1968c

THE CRITICAL APPRAISAL OF INFORMATION RETRIEVAL SYSTEMS

by

Cyril W. Cleverdon
The College of Aeronautics
Cranfield Bedford

To be presented at

THE INTERNATIONAL CONGRESS
of the
INTERNATIONAL FEDERATION FOR DOCUMENTATION

MOSCOW
SEPTEMBER 1968

The Critical Appraisal of Information Retrieval Systems

by

Cyril W. Cleverdon

College of Aeronautics, Cranfield

There are three separate aspects in the critical appraisal of an information retrieval system.

1. The design of a new system.
2. The evaluation of an existing operational system.
3. The continuous quality control of an operational system.

The basic aim of all such activities is to enable a system to operate at a performance level which will meet the requirements of the users of the system, and to do this at the lowest possible cost.

Before discussing these matters in detail, it is necessary to define what is meant by a "system". The Cranfield projects have generated many new terms; while most of these are now in general use, others still tend to be used in a special sense, and a glossary of such terms is included as Appendix 1. However, where we are proposing new definitions as part of the argument in this paper, they will be considered in the text. Such a term is "information retrieval system"; the main point to be emphasised is that the users must be considered part of the system. We, as well as many others, have been talking and writing of "user-system interaction" which implies that the user is outside the information retrieval system. One result of this has been that some people have argued that in an evaluation test one should not "penalise the system" for errors which are basically the fault of the user. This is an incorrect attitude, for the reaction

of the user will be influenced by the subsystems of indexing, of index language and the store, and therefore these subsystems cannot be evaluated (in an operational system) without involving the users. In this paper we shall take the term "system" as including the users, and the interface will be described as "user-subsystem interaction".

With this definition it becomes easier to differentiate between experimental tests and evaluation tests. An evaluation test is one where the whole system is involved, and this in turn implies that it must be an operational system. Even though only a subsystem may be the main objective of the evaluation test (e.g. comparative levels of indexing exhaustivity*) yet it would still be carried out in the real environment of an operational system, and must be related to the end-product of all systems, namely the reaction of the users to the documents which they receive. On the other hand, experimental tests would always deal with subsystems and have an artificial, created environment. An attempt might be made to simulate reality in this environment, but apart from any other reason, the lack of a user group able to give valid relevance judgements imposes a final restriction.

DESIGN OF A NEW SYSTEM

The design of a new system involves a combination of user surveys and experimental testing. The criteria by which the users judge a system can be listed as follows:-

1) Coverage, i.e. the proportion of the useful literature which is input to a system.

---

* Terms which are defined in Appendix 1 are indicated by an asterisk the first time they appear.

2) Recall, i. e. the ability of the system to retrieve documents which the user will consider relevant to his enquiry.

3) Precision, i. e. the ability of the system to hold back non-relevant documents.

4) Response time, i. e. the time between the question being put and the output being received.

5) Presentation, i. e. the format in which search results are made available.

6) Effort, i. e. the amount of effort which the user must make himself in obtaining an answer to his enquiry.

The purpose of the user survey must be towards finding the necessary information which will enable management to decide the most effective methods for designing the system, but before this can be done it is first necessary to delineate the user group. In some cases this may be comparatively simple, such as in a research organisation where there is a closed group of individuals for whom the system has to cater. The problem becomes more difficult when an open system is planned to operate on a national or international basis. Consider a system in transportation technology. This would reasonably be expected to cover all types of transport engineering, such as automobiles, trains, ships and aircraft, with probably special emphasis on new forms of transport such as hovercraft or hydrofoils. However, transportation cannot be separated from the environment in which it operates, and it might be necessary to cover not only the interaction of the design of cities and the means of transport, but also the living habits of the population. One has to consider the possible long term effects of developments in packaging, such as the use of containers, and in materials handling.

There are the economic consequences to the local communities of factors such as building new bridges or the construction of a tunnel under the English Channel to link France and Britain. For this it is necessary to know something concerning construction techniques, and by this time the system would be covering a subject field which would involve engineers of all forms of transportation, civil engineers, geologists, economists, social psychologists, town planners, architects and many other related activities.

No information system exists in a vacuum, for it will be surrounded by other information systems, some of which will have overlapping interests, and it is this that is likely to be the main factor in fixing the limits of a new system. Whatever way this is done, the first step is undoubtedly to define the purpose and scope of the system and determine who will make up the user group.

When this has been agreed, a user survey can be carried out to determine the requirements of the user group in relation to the criteria given above. Assume that the system is intended to operate on a national level, one needs to know what kinds of questions will be put to the system. Will they be general or very specific? Will the users require a high recall ratio* or will some users be satisfied with a relatively small number of relevant items? Will the questions require an immediate answer, or will the users tolerate a delay? What sort of output will the users require; will titles be sufficient or will abstracts be necessary? How much effort can be expected or demanded of the users?

On another level, one must find what type of services should be

provided. Allied to a retrospective search system, would there be a demand for a selective dissemination of information service? Would a weekly bulletin of titles satisfy some users? If the system is to be financially self-supporting, it is also necessary to determine what the users would be prepared to pay for the services. Supplied with information on such matters, the system managers can begin to plan the system.

Ideally a system might be expected to operate at 100% for all user criteria, which is to say that it should have complete coverage, give 100% recall and precision*, provide the output immediately, in hard copy, with no effort to the user. In fact one knows that certain of these aspects are impossible to attain, and it is the task of managers to decide, in the light of data gathered in the user survey, which aspects they will emphasise, and formulate their decisions so that the system may be suitably designed.

The user criteria are closely inter-related. Assume that the management decision is that the system should be capable of operating at an overall recall ratio of 60%. This can be achieved by having 100% coverage with a recall ratio of 60%. Alternatively it could be obtained with 60% coverage and a recall ratio of 100%. Equally a coverage of 80% and a recall ratio of 75% would give the same end result, and it might well be that this latter alternative proved to be best from the economic viewpoint.

The options which are open to the system designers are many and varied and will have different emphasis in different situations. It would be quite wrong to assume that the research done at Cranfield

or elsewhere has provided results which can be immediately applied in any given situation. What this work has done is to show the restrictions under which a system operates, to make clear the alternatives which are available to the system designers and to suggest which options are worth investigating. The most complete outline of these variables was published in a report prepared by Human Sciences Research Inc. (Ref. 1) and by kind permission of the Director, Dr. Dean Havron, a chart from this report has been reproduced as Appendix 2.

The ultimate criterion which management must apply is that of cost. For example, it is known from Cranfield II that there is an optimum performance level of indexing exhaustivity. If the average number of terms used in indexing a set of documents is plotted against the normalised recall ratio, then the position is as shown in Figure 1.

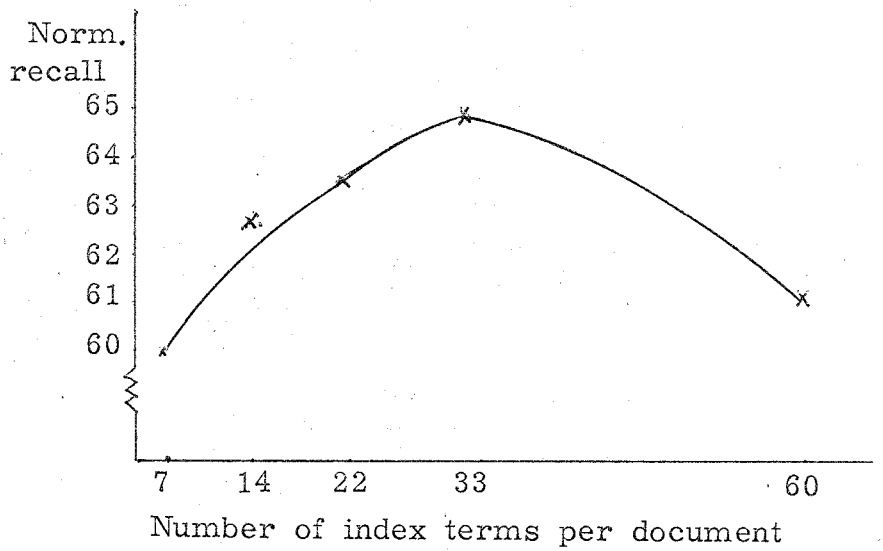

TABLE 1 Variation in Performance According to Number of Index Terms

This indicates that in the particular environment in which these tests were carried out, an average of 33 terms per document gave the best result; the use of 7 terms (representing words in a title) was too few, but the use of 60 terms (representing terms in an abstract) was too many. However, the increased input and search costs involved in using 33 terms as against, say, 14 terms, might be considered to more than offset the relatively small gain in performance. It can also be shown that there is improvement in employing professional indexers rather than automatically extracting key terms from the title, the abstract or the document, but again the economic factor may outweigh the improved performance. Decisions on these and a large number of similar matters can only be taken in relation to the costs involved; a measure which enables such cost/effectiveness comparisons to be made is discussed in a later section of this paper.

A number of tests might be necessary to determine the optimum procedures to be used in a new system, and some people might argue as to whether the cost and time involved in such testing could be justified. Clearly the design effort must be related to the ultimate size of a system, but we would argue that, with the techniques which are available for experimental testing, there is no justification for neglecting the design stage of a new system.

EVALUATION OF AN OPERATIONAL SYSTEM

The second aspect to be considered is the evaluation of an operational system which it is assumed has been operating for a number of years. The methodology that will be discussed was developed from the investigations at Cranfield (Ref. 2, 3 and 4) and will be illustrated by examples from the evaluation of the MEDLARS system at the National Library of Medicine in Washington for which the author was a consultant. The work was carried out by Mr. F. Wilfrid Lancaster, a former member of the Cranfield project group to whom I am indebted for much of the data given in this section. For the full account of the evaluation of MEDLARS, the reader is referred to the report by Mr. Lancaster (Ref. 5).

