ASLIB

CRANFIELD RESEARCH PROJECT

FACTORS DETERMINING THE PERFORMANCE OF INDEXING SYSTEMS

VOLUME 1. DESIGN

by

Cyril Cleverdon, Jack Mills and Michael Keen

Part 1. Text

An investigation supported by a grant to ASLIB
by the National Science Foundation

E. M. Keen

ASLIB CRANFIELD RESEARCH PROJECT

FACTORS DETERMINING THE PERFORMANCE OF INDEXING SYSTEMS
VOLUME I. DESIGN

by

Cyril Cleverdon, Jack Mills and Michael Keen

Part 1. Text

An investigation supported by a grant to Aslib
by the National Science Foundation

Cranfield
1966

ACKNOWLEDGMENTS

The project reported in this volume followed on from the first Aslib-Cranfield project, and was again financed by a grant from the National Science Foundation. The support of the Foundation was, as usual, not restricted to financial assistance, and I am most grateful for the advice and encouragement which we received, particularly from Mrs. Helen Brownson.

The full-time members of the project group were Mr. Jack Mills, from September 1962 to August 1965, Mr. Wilfrid Lancaster in the year of 1963, and Mr. Michael Keen, who commenced in January 1963 and is still engaged on the final stages of the project. In addition, some sixty-three other persons have worked part-time at some stage. To all these people, I have to express my appreciation for their efforts.

An essential requirement of the project involved co-operation of a large number of research scientists. The response to our request was most satisfactory, and I acknowledge with thanks the generous assistance of some two hundred scientists, many of whom are known to me only by name.

As before, Aslib administered the grant and also, on this occasion made accommodation available in their headquarters in London, and I am grateful for the help given by the Director, Mr. Leslie Wilson, and many members of his staff.

I would also express my appreciation to the Principal and Senate of the College of Aeronautics for agreeing to my taking part in this project while continuing my normal duties.

Finally, there are many friends and colleagues with whom, during the past three years, I have had the opportunity of discussing the Aslib-Cranfield projects. Their comments and suggestions have always been helpful, and I am most grateful for the interest which they have shown.

CONTENTS

PART 1

LIST OF TABLES AND FIGURES

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.

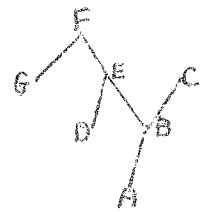
Possibly the point which has attracted most attention and criticism has been in regard to the assertion that there is an inevitable inverse relationship between recall and precision. This, in other words, implies that if an attempt is made to retrieve

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.

Chapter 2
TEST DESIGN

There has been a considerable amount of comment during the past few years about test design in general and the test design for Cranfield I in particular. That much of this has been, unfortunately, misinformed has been due both to a failure to appreciate the basic problems and purposes of an evaluation test, and also to a failure to distinguish between two main types of testing.

The first type of testing is that which is concerned with the evaluation of an operational information retrieval system, a sub-system of an operational system or a system or sub-system proposed for an operational system. In all such cases, there is no basic intention of advancing knowledge concerning information retrieval systems in general, although in the present state of fragmentary knowledge, this may well be a by-product. Basically such a test is designed to provide data for an analysis to be made which will show how the system can work more efficiently either in regard to operational or economic factors, in supplying the particular requirements of a given body of users. Such a test was that performed by Lancaster on the index of the Bureau of Ships (reference 5). Well designed on the basic Cranfield test procedure, with defined limited objectives, it produced, economically and quickly, data which enabled decisions to be taken on the optimum methods for the information retrieval system at the Bureau of Ships. As a 'research' pay-off, it revealed yet another situation where the use of roles was economically inefficient and operationally of doubtful value, and added to the growing body of data on the problems created by the use of roles of the type proposed by the Engineers Joint Council, in the Thesaurus of Engineering Terms.

There are many different variations of this type of test situation. One can, for instance, devise a new system or sub-system and test it while it is still comparatively small as effectively as one can test the performance of a long-established operational system, but the characteristic of all such tests is that they are made with a given situation in mind, their parameters are fixed by the pre-determined environment of the system being evaluated.

The second type of test - the type with which this report is concerned - is where one is dealing with an experimental situation. In such a case, the purpose of the test is to advance knowledge in some aspect of information retrieval without any particular operational requirement in mind. For this to be done, it is necessary to advance from a firm foundation of what is known. To make such an advance may require the use of unproved techniques, and, since the attempt is being made to investigate the unknown, there is always the possibility that, however meticulously the test has been designed, some unexpected factor will interfere with the objective of the test. If such a factor can be recognised early enough, it may be possible to adjust the design to take account of the new situation, but the risk has to be accepted that the weakness may only become apparent towards the end of the test.

A classical example of such a situation was the test carried out by Documentation Inc. Inc., where the objective was to compare the performance of a Uniterm index and the alphabetical subject catalogue compiled by the Armed Services Technical Information Agency. The first stage of the test involved the indexing of 15,000 documents by the Uniterm system, at the same time as they were also being indexed by the ASTIA staff. The second stage was for the two groups to carry out searches in their indexes for some ninety odd questions and then for each group to analyse the output of their searches to find which documents were relevant. Up to this point, everything appears to have

gone according to plan. The final stage was intended to be a comparison of the output of the two sets of searches, in order to find which system had been successful in obtaining more relevant documents.

The problem which arose at this final stage was that neither group was willing to accept the relevance assessments of the other group; rumour has it that at the end of the second day of discussion, the two groups were still arguing about the meaning of the first search question. No real blame can be fixed on those who organised the test; in 1952 it was not unreasonable to think that two groups of intelligent people would, without serious difficulty, be able to come to an amicable agreement as to which documents were relevant to a particular question. If any fault can be found, it only lies in the failure to make generally available either of the two reports which are said to have been prepared by the two groups taking part in the test. The only published account was a brief paper by Gull which appeared some years later in American Documentation (reference 10), and which dealt mainly with the results of the searches. Gull does, however, make the following very apt comment: "When one considers that a fairly thorough search of the literature indicates that this comparison of two reference systems is the first undertaken so far, it is not surprising that the results revealed clerical errors and an incomplete design of the test."

With the exception of a small test done in 1953 by Cleverdon and Thorne (ref. 11), this had been the only test of an I.R. system carried out before the test design for Cranfield I was prepared in 1956. While access to the complete reports of the ASTIA-Uniterm test might have revealed some more information, the only positive fact known in 1956 concerning test design of I.R. systems was that failure to have a firm agreement on question-document relevance could result in complete failure to realise the test objectives. Concerning information retrieval systems, however, nothing was known for certain. For any belief categorically stated by one expert, it was possible to find the exact opposite stated by another expert. Those were, in fact, the halycon days when one could argue all night without producing a shred of evidence for one's views, when Metcalfe, for instance, could write a fascinating book (ref. 12) proving in three hundred pages that an alphabetical subject catalogue was vastly superior to a classified catalogue without having to, or being able to, present one piece of experimental data to support any of his many assertions.

