THE

POSTGRADUATE

MEDICAL JOURNAL

London *February*, 1966

MEDLARS INFORMATION RETRIEVAL IN BRITAIN

A. J. HARLEY, Ph.D.,
National Lending Library for Science and Technology, Boston Spa, Yorkshire.

ELIZABETH D. BARRACLOUGH, M.Sc.,
The Computing Laboratory, The University, Newcastle-upon-Tyne.

Medicine and the Information Explosion

The words "information explosion" are now-a-days a commonplace. Scientists of all kinds are affected with feelings from ill-at-ease to near-panic, while continuing to add papers to the mounting pile. Evidence grows that many scientists are unable to find information which would be important to them in this flood of paper; for example, in the course of training groups of postgraduate students in the use of literature, one in ten found papers, previously unknown to their research supervisors, which dealt fairly exactly with their topic of research (Urquhart, 1964).

There are now about four thousand different periodicals published around the world, dealing with the broad field of medicine. A library collecting all of them will need perhaps a thousand feet of extra shelf space each year and it is quite plain that no one person can read it all.

The individual has, in fact, two problems. First, he must make some attempt to keep up with the work going on in his own field of investigation. He attends conferences, listens to the professional grapevine, and reads periodicals. A priori, one would suppose that the latter was the most efficient in terms of certainty and precious time. However, it is bedevilled by the scattering law (Bradford, 1937), which shows that, if he finds x papers of interest by scanning a small number of periodicals n, he will find 2x by scanning n^2 periodicals, 3x from n^3 periodicals, and so on. Indeed the only way of being quite sure of finding every paper is to scan every journal. It is thus a matter for each individual to strike some acceptable balance. He must look at the most likely journals and supplement this with one of the newer "current awareness" guides such as "Current Contents", or by looking at the likeliest headings in the monthly issues of the abstracting and indexing journals. Even then, he must resign himself to missing some papers of interest.

The second problem is that one must, from time to time, undertake a detailed retrospective search, to mine relevant papers out of the mountain of published literature. In this, the main devices have been the abstracting and indexing journals. All researchers will be familiar with some of these, but according to one list (N.F.S.A.I.S., 1963) there are now 433 dealing with aspects of medicine.

The N.F.S.A.I.S. list breaks medicine down into 33 subject subdivisions, using the U.D.C. classification system. According to the list no single abstracting/indexing journal covers them all. The analysis is as follows: —

Index Medicus	31	subdivisions
Biological Abstracts ...	28	„
Excerpta Medica	19	„
Meditzinskii Referativniyi Zhurnal	12	„
Chemical Abstracts ...	6	„

None of the other 428 guides covers more than two of the subdivisions.

Admittedly, this is a rough analysis, but one may conclude that no medical library should be without at least one of the major guides. The smaller guides may, on the other hand, be very useful indeed to a worker whose

field they cover exactly, particularly for the problem of "current awareness". No-one should feel comfortable who remains ignorant of whether there is one such in his field.

Index Medicus and MEDLARS

Index Medicus is the indexing journal produced by the American National Library of Medicine. This library was founded in 1836 as the "Library of the Surgeon General's Office": It became the "National Library of Medicine" in 1956, and in 1962 moved to new buildings at Bethesda, close to the research complex of the National Institutes of Health. Almost since its foundation it has published guides to the medical literature, and the present automated system was necessitated by the rapid growth of output.

The automated system is known as MEDLARS (Medical Literature Analysis and Retrieval Service), and it has two main outputs; Index Medicus, which is printed monthly with an annual cumulation; and the MEDLARS retrieval tapes.

The published indexes and the computer retrieval service are in a sense complementary: when the answer to a simple question is sought, it is certainly quicker and cheaper to take a printed index off a shelf and use it conventionally. The computer search may be more efficient when an exhaustive search of a complex question is necessary. No one yet knows exactly where the boundary between these may lie.

The Key: Control of Terminology

A computer is far more diligent than a human brain, but at the same time far more stupid. A human, using a conventional index makes many judgements quite unconsciously: he accepts "fetus" for "foetus" without much difficulty, and can readily decide to look under "antibiotics" if he does not find the kind of article he wants under "streptomycin". A computer must be taught laboriously to do these things.

The intellectual key to MEDLARS is therefore a list of the words which may be used. At present, this list contains about 6000 terms. It is published annually as part (2) of the January issue of "Index Medicus" and is known as MeSH ("Medical Subject Headings"). The published version consists of an alphabetic list, and a list broken into categories which displays some of the relationships between terms.