The term "evaluation" has tended to be used in a broad sense, but as discussed earlier, we feel that it would be more satisfactory if a term such as "experimental test" were applied to projects such as Cranfield II, and if, within the context of information retrieval, the term "evaluation" could be confined to those tests on complete systems where a detailed analysis of failures is made. It is in this sense that we use the term, and the discussion in this section relates to evaluation of operational systems.

The basic pattern of an evaluation test can be stated as follows:-

1) Statement of purpose
2) Preparation of test design
3) Test of system
4) Analysis
5) Interpretation.

Purpose of Test

Nothing can be done until it has been decided what questions the evaluation is intended to answer, for the design of the test will depend on the breadth and type of information that is required. There is no necessity for an evaluation to cover all aspects of a system; one might wish, for instance, to investigate nothing else except the effect on performance of having the enquirer interact with the subsystem in real-time instead of having search strategies prepared by system staff.

It is, presumably, agreed that an operational information service is established to serve a body of users, and that therefore the purpose of an evaluation must be in some way related to the requirements of the users.

On the other hand, as we have already discussed, management is interested in a different set of criteria, for management must know whether a procedure is meeting a required performance level and whether the activity is being carried out in an economic manner. For instance, the particular index language being used or the qualifications of the indexers are matters of no concern to the user so long as his requirements are being met. For the system managers, however, these matters are important, but they must be evaluated from the viewpoint of the effect which they have on user criteria.

It could be argued that the user might also be concerned with the cost factor, and this is to some extent true in those cases where the user has to pay for the service. However, such cases are relatively few, and even then, once a decision has been taken to pay for or

subscribe to a service, the techniques of operation are unlikely to interest the average user.

There appears to be no limit to the questions which can be asked of an evaluation test, but it must be emphasised that it is of paramount importance that these should be clearly stated before the test design is prepared.

A reasonably comprehensive evaluation would be one on which information was sought on most of the user criteria. The main objectives might be presented as follows:-

1. To study the demand search requirements of the users of the system.
2. To determine how effectively and efficiently the present service is meeting these requirements.
3. To recognize factors adversely affecting the performance of the system.
4. To disclose ways in which the requirements of users may be satisfied more efficiently and/or more economically. In particular, to suggest means whereby new generations of equipment and programs may be used most effectively in satisfaction of demand search requirements.

More particularly the test could be designed to answer specific questions as given below:-

Overall Performance

a. What is the overall performance level in relation to user requirements? Are there significant differences for various types of request and in various broad subject areas?

Coverage and Processing

a. How sound are present policies regarding indexing coverage?

b. Is the delay between the receipt of a journal and its appearance in the indexing system significantly affecting performance?

Indexing

a. Are there significant variations in inter-indexer performance?

b. How far is this related to experience in indexing and to degree of "revising"?

c. Do the indexers recognize the specific concepts that are of interest to various user groups?

d. What is the effect of present policies relating to exhaustivity of indexing?

Index Language

a. Are the terms sufficiently specific?

b. Are variations in specificity of terms in different areas significantly affecting performance?

c. Are pre-coordinate type terms and subheadings hindering the efficiency?

d. Is the need for additional precision devices, such as weighting, role indicators, or a form of interlocking, indicated?

e. Is the quality of term association in the index language satisfactory?

f. Is the present "entry vocabulary" adequate?

Searching

a. What are the requirements of the users regarding recall and precision?

b. Can search strategies be devised to meet requirements for high recall or high precision?

c. How effectively can searchers screen output? What effect does screening have on recall and precision figures?

d. What are the most promising modes of user/subsystem interaction?

(1) Having more liaison with information staff at the local level?

(2) Having more liaison directly with the search analysts?

e. What is the effect on response time of these various modes of interaction?

f. Are there significant differences in performance between different operational centers?

Input and Computer Processing

a. Do input procedures, including various aspects of clerical processing, result in a significant number of search failures?

b. Are computer programs flexible enough to obtain desired performance levels? Do they achieve the required checks on clerical errors?

c. What part of the overall response lag is attributable to the data processing subsystem? What are the causes of delays in this subsystem?

Economics

a. What are the cost factors involved in the various possible modes of satisfying user requirements?

b. What are the "payoff" factors or cost/effectiveness ratios for various policies relating to exhaustivity of indexing?

c. What are the payoff factors for various categories of materials input to the system, e.g. certain foreign languages or language groups?

Preparation of Test Design

It is on the basis of the objectives that the test design is prepared. A limiting factor with a fully operational system is that it will probably be working near to capacity of its available resources, and there must be the minimum of interference with its normal operation. No action must be taken that would delay or speed up the servicing of an enquiry, or which might influence its performance. In other words, no perturbations can be permitted which might affect the normal performance.

It is possible to recognise four main types of test requirements. First there are those which involve what we will call the main test. Examples from those given above would be "what is the overall performance level in relation to user requirements" or "what is the effect of present policies relating to exhaustivity of indexing". All the data obtained in the main test can be used to provide answers to such questions as these, and such data must be obtained within the framework of the main test. Secondly, there are those questions which can more satisfactorily be answered by a sub-test, since they are investigations which involve some change in the normal operational

Rolling suggested (Ref. 7) an approximation method which depended on the assumption, which might possibly be false, that a relatively high recall ratio (e.g. 80 - 90%) could be readily obtained. It is not, of course, the recall ratio which is important but the identification of relevant documents which the system does not retrieve, and, until recently, there was no tested satisfactory method of obtaining this.

The solution is relatively simple, and is to ask the enquirer, after his question has been put to the system, to name any relevant documents of which he already knows. In the case of the MEDLARS evaluation, it was possible for more than 70% of the questioners to do this; sometimes they only knew of one such relevant paper but three out of ten could list five known relevant documents. From this base, the recall ratio can be obtained by checking the output of each search to determine whether the known relevant documents have been retrieved.

However, to supplement this procedure it was found possible to locate other documents which appeared to be relevant from other sources such as the Science Citation Index or one of the many specialised abstracting services in the field of medicine. These were added to the collection of documents to be assessed for relevance by the questioner.

In the artificial environment of an experimental test, the evidence appears to suggest that the relevance decisions are, within reason, relatively unimportant, but in an operational test such as the MEDLARS evaluation, the relevance decisions are absolutely vital and it is essential that they should be as reliable as is possible. Obviously they must be made by the person asking the question, for he is the only

person who can be certain of what documents are relevant to his real information need. To obtain reliable relevance assessments, it appears that four requirements must be satisfied. First, there is a limit to the amount of effort that can be demanded of a user; if there are 200 documents retrieved in answer to a question, it is quite unrealistic to expect the average user to be willing to spend the time required to make reliable relevance judgements on this number of documents. Secondly, the decision can only be taken on the basis of the complete document; to ask for these judgements to be made on the basis of titles or abstracts is to introduce error, for some papers which appear to be relevant from the abstract will be found to be non-relevant with the complete document, and vice versa. Thirdly, to ensure that the task is done thoroughly, the reason for the decision in regard to each document must be recorded. Finally, it is essential that the instructions sent out to the questioner must be clearly stated, not subject to misinterpretation and capable of being applied consistently by a large number of different people.

The suggested procedure would therefore be as follows. Following the submission of a question to the system, the questioner is asked to name any relevant papers already known to him. These papers are then checked to ensure that they are in the data base of the system; assume that there are two such known relevant papers, which we will designate as R1 and R2. Searches are then carried out in the various published indexes, and we will assume that six others papers designated R11, R12, R13, R14, R15 and R16 are located which appear to be relevant.

Assume that when the search is carried out, a total of 200 citations are retrieved in answer to the question. It is suggested that the average user would be willing to assess some 25 papers for relevance, so random sampling is made of the 200 citations to obtain the 25 citations required. The originals of these 25 citations are then obtained and copied, and it is this set of documents that will be used for determining the precision ratio for the question. We will designate this set of documents as P1 - P25. It is now necessary to check whether any of the set R11 - R16 are included in the set P1 - P25. Assume that R13 is the same document as P20. In this case, copies would be obtained of the remaining five possibly relevant documents, R11, R12, R14, R15 and R16, and these will be sent to the questioner, together with documents P1 - P25 for him to determine relevance. No indication will be given to the questioner as to the manner in which these two sets of documents were obtained. Attached to each document will be a relevance assessment form. This might include a number of matters, but vital to the relevance assessment is that the questioner should be asked to state whether the document is relevant or not relevant to his information requirements and, most important, that he should be asked to give the reasons for his decision.

These forms are returned, and we will hypothesise that decisions have been taken as follows in regard to the relevance of the documents in the precision set.

R Relevant N-R Not-relevant

P1	R	P6	R	P11	N-R	P16	N-R	P21	R
P2	N-R	P7	N-R	P12	N-R	P17	R	P22	N-R
P3	N-R	P8	R	P13	N-R	P18	R	P23	N-R
P4	N-R	P9	R	P14	R	P19	N-R	P24	R
P5	R	P10	N-R	P15	R	P20	R	P25	N-R

This would indiate that 12 of the 25 documents were considered to be relevant, so the precision ratio would be $\frac{12}{25} \times 100 = 48\%$. In regard to the documents to be used in the recall base, we will hypothesise that the following decisions were made:

R11 R

R12 N-R

R13 R (same as document P20)

R14 R

R15 N-R

R16 R

This means that there is now a total of six documents available for determining recall, namely the original two relevant documents (R1 and R2) which were known to the questioner, plus the four documents found by consulting another system and judged relevant by the questioner. It is now necessary to ascertain which of these six documents are included in the total set of 200 documents retrieved by the test system. If, in fact, four of these documents were included in the total set, we can say that the recall ratio was $\frac{4}{6} \times 100 = 66.6\%$. While the above procedure may appear, in print, to be rather complex, it is in practice perfectly straightforward and worked quite

satisfactorily in the MEDLARS test. An apparent weakness is that it is possible to obtain a positive precision ratio with a zero recall ratio, which is, of course, a logical absurdity. This is due to the fact that two different sets of documents are used for the recall base and the precision base. However, it must be emphasised that the performance figures for individual questions are relatively unimportant in an evaluation of an operational test (as will be discussed in detail later) so any aberrations of this nature do not seriously affect the overall value of the test.