The field of investigation for Cranfield I was therefore wide open, in the sense that it would prove or disprove some conflicting beliefs. Since it was uncertain as to what was of major importance, the decision was deliberately taken to plan the test over a wide range of aspects. Not only index languages but qualifications of indexers, indexing time, categories of documents, search tactics and search capability, optimistically (over-optimistically some might argue) all were incorporated in the test design. Any knowledge would be new knowledge and there was practically no limit to what could be attempted, although there were certainly definite but unknown limits as to what could be achieved. From a personal viewpoint, however, one limitation was essential in the design; actual questions could not be used if these involved relevance assessments by other people than the questioners. This restriction had to be accepted, and the result was the adoption of the technique of using prepared

questions based on source documents. Although this technique has been strongly attacked in many papers, no-one has suggested any other method which would have permitted so much reliable data to be obtained so economically.* However, by the time the design of the present project was being considered, the position had changed radically. The conclusions coming from Cranfield I, supported by other smaller investigations, had deliniated more sharply the problem areas for investigation; equally important was the realization that progress would be dependent on the use of more refined test methodology.

As outlined in the previous chapters, the new project was to deal with index language devices; the first objective was the precise measurement of recall and precision ratios. The essential prerequisite to obtaining these measures (in an experimental situation) is the determination of the sets of documents which are and are not relevant to each of a set of test questions. Before proceeding to discuss the various ways of determining this matter, it may be helpful to consider a recent paper by the late Dr. Taube 'The pseudo-mathematics of relevance' (ref.13), which is being widely quoted as discrediting the results of the Cranfield investigations.

Any paper by Dr. Taube merited serious consideration, and in particular any paper dealing with the question of relevance, since this was the critical problem in the original test carried out by Documentation Inc. While the paper presents what at first sight appears to be a plausible argument, it is, in fact, based upon a confusion and distortion of meaning of two uses of the term 'relevance'. First there is the use of the term on its own where it denotes, in a true life situation, the subjective assessment of an individual in relation to a document or a set of documents which he receives in answer to a search question, so that he says "these documents are relevant to my questions, those other documents are not relevant". The second use of the term is in 'relevance ratio', which is the manner of expressing the proportion of relevant documents retrieved to the total of documents retrieved in a search. As such, 'relevance ratio' has nothing to do with the determination of relevance, but merely involves a numerical calculation of those documents which have been previously allocated to one of the two sets of relevant and not relevant.

At a meeting in Washington in 1964 of a group of some thirty people concerned, to a greater or lesser degree, with evaluation of I.R. systems, the paper in question, (which was originally written in March 1964) was amongst the documents circulated. Since it was clear from the discussion that Dr. Taube was still confusing the two meanings, Cleverdon agreed that in future we would cease to use the term 'relevance ratio' and substitute another term. Possible alternatives were 'acceptance rate' or 'precision ratio', both of which were being used by other groups with the same meaning as 'relevance ratio'. As stated earlier, 'precision ratio' was selected, and if one substitutes this term in those cases where Taube

---

*In these days when large grants are common for small investigations, it is of interest to recall that the five years' work of Cranfield I, including the test of the Metallurgical Index of Western Reserve University, was covered by two grants from the National Science Foundation, totalling $44,000.

Bibliography #: 453 Title: Use of Fluorine for Rocket Propulsion

A. Terms/Hits

a. Total Search Terms	37
b. Maximum Hits Possible	906
c. Anticipated Hits	500

B. Most Heavily Posted Terms:

	Terms	Postings
1.	PROPELLANT	2830
2.	FUEL	1765
3.	OXIDIZER	384
4.	FLUORINE	300
5.	FLUORIDE	259

C. Type of Logical Equation Specified:

____ a. Loose. High output. Irrelevant material expected.
____ b. Moderately loose. Some irrelevant material.
X c. Moderately tight. Very little irrelevant material.
____ d. Tight. No irrelevant material expected.
____ e. Analog. Analog measure: ____

D. Initial Search Results:

a. Hits (Total Output = T)	441
b. Accepted Hits After Editing (Accepted Accessions = A)	379
c. Acceptance Ratio, A/T x 100 =	86.5 %

E. Auxiliary Search Results:

a. Hits (T')	10
b. Accepted Hits (A')	10

F. Reject Analysis:

a. Rejects on Initial Search	63
b. Rejects Attributed to Type of Equation, i.e., out-of-scope or marginal upon examination	62
c. Rejects Attributed to "Noise", "False Drops", etc.	
d. Other Rejects, e.g., Indexing Errors	1

e. Total Rejects Considered
Excessively high ___ High ___ Average ___ Low X

G. Miss Analysis:

a. Misses detected, overall 2
b. How were misses detected? STAR CUMULATIVE INDEXES — PUBLISHED SUBJECT INDEXING

H. Analyst's Comments and Recommendations (i.e., search strategy, reject analysis, new terms, delete and transfer suggestions, indexing errors, etc.):

1. THE 62 REJECTS ALL DEALT WITH FLUORINE COMPOUNDS AND FLUOROCARBONS, BUT WITHOUT SUFFICIENT RELATIONSHIP TO THEIR USE IN PROPELLANTS.

2. AUXILIARY SEARCH: CHLORINE TRIFLUORIDE AND OXYGEN DIFLUORIDE (OXYFLUORIDE) WERE SEARCHED INITIALLY UNDER THEIR "PRE-COORDINATED" NAMES. THE AUXILIARY SEARCH UTILIZED (1) CHLORINE INTERSECT FLUORINE UNION FLUORIDE, AND (2) OXYGEN INTERSECT FLUORINE UNION FLUORIDE. INDEXERS SHOULD BE ENCOURAGED TO USE "PRE-COORDINATED" PROPELLANT NAMES.

3. TERM "OXYGEN DIFLUORIDE" SHOULD BE USED IN LIEU OF "OXYFLUORIDE".

4. N62-10209: DELETE POSTING UNDER "FUEL"

5. N63-10406, N64-11223: POST UNDER "PROPELLANT".

FIGURE 2.1 NASA SEARCH SYSTEM ANALYSIS SHEET

used the term 'relevance' with this meaning, it is immediately apparent that the whole argument is defective. The argument in the paper starts with a quotation from a Cranfield paper written before this decision to change to the term 'precision ratio' had been taken. Substituting this term, but not in any way changing the original meaning, we would now have written, "With the aid of the set of documents and the set of questions [for which the document/question relevance assessments have been previously made by the questioner] it will be possible to test each index language device in turn and so get precise figures for the effect on recall and precision ratios."

Taube's comment on this was 'some way or another a vague or hardly recognisable and admittedly difficult notion [i.e. relevance] has turned out to be precisely measurable". It is not, of course, relevance which is being measured, but the decisions regarding relevance which have already been taken. As Salton says (Ref. 14), "once acceptable relevance judgements are available for all documents with respect to all search requests, the calculation of recall and precision becomes perfectly straightforward and unambiguous."

