CHAPTER 1

General Considerations

The original Aslib-Cranfield investigation on the efficiency of indexing systems (refs. 1, 2 and 3) did not, by itself, produce firm answers to what is one of the basic problems in information retrieval, namely the decision as to which index language should be used. Certainly it did not, as some people had anticipated, demonstrate that one system was 'better' than another, either generally, or in any given situation. The positive contributions of Cranfield I can be grouped into four areas:

1. It swept away a number of popular misconceptions concerning indexing and index languages that were extant in 1957 when the project commenced. Every index language had its passionate adherents and opponents. The modernists against the traditionalists, those arguing for natural language against controlled vocabulary, those preferring alphabetical as opposed to classified arrangement, all could find both comfort and dismay in the results of Cranfield I. It was shown to be not true that postcoordinate indexing was vastly superior to precoordinate indexing, it was not necessary to put 120 entries into a card catalogue to retrieve a document covering five concepts, yet on the other hand it was not true that a postcoordinate system (at that time usually associated with the Uniterm system) necessarily need have weaknesses due to lack of term control; the chain index did not provide a satisfactory means of entry into a single order facet classified catalogue nor, on the other hand, did engineers find any particular difficulty in using the long numerical notations of the Universal Decimal Classification. Such were only some of the viewpoints which had been endlessly argued without any experimental evidence to justify either side.

2. With the test of the index of metallurgical literature of Western Reserve University, it was shown that an evaluation could be made of an operational system with comparatively little effort and by using only a small sample of the collection. Since that time improvements have been made in the methodology, and experience has shown in what respects improvements are still necessary, but the general methods first tried in 1962 have been successfully used in a number of different applications (e.g. Refs. 5 and 6).

3. It stimulated a considerable amount of discussion (see, for instance, the bibliography in ref. 4) which has helped to clarify the problems of information retrieval, and created an interest in the methodology of evaluation.

4. It provided sufficient data to enable provisional statements to be made covering a number of aspects of information retrieval systems.

It was in the new hypotheses which could be formulated that the earlier project is of main importance in regard to the present work. Swanson (Ref. 4), in the most exhaustive and scholarly review of Cranfield I that has been made, has listed the following points which appeared to him as being significant.

1. No significant improvement in indexing is likely beyond an indexing time of four minutes, (which is taken to be equal to about seven minutes in a real-life situation).

2. Trained indexers are able to do consistently good indexing although they lack subject knowledge.

3. Indications are that information-retrieval systems are operating normally at a recall ratio between 70% and 90% and in the range of 8% to 20% precision ratio.

4. There is an optimum level of exhaustivity of indexing. To index beyond this limit will do little to improve recall ratio but will seriously weaken the precision ratio.

5. There is an inevitable inverse relationship between recall and precision.

6. Within the normal operating range of a system, a 1% improvement in precision will result in a 3% drop in recall.

7. The most significant result of the main test program was that all four indexing methods were operating at about the same level of recall performance.

In some published comments on Swanson's paper (Ref. 4A) it was suggested that the following points should be considered in addition to those listed above.

8. The most important factors to be measured in the evaluation of information retrieval systems are recall and precision.

9. The physical form of the store has no effect on the efficiency of the system with regard to recall and precision.

10. The index language has a relatively minor effect on the operational performance of an information retrieval system. The main influence is the intellectual stage of concept-indexing.

11. Given the same concept-indexing, any two or more kinds of index languages will be potentially capable of similar performance in regard to recall and precision.

12. The more complex an index language (i.e., the more devices it incorporates), the greater the range of performance in regard to recall and precision.

13. Maximum recall is dependent on exhaustivity of indexing; maximum precision is dependent on the specificity of the index language.

Of the above, numbers 1, 2, 3 and 6 were presented with the qualification that they only applied to the set of documents and set of questions that were investigated, namely a collection in the general subject area of engineering, metallurgy and physics. The remainder appeared to be of general application, and numbers 4, 5, 7, 8, 11, 12, and 13 in particular formed the basis of the present work. It is not suggested that all these hypotheses were new; it was merely that, with the results from Cranfield I, experimental data were now available which appeared to justify them.