The other difficulty arises when a question is asked for which there is no relevant literature. This situation is to be expected; the percentage of times that this will happen is likely to vary with different systems. With MEDLARS, such questions formed approximately 1% of the total. The problem lies in determining how to measure the performance of the system when it provides a "perfect" response of not retrieving a single document in answer to such a question. In these circumstances, it appears that it is possible to consider $\frac{0}{0}$ as equal to 1, so one can mark the perfect search with a perfect score of 100% recall and 100% precision.

Although not essential, experience shows that it is useful to carry out a small preliminary test before the main test design is finalised. This will give some idea of the general performance of the system; it should ensure that the forms are satisfactory, and will also provide an indication of the amount of effort that will be necessary.

The Test

When the design has been completed, the actual carrying out of the test can proceed as a routine operation. The only comment necessary is to emphasise the importance at this stage of maintaining complete and accurate clerical records.

Analysis

This will be considered from two aspects, namely the statistical analysis which produces the test figures and the failure analysis which provides the basis for system improvement.

The main sets of figures will relate to the performance of the system in regard to recall and precision. These measures, although now widely used, still meet with opposition in certain quarters. A discussion is presented in the study on evaluation methodology, commissioned by the National Science Foundation and undertaken by Human Sciences Research Inc. (Ref. 1). This study accepts that 'as types of ultimate criteria of system operation, these two measures [recall and precision ratios] are valid and necessary' but suggests that they are 'relatively insensitive to the manipulation of independent variables' and 'often not very appropriate to measure the specific treatments being tested'. Their proposed solution is to develop a set of intermediate performance criteria which it is claimed would:

1. 'Be used to measure the effect of manipulated variables at the first point where such manipulation affects the output'.
2. 'Enable the researcher to determine which part of the system is behaving improperly'.

These requirements are reasonable and desirable; the ability of the Cranfield measures, combined with the technique of failure analysis, to meet these requirements is considered in the following paragraphs.

To illustrate these points, we shall present the test results obtained in the recent evaluation of MEDLARS, but will not be commenting on these from the viewpoint of the performance of MEDLARS, but only in relation to the aspects of methodology with which this paper is concerned. For this reason, only certain sets of figures are being used, and these are presented without the comment and explanation which will be found in the main report by Lancaster (Ref. 5).

The average recall ratio for the complete set of 302 test searches was 57.7%, while the average precision ratio was 50.4%. These figures are based on the average of the ratios* obtained in the individual searches, as are the remaining figures presented in this paper.

There were various situations where different sets of figures could be obtained by manipulating one variable at a time. It was, for instance, possible to calculate the effect on precision and recall by searching on Index Medicus terms only; as against the average of 9 terms used for input to MEDLARS, only 2.6 terms on average are used per document in the printed Index Medicus, so this represents a significant variation in the level of exhaustivity of indexing. Based on a sample of 88 searches the results are as in Table 1.

	Recall Ratio	Precision Ratio
Complete Indexing	60%	52%
Index Medicus terms	44%	60%

TABLE 1 Effect of Exhaustivity

The users judged documents as being of major or minor importance; considering the performance figures from this aspect of level of relevance, the results are as in Table 2. It will be noted that although with the documents of major value there is the small expected increase in the recall ratio, there is a large and very significant drop in the ~~recall~~ precision ratio. The figures represent an interesting example of the effect of the generality number*, which is considered in the report by Cleverdon and Keen (Ref. 4, Vol. 2 chapter 3).

	Recall Ratio	Precision Ratio
Major value	65.2%	25.7%
Major and Minor value	57.7%	50.4%

TABLE 2 Performance by Relevance Levels

Five different MEDLARS Centres took part in the evaluation, with each Centre being responsible for preparing the search strategies for the test questions which it received. There were, of course, differences in the questions serviced by each Centre, but in whatever way the figures as given below are broken down, the general effect is the same (see Ref. 5), so that the manipulated variable in Table 3 can be considered to be the general policy at each Centre with regard to the preparation of search strategies. While one would forecast that there would be an inverse relationship between recall and precision, it is surprising to find that this is so consistently the case, particularly in view of the relatively small number of searches on which some of the figures are based.

	Number of Searches	Recall Ratio	Precision Ratio
Centre A	11	69.2%	40.7%
Centre B	41	64.6%	43.2%
Centre C	198	57.9%	50.9%
Centre D	21	55.5%	55.6%
Centre E	28	43.3%	57.2%

TABLE 3 Results According to MEDLARS Centre

The next set of figures relates to 118 questions for each of which they were in essence three separate levels of search strategies, ranging from a specific search to a general search. The expectation that recall and precision would vary in these conditions is confirmed by the test figures of Table 4.

	Recall Ratio	Precision Ratio
General search	62.7%	51.3%
Medium search	48.3%	59.7%
Specific search	32.3%	65.7%

TABLE 4 Effect of Variations in Search Strategies

If the performance figures as given in Tables 1 - 4 are plotted, the result is as Figure 2. While no data exists above 70% recall or below 32% recall, the curve has been continued at both ends so that it may be used in later discussion.

It will be noted that the majority of points fall close to the curve; the exception is the figure for the retrieval of major relevance documents (all other figures are based on major and minor relevance documents)

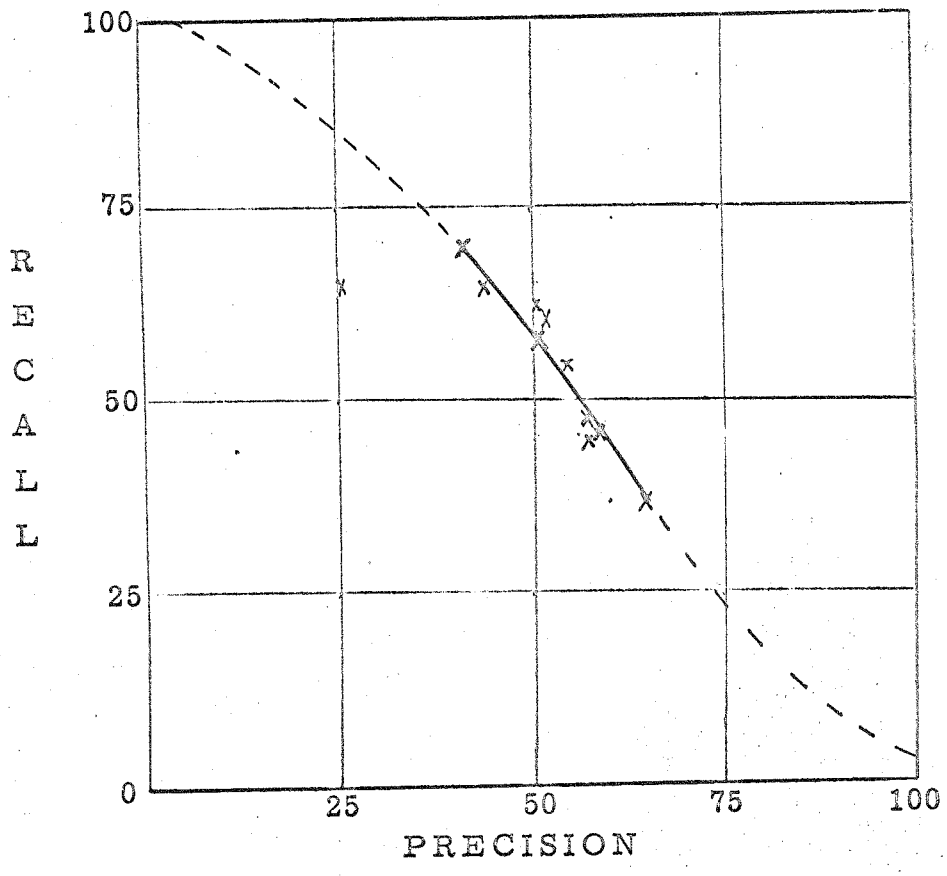

FIGURE 2. Plot of Test Results From Tables 1 - 4

and shows the powerful effect which generality has on performance, as has already been illustrated in the results of Cranfield II (Ref. 4, Vol. 2).

We can now consider two sets of results where a different situation prevails. It has previously been hypothesised (Ref. 4, Vol. 1) that different subject areas would have different performance levels, this being influenced by the subject languages which may vary from "hard" to "mushy". The questions were grouped into broad subject areas, and the results of searches for four such subjects where there were at least 25 questions are given in Table 5. There is a significant difference in performance between the best and the worst subject areas, with Drugs/Biology showing a drop of 10% in both recall and precision as compared to Technics, and this result would appear to be a further confirmation of the original hypothesis.

	Recall Ratio	Precision Ratio
Technics	63.4%	53.7%
Disease	59.7%	48.1%
Preclinical Sciences	59.0%	53.7%
Drugs/Biology	52.1%	43.1%

TABLE 5 Results According to Subject

A final set of results, however, presents a quite unexpected situation. One matter to be investigated was the various methods of user-subsystem interaction, three of which could be recognised in the MEDLARS situation. First there were the occasions when an enquirer

wrote direct to the MEDLARS Centre and the search strategy was prepared on the basis of his stated need without any further communication. Next there were the cases where the enquirer interacted with a member of the information staff in his own organisation, and the query was then passed to a MEDLARS Centre. Finally there were the occasions when the enquirer visited the MEDLARS Centre and discussed his problem with the person preparing the search strategy. The original hypothesis was that moving from the first to the third method would result in a positive gain in performance, that is to say that both the recall ratio and precision ratio would be increased. Such a change did, in fact, take place, but in the reverse way to which had been expected, as is shown in Table 6.