It is interesting to find, in the issue of American Documentation for April 1965, that there is a brief note (ref. 15) by two members of the staff of Documentation Inc., in which they discuss a NASA Search System Analysis Sheet. The example which they presented has been reproduced on page 12, and from this it can be seen that these members of the staff of Documentation Inc. have been able to derive, for this particular search, an acceptance rate (i.e. precision ratio or relevance ratio) of 86.5%.* It is interesting to note that, on the Analysis Sheet, the phrase used is 'accepted hits after editing'. This implies that the determination of the relevance of the document to the question has been by a member of the staff of Documentation Inc., and his standard for relevance might be very different from that of the questioner. This leads us back to the point we had reached before the diversion to consider briefly the matter of relevance. As we argued earlier, there were sound, compelling reasons for the use of source-document questions in Cranfield I, because they gave, simply and economically unequivocal relevance assessments. More particularly, it still remains probably the most effective and economical method of establishing the general recall ratio in many test situations. By 1961, however, it was quite unacceptable for an experimental investigation of the type we had in mind. What were the alternatives? These can most simply be tabulated under various aspects as follows.

Types of search questions

1. An actual question that is put to an information retrieval system and searched at the time it is required.

2. An actual question that has been put to an I.R. system. In other words, one obtains questions that have been used previously, either with the system being tested or some other system.

---

*To save misunderstanding, we would point out that an error has been made in calculating this figure. It should, of course, be 85.9%.

3. A prepared question, that is a question which has been composed specifically for the purpose of the test and is not a question which meets an actual need of the questioner. Such prepared questions may or may not be based on a particular document or documents.

Method of Relevance Assessment

I By the questioner
II By the consensus of opinion of a group of people
III By an individual, not the questioner
IV By matching the indexing with the search programme.

Type of Individual(s) Involved

A User of a system
B Scientific or technical staff, not users of the system
C Librarians or other information staff.

If we now chart Type of Question against Method of Relevance Assessment, the various possibilities can be shown

Method of Relevance Assessment

Type of Question	I	II	III	IV
1	A / A	A / BC	A / BC	A / -
2	A / A	A / BC	A / BC	A / -
3	ABC / ABC	ABC / ABC	ABC / ABC	ABC / -

In the chart, the upper half of each box represents the type of person asking the question, the lower half represents the type of person making the relevance assessment.

An additional variable concerns the type of document on which the reference decision is based, for this can be either

α The complete text
β An abstract
γ The title.

It can be seen that the Documentation Inc. example discussed above was, presumably, the use of an actual question (1A) where the relevance assessment was made by an individual, not the questioner (III) who was a member of the information staff (C), probably basing his decisions on document titles, making up the code (1A)(IIICγ). For Cranfield I the code would have been (3B)(IBα), which is to say that prepared questions were used (3), based on complete documents (α), this resulted in the relevance being determined by the questioner (I) and the individuals involved were technical staff not concerned with the system (B).

The theoretical ideal is (1A)(1Aα) that is the use of actual questions with a relevance assessment made at the time by the questioner from complete texts. This cannot be achieved in an experimental situation since there is no body of users who can ask questions, nor would the experimental collection normally be of sufficient size to justify actual searches. For this project, it was considered that the nearest to the ideal would be the combination (2B)(1Bβ+), that is questions which had been asked, with a relevance assessment being made by the questioner who would be a scientist. (β+) implies that nothing less than abstracts would be used; the expectation would be that full texts would also be used. The wisdom and implications of this choice will be considered in relation to the test results. What can be stated here is that the operational performance characteristics of the system being tested will almost certainly change depending on the combination of questioner and relevance assessor used, and care should be taken in interpreting figures which do not define how they have been obtained in this respect. A few illustrations of what can happen may help to clear up this point. In the Documentation Inc. example previously quoted, the precision ratio of 86.5% is very high. A probable reasin is that it is based on the relevance assessment of a member of the information staff; when the set of documents is sent to the questioner, his relevance standards may be such that he will grade the large majority as non-relevant, so the relevance ratio would then drop considerably.

As another example, in a report of the evaluation of the EURATOM information retrieval system (ref. 16), a precision ratio of 65% is given. The key to this high figure is in the following sentence taken from the text of the paper. "Finally, the computer's answers have to be checked, since it would be unreasonable to expect them to be 100% complete and correct".

What has happened in this case is something rather different. The precision ratio is not being calculated on the actual search output but on the search output after technical information staff have rejected the documents which they considered non-relevant. A somewhat similar reason was the cause of some confusion at the NATO Advanced Study Institute on evaluation of information retrieval systems, when Altmann, in presenting the results of a test on the information retrieval system of the Harry Diamond Research Laboratories (ref. 17) gave figures of 80% for precision ratio. In this case, it appeared that the procedure was for the questioners, who were also making the searches, to eliminate documents which, from title or abstract, appeared to be non-relevant; this maybe gives interesting information about the ability of users to eliminate non-relevant information on the basis of the title but, as with the EURATOM test, gives no information at all on the performance of the system in regard to precision.

The discussion so far has been dealing with precision ratios; while there is still considerable doubt as to the most useful way, in an experimental situation, of obtaining relevance assessments, once that assessment has been made the determination of precision ratio is a straightforward matter. The same is not, however, true of recall ratio, because this is dependent on the number of relevant documents which have not been retrieved. This problem was effectively side-tracked in Cranfield I by the use of source-document questions; since this method had been ruled out for the present test, there was only one apparent alternative, namely to look at every document in relation to every question. This decision automatically placed a restriction on the size of the test collection and the number of questions to be searched. This was not considered a serious handicap, since the W.R.U. test had shown that a collection of only one thousand documents was sufficient to provide a considerable amount of data for analysis. There seemed to be some advantage in having a larger number of questions

in relation to the number of documents in the collection than had previously been used, and the decision was to aim at 1,200 documents with 300 questions.

There was no readily available collection of questions which had actually been used on some previous occasion. Even if there had been, it would not have been possible to have the originators of the questions check the documents for relevance. The method adopted, therefore, to obtain the documents and the questions was to select a number of recently published research papers, mainly dealing with high speed aerodynamics, but about 20% of which covered aircraft structures. The author of each paper was to be requested to provide the basic problem, in the form of a search question, which had been the reason for the research being undertaken, and also to give some additional problems which had arisen in the course of his work. At the same time he would be asked to state which papers in his list of references were relevant to the various questions he had provided. It was intended that the document collection would be made up of the papers that had been included as references.

'Relevance' is obviously a matter of degree. The problem in arranging for relevance assessments to be made is to decide how many degrees of relevance can be consistently recognised. In the test of the index of Western Reserve University, two levels of relevance were used; previously , Swanson (ref. 18) had attempted ten levels. The decision in this test was to use four levels of relevance; details of this and the whole procedure of obtaining the questions and document collection are given in chapter 3.

