in natural
language — what in ordinary library practice makes cataloguing a profes-
sional job needing skill and experience. The elements 2, 6 and 7, 9 to 11
can be used either as guides or in themselves for assigning indexing terms
or selecting keywords or descriptors in the whole process of subject spec-
cification. Elements 1 and 5 also provide the entries for citation indexing (8, 9).
Sharp (10) describes a citation index as a «listing of documents, usually ar-
ranged by author, with a listing against each entry of other documents which
have cited the document represented by the entry».

However it is subject indexing that is the central problem of IR. Here
there are a number of techniques based on what are called indexing language
devices. These fall into two classes; those that are «derivative», that is
taking only words used in the document (elements 2, 10 and 11), and those
obtained by «assignment» from a «controlled vocabulary» (an alphabetical
list of subject headings, keywords from a thesaurus, or class headings from
a hierarchical classification scheme). Traditionally hierarchical classification
has been the common method for subject control. For reasonably effective
retrieval it pre-supposes indexers with a professional training in classification
and a good general knowledge of science and technology, as well as a working
knowledge of the common European languages. In the hope of reducing
personnel costs, co-ordinate indexing based on thesauri has become in-
creasingly popular since the fifties, especially in industrial libraries. It
was assumed that working with keywords, which are recognisably part of
natural language, indexing would need less highly trained staff. We know
now that, for large collections of documents, these hopes are not really
justified.

Above all, the searching operations involved in co-ordinating keywords,
so as to match document content with question, can be done with speed by
computers. But however attractive computer methods may be for the
storage and matching of terms assigned to documents and to the questions
posed to the system, the number of relevant documents retrieved depends
essentially on the quality of the original intellectual effort of the human
indexers. In practice it does not matter much whether you are using classi-
fication or co-ordinate indexing, the final results is on the average much
the same and not very good. For under the best conditions that can be
achieved in the general retrieval situation in a large collection of documents
covering many subject disciplines, three out of four documents provided
in answer to a question turn out to be of little or no interest to the inquirer.

It is thus not surprising that derivative indexing based only on the
words in the document is attractive once juggling (searching, matching,
ordering, printing-out) with words can be done so quickly and efficiently with a computer. Already at the end of the fifties H. P. Luhn (11), the pioneer of computerised documentation, developed a fairly simple computer program to rotate the words in the title of a paper to produce his so-called Key Word In Context (KWIC) index. Here the informative words in the title become indexing terms. This type of index is very suitable for computer operation and has become popular since the American Chemical Society started publishing its «Chemical titles» in this way in 1961.

While the KWIC index is fairly useful for subjects like chemistry in which the vocabulary is rather precise and unambiguous, such an artificial restriction of indexing terms, to only these words which happen to occur in the title of a paper, is only a crude first approximation to satisfactory IR. Clearly if the whole text of a document could be made machine-readable and stored in the computer, operations on all the words would give a better indexing description of the document. Actually over more than a dozen years, billions of dollars have been spent in the U.S.A. on research in «automatic indexing». The objective is complete automation. There should be no intellectual intervention at any point in the whole cycle from when the document has been entered into the computer store until the answer to a question has been printed out on the line printer.

The techniques used in automatic indexing are mainly statistical, the frequency count of certain words and combinations of words in the whole text, supplemented by the semantic and grammatical relationships between these words. In 1965 the U.S. National Bureau of Standards published a survey report (12) which noted a few experiments with encouraging results. In Europe some research in this field has been carried out at the Centre européen de Traitement de l'Information scientifique (CETIS) of the Euratom research establishment at Ispra (13). None the less as one American authority has put it, «Each time one of these indexing systems has been taken out of the laboratory and subjected to the real world the results have been uniformly bad».

During the last few years free text searching on the words of an abstract, or preferably the whole text of the document, is being used operationally by some large systems. Retrieval of documents is based on the conjunction in the texts of suitable words which are programmed from the question of the inquirer. It is claimed that free text searching by words is no less efficient than the use of authority lists or classification schedules. However most of the evaluation has been done in areas like chemistry or aerodynamics where terminology is fairly precise.