	Recall Ratio	Precision Ratio
No interaction	60.8%	53.9%
Local interaction	55.0%	46.9%
Personal interaction	56.4%	49.3%

TABLE 6 Results According to Mode of Interaction

While asserting a belief that these measures of recall and precision are in themselves completely adequate for presenting the performance of operational information retrieval systems, it is necessary to repeat what we have often said before, namely that any given set of such performance figures has no direct relation to a set of figures obtained with any other information retrieval system. They are unique to the particular environment in which they were obtained, and different subject fields, different document collections

or different user groups will all make direct comparison impossible. This is, of course, as of now; as more experience is gained with the evaluation of different systems, as research showing more clearly the effect of the variables mentioned above is carried out, then it may be possible to make direct comparison, but even so it is unlikely to be on the basis of recall and precision ratios. It will be necessary to develop new measures for this purpose; one such possible measure is considered later.

An interesting result is presented in Figure 3. This is a scatter diagram of the individual performance figures for the 302 searches. The wide scatter might be put down to the relatively small number of documents in the recall base, but even with those searches with at least ten known relevant documents, the scatter is very similar, so it appears that these results (based, of course, only on a sample of the total outputs) are representative of the true position. This diagram emphasises the necessity for considerable caution in interpreting performance figures; when, for instance, we give the performance as 57.7% recall and 50.4% precision, these figures represent the averages for a large number of searches. The figures do not represent an average search, for, as can be seen from Figure 3, very few searches come anywhere near this precise point.

The measures of recall and precision have been criticised, mainly by individuals who have never carried out tests of either experimental or operational systems. Certain other measures have been investigated at Cranfield, and for experimental testing the fallout ratio*, which is the number of non-relevant documents retrieved over the total number

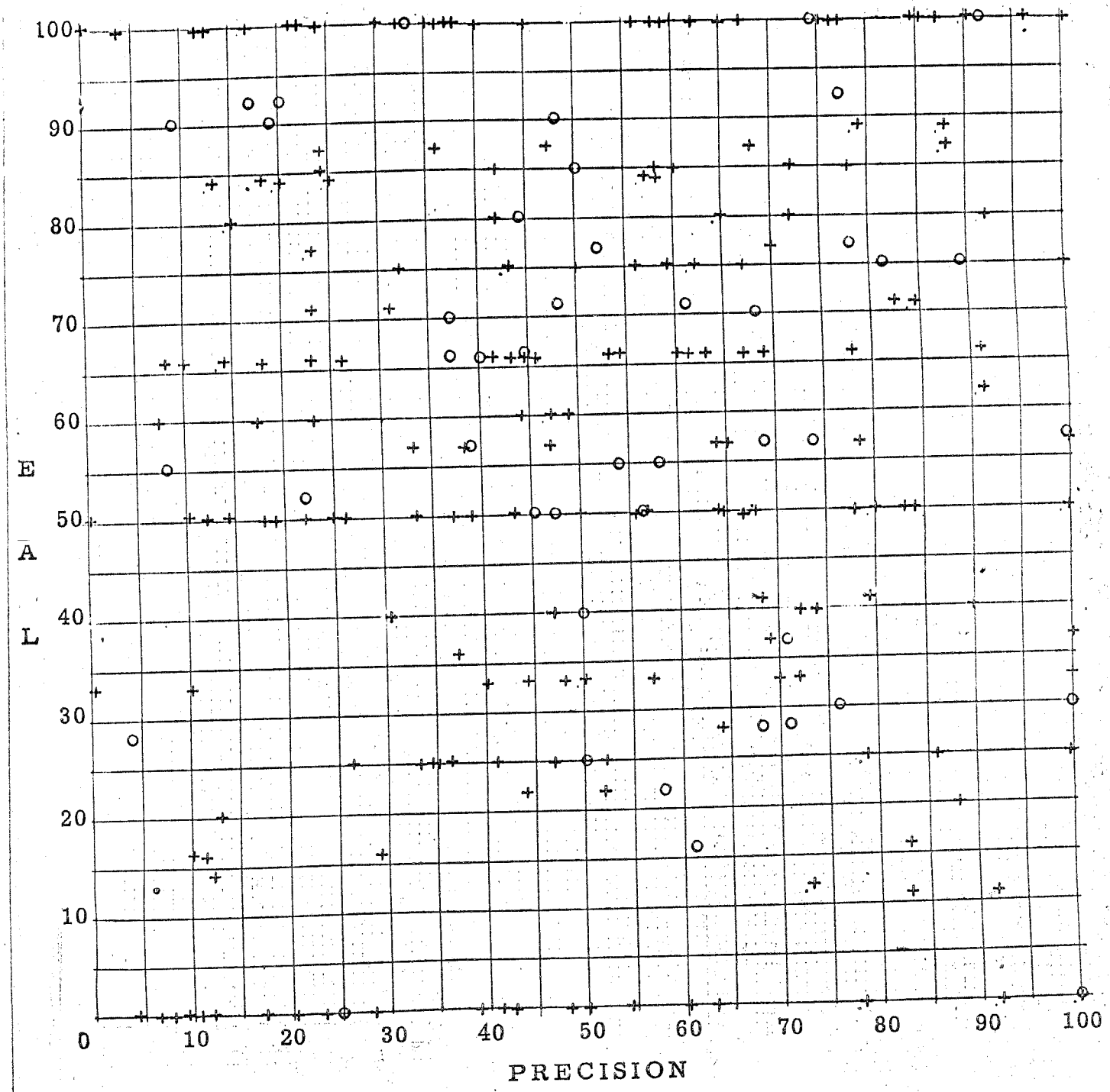

FIGURE 3 Scatter Diagram of Recall/Precision Ratios for 302 Questions

\+ 1-9 Documents in recall base

o 10+ Documents in recall base

of documents in the collection, has proved to be of value, but in operational tests, recall and precision ratios have been shown to be entirely satisfactory and far more sensitive than expected.

However, the statistical analysis is of minor importance as compared to the failure analysis. The figures already presented - and many others that could be so derived - have a limited value for management. It is of interest to know the general performance figures; the differences in performance between different centres might be considered worthy of special investigation, and the results with different modes of interaction give a lead as to the possible methods of optimising this activity. Even so, there is nothing in the figures which shows management how the performance can be improved. This is the area of failure analysis, a time consuming but rewarding operation. This activity was pioneered in the Cranfield I project (Ref. 2) and the WRU-Cranfield test (Ref. 3). In the evaluation of MEDLARS Mr. Lancaster dealt with 3835 failures, made up of 797 known-relevant documents which were not retrieved, and 3038 documents which were retrieved but which were judged to be non-relevant.

The technique of failure analysis is in itself relatively simple. For each failure, whether a failure to retrieve a relevant document (i.e. a recall failure) or failure of retrieving a non-relevant document (i.e. a precision failure) an examination is required of the following:

1. The complete text of the document concerned.
2. The indexing record for this document.
3. The request statement as made by the user of the system.
4. The search programme prepared from the statement and used in conducting the search.

and, for precision failures only,

5. The assessment form completed by the user giving his reason for articles that he has decided are non-relevant.

It is on the basis of all these records that a decision is made as to the main reason or reasons for the particular failure under review.

Examples of various kinds of failure analysis will now be given. In one search related to the tubular secretion of creatine, about 80% of the items retrieved were assessed as non-relevant. Analysis showed that the reason for most of the failures could be put down to over-exhaustive indexing, where the term "creatine" had been used in indexing although the articles contained little directly about it; for example, the article might refer to a creatine value obtained in a routine kidney function test. On the other hand, a low level of exhaustivity can cause recall failures, as in a search relating to the transmission of viral hepatitis by parenteral inoculations of materials other than blood or blood products or during venipuncture. One major value article that was not retrieved deals with hepatic inflammation in narcotic addicts. The fact that viral hepatitis is transmitted by contaminated injection equipment was mentioned in the text but was not covered by the indexing.

Other failures were due to inadequate search formulation; in a search on potassium shifts in isolated cell preparations, no use was made of the term CELL MEMBRANE PERMEABILITY. Used in conjunction with POTASSIUM or POTASSIUM CHLORIDE, it would have brought out several major value articles.

Alternatively relevant documents might not be retrieved because the search formulation was too exhaustive; consider a request for "influence of the styloid process on facial and head pains". The searcher required that some term indicating "face" or "head" be present, as well as a term indicating "pain" and the term for site of the "styloid process" (TEMPORAL BONE). This was unnecessarily exhaustive because it is reasonable to assume that pain relating to the temporal bone would involve face or head. The simple, less exhaustive formulation TEMPORAL BONE and PAIN would have resulted in the retrieval of several more relevant documents.

On the other hand, precision failures could result by using terms that were too general. In a search on electrical brain stimulation, the searcher generalized to BRAIN ELECTROPHYSIOLOGY. This led to the retrieval of 533 citations, of which only 17% were relevant.

In other cases the index language was judged to be inadequate. To express perceptual completion phenomena, the searcher was forced into very general combinations (e.g., VISION and ILLUSIONS) which, although they retrieved 173 citations, achieved only 17% recall at 10% precision.

Inadequate interaction between the user and the subsystem can produce many failures. There are three possible situations as shown in Figure 4.