Such a list cannot be static, and terms are constantly being added to it by the indexers at N.L.M. These new terms begin as "provisional headings" which are then available to the indexers and the searchers who use the tapes. Some of them are eventually admitted to full status. Only the terms appearing in the annual published MeSH are used as headings in Index Medicus.

The Input to MEDLARS

It is important to see what goes into the MEDLARS system, because this determines what can be obtained from it.

For each article in every journal processed, a unit record is produced, which contains: —

The names of the authors
The title of the article
A translation of the title into English if necessary
The journal reference (journal title abbreviation, year, volume, pages)
An abbreviation indicating the language
A list of subject headings from MeSH, each designated "print" or "non-print".

Allocation of the subject headings is performed by the indexers. They skim through each article, and apply words or combinations of words to describe each concept in the article, always using the most specific headings available.

This rule of specificity is important, for it has consequences when the tapes are searched. By way of an example, consider an article in which the author mentions treatment with streptomycin and neomycin. The indexers would apply the terms "streptomycin" and "neomycin", but not the more general term "antibiotics". This latter would only be applied if the author did not specify what antibiotics he used, or if he specified one for which a precise MeSH term did not exist.

In some cases, a concept must be covered by co-ordinating two terms. For example, at one time the concept of pancreatectomy had to be indexed as "pancreas" and "surgery, operative". In this particular case, the term "pancreatectomy" was provisionally introduced (April, 1965). It can be used by the indexers and searchers, but cannot be used as a "print" heading for Index Medicus until it is given full status and published in the annual "Medical Subject Headings" issue of "Index Medicus".

Many headings are already pre-coordinated. For example, "glucose metabolism" is always used in preference to "glucose" and "metabolism" when the concept of glucose metabolism is required.

In addition to specifying an average of ten headings for each article, the indexers designate about three of these as "print" headings, and the article is recorded under these in the published "Index Medicus". In general, the "print" headings are the most important specific headings.

The product of this indexing effort is fed into a computer (Honeywell 800 with satellite Honeywell 200). Complex sorting operations are performed, and two magnetic tapes emerge. One is the "Compressed Citation File" (C.C.F.) which is available to the searchers; the other controls "GRACE", which is a photocomposing machine, capable of producing photographic master copies for the monthly and annual issues of "Index Medicus", at a rate of over 400 characters per second. It is to be noted that all the typesetting for "Index Medicus", including typeface selection, indentation, and insertion of subject headings, is performed by the computer, and recorded on the "GRACE" tape. This data is, however, absent from the C.C.F. tape, which contains the basic information about the article, coded into a compact form.

MEDLARS Searching

There are now about 300,000 unit records on the C.C.F. magnetic tapes, available for search. In principle, the searcher specifies what characteristics a document must have in order to provide a possible answer to his problem: the machine then reads every unit record on the tapes, and copies those which meet the specification.

In practice, it is convenient to process searches together in batches, rather than run through all the tapes separately for every search. The computer at N.L.M. handles batches of about twenty searches at a time; the British program, for an English Electric KDF9, will probably handle up to fifty.

The most important specifications which the searcher makes concern the subject headings which are to have been applied to the desired citations. However, he can also specify any of the following characteristics: —

Authors' names
Geographic headings
Computer entry dates
Journal titles
Languages
Places of publication
Years of publication

Specification of a subject heading can use any heading selected from MeSH, or any provisional heading; it can also specify that the citation should only be considered if that heading had been used as a "print" heading for the citation in "Index Medicus". Alternatively one can specify that the heading was "non-print", or be non-committal on this point.

Now, suppose we wish to make a search for articles about bone fractures arising in automobile accidents. Reference to a copy of MeSH shows that the term "accidents" is put in section I, (Social Science). There we have "accidents" and, listed under it, the more specific terms "accident prevention", "accidents, aviation", "accidents, industrial" and "accidents, traffic". Under "accidents, traffic" in turn, is listed the still more specific term "automobiles". It is reasonable to suppose that papers of interest to us, will be indexed with either "accidents, traffic" or "automobiles".