If these are the things that cannot be done well so far by computer methods let us look at the real achievements and their implications for the future of scientific documentation. At present they lie mainly in scienti-
fic publication. Computer-aided printing has recently made possible an important integration of two functions in communication—primary publication and secondary retrieval. For example, in a study for the American Institute of Physics, Buckland (14) has shown how the machine-recording of the text of scientific papers during the publication of periodicals can be organized to produce the various indexes, abstracts and so on as a simultaneous by-product at a purely nominal cost. Thus each article after editing is typed on a tape-typewriter with all the retrieval elements (author, title, author abstract and so on) tagged with non-printing codes. All printing control instructions (type face, type size, leading, justification) are added at this input stage. This coded text on paper tape is stored in the computer and by suitable processing provides output in the same printed format as that in which the periodical has always appeared, as well as the regular cumulations of indexes and abstracts» (15). A typical example of the application of these techniques is «Psychological abstracts» which has been assembled and printed in this way since January 1966 (16).

The most recent technical advances, mainly in the hardware, are the result of pressure from newspapers for whom speed, large scale input and multiple editions in various large cities are important considerations. Once the problems of automatic justification of the print line and long distance, even across the Atlantic, transmission with on-line terminals, were mastered the rate of progress of computer-aided typesetting in commercial publications has grown rapidly. Thus a survey, covering the U.S.A. and 26 countries, published in October 1968 shows that the number of computers used for this purpose has reached 821, representing a 54% growth over the previous year.

Although newspapers are not concerned with a wide range of characters, periodicals in science, especially the physical sciences, need large character sets. This used to be a limiting factor, but during the sixties the crudities of the computer line printer have been completely overcome. This has been shown in the most recent achievement, the six publications which constitute the Information Service in Physics, Electrotechnology, Computers and Control (INSPEC) of the Institution of Electrical Engineers in London (17). They became entirely computer-produced from January 1969. The printing unit, the Lumitype 713 Filmsetter, is driven by a computer-created magnetic tape and can handle over 700 different characters and symbols at speeds of 1000 characters per minute. In this way not only are «Physics abstracts>, «Electrical and electronic abstracts» and «Computer and control abstracts» as well as the corresponding Current Papers produced in high quality print, but a wide range of indexes for each issue (author, report number, conference proceedings, subject etc.) as well as their cumulations are automatically printed from the computer-stored references.
Thus it has been established on an operational basis that periodicals and printed lists (indexes, abstracts, bibliographies) can be printed economically up to the best standards if the scale of operation is large enough. The store on magnetic tapes can be used for organizing and «re-packaging» texts and printing them out in different presentations, thus adding considerably to the flexibility and usefulness of the stored records for the dissemination of information.

As has already been mentioned, mechanised IR does not show up the computer at its best, but that is not so much the fault of the computer, as the inherent difficulties of IR, whether mechanised or manual. Within certain limits the Euratom Nuclear Documentation System (ENDS), the pioneer and the largest system in Europe, has shown what can be done in a somewhat special situation. It started with two advantages: a subject, applied nuclear science, in which the documentation started in 1946 and the existence of Nuclear Science Abstracts which fundamentally covers the whole field with highly competent abstracts. Furthermore the rate of growth of nuclear literature (by the end of 1968 it had reached 800,000 references) was such that computer control could pay off. The two main concurrent jobs were the recruitment and training of a team of highly qualified subject and indexing specialists and the construction of a thesaurus suitable for the interlocking subject disciplines. The Euratom Thesaurus (20) developed over many years is in some ways a model for this type of compilation. It consists of some 4700 terms, including only about 1200 general purpose keywords, 1760 inorganic compounds with the remainder for nuclides and alloys. The method for searching in this co-ordinate indexing system can best be illustrated by following through an actual run on a real question.

What papers exist on the joint occurrence of rhenium and osmium anywhere in the solar system? (The abundance of Rh 187 and Os 187 is used for age estimation as Rh 187 has a half life of sixty thousand million years producing Os 187 through beta decay)

**KEYWORDS**

A1 Rh 187  
A2 Rh isotopes  
A3 Rh

B1 Os 187  
B2 Os isotopes  
B3 Os

**C1** age estimation  
**C2** abundance

**SEARCH STRATEGY**

S1 (tight) A1 AND B1 AND (C1 OR C2)  
This formulation will «pull out» only those papers which have been indexed by 2 compulsory and 2 optional keywords.
S2 (medium) (A1 OR A2 OR A3) AND (B1 OR B2 OR B3) AND (C1 OR C2)
This less rigid formulation demands the co-existence of 3 keywords, but the first two have 3 options and the third one has 2 options.
S3 (loose) (A1 OR A2 OR A3) AND (B1 OR B2 OR B3 OR C1 OR C2)
This formulation asks for the simultaneous presence of 2 keywords; one with 3 options and the other with 5 options.
These 3 searches, which represent essentially a matching of these combinations and permutations in all the documents (in this case the store contained 100,000 references), finally provided a print-out by the computer of the abstract numbers in Nuclear Science Abstracts.