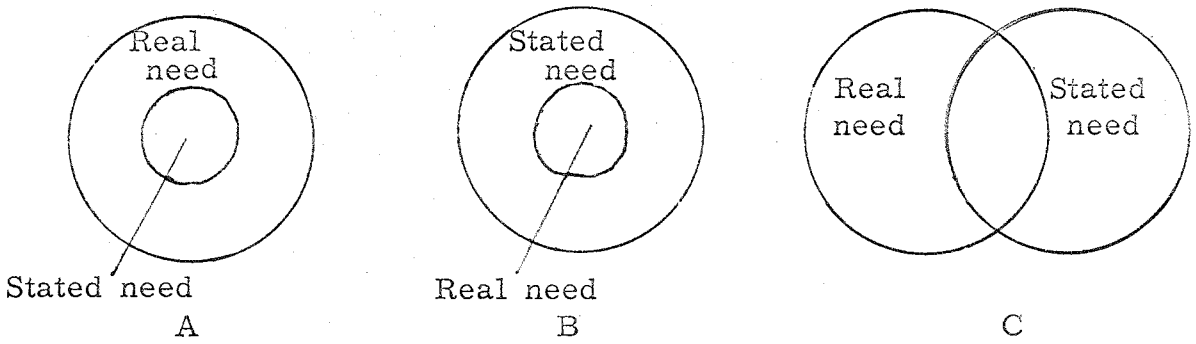

FIGURE 4 Stated and Real Needs of Users

In Figure 4A, the statement has been too specific, and therefore there will be a failure to retrieve relevant documents, whereas in Figure 4B, the statement has been too general and this will result in the retrieval of many non-relevant documents. Finally in Figure 4C there is overlap between the real and stated need; this is likely to result in both recall and precision failures. An example of the first is where the request is for "crossing of fatty acids through the placental barrier; normal fatty acid levels in placenta and fetus". Analysis showed that the requester was interested in the broader area of lipid transfer and also in lipid levels in the newborn infant.

A too general request was for structure and function of the lymphatic system of the lung of any animal. The fact that pathological conditions were not of interest was not made explicit.

The complete analysis for recall and precision failures in the MEDLARS test is set out in Tables 7 and 8. The figures presented in these Tables show some interesting points. Consider, for instance, the fact that MEDLARS, at the time of the evaluation, consisted of entries for more than 600,000 documents. Accepting the impossibility

Source of Failure	Number of Missed Articles Involved		Percentage of Total Recall Failures Involved
Index Language			
Lack of appropriate specific terms		81	10.2%
Searching			
Searcher did not cover all reasonable approaches to retrieval	171		21.5%
Search formulation too exhaustive	67		8.4%
Search formulation too specific	20		2.5%
"Selective printout"	13		1.6%
Use of "weighted" terms	2		0.2%
Other searching failures due to sorting, screening, clerical error	6		0.8%
TOTAL FAILURES ATTRIBUTED TO SEARCHING		279	35.0%
Indexing			
Insufficiently specific	46		5.8%
Insufficiently exhaustive	162		20.3%
Exhaustive indexing (searches involving negations)	5		0.6%
Indexer omitted important concept	78		9.8%
Indexer used inappropriate term	7		0.9%
TOTAL FAILURES ATTRIBUTED TO INDEXING		298	37.4%
Computer Processing		11	1.4%
Inadequate User-Subsystem Interaction		199	25.0%
		868*	

* 868 factors contributing to 797 recall failures.

TABLE 7 Reasons for 797 Recall Failures

Source of Failure	Number of Unwanted Articles Involved		Percentage of Total Precision Failures	
Index Language				
Lack of appropriate specific terms	534		17.6%	
False coordinations	344		11.3%	
Incorrect term relationships	207		6.8%	
Defect in hierarchical structure	9		0.3%	
TOTAL FAILURES ATTRIBUTED TO INDEX LANGUAGE		1094		36.0%
Searching				
Search formulation not specific	462		15.2%	
Search formulation not exhaustive	356		11.7%	
Searcher used inappropriate terms or term combination	132		4.3%	
Defect in search logic	33		1.1%	
TOTAL FAILURES ATTRIBUTED TO SEARCHING		983		32.4%
Indexing				
Exhaustive indexing	350		11.5%	
Insufficiently exhaustive (searches involving negations)	5		0.2%	
Indexer omitted important concept (search involving negations)	1		0.03%	
Insufficiently specific	1		0.03%	
Indexer used inappropriate term	36		1.2%	
TOTAL FAILURES ATTRIBUTED TO INDEXING		393		12.9%
Inadequate User-Subsystem Interaction				
Explicable	464		15.3%	
Inexplicable	39		1.3%	
TOTAL FAILURES ATTRIBUTED TO INADEQUATE INTERACTION		503		16.6%
Computer Processing		3		0.1%
Value Judgement		71		2.3%
"Inevitable" retrieval		4		0.1%
		3051*		

* 3051 factors contributing to 3038 precision failures

TABLE 8 Reasons for 3038 Precision Failures

of re-indexing this collection (which represents some four years input), it means that an equal amount of documents must be input to the system before any improvements in the index language or the indexing can give 50% of their potential benefit. On the other hand, a new policy in regard to user-subsystem interaction can have an immediate beneficial effect on performance.

Another point to be considered relates to the area of inevitable failure. This is illustrated by the failure analysis of 10 non-relevant documents retrieved in a search relating to "Slipped upper femoral epiphysis", in which the questioner was particularly interested in articles dealing with epiphysiolysis in humans when racial factors were discussed. Analysis showed that the only satisfactory way to have prevented the retrieval of the 10 non-relevant documents would have been if an additional requirement of the search programme was that ethnic terms, such as 'Caucasian race', 'Negroes' etc. should have been present. On the other hand, if this action had been taken, of the four known relevant documents, three would also have been eliminated. This problem of the effect of exhaustivity - whether it be in the indexing or the search programme - is shown by the failure analysis of Tables 7 and 8.

20.3% of the recall failures were considered to be due to the indexing being not sufficiently exhaustive. On the other hand, 11.5% of the precision failures were because the indexing was too exhaustive. With searching 8.4% of recall failures were because the search formulation was too exhaustive and 11.7% of precision failures because the search was not exhaustive. Very exhaustive indexing could presumably have eliminated the recall failures attributed to the low

level of exhaustivity; what this level would have had to be cannot be stated exactly, but there was some evidence from the analysis to indicate that it might have been as high as 25 terms per document, which is nearly three times the present level. Were indexing to be at this level, the number of non-relevant citations retrieved must have been very much higher.

Although statistical analysis of results is necessary, it is the failure analysis which is really important in an evaluation test, for it enables management to determine which parts of the system are inefficient and suggests methods by which the overall performance can be improved.

Interpretation

In the preceding section the matters discussed have been reasonably straightforward in regard to the necessary action and the methods of carrying them out. It might seem that the failure analysis would be prone to variation of decisions; to some extent this is true, but experience has shown that in the very large majority of cases, the decision was clear and unambiguous and there was a higher correlation than might be expected between different people doing the task. This, however, is less likely to be true in the final stage of interpretation of the test results. What, for instance, is the meaning of the reversal of the expected result in regard to the mode of interaction? How does it come about that such different search policies should develop amongst groups that have all had a common training in the method of using the system? Why do the individual search results show such a completely random scatter?

In such circumstances, it may be shown that there is a necessity for more detailed analysis or even for additional testing to be done. One might, for instance, suspect that the relatively bad result produced by

direct interaction between the user and the search compiler was due to a small percentage of particularly inefficient searches. This could probably be ascertained, either by checking the scatter diagram to see whether a higher proportion of the bad results were with this interaction mode, or by a more detailed breakdown of the search fai lures. In order to obtain more information concerning the search policies of different centres, it might be desirable to have search strategies for, say, 10 questions prepared at each centre, and then for the output to be checked.

Whatever may be required, the initial process of the evaluation can be considered to be completed when satisfactory answers can be given to the questions originally proposed. The answers may, in fact probably will, generate fresh questions, but whether or not it would be possible, within the framework of the original test, to provide answers for all such further questions cannot be foretold. To take the situation regarding the questions posed by the management in the MEDLARS evaluation, some of the questions have been answered by the results included in this paper; the others are dealt with in the additional figures and analyses included in Mr. Lancaster's main report. The exception to this statement relates to the final set of questions which deal with economics; in many ways these questions presented the real challenge of the evaluation of MEDLARS, for nothing of this kind had been previously attempted. One could ascertain the exact effect of variations of indexing exhaustivity, knowing in advance that it would result in an inverse relationship between recall and precision ratios. Alternatively, one could calculate the cost of input depending on different levels of indexing exhaustivity. What was

not known was the way by which these matters could be brought into relation with each other and provide the answer to such a question as "what are the cost-effectiveness ratios for various policies relating to exhaustivity of indexing?" In the following section is put forward a measure which relates operational performance to the economic aspects of the system.

Cost-Effectiveness Measure

The Director of the National Library of Medicine, Dr. M.M. Cummings, presented a paper (Ref. 8) giving details of the cost of the information service for the Fiscal Year 1966. This showed that the total cost for some 3,000 demand search bibliographies amounted to $455,614. In the course of these 3,000 searches, 615,700 documents were retrieved, an average of 203 for each search. Although the test results presented earlier in this paper did not relate to exactly the same period of time, we are, for the purpose of demonstrating the use of this measure, assuming that the performance and cost figures are compatible.

The measure proposed (C_R) is one which determines the cost of retrieving a relevant citation. It is reasonable to argue that this is the basic purpose of an I.R. system, and that therefore the lower the cost of providing a relevant citation, the more efficiently the system is operating.

The basic figures on which the calculations are being made can be summarised as follows.