The references in any given paper might be expected to give a high proportion of relevant documents to any question arising in connection with that paper, but at the same time there was the probability that other documents in the test collection would also be relevant. The author might have known about these documents but have decided not to use them. Alternatively, he might not have been aware of their existence; possibly they might have been published after he had finished his work. While it was essential that ther should be a complete cross-check of every document and of every question, it was impracticable to send 1,200 documents to each of 200 or so authors for them to make the assessments individually, so a screening process was first necessary. This was to be done by recruiting a number of postgraduate students who would (hopefully) be able to eliminate most of the non-relevant documents for each question. Then it would only be necessary to send to each author those papers which had a reasonable possibility of being relevant, for each author to make a final decision concerning relevance.

We would forestall criticism of the method outlined above, by admitting immediately that it includes nothing which overcomes the basic problems of the meaning and determination of relevance. No-one is more aware that relevance is a shifting notion, certainly between individuals and often for the same individuals at different times. Is there, then, justificiation for the comments by Taube that any attempt to measure system performance is useless, since such measurement must be based on relevance decisions. We would strongly argue against this, for it is the very situation which an information retrieval system has to face. Users do ask questions and then accept or reject the search output in what might seem an arbitrary manner. The objective of the methods used in this test was to get as near as is possible in an experimental test to a true life situation in relation to relevance decisions. While they certainly represented an advance on the methods in Cranfield I, it is not intended to suggest that the design was perfect; again it is necessary to go back to the time when the test was designed, and say that in 1961 it appeared to be the best technique that could be adopted for the particular requirements. The experience of this test has shown not only its advantages, but also some disadvantages, and these are briefly discussed in Chapter 8.

So far the discussion on the test design has been entirely concerned with the methods to be used in obtaining a set of test documents and questions, and establishing the relationship between the documents and the questions. All such activity was an essential preliminary to the investigation itself, the general background of which was considered in the previous chapter. To summarize this briefly, we started from the belief that all index languages are amalgams of different kinds of devices. Such devices fall into the two groups of those which are intended to improve the recall ratio and those which are intended to improve the precision ratio. In other words, there are some devices which will always enlarge the class and thereby retrieve more documents, with the probable result that more relevant documents will be retrieved. On the other hand, the precision devices will always act in the reverse manner by narrowing the class, thereby retrieving fewer documents, with the probable result that some relevant documents will be eliminated. The purpose of the test was to investigate the effect which each of these devices, alone or in any possible combination, would have on recall and precision.

To enable this to be done, it was essential that it should be possible to hold everything constant except the one variable being investigated. The organization of the file, with its completed matrix of document/question relevance assessments, was the first step towards this. The next stage was to determine and fix, once and for all, the concept-indexing of the documents and the relationships of the concepts. By concept-indexing is meant the decision as to which concept and groups of concepts are significant from the viewpoint of retrieval. Such concept indexing can only be in the terminology of the document. As soon as there is any 'translation' of the document terminology to any kind of formalized language, then one of the index language devices must have been brought into use. Therefore the decision was to concept-index, at a high level of exhaustivity, the documents in the collection so that they might be translated into any type of index language which it was desired to test. Details as to how this was done are given in chapter 4.

The original proposal to the National Science Foundation contained the following statement. "At this stage it should be possible to decide which technique appeared to have the most satisfactory characteristic for adaptation to automatic indexing. Dr. J. O'Connor has explained the techniques which can be used to investigate methods of automatic indexing without actually using computers. (ref. 19). Our approach would be partly to investigate new techniques, but might as usefully be concerned with testing methods proposed by others and measuring the performance of such methods against the results from human indexing."

The possibility and the hope that the test collection could be used by other groups and provide direct comparison with the Cranfield results was partly responsible for the decisions concerning the indexing technique and also the searching method. This permitted starting from the absolute basic point of matching any actual word in the question with any term used in the concept-indexing and then to introduce all the devices by stages. It is agreeable to be able to record that it will be possible to compare the results of the Cranfield tests with two experiments using computers. In England, at the Cambridge Language Research Unit, the complete set of Cranfield indexing is being processed on the Atlas computer, and it will be possible to measure and compare the effectiveness of the 'clumping' process which Dr. Needham has been investigating(ref.29) In the United States, at the Harvard Computation Laboratory, a sub-set of the indexing has been processed by a number of the options of the SMART programme which Professor G. Salton has designed. There is particular interest in this work, in that,

in addition to the searches based on the Cranfield indexing, searches have also been made on the abstracts taken from the documents. This work is discussed in more detail in Chapter 7.

It was, of course, known that decisions would have to be made concerning the physical methods which would be used for carrying out the searches. Fortunately, no firm decisions were taken on this point; the methods ultimately used are discussed in Chapter 6.

Finally, there would be the necessity to present the results in a meaningful manner. The recall/precision ratio figures and curves of Cranfield I have undoubtedly taken a hammering over the past few years, and there are many who have sought the elixir to change them into the pure gold of a single figure. Far from being able to do this, it was by 1961 clear to us that, if there was to be any comparison of experimental results, it was necessary first to investigate the effect on performance of the generality ratio, namely the relationship between the number of relevant documents and the size of the collection. The first tentative ideas on this had been put forward on page 101 of Ref. 2; in this project it was planned to attempt to measure the effect of this factor on recall and precision.

CHAPTER 3

Documents and Questions

To provide the necessary basis for the test, we required a collection of documents, a set of search questions, and a complete assessment to determine the documents relevant to each question. These aims were accomplished in three main stages:

Stage 1. A letter was sent to authors of research papers, requesting search questions and a relevance assessment of the papers they cited.

Stage 2. Using the collection of documents and a set of questions made up from the replies to stage one, technically competent people examined every document in relation to every question to find any relevant documents in addition to the authors' cited documents.

Stage 3. The additional documents judged relevant in stage two were submitted to the authors, requesting their final assessment of relevance.

First will be given details of the methods used in these three stages, and the response made by the authors. Then will follow a more detailed examination of the question-document assessment of relevance, and finally a brief analysis of the questions.

Methodology and authors' responses

271 recent papers on the subject of high speed aerodynamics and aircraft structures were obtained. Although high speed aerodynamics had been chosen as the main subject for the test, a small set of documents dealing with aircraft structures was introduced in order to examine the effect of including two dissimilar subjects in one collection. These papers were referred to as base documents, and in order to be accepted for the test a base document had to satisfy certain criteria; it had to be a paper published in the English language containing at least two references in a bibliography, these references being in English, dated 1954 or later and likely to be readily obtainable. Since aerodynamic papers contain on average about twelve references, neither this nor any of the other requirements caused the rejection of many papers. Most of the selected papers were published during 1962, and the first half of 1963; the articles from one prominent journal predominated, but some research reports were included. A list of the different sources of those which were finally used is given in Table 3.1. 76.9% of the papers are American publications, 22.5% British and 0.6% Swedish.

To the author of each of these papers was sent a form, quoting the title and reference of his own paper, and also listing up to ten of the papers which had been included as references. The authors were asked to do two things.

1. To state the basic problem, in the form of a search question, which was the reason for the research being undertaken leading to the paper, and also to give not more than three supplementary questions that arose in the course of the work, and which were, or might have been, put to an information service.