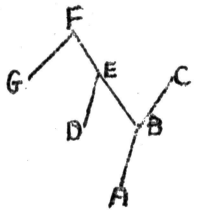
Possibly the point which has attracted most attention and criticism has been in regard to the assertion that there is an inevitable inverse relationship between recall and precision. This, in other words, implies that if an attempt is made to retrieve

more relevant documents, one is forced to accept a proportionately larger number of non-relevant documents. Alternatively, if it is desired to restrict the number of non-relevant documents, this can only be done at the cost of also missing some of the relevant documents. Our experience, backed by the results of tests carried out by a number of other investigators, leads us to believe that this is a fact. However, until in a later volume the further evidence of some 120,000 searches has been published, we will, to avoid argument, call it a hypothesis.

Instead of the form in which it is stated in (5) above, it would be more precise if it were stated as follows: Within a single system, assuming that a sequence of sub-searches for a particular question is made in the logical order of expected decreasing precision and the requirements are those stated in the question, there is an inverse relationship between recall and precision, if the results of a number of different searches are averaged.

There are here four qualifications to the original statement. Concerning the logical order of sub-searches, assume the request is for information on Siamese cats. A reasonably logical order of sub-searches might be

A Siamese cats
B Domestic cats
C Domestic pets
D Wild cats
E Cats
F Felidiae
G Lions



In such a case the inverse relationship would be expected to hold. However if one first searched under 'Lions', it might reasonably be expected that the recall ratio and the precision ratio would be very low, so that going next to 'Siamese cats' would improve both recall and precision. This qualification is therefore only put in to cover the somewhat absurd situation suggested, and can hardly be said to weaken the basic assertion, any more than can the point that the requirements are those stated in the question. This is to cover the situation when the questioner asks for information in Pekenese dogs and, when presented with the output, says that he really required information on Siamese cats. In a very much more subtle way, this situation frequently occurs in operational systems; what is really happening is that a new question is being put to the system.

In single cases there may be exceptions to the general rule, particularly in the case where, although there is at least one, there are relatively few relevant documents. In such a situation, the first sub-search may well fail to produce a relevant document, so at this stage the recall can only be described as 0% recall and 0% precision. The finding of a single relevant document in a later sub-search will obviously improve both relevance and recall so, for complete accuracy, it is necessary to add the qualification that the results of a number of searches should be averaged.

The final qualification "within a single system" is more difficult to discuss at present, for the question of what is a "single system" is fundamental to the project considered in this volume, for it could be said that we have been endeavouring to find how the changing of a component (e.g. any variable) in a sub-system (e.g. an index language) of a complete I.R. system can improve both recall and precision. This point also came to the fore in connection with the test results obtained by Professor Salton with the SMART Programme (ref. 30) where a number of different "options" -

(which correspond to the devices being investigated in this project) - are used. The question of exactly what constitutes a different system will therefore be discussed later.

Considered separately, certain conclusions drawn from Cranfield I may be difficult to justify, since it is possible for different interpretations to be placed on the evidence. Consider the matter of the relatively equal performance that could be obtained by the four systems. It has been argued, quite reasonably, that the unnatural relationship between the questions and their related documents was such that there was little difficulty in locating the source document by whichever method it was indexed, and that this was the reason for the level performance. The advantage of an experimental situation is that, in a well-designed test, it is reasonably simple for different hypotheses to be tested, and, by analysis of the search failures, it was simple to show that recall (which was the main objective in Cranfield I) is far more dependent on the concept indexing than on the index language. Therefore, since the concept indexing was in general the same for all four systems, the first step had been taken to ensure that the performance should be much the same for all four systems.