On looking at section C14 of MeSH, "injury, poisoning, allergy, shock and related conditions" we find the term "fractures". Listed under it are "femoral fractures", "fractures, spontaneous", "fractures, ununited", "humeral fractures", "radius fractures", "rib fractures", "shoulder fractures", "skull fractures" and "tibial fractures". Under 'femoral fractures" we find "femoral neck fractures". Some of this list we can obviously ignore from our point of view, and in fact we are likely to be satisfied with any paper that has been indexed with: —

fractures
or femoral fractures
or femoral neck fractures
or humeral fractures
or radius fractures
or rib fractures
or shoulder fractures
or skull fractures
or tibial fractures

and { accidents, traffic or automobiles

Note that, because indexing policy is always to use the most specific term available, we have to include all the relevant specific terms. In addition, we use the more general term "fractures" to cover fractures of bones that do not have specific pre-coordinated fracture terms. Thus a paper dealing with a fracture of the pelvis, which would have been indexed with "pelvis" and "fractures", would still be selected. We do not have to specify the term "pelvis".

Very probably the above example would not justify a MEDLARS search, because all the articles likely to be found, could also be found much more quickly by looking under "accidents, traffic" and "automobiles" in "Index Medicus".

Suppose, however, we require something a little more complex. We will search for papers dealing with the metabolism of sulphur-containing amino acids in tumours. We are primarily interested in tumours in man, but will accept papers on primates or, more reluctantly, other vertebrates. We are only interested in papers in English, French or German. Fig. 1 shows what the computer input will probably look like (for a search on the English Electric computer).

The first eight lines of the input give a search number, title and various instructions to the computer. In this instance the important instructions in line eight are; to print on paper (3 in. x 5 in. cards may be available later); to sort output by alphabetical order of author; and to print out all the index headings with each citation.

This is followed by the list of terms relevant to the enquiry. The function of the "sum" terms (M4, M9 and L4) deserve explanation. Briefly, where M4 appears in the "search statement" (see below), it is to be read as "M1 *or* M2 *or* M3". The "category terms" C1, C2 and C3, likewise instruct the computer to look at citations labelled with any term in a whole MeSH category. Thus "C1 = C2" means that where "C1" appears in the search statement, it is read as meaning any of the 300-odd terms in category C2, "Cysts, Neoplasms and Granulomatous Diseases" of MeSH. Similarly "C2 = B1" looks for category B1, "Invertebrates" and "C3 = B2" for category B2 "Vertebrates". Note, however, that the term "Man" does not appear in the category "vertebrates" nor indeed at all in MeSH. This has consequences which are potentially useful to the searcher.

The last three lines are the "search statement", divided into three subsearches, the most general appearing first. This instructs the computer to select any citation with the following characteristics. It must have: —

(a) one or more of the terms in M4 (the amino acids); and (b) the term "amino acid metabolism"; and (c) any term from category C2 (cysts, neoplasms etc.); and (d) *no* term from category B1 (invertebrate animals); and (e) one of the specified languages.

When the computer has scanned through the whole tape file, it will have accumulated a collection of citations meeting these requirements. The second search statement selects, from this primary collection, all those which were *not* labelled with one of the terms in M9 (primates, apes and monkeys). The latter remain in Ra.

FIG. 1

```
N 10064
1  0  0
J. Doe, M.D.
Pathology Dept., Blankshire County Hospital.
Tumours and the metabolism of sulphur amino acids.
→
1
P;  A;  X
M1 = Methionine
M2 = Cystine
M3 = Cysteine
M4 = Sum M1-M3
M5 = Amino acid metabolism
M6 = Primates
M7 = Apes
M8 = Monkeys
M9 = Sum M6-M8
C1 = C2
C2 = B1
C3 = B2
L1 = Eng.
L2 = Fr.
L3 = Ger.
L4 = Sum L1-L3
Ra: = M4 and M5 and C1 and not C2 and L4;
Rb: = not M9;
Rc: = not C3;
→
```

The third statement Rc selects from those in Rb, ones which have not been labelled with a "vertebrate" term. This eliminates all references except those dealing exclusively with work on man.

The final printout will be divided first into three sections, consisting of: —

(1) citations which satisfy Rc (i.e., man).

(2) citations which satisfy Rb but not Rc (i.e., non-primate vertebrates);

(3) citations which satisfy Ra but not Rb or Rc (i.e., primates).

Within each group, they are sorted by alphabetical order of author.

It is clearly impossible to explore in this paper all the possibilities available to the searcher in this highly flexible system. In this example, the use of subsearches was a device to separate out papers on three groups of subjects. It can equally be used to produce three groups of progressively more stringent requirements.