	Total
U.S.A.	
Journal of the Aerospace Sciences (later A.I.A.A. Journal)	102
National Aeronautics and Space Administration Technical Notes	38
Great Britain	
Royal Aircraft Establishment Reports and Notes	22
Aeronautical Research Council Papers	6
National Physical Laboratory Reports	3
National Gas Turbine Establishment Reports	1
Southampton University Reports	1
College of Aeronautics Reports	3
The Aeronautical Quarterly	3
Journal of the Royal Aeronautical Society	2
Sweden	
Aeronautical Research Institute Reports	1

TABLE 3.1 BIBLIOGRAPHICAL ORIGIN OF BASE DOCUMENTS USED IN THE TEST

	Totals
Australia	1
France	1
Great Britain	49
India	1
Israel	2
Japan	3
Sweden	1
Switzerland	1
United States	123

TABLE 3.2 COUNTRY OF RESIDENCE OF AUTHORS OF BASE PAPERS

	U.S.A.	G.B.	Other
AUTHORS' COUNTRY	67.6%	26.0%	5.5%
	(123)	(49)	(10)
COUNTRY OF PUBLICATION	76.9%	22.5%	0.6%
	(140)	(41)	(1)

TABLE 3.3. COMPARISON OF AUTHORS' COUNTRY OF RESIDENCE AND COUNTRY OF PUBLICATION

2. To assess the relevance of each of the submitted list of papers which had been cited as references, in relation to each of the questions given. The assessment was to be based on the following scale of five definitions:

(i) References which are a complete answer to the question. Presumably this would only apply for supplementary questions, since if they applied to the main question there would have been no necessity for the research to be done.

(ii) References of a high degree of relevance, the lack of which either would have made the research impracticable or would have resulted in a considerable amount of extra work.

(iii) References which were useful, either as general background to the work or as suggesting methods of tackling certain aspects of the work.

(iv) References of minimum interest, for example, those that have been included from an historical viewpoint.

(v) References of no interest.

An example of a completed sheet was included with each letter; this, the covering letter and example of material sent, are shown as Appendix 3 A.

It was originally expected that half the authors would complete the form to our requirements, and that there would be an average of two questions with each reply. During early March, 1963, 82 letters were sent out and by the end of that month 47 replies had been received with an average of $3\frac{1}{2}$ questions. Further letters were despatched up to the middle of July, and then later one chase letter was sent to those who had not replied. By the end of September we had received the excellent response of 182 completed forms of the 271 sent (67.2%). Some authors wrote to say that they could not spare the time; many other letters were returned because change of address prevented delivery. The authors continued to supply an average of $3\frac{1}{2}$ questions, and the total of those received was 641.

Most of these authors, 67.6% lived in the U.S.A., with 26.9% in Great Britain and 5.5% in other countries. Table 3.2 shows the figures from each country, based on the 182 authors with whom we corresponded. A complete list of the authors is given in Appendix 3F. It is an interesting sidelight on publishing habits to notice that eight of the British authors published in American sources, and nine out of ten of the other foreign authors did the same, but all the authors residing in the U.S.A. published there. Figures are given in Table 3.3. Some of the authors had changed their country of residence by the time of the test, and the figures are based on the country of residence in which their particular research paper was written.

As the forms were being received, the document collection was being made up, and 1,018 unique documents resulted from the cited papers. The base documents themselves were also included in the collection, adding 173 more documents (9 were already included as cited papers), but in order to avoid any possible bias in the results, these base documents are always completely deleted from the results when the questions to which they gave rise are being tested, 209 further documents, taken from similar sources, brough the whole collection to its final 1,400 documents. For the indexing, which was proceeding during this time, single xerox copies of the documents were made. Full bibliographical information concerning the document collection is given in Appendix 3C.

To prepare for the next stage, 361 of the 640 questions were selected for use in the test. The basis for this selection was questions that had two or more documents assessed as relevance grade 1, 2 or 3, and questions that were grammatically complete were selected first. Some questions were received abbreviated, although the missing idea was quite clear from another of the author's questions. For example,

Q.247 when received was worded "Can the hypersonic similarity results be applied to the technique". By examination of the other supplementary and basic questions, (Q.13, Q.12) it is seen that the technique under investigation is methods for predicting surface pressures of an ogive forebody at angle of attack, so question 247 was rewritten to include this. When, as in the example, the meaning was quite obvious, we inserted the missing words, and the re-submission of the question to the authors in stage three revealed no disagreement with the amendments.

The next task was to find whether there were in the collection documents other than those which had been in the list of citations, which were also relevant to any of the questions. This was done by examining every document in relation to every question, noting any new documents that were judged as possibly relevant, and then submitting these documents to the original authors for their final assessment of relevance.

The task was performed by students, with a knowledge of aerodynamics, who were engaged in post-graduate study at the College of Aeronautics. Over 1,500 man-hours of effort during the 1963 summer vacation were put in by five people. The job involved in theory over half a million individual judgements, and was an extremely onerous task. The questions were supplied on individual slips, with space given for recording the file number of any document judged as relevant. Access was also given to the original forms giving all the questions supplied by the author, the source document, and the authors' relevance assessment of the cited papers. Details of the document collection were supplied in the form of typed sheets, listing the documents in file order, and giving authors, titles and bibliographical details. Complete copies of all the documents were readily available to the students.

The ultimate procedure adopted was to work on sections of the document collection, ranging from 100 to 400 documents, depending on the number of people working at the same time. The questions were first sorted into broad subject groups, and small batches of very similar questions were done together. Thus some of the prominent features and subject areas of sections of the documents were soon committed to memory, to assist fairly rapid scanning of the document lists. The document titles were examined first, and any documents that could remotely contain material connected with the question were recorded on the question slip, so that at the end of a 'scan' of the titles, the documents themselves could be examined. The students were instructed to be quite liberal in their judgements, and to include documents that they considered were only possibly relevant. An initial attempt was made to grade their decisions for relevance, but this was found to be too difficult to do consistently, and so was given up.

The task was tedious, particularly for people of intellectual capability, but 361 questions were finally completed. Those who carried out the task would not claim to have found every possibly relevant document, since question interpretation would not always agree completely with the authors' real need, and since human error was inevitable. Some figures giving information on the number of relevant documents missed by the students is given later in this chapter. Documents judged as relevant, which really were not, did not cause any difficulty, since the original author of the question was taken as the final arbiter of relevance. For 86 of the 361 questions, no other documents were considered to be relevant; for the other 275 questions, there was found at least one document judged as possibly relevant, with an average of 3.3 per question.

When submitting these documents to the authors, it was decided to add some extra documents which had been suggested as a result of a test of the questions by the technique known as bibliographic coupling. A description of the processing of the cited papers in the documents of the collection, which resulted in a citation index and bibliographic coupling groups, is given in chapter 7. In the theory of bibliographic coupling, as worked out by Dr. M. M. Kessler, (Ref. 20) it is shown that, as the coupling strength increases, so also does the probability of the document being relevant to the question. It was therefore decided to include all documents retrieved by bibliographic coupling at a coupling strength of 7 or more (i.e. documents that had seven or more references in common with one of the author's cited relevant papers of grade (1), (2) or (3)). Of the 213 documents produced in this way, only the unexpectedly small number of 15 had already been assessed as possibly relevant by the students. The balance of 198 were submitted, along with the student assessed documents, in the second communication to the authors. This time the authors were requested to do three things; for reasons considered later.