It is not intended to re-argue the conclusions listed above. It is sufficient to say here that they appeared reasonable as a basis for future work. All hinged on the twin factors of recall and precision. Why this should be the case has aroused a considerable amount of argument, and many different suggestions have been made regarding the criteria that are of importance in the evaluation of an information retrieval system. Bourne (ref. 7) presented a long list of such possible criteria and asked, "It is not clear why so much attention has been given to recall and relevancy. Should these be regarded as better criteria than any of the others proposed?"

We would suggest that all criteria fall into one of two groups. The first group, which we call user criteria, is made up of those factors which are of concern to the users of a system. Such criteria are related to the operational performance of the system and can be listed as follows:-

1. The ability of the system to present all relevant documents (i.e. recall)
2. The ability of the system to withhold non-relevant documents (i.e. precision)
3. The interval between the demand being made and the answer being given (i.e. time)
4. The physical form of the output (i.e. presentation)
5. The effort, intellectual or physical, demanded of the user (i.e. effort).

The second group is made up of criteria in which the ordinary user is not directly interested and which are therefore the sole concern of the managers of the system, that is to say all those who decide the policy, finance the system, or are in any way responsible for or participate in the actual operation of the system. The user is not normally concerned with the intellectual methods that are adopted to achieve a particular result, nor is he interested in the economics of the techniques used. Such matters are, however, of major concern to the management, but, on the other hand, they cannot be considered in isolation or as an end in themselves. It is a reasonable assumption that an I.R. system basically exists for the purpose of meeting the requirements of the user group, and any evaluation of management criteria must always be made in relation to the effect which they have on the user criteria. It cannot, for instance, be argued that one indexer is better than another without relating their indexing to the requirements of the users of the system.

To consider these five user criteria from the viewpoint of their evaluation, 'time' and 'presentation' offer few problems, for both are mainly influenced by management decisions concerning hardware. To find the time factor it is only necessary to record the time lapse between the request and the receipt of the output for a statistically valid number of cases. To evaluate the presentation, one has merely to observe whether the user receives a list of document numbers, a list of bibliographical references, a list of titles, a set of abstracts or a set of complete documents, either readable text or microform. To evaluate the effort demanded of the user in obtaining an answer to his query is only slightly more complex because of the possibility, in certain systems, that the effort can vary from the minimum of expressing the query in natural language to the maximum of conducting the complete search unaided in, for instance, a citation index. However, in any single system, evaluation of this point appears only a straightforward observation of a number of cases.

This only leaves recall and precision and the comment and the question by Bourne can now be answered. The reason why so much attention has been given to recall and precision is that these are the only two user criteria which demand any serious intellectual effort in their measurement. They are concerned with whether the system is capable of locating what is sought and are so fundamental that they can be said to be on a different level to the other criteria. Whether they are "better" than any of the other proposed criteria does not enter into the argument; it is certainly not suggested that they are the criteria which are always uppermost in the mind of a user. The unarguable fact, however, is that they are fundamental requirements of the users, and it is quite unrealistic to try to measure how effectively a system or a subsystem is operating without bringing in recall and precision.

Cranfield I had attempted, as its original objective, to establish the, at that time, generally accepted hypothesis that there were significant differences in the operational performance of various types of index languages, but this it had most definitely failed to do. It had appeared to show that all four indexing languages were operating at about the same level of recall performance; more positively, it had shown, by the analysis of search failures, that the decisions by the indexers in recognising significant concepts in the documents were far more important than any variations in the structures of the various index languages. The test of the Western Reserve University index appeared to indicate that there was an optimum level of exhaustivity of indexing, for a higher level of exhaustivity did not significantly improve recall but it weakened precision, while a low level of exhaustivity inhibited maximum recall. In these matters, the index language appeared to play a relatively insignificant part, for these were intellectual decisions by the indexer and were made in complete independence of the index language being used.

It was then realized that theoretically there was no reason why, given the same concept indexing, there should be any difference in the performance of two index languages. It was recognised that in practice the physical form of the index might affect the operating efficiency - and still more, of course, the economic efficiency - but theoretically there is a possibility of matching performance. To understand this, it is necessary to consider the fundamental aspects of index languages.