S1 gave 2 references
NSA 08 06872 (08 = volume 8)
NSA 16 10751 (16 = volume 16)

S2 gave 6 references (2 + 4)
NSA 01 1192
NSA 11 1917
08 6872
11 7242
08 6873
16 10751

S3 gave 62 references (2 + 4 + 56)
The search analyst now goes to the collection of numbered printed abstracts and examines each abstract. For S1 the two references are found to be «hits», i.e. entirely relevant. For S2 there is an additional hit (NSA 08 6873) and one possible hit (NSA 01 1192), the abstract not being detailed enough for certainty. The other three are irrelevant. For S3 in the same way they establish two further hits and 7 possibles.
Photocopies of all the abstracts of the hits and the possible hits are sent to the questioner.
The above shows that the whole organisation needed (computer, abstracts and highly trained staff) is considerable and is only justified if a large and varied clientele, in this case basically the technical community of the 6 countries, can be served. In other words, mechanised information retrieval of this type is still a very expensive operation and is probably beyond the range of the average documentation service in industry or in a research institute. Certainly for the main disciplines like bio-medicine, chemistry, physics, etc., one national centre for each, in countries of the size of the U.K., is likely to be viable under present conditions, and it may need to have some government subsidy at that.

Summary. — The origin of the term «information retrieval» is recent and arose somewhat erroneously, as the computer answer to the impact after World War II of the exponential growth of the literature of science. This misnamed concept is not as serious as the popular press has painted it and must be seen in perspective. Exponential growth (multiplying by
some fixed amount in equal periods of time) is one of the statistical laws of community life. Thus the accuracy of measuring instruments doubles on the average every 20 years, the number of scientific abstracts published annually doubles every 15 years. A major product of research and development is new information as recorded in documents and therefore there is a direct correlation between research expenditure and publication.

IR involves 2 distinct and overlapping processes — data-providing systems (fact retrieval) and reference-providing systems (document retrieval). The latter is the more difficult as the provision of a set of documents answering a specific question is a complex task involving indexing, classification, semantics and linguistics.

The examination of a typical article in, say «Nature», shows the elements for identification and some of the clues to its subject content. Indexing techniques are still very crude and inefficient. Devices for the indexing control of documents vary from the free word indexing of the KWIC (Keyword in context) through controlled vocabularies (thesauri in co-ordinate indexing and standard lists of subject headings) to hierarchical and faceted classifications, and more recently citation indexing. However in a large collection of documents covering many subjects retrieval, under the best conditions, is such that 3 out of 4 documents, provided in answer to a question, turn out on the average to be of no interest to the inquirer.

Since 1954 computer techniques have been applied to achieve «automatic indexing» whereby there is no intellectual intervention (human indexing) at any point in the whole cycle from the document entry into the system to the final printed-out answer to the requester for specific information. In spite of some hopeful results in pilot projects no practical systems of general validity have emerged. The difficulties of automatic indexing are related in principle to those of the machine translation of natural languages. So far the main achievements are in computer-aided printing and the utilisation of the tagged and stored text for retrieval centrally or in a decentralised manner by the provision of duplicate magnetic tapes.

The only computer systems for large scale indexing and IR actually operating are based on humanly assigned index terms, e.g. ENDS (the Euratom Nuclear Documentation System), MEDLARS (for bio-medicine) and INSPEC (Information Service in Physics, Electrotechnology, Computers and Control). A typical computer search to answer an actual question (what papers exist on the joint occurrence of rhenium and osmium anywhere in the solar system?) is described and discussed.

Mechanised information retrieval is still a complex and expensive operation and therefore even for disciplines like bio-medicine, chemistry and physics one national centre for middle-sized countries, like the U.K., is all that is justified.