1. To make a relevance assessment of the new documents submitted, in relation to their search questions, using the same relevance scale as before.

2. To examine the selected questions (which they themselves had originally asked), and to indicate the relative importance of each term or concept in the question by marking with a 'weight' from the following scale:-

(i) A paper that did not cover this term would be of no use.
(ii) It is desirable that this term should be covered by the document.
(iii) This is a term which is not absolutely essential to the enquiry.

3. To list any alternative terms or concepts that might be used in a search programme for the questions and, if necessary, to include a completely rephrased version of the question.

A xerox copy of the questions as he originally wrote them was sent to each author, together with a list of the new documents submitted, giving authors, titles and bibliographical references. Against each such document submitted was indicated the question to which the document was thought to be relevant, and to assist the relevance assessment a xerox copy of each document abstract was included. Each of the questions was re-submitted on a separate sheet, with space provided for alternative words to be added, either against each single term, or the concepts of the questions. Examples of the above are given in Appendix 3B.

Most authors received a total of at least eleven sheets for examination, which together with the abstracts of the documents submitted, made a somewhat daunting package. In spite of this, 144 out of 182 authors (79.1%) returned completed forms, with yet others being unable to help and some having changed addresses as before. Our main concern was to obtain the relevance assessments, which were needed for 283 of the questions and the authors' responses provided assessments for 201 of these. 78 questions had not been resubmitted to the authors because no possible relevant documents had been noted; adding these to the 201 questions where the relevance assessments had been completed meant that there was a total of 279 questions which could be used. This fell slightly short of the 300 questions originally planned; as will be considered later, we were by this time beginning to suspect that the test would provide more data than could be handled or would be required, and therefore no effort was made to bring the total number of questions back to 300.

Most authors included the weighting of the questions in their reply, over half of the questions had some alternative terms added, and 28 of the questions were submitted in rephrased form. (See Appendix 3B).

A summary of the position regarding the questions is as follows:-

1. Total of questions received	641
2. Questions discarded for various reasons	280
3. Questions matched against complete document collection for relevance ((1) - (2))	361
4. Questions having no additional relevant references	78
5. Questions resubmitted to authors for relevance decisions ..	283
6. Questions returned by authors from stage (5)	201
7. Questions available for test ((4) + (6))	279

The relevance assessments

The basic data on the authors' relevance assessments is given in Tables 3.4, 3.5, 3.6 and 3.7. These tables highlight various aspects of the relevance assessments, and the figures given are taken from the 279 usable questions obtained. In each table, the documents that were submitted to the authors are split into three categories:-

1. Those cited in the author's own original paper;
2. Those the students found and judged as being relevant;
3. Those retrieved by bibliographic coupling at a strength of 7 plus, and which were additional to the two categories above.

Each table also gives a figure for the total of all categories, the four divisions being shown as the left hand parameter in each table. The relevance assessments made are given in the body of the tables, these being split into several categories:-

1. Documents submitted (Tables 3.4 and 3.6)
2. Documents assessed as relevant, i.e. accepted:-
   (a) Totals (Tables 3.4, 3.5, 3.6 and 3.7)
   (b) Details of the four grades of Relevance (Tables 3.5 and 3.7)
3. Documents assessed as not relevant, i.e. rejected (Tables 3.4 and 3.6)
4. Total documents assessed as relevant expressed as a percentage of documents submitted. (Tables 3.4 and 3.6).

The figures given are in two forms in each table:-

1. Grand totals of documents, resulting from the whole set of questions involved.
2. Figures for one average question, calculated by the arithmetic mean. These averages are correct to one decimal place, but in a few cases a slight adjustment has been made to preserve the correct totals.

Tables 3.4 and 3.5 giving the figures for the whole set of 279 questions will be examined first. The bottom section of Table 3.4 shows that 3,087 documents were submitted to the authors of which 1,126 were rejected as not relevant, and 1,961 (i.e. 63.5%) were accepted as relevant. Table 3.5 gives a breakdown of the 1,961 documents accepted, showing that 171 were graded relevance (1), 461 were relevance (2),

Origin of documents	Submitted to the authors for assessment	Accepted as relevant	Rejected as non-relevant	% accepted as relevant
1. Cited in authors' papers	1972 (7.1)	1250 (4.5)	722 (2.6)	63.4%
2. Additional documents selected by students	917 (3.3)	592 (3.1)	325 (1.2)	64.6%
3. Additional documents by bibliographic coupling	198 (0.7)	119 (0.4)	79 (0.3)	60.1%
4. Complete total	3087 (11.1)	1961 (7.0)	1126 (4.1)	63.5%

TABLE 3.4 RELEVANCE ASSESSMENTS OF DOCUMENTS AS DECIDED BY AUTHORS IN RELATION TO THEIR SEARCH QUESTIONS

The total for all 279 questions is shown, with the average for each question in brackets.

Origin of documents	Relevant documents	Grades of Relevance 1	2	3	4
1. Cited in authors' papers	1250 (4.5)	158 (0.6)	348 (1.2)	492 (1.8)	252 (0.9)
2. Additional documents selected by students	592 (2.1)	12	97 (0.4)	344 (1.2)	139 (0.5)
3. Additional documents by bibliographic coupling	119 (0.4)	1	16 (0.1)	66 (0.2)	36 (0.1)
4. Complete total	1961 (7.0)	171 (0.6)	461 (1.7)	902 (3.2)	427 (1.5)

TABLE 3.5 GRADES OF RELEVANCE AS DECIDED BY THE AUTHORS

The total for all 279 questions is shown, with the average for each question in brackets. It will be noted that this table represents a breakdown of the figures as given in the second column of Table 3.4.

Origin of documents	Total submitted to the authors for assessment		Accepted as relevant		Rejected as not relevant		% accepted as relevant	
	Basic	Supp.	Basic	Supp.	Basic	Supp.	Basic	Supp.
1. Cited in authors' papers	820 (7.0)	1152 (7.2)	589 (5.0)	661 (4.1)	231 (2.0)	491 (3.1)	72.0%	57.3%
2. Additional documents selected by students	351 (3.0)	566 (3.5)	258 (2.2)	334 (2.1)	93 (0.8)	232 (1.4)	73.5%	59.0%
3. Additional documents by bibliographic coupling	85 (0.7)	113 (0.7)	59 (0.5)	60 (0.4)	26 (0.2)	53 (0.3)	69.4%	53.1%
4. Complete total	1256 (10.7)	1831 (11.4)	906 (7.7)	1055 (6.6)	350 (3.0)	776 (4.8)	72.2%	57.6%

TABLE 3.6 THE RELEVANCE ASSESSMENTS GIVING A COMPARISON OF BASIC AND SUPPLEMENTARY QUESTIONS

This table gives the same data as Table 3.4 except that the 279 questions are divided into the two groups of 118 basic questions and 161 supplementary questions, the figures in brackets representing the average for each question.