Annual cost $455,000

Number of citations retrieved 615,700

Recall ratio 60%

Precision ratio 50%

Number of relevant citations retrieved $\frac{615,700 \times 50}{100} = 307,850$

Total number of relevant citations in data base

$$\frac{307,850 \times 100}{60} = 513,090$$

Cost per relevant citation retrieved $\$ \frac{455,000}{307,850} = \$1.48$

From this starting point, it is possible to calculate the effect of performance changes on C_R, and hypothetical situations are considered and comparison made with the present position as given below.

(a) Present position of 60% recall and 50% precision.

(b) Improved recall. 70% recall and 50% precision.

(c) Improved precision. 60% recall and 60% precision.

(d) Improved recall and precision. 70% recall and 60% precision.

(e) Greatly improved recall and lower precision. 80% recall and 40% precision.

It is assumed in the first place that these changes are made without any changes in the operational costs of the system apart from the cost of printing citations, which is taken to be 5 cents per citation. To consider (b), the effect of improving recall to 70% would result in the retrieval of $513,090 \times \frac{70}{100} = 359,159$ relevant citations, and therefore with precision remaining at 50%, the total number of citations retrieved would be 718,318. Allowing for the printing cost of the citations that are additional to those retrieved with (a), the cost figure is shown to be $\$455,614 + \$(10218 \times 0.05) = \$460,745$. In the situation C_R will now be \$1.28.

With (c) precision has been improved, so the total number of citations retrieved is reduced to $307,850 \times \frac{100}{60} = 513,090$; this results

in a decrease in printing costs, so the total cost is now $450,484, making C_R = $1.46.

The figure for C_R in the five different cases is shown in the first line of Table 9, and reveals some interesting features. An improvement in recall of 10% reduces C_R by 21 cents, whereas an improvement in precision of 10% only results in a reduction of 2 cents. A major increase in recall of 20% combined with a drop in precision of 10% gives a reduction of 35 cents.

	60% R 50% P	70% R 50% P	60% R 60% P	70% R 60% P	80% R 40% P
1. C_R	$1.48	$1.28	$1.46	$1.27	$1.13
2. C_R	$1.58	$1.38	$1.55	$1.33	$1.31

TABLE 9 C_R for Various Performance Levels

These figures appear to indicate that an improvement in recall is more effective in reducing C_R than an improvement in precision. The reason for this is that, apart from the relatively low cost of printing additional citations, the retrieval of non-relevant documents is not directly a charge against the system, but it is a charge against the user, since he is involved in additional effort in rejecting non-relevant citations. The more non-relevant citations, the more time must be spent by the users in discarding them, and it appears reasonable to argue that these user costs, which result from system performance, should be taken into consideration in assessing the cost-effectiveness of the system. Such user costs are likely to vary in different situations. While many citations can be rejected on the basis of the title, in

other cases it might be necessary to consult the complete document before deciding that it was non-relevant, and in such cases the cost would vary according to the ease or difficulty of obtaining complete documents. The second line of Table 9 shows the effect of adding a charge of 10 cents for every non-relevant citation retrieved. It will be seen that even in this case it is more effective to improve recall and precision, and it would be necessary to penalise the system with a charge of at least 50 cents for every non-relevant citation before an improvement in precision was shown to be more effective.

It is now possible to translate some of the test results to C_R. Consider Table 1, dealing with variation in indexing exhaustivity. For Index Medicus terms, with a recall ratio of 44% and precision ratio of 60%, 225,720 relevant documents are retrieved, the total costs come to $457,304, and C_R = $2.02 as against $1.58 for the complete indexing of MEDLARS. However, there would be an obvious saving in costs by reducing the average number of terms from 9 to 2.6; for this saving to reduce C_R to a comparable level, the total costs of the system would not have to exceed 225,720 × $1.48 = $334,006, which would require a reduction of some $123,000 on the present costs.

Another example relates to the figures in Table 3 showing the performance according to MEDLARS Centres, where the performance ranged from 69.2% recall and 46.7% precision to 43.3% recall and 57.7% precision. Converting this set of results into C_R is shown in the following figures.

	C_R
Centre A	$1.44
Centre B	$1.49
Centre C	$1.59
Centre D	$1.67
Centre E	$2.09

As has often been said before, an information system is a matter of compromise. Everyone with experience knows that it is impossible to operate an information retrieval system and give 100% performance in respect of each of the user criteria, and it has been argued by Bourne (Ref. 9) that above a certain level, which he puts at 90%, costs will rise out of proportion to the extra benefits gained. So the first compromise is between performance and costs. However, further options exist; the input efficiency can be maximised (and the proportion of the total cost for this operation thereby increased) in the hope that the search strategy can be relatively straightforward and simple but still give effective performance. This, in general, can be said to be the policy for most present-day operational systems, where highly trained staff apply their intellect to the process of indexing and full-time staff are engaged on the control of the index language. On the other hand, the input costs can be reduced (thereby accepting a lower input efficiency) and a more complex (and expensive) matching programme used so that the output can be ranked and the overall performance maintained. Such a procedure is that advocated by Professor Salton with the SMART system (Ref.10) where no intellect is applied at input, but where the far more invdved matching results in an output of documents ranked according to

presumed relevance.

The above discussion has given just a few examples of the methods and the situations in which C_R can be used. It is a measure for management, and is thereby quite different from recall and precision ratios, or any other performance measure which has been used, for management must concern itself with cost as well as performance. This cost-effectiveness measure which pre-supposes, of course, that an evaluation of system performance and a calculation of system costs have been made, permits management to take decisions based on full knowledge of the implications, and as such would appear to be of importance and value. A more complete analysis together with a mathematical treatment of the measure, will be published in Ref. 11.

Comparative Evaluation

In the preceding discussion, where we have been considering the evaluation of an operational system, it has been stressed that, since every system operates in its own unique environment, direct comparison of results obtained with different systems is not possible. However, a new situation is now developing, with computerised information services which cover the same field being offered on a subscription basis. For example, in the field of chemistry, the Chemical Documentation Research Unit at the University of Nottingham is offering a Selective Dissemination of Information service based on Chemical Titles. It is also possible to obtain in this subject field an S. D. I. service (ACSA) from the Institute of Scientific Information. The result is that there is now a situation where two systems can be evaluated not for their intrinsic merits but for their effectiveness in a single environment. It must be emphasised that in such a case the results can only be taken to apply to the particular organisation

in which the evaluation is made. However, if a number of different organisations made such a test and consistently agreed that system A was more efficient than system B, this might be taken to be a reasonable indication of comparative merit.

An interesting example of such a comparison evaluation is presented by Abbot, Hunter and Simkins in Ref. 12. The task is reasonably straightforward, involving an analysis of the output received from each system for the same search request or, as in this particular case, user profile. Different organisations will place a varying emphasis on coverage, on recall, on precision, on timeliness, on cost, on presentation or on other aspects that may concern them, but if the work is carried out and reported with the same care as has been shown by Abbot and his colleagues, it is to be expected that our knowledge of system performance will rapidly increase.

CONTINUOUS QUALITY CONTROL

The final type of appraisal we have suggested relates to the continuous quality control of an operational system. Because of the existing store of an operational system, improvement in performance must of necessity be slow; it should, however, be continuous, and this can only be the case if there exist techniques for continually assessing performance.

Consider the failure analysis shown in Tables 7 and 8. 81 recall failures and 534 precision failures were attributed to the lack of an appropriate specific term in the index language. In the course of the test, certain weaknesses in the index language have been high-lighted by the evaluation and can be corrected, but the cases

that have been found are only those which happened to be revealed by 302 test questions. There is a strong probability that there are other terms in the index language which could be improved, but these will only be shown by further failure analysis.

Lancaster (Ref. 5) has suggested that continuous quality control would cover at least the following functions:

1. Recognise requests that cannot be adequately covered because of present indexing policies or vocabulary inadequacies.
2. Recognise, post facto, searches that have, relatively speaking, a poor performance.
3. Recognise, in the indexing operation, items of subject matter that cannot be specifically expressed in the present index language.

It is the second of these requirements that raises the main problem, and that also relates to the other matter we wish to consider. As of now, a user of the large majority of systems is not normally aware - nor can he be aware- of either the performance he can reasonably expect, the implications of his performance requirements, or the actual performance he obtains. Unless and until such knowledge is available, then we can hardly consider that information retrieval systems have come of age.

The assumption is frequently made that a recall ratio of 100%, though it may be impossible to achieve, is the ideal, but any practising librarian in an industrial or research organisation will know that in the majority of cases, the user has no wish for comprehensiveness on this scale, but only requires a few relevant documents on the subject of his enquiry. In fact, to give such an

enquirer 100 documents might result in his rejecting 95 of them, not on the grounds that they were not relevant, but because they were of no value to him. His requirements would be satisfied by the information contained in the first five documents he had read, and probably any other sub-set of five documents would have been equally relevant and useful.

Keen (Ref. 13) has already proposed a measure for recall based not on the total number of relevant documents which are in the collection, but on the number of relevant documents which the user requests. He calls this measure "relative recall" and defines it as

$$\frac{\text{total number of relevant documents seen by user}}{\text{total number of relevant documents user would like to see}}.$$

As Keen implies, this is meant for a test of an operational system, and appears to be a perfectly reasonable measure.

In one particular case in the MEDLARS evaluation, it was found that the enquirer was satisfied with a search output which provided him with 17 relevant documents. It was estimated that there was possibly a total of 50 documents which would have been relevant, and the recall ratio was therefore given as 34% which is well below the average recall performance. However, relative recall (as expressed above) would be 100%, so this search could be considered entirely satisfactory.