Riassunto (Tendenze attuali e future nel campo del ricupero dell'informazione). — Il termine «ricupero dell'informazione» alquanto impreciso ed improprio, è stato creato di recente quando, a causa della crescita esponenziale della letteratura scientifica verificatasi dopo la Seconda Guerra Mondiale, sorse il problema della cosiddetta «esplosione dell'informazione» e si ritenne di poter risolvere tale problema con l'impiego dei calcolatori. Il concetto di «esplosione dell'informazione» non va preso così sul serio come la stampa divulgativa l'ha dipinto, ma deve essere visto in prospettiva. La crescita esponenziale (il moltiplicarsi per un fattore fisso in eguali periodi di tempo) è una delle leggi statistiche della vita di una comunità. In tal modo l'accuratezza degli strumenti di misura raddoppia in media ogni 20 anni, mentre il numero dei riassunti scientifici pubblicati annualmente si raddoppia ogni 15 anni. Uno dei principali prodotti della ricerca e dello sviluppo è la nuova informazione registrata nei documenti; vi è quindi una correlazione diretta tra la spesa per la ricerca e la pubblicazione.

Il ricupero dell'informazione comprende due processi distinti e sovrapposti: i sistemi che forniscono dati (ricupero di fatti) e i sistemi che forniscono riferimenti (ricupero di documenti). Quest'ultimo ricupero è il più difficile poichè la preparazione di un documento che possa rispondere ad una richiesta specifica è un lavoro complesso che comprende l'indicizzazione, la classifica, la semantica e la linguistica. L'esame di un articolo tipico, per es. in Nature, mostra gli elementi per l'identificazione ed alcuni riferimenti agli argomenti in esso trattati. Le tecniche della indicizzazione sono molto rudimentali ed insufficienti. Gli strumenti che permettono di controllare i documenti mediante indicizzazione si estendono dalla indicizzazione basata su parole indipendenti KWIC (Keyword in Context) attraverso vocabolari controllati (thesauri nell'indicizzazione coordinata, ed infine soggetti standard) fino a classificazioni gerarchiche e a faccette più recentemente ad indicizzazione per citazioni.

Tuttavia in una vasta collezione di documenti estesa a molti soggetti, il ricupero nelle migliori condizioni è tale che 3 su 4 dei documenti forniti in risposta ad una richiesta di informazione si rivelano in media di nessun interesse per il richiedente. A partire dal 1954 sono state applicate le tecniche dei calcolatori per realizzare un'indicizzazione automatica, mediante la quale si esclude l'intervento intellettuale (un'indicizzazione umana) in tutti i punti dell'intero ciclo, dall'immissione del documento originale nel sistema sino alla stampa della risposta fornita all'utente che ha chiesto un'informazione specifica. Nonostante qualche risultato positivo ottenuto nei progetti pilota, non è emerso nessun sistema pratico di generale validità. Le difficoltà di una indicizzazione automatica sono connesse in linea di principio a quella della traduzione meccanica di lingue naturali. Finora i maggiori successi si sono verificati nella stampa controllata dal calcolatore e nella

utilizzazione di un testo codificato e registrato per il ricupero in sede centrale o decentralizzata attraverso la fornitura di duplicati dei sostegni magnetici. Gli unici sistemi attualmente in opera per la indicizzazione su vasta scala e per il ricupero dell’informazione a mezzo di calcolatori sono basati su termini per l’indicizzazione assegnati dall’uomo ad es.: la ENDS (Enaratom Nuclear Documentation System), il MEDLARS (Medical Literature Analysis Retrieval System) e l’INSPEC (Information Service in Physics, Electrotechnology, Computers and Control). Per illustrare le possibilità offerte dal sistema ENDS, viene descritta e discussa una ricerca tipo compiuta per rispondere alla domanda: quali articoli esistono sulla presenza contemporanea del renio e dell’osmio in qualsiasi punto del sistema solare?

Il ricupero meccanico dell’informazione è tuttora un’operazione complessa e costosa e quindi persino per discipline quali la biomedicina, la chimica, la fisica un unico centro nazionale per un paese di media grandezza quale il Regno Unito è il massimo che si possa giustificare.

REFERENCES

(7) CAS today. Facts and figures about Chemical Abstracts Service. American Chemical Society, Columbus, Ohio, 1967. 60th anniversary ed.