Origin of documents		Relevant Total	Grades of Relevance			
			1	2	3	4
1. Cited in authors' papers	(B)	589 (5.0)	12 (0.1)	159 (1.3)	273 (2.4)	145 (1.2)
	(S)	661 (4.1)	146 (0.9)	189 (1.2)	219 (1.4)	107 (0.6)
2. Additional documents selected by students	(B)	258 (2.2)	1	53 (0.5)	144 (1.2)	60 (0.5)
	(S)	334 (2.1)	11 (0.1)	44 (0.3)	200 (1.2)	79 (0.5)
3. Additional documents by bibliographic coupling	(B)	59 (0.5)	0	6 (0.1)	30 (0.2)	23 (0.2)
	(S)	60 (0.4)	1	10 (0.1)	36 (0.2)	13 (0.1)
4. Complete total	(B)	906 (7.7)	13 (0.1)	218 (1.9)	447 (3.8)	228 (1.9)
	(S)	1055 (6.6)	158 (1.0)	243 (1.6)	455 (2.8)	199 (1.2)

TABLE 3.7 RELEVANCE ASSESSMENTS GIVING A COMPARISON OF BASIC AND SUPPLEMENTARY QUESTIONS FOR ALL GRADES OF RELEVANCE

This table gives the same data as Table 3.5 except that the 279 questions are divided into the two groups of 118 basic questions and 116 supplementary questions, with the average for each question in brackets.
(B) = Basic question (S) = Supplementary question.

902 were relevance (3), and 427 were relevance (4). In terms of an average question, one can read off the figures as 11.1 submitted, 4.1 rejected, 7.0 accepted, and so on.

Examining the different origins of the documents in turn, the cited papers are seen to exceed all the other categories in size. From this group 4.5 documents per question were assessed as relevant; the additional groups of documents added another 2.5, making an average of seven relevant documents for each question. 63.4% of the cited documents submitted were accepted as relevant, and this seems satisfactory when it is remembered that all the references cited would not be relevant to all the questions given. In many cases some references are relevant to one of the questions only, and not relevant to the other questions at all. Table 3.5 shows that 14% of the relevant documents were graded as relevance (1), and some more details concerning this will be given when considering Table 3.7.

The additional papers that the students judged as relevant totalled 917. These are not, of course, 917 unique documents, as one document might be relevant to several questions. The acceptance rate was 64.6%, and this may be taken as a clue to the success of this difficult task, but further details are given when Tables 3.6 and 3.7 are examined, and when comment is made on the success of the students' judgements. Of the 592 accepted, only 12 (2%) were graded at relevance (1), so in most cases the authors considered these additional papers submitted were not as relevant as the cited ones about which they already knew.

The additional bibliographic coupling documents, submitted because they had seven or more of their references in common with the cited papers of relevance (1), (2) or (3), were only those which had not already been selected by the students as possibly relevant (see chapter 7). Table 3.8 shows that of the 312 documents retrieved by bibliographic coupling, 87 were cited papers and 12 were base documents; of the remainder only 15* had been selected by the students as possibly relevant, leaving a balance of 198 further documents to be submitted to the authors. The acceptance rate of these was 60.1%, a little lower than the acceptance of the students' documents, and only a single document of the 110 accepted was graded relevance (1).

In assessing all the additional relevant documents submitted, the authors did not know which had been selected by the students and which were retrieved by bibliographic coupling. The small variations in the acceptance rate (see final column of Table 3.4) by the authors for the different categories are so slight that they are not statistically significant. However there is significant difference in the proportion of documents put into the various relevance grades. From Table 3.5, it can be seen that with the cited papers 41% were included in grades (1) and (2); of the additional relevant papers found by the students only 18% were put in those grades and 15% of these revealed by citation indexing. The fact that so many of these additional references were placed in relevance grades (3) and (4), may be due to the fact that the authors did in fact know of the existence of many of those additional papers, but had selected the cited ones as being the most relevant to include in this paper.

So far the figures have been derived from the total set of 279 questions, but, as previously stated the questions fall into two groups. The authors had been asked to give the one basic question that gave rise to their work, and then to give any supplementary questions that came up during the progress of the work. Of the 279 questions, 118 are basic, and 161 supplementary. In order to discover whether the authors' assessments of their basic questions were in any way different to the supplementary questions, the same figures from Tables 3.4 and 3.5 are set out again in Tables 3.6 and 3.7, now divided into the two categories of questions.

---

*The 15 documents which were both selected by the students and retrieved by bibliographic coupling might be expected to have a higher acceptance rate by the authors, but in fact only 10 of them were accepted.

Table 3.6 shows a considerable difference between the basic and supplementary questions. 72.2% of all documents submitted to the basic questions were accepted as relevant, but for the supplementary questions acceptance was 57.6%. Such a difference might be expected in the case of the cited documents, since more of the references in an author's paper are likely to be included as relevant to the basic question, but the difference in acceptance shows the same proportional difference in all the additional documents submitted as well, (see Table 3.6). A possible explanation of this is the probably different attitude of the authors regarding the basic and supplementary problems. In the case of the basic problem no one complete answer would be available, and any document that shed some light on the problem, even if only remotely, would be likely to be accepted. The supplementary problem had more often been solved satisfactorily some time previously, and the author would therefore want to accept only those documents which dealt with the problem in a way that met his particular requirements.

Individual relevance assessments, done by 182 different people, and with no personal interaction with the project staff, cannot be entirely consistent. However the assessments were made by experts in their subject, and represent the individual and personal needs of the people concerned - the situation in which every information retrieval system has to operate. The evidence appears to show that the assessments were carefully done, although the task was sometimes difficult; as one author said:-

> "Relevance assessment is not easy, but I have done the best I can. In the case of this subject matter, the literature is so extensive that the chances of a relative newcomer picking out what mattered would be very poor; much of what are, in this connection, significant details have not been published anyway; even more important perhaps is that only long association with such a subject, both academically and experimentally, can enable one to appreciate what is useful and to judge what is misleading, unreliable or definitely faulty."

The use of four relevance grades might appear to be too precise a distinction to be able to make in practice, but quite a number of the authors indicated '$\frac{1}{2}$' grades, i.e. (1-2), etc. For the testing stage we accepted these documents at the lower grade. The definitions of the grades was a problem to one author:-

> "Actually ... none of your definitions (1), (2), (3), (4), (5) fits my attitude toward the references. All of the references were of considerable interest to me because they showed me what people had done so far, how recently, and by what methods. None was useful in suggesting methods of tackling the problem. I already knew all of the mathematical procedures that had been used in the papers, and several that had not been employed. To a large extent, it was interesting to find how little had been done, and in some cases, how inadequately."

Another author suggested that papers containing new or original answers to a problem should have a separate grade, and several authors indicated that a given document was a complete answer to their question, but an incorrect one. One new idea for assessing relevance was suggested:-

> "... the 'assessment of relevance' categories seemed particularly difficult to interpret in relation to most of these additional documents. I believe that I have 'scored' the documents roughly in proportion to the degree of irritation I should feel if a librarian produced them in response to my original query. Whether this is a proper basis for measurement of relevance may be arguable!"