In order to meet the second requirement for quality control, it is necessary to obtain detailed information from the questioner concerning his needs, not only in relation to the true subject of his enquiry, but also in relation to the type of output he might expect or require. To do this he must be provided with reliable data

concerning the probable performance of the system, and this must be translated into meaningful figures, for recall and precision ratios can mean little to the ordinary user. The first requirement is that he should give an indication of the probable number of documents relevant to his question; this information being given, he can then be shown a chart, such as that in Table 10. This now presents him with a somewhat cruder version of the information a user receives with a conventional classified card catalogue. Assume that a user approaches a card catalogue for information on fatigue of materials. Knowing that, for instance, the appropriate Universal Decimal Classification number is 539.388 he finds that there are three drawers of cards, representing 2,000 or more references at this particular number, so he probably makes a mental re-adjustment and decides (in effect though perhaps not consciously) that he will forego maximum recall for the sake of improved precision by limiting his search to that section which deals with the fatigue of light alloys. Again this section might present him with several hundred cards and be too forbidding, so he places a new restriction on his search by selecting that subsection dealing with the fatigue of light alloys at high temperatures. By this time he will probably have lost considerably in recall, but as regards precision (which is, in effect, the number of citations he is willing to scan) a tolerable level has been reached.

Similarly, although in different conditions, with the situation hypothesised in Table 10. Given that the user expects that about 100 papers have been written that are in the system and relevant to his request, then if he insists on 100% recall he may be faced with the same situation as the user of a card catalogue and have to scan

some 2,000 citations. If this is more than he is prepared to accept, he reduces his requirements so that he may expect to come closer to what he considers to be a tolerable level. Clearly this tolerance level will be different for users with different requirements; the person who is writing a book or perhaps a person commencing a major research project would possibly be prepared to accept a very low precision ratio in order to ensure having complete coverage of the relevant literature. The fact that individual users will have varying requirements is to be expected; the important point is that they should be aware of the probable outcome of their requests.

Recall Ratio	No. of Relevant Documents							
	5	10	15	30	50	100	200	400
100%	50-100	100-200	150-300	300-600	500-1000	1000-2000	2000-4000	4000-8000
80%	20-40	40-80	60-120	120-240	200-400	400-800	800-1600	1600-3200
60%	5- 9	9-18	14- 27	27- 54	45- 90	90-180	180-360	360-720
40%	3- 5	6-10	9- 15	18- 30	30- 50	60-100	120-200	240-400
20%	1- 3	3- 5	4- 8	9- 15	15- 25	30- 50	60-100	120-200

TABLE 10 Probable Total Retrieval Related to Total of Relevant Documents and Required Recall Ratio

For any given system a performance estimate such as that in Table 10 can only be prepared when a systematic evaluation test has been made. As more information is known concerning the system, the table can be refined; for instance, to consider the results from Table 5, if the subject of the question were Technics, the performance estimates would differ from those of a question in the field of drugs/biology.

The requirements of the user must be known (insofar as he can give them) before planning the search strategy. Clearly there is no point in writing a specific search programme for an individual who insists on 100% recall, nor in reverse to write a general search programme for an individual who is going to be satisfied with 20 relevant documents out of an estimated total of 100 relevant documents.

However, it is certain that many of the search results will not match the requirements of the users, and it is a main task of the quality control unit to recognise this immediately. There is no point in this paper in speculating how this might best be done, for the techniques must be developed in an operational situation. Clearly it will demand more effort from the users, but more particularly it seems that it will require a completely professional approach by the system operators. They must know everything possible about the system, how it will operate in different circumstances, be able to recognise the strength and weaknesses of the system, know that one question is straightforward while another will present serious difficulties. The by-product of such an activity will be the gradual elimination of the weaknesses of the system with the result that system performance should continually improve.

CONCLUSIONS

During the past five years many papers have been published concerning the methodology of testing information retrieval systems, and the writers fall into three main groups. There are those who report experimental work, those who comment on the work of the experimenters and those who advance theoretical viewpoints. The

first group is relatively small for only about 20 experimental or evaluation tests have so far been reported, while it would be easy to list several hundred papers in the latter two categories. In this text it is worth quoting at length from the chapter on evaluation by Rees in the "Annual Review of Information Science and Technology Vol. 2" (Ref. 14).

> "Much of the discussion in 1966 has continued to revolve around methodological issues and has consisted largely of a repetitious dissection of a very limited amount of experimental activity with little theoretical basis. Refutations of rebuttals are often interesting but they typically generate little additional knowledge. Thus, the dialogue between those who accept the Cranfield methodology and those who, for a variety of reasons, are critical of it, has not been particularly productive from the point of view of advancing the state of the art. There is, and can be, no one way to test and evaluate retrieval systems, and it is absurd to imagine that any particular testing technique, or set of measures, will solve the problem of evaluation. Rather than engage only in carping criticism of the deficiencies of any one research project (thus giving rise to a new round of justifications of the procedures employed), it would be more desirable to devise and test alternate methodologies. Unfortunately, only scattered instances of this more positive approach can be found."

With most of this paragraph I am in entire agreement, the exception being the sentence concerning failure of dialogue in advancing the state of the art. It is true that dialogue as such does not directly bring about improvement, but the informed criticism of

colleagues has played a vital part in the continued development of the Cranfield methodology. The only critics with whom we completely disagree are those who argue that no experimental or evaluation tests should be done until a perfect methodology is available, for it has been shown that every test that is carried out will, if it uses the best methodology currently available, increase our knowledge both of testing and of information retrieval systems. On the other hand, of course, some naive attempts to use different techniques or to devise new names for existing measures have done little except to create confusion.

This paper has reviewed one set of methods which can be used in the critical appraisal of various stages of an information retrieval system. As Rees implies, it would be absurd to claim that there is only a single test technique or a single set of measures, and we certainly do not wish to give the impression that the techniques discussed in this paper represent the ultimate in experimental or evaluation testing. The methodology has changed considerably since the days of Cranfield I, and there is no reason to suppose that it will not continue to develop in the future. A number of different groups are now actively engaged in experimental or evaluation tests using procedures similar to those presented here, and it is certain that in time ways will be found of refining many of the crudities of the present methodology.

Many tasks still remain to be done; not only has the work been so far limited to a few subject areas, but no attempts have yet been made to consider the language problems at an international level. It is in this particular area that the International Federation of Documentation

could well concentrate their activities of experimental and evaluation testing. While superficially the results recently presented by Freeman (Ref. 15) do not give much encouragement to those who advocate the use of the Universal Decimal Classification, there is evidence to suggest that the U.D.C. could be made to operate as well as any other index language. Yet if there is one lesson which we should have learned from the work of the past ten years, it is that no subsystem can be considered in isolation. We cannot take an index language, such as the U.D.C., and expect that every system in which it is used will operate at the same performance level. An index language is only one of the links in the chain of producing a given document in response to a user need, and it will be directly affected by such matters as indexing decisions, the level of exhaustivity and search strategies. Even though one may evaluate a system which uses U.D.C. as the index language, this would reveal very little concerning the practical difficulties that would be involved in using the U.D.C. in an international information retrieval system, with different countries responsible for input. Expressing a personal opinion, I feel that the many groups of enthusiasts who give so generously of their time and effort in preparing revisions of various sections of the U.D.C. deserve to have more information on the probable effect of their decisions. At present, experienced though they will be in the subject area and in the theory and practice of classification, they have no real means of knowing whether or not their efforts significantly improve performance. My view is that for F.I.D. an evaluation programme should go hand in hand and have at least equal importance with the activities of the Federation in regard

to the revision of the U. D. C., and I am encouraged in this opinion by the prominence given to the matter by the organisers of the International Congress in Moscow.

Finally, I must acknowledge the debt which this paper owes to the National Library of Medicine in Washington, and in particular to the Director, Dr. M. M. Cummings, and to the evaluator, Mr. F. Wilfrid Lancaster. Anyone who reads the report of the MEDLARS Evaluation will realise that Mr. Lancaster was responsible for an excellent piece of work, and his final report contains a vast amount of information, not only on MEDLARS but on the operation of large information retrieval systems in general. Throughout the test the encouragement and support of the Director was invaluable, and both in his agreement that the test should be done and in his intention that it should be fully and publicly reported, he has set an example which it is to be hoped others will follow.

REFERENCES

1. SNYDER, M. B. and others — Methodology for Test and Evaluation of Document Retrieval Systems: a Critical Review and Recommendations. Human Sciences Research Inc., 1966.

2. CLEVERDON, C. W. — Report on Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems. Cranfield, 1962.

3. AITCHISON, J. and CLEVERDON, C. W. — Report on a Test on the Index of Metallurgical Literature of Western Reserve University. Cranfield, 1962.

4. CLEVERDON, C W., MILLS, J. and KEEN, E. M. — Factors Determining the Performance of Indexing Systems. Vol. 1 Design, Vol. 2 Test Results. Cranfield, 1966.

5. LANCASTER, F. W. — Evaluation of the MEDLARS Demand Search Service. National Library of Medicine, 1968.

6. CLEVERDON, C. W. — The Testing and Evaluation of the Operating Efficiency of the Intellectual Stages of Information Retrieval Systems. Classification Research, Munksgaard, 1965.

7. ROLLING, L. N. — A Computer-Aided Information Service for Nuclear Science and Technology. Journal of Documentation, 22, 1966, pp. 93 - 115.

8. CUMMINGS, M. M. — Needs of the Health Sciences. Proceedings of the Conference on Electronic Information Handling. Pittsburgh, 1967.

9. BOURNE, C. P. — Some User Requirements Stated Quantitatively in Terms of the 90% Library. Stanford Research Institute, 1964.

10. SALTON, G. — Progress in Automatic Information Retrieval. I. E. E. E. Spectrum 20, 1964, pp. 5 - 15.

11. CLEVERDON, C. W. — Factors Determing the Performance of Indexing Systems Vol. 3. Methodology for Experimental Testing and Evaluation. Cranfield (to be published).