It should be made quite clear that we are concerned with index languages only in their theoretically perfect form; even in Cranfield I, we endeavoured to optimise each index language that was being used. Although in this process nothing was done which any person or organization using a particular index language could not equally well have done, this did not prevent a number of people from sending in critical comments on this score. To quote from some of the letters,

"You had no right to be so intelligent with the uniterm system; it is meant to be used by people of low intellect."

"The UDC had an unfair advantage because of the detailed alphabetical index which you compiled."

"If you had not used the colon device (of the UDC) so much, it would not have performed so well."

"Subject headings are not meant to be so specific as those you used, and that is why the alphabetical subject index performed so much better than it would normally have done."

Although such comments seemed amusing, they were understandable in that in 1961, the results coming from Cranfield I were contrary to firmly held beliefs, and the implications of the test results had not been appreciated. However, in a recent paper (Ref. 8) Richmond writes "... systems designed with a universal approach to the intellectual organization of information and those designed for limited use in parts of the whole. The former, when one comes to a specialized field like aeronautics, is a dilute approach, while the latter is a concentrated one. At Cranfield, the dilute approach was made through the UDC and through alphabetical subject headings, which are generalized concept terms. The concentrated one was made through a faceted classification, tailor-made for the subject and through uniterms, which had a vocabulary of words taken directly from documents dealing with the subject".

Here is shown the same categorical assertions as are contained in the earlier quotations, that the UDC and alphabetical subject headings are only for universal application,that they must not be used in a specialized subject field, and that if so used, they cannot possibly be as efficient as the "concentrated systems". The fact that all the experimental evidence is to the contrary appears to mean nothing, nor does the fact that probably 90% of the operational UDC systems are concerned only with a "concentrated" subject area. The UDC schedules used in Cranfield I were no exception, having been developed over a long period by workers in the United Kingdom concerned with highly specialised collections in the fields of aerodynamics and aeronautical engineering.

Again in the above quotation, there is the same confused thinking when it is said of the uniterm system that it has a "vocabulary of words taken directly from documents dealing with the subject," the implication being that the words found in the other systems had come from some source outside of the documents. This is, of course, untrue, for the facet classification, as is reported in ref. 1, was prepared by taking the terms used in the literature and arranging them in categories and facets. Equally so, there is no single term in the alphabetical index to the UDC or in the alphabetical subject headings which is not found in the list of uniterms, or in its lead-in vocabulary.

Unconsciously (because the significance of what was being done was not then realised) we were providing an additional basis for a similar performance in regard to recall by providing all four systems investigated with an equally effective lead-in vocabulary, which is the first basic requirement for all index languages. By 'lead-in vocabulary' is implied a complete list of all the sought terms including all necessary synonyms, that are used in the set of documents being indexed or in the set of questions that is put to the system. While some - in fact, probably most - operational index languages are deficient in this respect, this is an incidental as

apart from a fundamental characteristic, and whatever the type of index language, it can readily be provided with a complete list of sought terms, that is a "lead-in vocabulary".

The second requirement for index languages is a set of index terms, while a third requirement is a set of code terms. Before attempting to explain the differences, it must first be said that in many index languages there will be some terms which will occur in the triple role of a lead-in term, an index term and a code term. Further, all index terms must be lead-in terms, and frequently the set of index terms will be the same as the set of code terms. For examples of the three types of terms, the Thesaurus of the Engineers Joint Council can be considered (Ref. 27).

A lead-in term represents a concept which is described by another term than itself. This may represent a synonym, e.g. Speed use Velocity, or may be a subordination of a specific term to a more general term, e.g. Hexagonal use Shape.

Code terms are those terms which are actually used in indexing, examples being Velocity, Rotation, Engine noise, Jet engines.