The example was chosen partly because it illustrates an inherent danger. It will be seen that negation is a dangerous weapon: suppose a paper dealt with the contrast between amino acid metabolism in vertebrates and invertebrates. This might be highly relevant to our search, yet, since it would almost certainly be indexed with some term for an invertebrate animal, it would be rejected by the "*and not* C2" in Ra.

Plainly, a certain degree of skill is required on the part of the searcher, and very close

contact with the research worker for whom the search is being carried out. Yet, at the present level of technology, the processing must be to some extent centralised.

MEDLARS in Britain

Early in the planning stages, it was decided that decentralised MEDLARS search services would be set up and provided with copies of the magnetic tape files. A number of American centres are already well advanced, and it is hoped to have a service available in Britain in the first half of 1966.

It will be operated jointly by the National Lending Library for Science and Technology, and the Computing Laboratory of the University of Newcastle upon Tyne. The former will supply library, clerical and distribution facilities (it collects all the periodicals indexed at N.L.M.), and the latter computing expertise. The Computing Laboratory is being financed by a three-year contract from the Office of Scientific and Technical Information (part of the Department of Education and Science) which is also responsible for the N.L.L.

The first stage of the operation is to acquire expertise and write the necessary computer programmes. This is no minor effort; the whole MEDLARS program, including printing "Index Medicus" on GRACE, is reputed to contain 30 man-years of programming effort. Fortunately, the search programme is a fairly small part, but because of the differences between the Honeywell and the KDF9 computers, it must largely be re-written.

When this is done, a complete file of magnetic tapes will be obtained from the National Library of Medicine, and will subsequently be kept up to date with monthly additions.

Two services are possible: retrospective searches of the whole file, and searches of the monthly additions on a routine basis. In either case, searches will generally be received and edited at the National Lending Library before being sent to Newcastle for processing once a week. It is planned to experiment with transmission of the searches by telex at a later date. The output will be sorted and dispatched from the N.L.L., which has facilities for very rapid processing of this kind.

During the period of the present contract with the Computing Laboratory, the service will be free, but with strings attached. It is held to be most important to try to measure the value of this service, and therefore it will be a condition that the user provide "feedback" about the output of the computer search. Fortunately, the N.L.L. have acquired some expertise in conducting surveys, and we hope that the feedback can be acquired painlessly.

Intellectual Devolution

One problem remains to be solved. It is to bridge the gap between the searcher working with the computer, and the ultimate user who alone knows just what he wants to know. It is almost inevitable that in most cases there will have to be a middle-man, capable of discussing the problem with the user on his own ground. At the same time, the middle-man will need a clear idea of how MEDLARS works: he will have to filter off the simple questions, that can be answered by half an hour with "Index Medicus", and also the questions that would be answered more sensibly in some other guide, ("Chemical Abstracts", for example). Once accepted, he will have to do at least the first stage of translating the question into a list of terms and a search statement.

Choosing the person best qualified to perform this service for a research group is a matter for careful thought. In many cases, where there is an adequate library, an intelligent and well qualified librarian will be the best choice, but this is to be decided in individual cases. It is hoped to run a series of appreciation courses for such people, beginning just before the system is fully operational.

MEDLARS is the first large-scale, widely available computer information retrieval service. It promises to be a valuable addition to existing bibliographic tools; but its greatest value may be in its catalytic effect. It certainly promises to oblige medical researchers to think clearly about the literature which they must use in helping to produce yet more literature.

Information about services and courses may be obtained from the Computing Laboratory, University of Newcastle, or the National Lending Library for Science and Technology, Boston Spa, Yorkshire (Telephone: Boston Spa 2031).

REFERENCES

BRADFORD, S. C. (1937): "The Extent to which Scientific and Technical Literature is covered by Present Abstracting and Indexing Periodicals" Report of the Proceedings of the 14th Conference of the Association of Special Libraries and Information Bureaux, London.

N.F.S.A.I.S. (1963): "A Guide to the World's Abstracting and Indexing Services in Science and Technology", Report No. 102 of the National Federation of Science Abstracting and Indexing Services, 324 East Capitol Street, Washington D.C.

URQUHART, D. J. (1964): Use of Scientific Literature by Research Students, *Nature* (*Lond.*), **202**, 732.

A fuller account of the MEDLARS system as developed at the N.L.M. appears in "The MEDLARS Story at the National Library of Medicine", published in 1963 by the U.S. Department of Health, Education and Welfare, Public Health Service.