12. ABBOT, M. T. J., HUNTER, P. S. and SIMKINS, M. A.
Current Awareness Searches on CT, CBAC and ACSA.
Aslib Proceedings, 20, 1968, pp. 129 - 143.

13. KEEN, E. M.
Evaluation Parameters.
Information Storage and Retrieval. IRS-13.
Cornell University, 1968.

14. CUADRA, C. (Ed.)
Evaluation of Information Systems and Services.
Annual Review of Information Science and Technology, 2, chapter 3. Wiley 1967.

15. FREEMAN, R. R.
Evaluation of the Retrieval of Metallurgical Document References using the Universal Decimal Classification in a Computer-Based System.
American Institute of Physics, 1968.

APPENDIX 1

DEFINITIONS

Index Language

Basically the index language is the set of index terms which are used in indexing a collection of documents. While in its simplest form an index language would consist only of such a set of terms e.g. uniterms, a more complex index language will indicate the relationships between terms, will incorporate various devices, will have a lead-in vocabulary and a set of roles governing its use, as, for example, the Universal Decimal Classification.

Store

A store is that part of a system where the indexing decisions are recorded for subsequent retrieval, and would include, for instance, either a card catalogue or computer tapes.

Exhaustivity and Specificity

One frequently reads of "indexing in depth", but it is often not clear as to which of two meanings is intended, and therefore the terms "exhaustivity" and "specificity" are used.

Exhaustivity in relation to indexing is a comparative term which refers to the number of index terms which are assigned to a given document. A very high level of exhaustivity of indexing would be indicated if 50 terms were assigned to a document; in comparison the lowest possible level would be if only a single term were assigned.

On the other hand, specificity, which is also a comparative term, relates to the generic level of a selected index term. A concept in a document can be translated in such a way that the index term is

and also are a division into those documents which are or are not retrieved. This is usually represented as shown below.

	Retrieved	Not Retrieved	
Relevant	a	b	a + b
Not Relevant	c	d	c + d
	a + c	b + d	a+b+c+d = N

From this table, the measures used in the Cranfield work are as follows.

Recall Ratio $100\left(\frac{a}{a+c}\right)$

Precision Ratio $100\left(\frac{a}{a+b}\right)$

Fallout Ratio $100\left(\frac{b}{b+d}\right)$

Generality Number $1000\left(\frac{a+c}{N}\right)$

Recall ratio can also be plotted against either precision ratio or fallout ratio. Generality number is mainly of use in experimental situations where document collections of different sizes are being compared.

There are two methods of obtaining the average performance of a set of questions. The total number of documents retrieved for each question in the test set can be totalled, and the recall and precision ratios calculated from these figures. Alternatively the recall and precision ratios can first be calculated for each question and the average of these ratios can then be calculated. The former method is known as the average of numbers; the latter method as the average of ratios.

For a full discussion on performance measures, see Vol. 2, Chapter 3 of Ref. 4.

CHART 2: MODEL DOCUMENT RETRIEVAL SUBSYSTEMS, PROCESSES, AND VARIABLES

FUNCTIONAL SUBSYSTEM: DOCUMENT ACQUISITION

BASIC ACTIVITY: Development of a potential store of information.

PROCESS	INPUTS & INPUT VARIABLES	SYSTEM VARIABLES			OUTPUTS & OUTPUT VARIABLES
		PERSONNEL	PROCEDURES	EQUIPMENT	
1. Determination of user needs for information	User needs for information Range of subject matter Number of users	Knowledge of user needs	Source of determination of needs (i.e., user, librarian, manager)	Channels and vehicles for expressing needs	Requests for documented information Range of subject matter Volume Type of document
2. Acquisition and accumulation of store of documents	Requests for documented information Range of subject matter Volume Type of document	Knowledge of sources of information Knowledge of subject matter	Constraints on ready acquisition of information (e.g., finances, availability, storage capacity)		Acquired documents Breadth and depth of coverage Number of documents Format of documents Form of information

FUNCTIONAL SUBSYSTEM: (DOCUMENT) ACQUISITION SELECTION

BASIC ACTIVITY: Selection of documents to be stored in system; preparation for analysis and further processing.

PROCESS	INPUTS & INPUT VARIABLES	SYSTEM VARIABLES			OUTPUTS & OUTPUT VARIABLES
		PERSONNEL	PROCEDURES	EQUIPMENT	
1. Screening of acquired documents for acceptability and suitability for inclusion in the system	Acquired documents Breadth and depth of information Number of documents Format of document (e.g., report, journal, book) Form of information (e.g., linguistic, analogue, tabular)	Knowledge of user requirements Knowledge of system capabilities and techniques, especially re: indexing language, file structure Awareness of significance or value of document	Explicit/implicit criteria of acceptability/suitability Range of information formats and types that can be handled by system		Document selected for inclusion in system Number of documents Range of document formats Range of information types Documents rejected for inclusion
2. Preparation and processing of chunk of information to be analyzed ("base object")	Documents selected for inclusion in system Number of documents Range of document formats Range of information types		Degree of standardization of format, size, type of information required for analysis or storage Constraints on locus of information to be analyzed (e.g., title only, abstract, full text)	"Standardization" machines, (e.g., microfilming, taping, copying, printing)	Base objects Degree of standardization of format, size, type of information Range of "base" (e.g., title only, abstract, full text, etc.)

FUNCTIONAL SUBSYSTEM: (DOCUMENT) INDEXING PROCESSING AND FILING

BASIC ACTIVITY: Preparation of information for storage: filing.

PROCESS	INPUTS & INPUT VARIABLES	SYSTEM VARIABLES			OUTPUTS & OUTPUT VARIABLES
		PERSONNEL	PROCEDURES	EQUIPMENT	
1. Transformation of index terms into symbolic/machine-readable code	Index Terms Number of terms Number of words per term Number of explicit interrelations	Ability to perform clerical tasks, especially matching, pattern recognition, transcribing	Degree of post-encoding quality control efforts	Symbolic code Degree of rigidity of code (open vs. closed) Number of terms in code Number of codes per index term Number of symbols per code entry	Symbolic code terms (and) interrelations Number of code terms
2. Preparation of file element; affixing term/item information on file medium	File terms (i.e., index terms or code terms) Number of terms (and) interrelations Bibliographic information (i.e., "identifiers") Number of identifiers		Number of operations per term and identifier	File medium Term-on-item vs. item-on-term Number of entry positions per unit	A set of file elements Number of elements Number of elements per term Format of file element
3. Insertion of file element into the file	File elements Number of elements		Number of operations per element	File Number of elements in file Number of distinct files	A set of filed elements Number of elements Number of terms Number of identifiers Number of internal linkages File structure Term-on-item Item-on-term

FUNCTIONAL SUBSYSTEM: (QUERY) SELECTIONS ANALYSIS

BASIC ACTIVITY: Identification and description of significant concepts to be indexed for searching.

PROCESS	INPUTS & INPUT VARIABLES	SYSTEM VARIABLES			OUTPUTS & OUTPUT VARIABLES
		PERSONNEL	PROCEDURES	EQUIPMENT	
1. Identification of significant concepts to be indexed	Query Number of concepts Number of implied logical products, sums, differences Level of specificity/ exhaustivity Level of specificity/ generality	Knowledge of subject matter Academic training Experience in profession/ research Bias toward subject field Ability to recognize and discriminate among concepts Ability to discriminate significant from nonsignificant concepts Familiarity with the literature of the subject field	Formalization of instructions re: type and number of concepts to be recognized Prescriptive force of instructions Depth of analysis prescribed	Analytical aids Extent to which desired concept categories are made explicit	Concepts Number of concepts Scope of concepts Type and number of interrelations
2. Description, in "lead-in" terms of significant concepts to be indexed		Knowledge of subject matter Academic training Experience in profession/ research Bias toward subject field Ability to recognize and discriminate among concepts Ability to discriminate significant from nonsignificant concepts Familiarity with the literature of the subject field Ability to verbalize Linguistic facility	Degree to which concepts are to be described in inquirer's or analyst's terms		"Lead-in" terms Number of terms Length of term Extent to which terms are made explicit
3. Determination of semantic/ syntactic/logical relations among concepts	"Lead-in" terms Number Length Explicitness	Knowledge of the association of ideas and relations expressed in the query			"Lead-in" terms Type and number of interrelations Degree of explicitness

FUNCTIONAL SUBSYSTEM: (QUERY) INDEXING PROCESSING AND MATCHING

BASIC ACTIVITY: Preparation of query for searching; searching.

PROCESS	INPUTS & INPUT VARIABLES	SYSTEM VARIABLES			OUTPUTS & OUTPUT VARIABLES
		PERSONNEL	PROCEDURES	EQUIPMENT	
1. Transformation of index terms into symbolic/machine-readable code	Structured query Number of terms Number of subqueries Number of interrelations	Ability to perform clerical tasks, especially matching, pattern recognition, transcribing	Degree of post-encoding quality control efforts	Symbolic code Degree of rigidity of code (i.e., open vs. closed) Number of codes in vocabulary Number of symbols per code entry	Structured query (encoded) Number of subqueries Number of terms/codes per query Number of operations per query
2. Preparation of file for searching (matching)	Structured query Number of subqueries Number of terms per query Number of operations per query	Ability to perform clerical tasks, especially matching, pattern recognition, transcribing	Number of elements of file that can be handled simultaneously Number of logical products, sums, differences that can be handled simultaneously		Elements of the file corresponding to requirements of search Number of elements Structured query
3. Matching; recording of matches	File elements corresponding to search requirements Structured query Number of operations Number of subqueries	Ability to perform clerical tasks, especially matching, pattern recognition, transcribing	Number of elements of file that can be handled simultaneously Number of logical products, sums, differences that can be handled simultaneously		Documents or document addresses fulfilling search requirements System output format