Index terms are all Code terms, and additionally any combinations of Code terms which make up and express new concepts. For instance, the Index term 'Peripheral speed' is expressed by the use of the two Code terms Rotation and Velocity, while the Index term 'Jet engine noise' is expressed by the use of the Code terms Jet engines and Engine Noise.

While these three types of terms, i.e., lead-in terms, index terms and code terms, are normal ingredients of an index language, most index languages also make use of auxiliary devices or aids. In a completely simple system, lead-in terms would always be the index terms and the code terms, which is to say that terms would be used exactly as they appeared in the literature. As soon as the set of index terms is fewer in number than the set of lead-in terms, then a measure of control has been introduced. This normally takes the form of combining terms which are synonyms, and is only the first of many devices which are used in various ways to make up different index languages. There is nothing exclusive about such devices which restrict their use to any particular type of index language; precoordinate or post-coordinate, alphabetical or classified, any type of index language can potentially be given the same devices and thereby have the operational performance of any other index language.

In his book "On retrieval system theory", (ref. 9), Vickery identified seventeen devices, and acknowledgement must be made that in the original project proposal, these formed the basis of our argument. Vickery lists these devices as follows.

Means of control	Field of use
1. No control.	Some amateur alphabetical indexes.
2. Rigid control - fixed vocabulary of descriptors.	Some mechanized systems with limited coding capacity.
3. Confounding of variant word forms.	Professional alphabetical indexes, including Uniterm, and most other systems.
4. Confounding of true synonyms.	Ditto.
5. Confounding of near synonyms.	Some subject heading lists, some classifications, and systems based on thesauri.

Means of control	Field of use
6. Generic descriptors.	Many mechanized systems.
7. Specific and generic descriptors linked hierarchically.	Classifications, thesauri, some subject heading lists, some mechanized systems.
8. Multiple generic links for each specific descriptor.	Some classifications, subject heading lists, and thesauri, a few mechanized systems.
9. Categories of descriptor, forming facets.	Faceted classifications, some mechanized systems.
10. Semantic factors to represent subject terms.	To some extent in faceted classification, the W.R.U. system, mechanized patent office systems.
11. Correlation of descriptors.	Many alphabetical indexes, some classified catalogues, all mechanized systems.
12. Weighted descriptors.	Some experimental computer systems.
13. Interlocking sets of descriptors.	Alphabetical indexes, classified catalogues, computer systems.
14. Regulated sequence of descriptors.	Alphabetical indexes, faceted classifications, fixed-field punched cards, some computer systems.
15. Interfixing descriptors.	Mechanized patent office systems.
16. Role indicators.	Some faceted classifications, some mechanized systems.
17. Relational terms.	Alphabetical indexes, some faceted classifications, some mechanized systems.

All the results of Cranfield I pointed to only one conclusion. Whereas one could evaluate the performance of an operational information retrieval system and find how the index language being used affected the performance of the particular system under investigation, it was not possible to do any basic research on index languages by this method, for there are so many uncontrollable variables in any operational system that comparison of index languages is impossible.

It has to be admitted that this view is not generally held, since one finds a new investigation which has the objective of comparing various UDC operational systems with other operational systems using different types of index languages. In that this results in even more variables than existed in Cranfield I, it is difficult to see how any valid data concerning the UDC can be obtained. On this point Richmond is in complete agreement, for she writes (ref. 8) "System evaluation by comparison testing is essentially a negative operation", and again, "Comparison with other systems does not answer problems arising from the weaknesses of this system. In each case, the faults are internal and only obliquely subject to evaluation by comparison with other systems".

To make advances in knowledge regarding index languages, what was now required was a laboratory-type situation, where, freed from the contamination of operational variables, the performance of index languages could be studied in isolation. While such an approach was unusual in 1961, at least two other organizations have also established similar conditions, namely the Centre for Documentation at Western Reserve University and the Computation Laboratory of Harvard University. The methods used at Cranfield to establish this situation are considered in the following chapters of this volume.