The relevance assessments that the authors made of their own cited papers reveal some information on the citation habits of authors, but any observations can only be made within the limits of this situation, in which in most cases only a selection of the cited papers was used.

A few of the authors assessed all their cited papers as not relevant to the basic questions, and one explicitly stated that he did not find any relevant at all. An analysis of 174 of the basic questions, more than was ultimately used, shows that 36% of the cited papers submitted were assessed as not relevant, and if marginally relevant papers graded (4) are included, the figure is 52%. The results from the 118 basic questions in Table 3.6 give results of 28% and 46% respectively. It may be concluded that about half the references in an author's paper are not included in connection with the main problem of the paper, a fact which may assist examination of the possibilities, and limitations, of bibliographic coupling and citation indexing.

There were some cases where a cited document was not strictly relevant to any of the search questions at all, as one author honestly explained:-

> "I have had some difficulty in classifying some of my references into the required categories: chiefly those which occur at the beginning of the report when I attempt to relate this report to my own previous work. It is difficult to know whether they should be categorised as 3, 4, or 5: from the librarian's point of view they should probably be in category 5, but it is not easy to admit that several of one's references are, strictly, irrelevant to all the questions discussed."

Another good explanation for this case was:-

> "In the particular paper of mine a number of references are included, not to give information on the basic search question, nor do they arise from any subsidiary questions; rather they are included to amplify certain details in the text. For example the first three references of my paper are included purely to save time and words in the report, as I felt it completely unnecessary to describe experimental equipment which had been described fully elsewhere. Thus the first three references merit a 'five' rating."

One author supplied us with his reasons for inclusion of six of his references.

> "My assessments of reference 3, 6 and 9 refer really to many papers of which these are typical examples; No. 8 was not located - it just happened to turn up at the right time; No. 4 did not come to hand until after the work was completed and the report nearly so; No. 11 was included merely in order to satisfy anyone who wanted a long list."

A separate investigation, to extract similar information more thoroughly, might be of value, particularly if the author supplied reasons why each paper was, or was not, relevant. Comments on relevance itself, in the match between the questions and the documents, is made later.

The authors' assessments of the additionally submitted documents might be expected to have suffered a little in reliability, due to the time lag between the first letter and the second, and due to the additional documents being supplied as abstracts only. However some authors would be expected to have been aware of some of the additional documents, and, having the full bibliographical details, could examine the full text if they wanted to. Of the 201 questions for which additional documents were submitted, 39 were returned with all the additional documents assessed as not relevant, leaving 162 questions which had one or more of the documents relevant. Several authors indicated a continuing interest in the problem of their own paper, and the quick response to the second questionaire may indicate that the time lag was not a problem.

The large and difficult task undertaken by the students must next be examined. Some error would be expected of any job like this, and two pieces of evidence may indicate the magnitude of the documents missed.

1. Of the 198 documents found only by Bibliographic Coupling, 119 were assessed as relevant by the authors, (see Tables 3.4 and 3.5). There was only one graded as relevance (1), and the majority were graded (3).

2. In cases where an author had given more than one question that we were using, and also where we submitted additional relevant documents in relation to more than one of the questions, all documents submitted were listed together on a sheet with an indication given against each document of the question to which that document was judged to be relevant (see Appendix 3.2). However, there were cases when an author considered that a document which had been submitted in relation to one of the questions only was also relevant to another of his questions. This occurred in 32 questions, and involved a total of 75 documents.

This last fact means that the figures in Table 3.4 referring to the additional student assessed papers include these 75 documents, and the corrected figure for documents selected by the students is 842. Of these, 517 were accepted as relevant, giving an acceptance rate of 61.4% as against the previous figure of 64.6%.

Together with the Bibliographic Coupling documents that were accepted, a total of 194 relevant documents were missed by the students, which means that they found 517 of the 711 that were assessed as relevant, i.e. 73%. Reasons for failing to find the known loss of 27% may be:-

1. The students' interpretation of the question was more strict than that of the author, resulting in the students rejecting what the authors may have accepted.

2. The enormity of the task and inevitable occurrence of human error.

We may hypothesise that if the students' interpretation of the question had been more liberal, a large number of possibly relevant documents would have been selected, resulting in a difficult task of assessment for the authors, and thereby perhaps

resulting in a much lower acceptance rate. Had more been submitted, more would probably have been accepted, but absolute perfection could not be achieved unless each author examined every document in the collection himself. The relevance assessments in relation to each question are given in Appendix 3G.

The questions

The authors apparently found no difficulty in preparing the search questions, and the number received was greater than expected, with each author supplying an average of $3\frac{1}{2}$ questions. Space was provided on the form for four questions, and of the 182 authors who replied, 120 supplied four questions. 40 supplied three questions, 18 supplied two questions, and 4 authors only submitted the basic question. The high average, together with the fact that two-thirds of the authors supplied four questions, suggests that some authors could have written more questions, if space had been provided. However, since in practically every case all of the cited papers submitted were assessed as relevant to one of the questions given, so implying that none of the references was included specifically to answer a question which they had not supplied because of lack of space, it is reasonable to assume that four questions represented a near maximum for these authors.

The requested distinction between basic and supplementary questions clearly fitted the authors' view of their different problems, and only in one case did an author indicate that two of his questions were equally concerned with his basic problem. The 279 questions finally available for testing comprised 118 basic and 161 supplementary questions. There appeared to be no fundamental difference between the basic and supplementary questions. The set of questions is given in Appendix 3D.

The subject areas of the base documents were high speed aerodynamics and aircraft structures. The questions mainly fall into these two areas, but some of the supplementary questions in particular concerned subjects away from the centre of the two subject fields chosen. In aerodynamics, some questions dealt with chemistry of gases, sonic boom, flow in compressors, stability and control, spaceflight re-entry, and heat conduction. The structures questions mostly involved thermal and mechanical deformation and loading, with a few on vibration, effects of noise, and material properties. Some questions involved both subject areas, namely on aeroelasticity and flutter, while there were also some purely mathematical requests.

The generality of search questions is largely a matter of degree, but we would say, in the context of an aeronautical research organization, that most of the questions are reasonably precise, asking for a clearly defined part of the subject. There are a few broader questions (e.g. Q.41 "What progress has been made in research on unsteady aerodynamics"): there was one question which was not used in the tests because we considered it might have a hundred relevant documents, and would probably have retrieved the whole collection.

As previously stated, 279 questions were available for searching. Of these, 58 were really two or more questions stated in one, since they had a logical sum relationship. (e.g. Q.129 "What experimental measurements exist of spanwise and chordwise loadings on swept wings at low subsonic speeds and small incidence") For this reason, most of the tests were made with the remaining 221 questions, although at later stages in the tests, various subsets of thirty to forty questions were used for various purposes. The composition of these various groups of questions is given in Appendix 3E). Questions varied in length; the search terms ranged from 2 to 15 and the average number of individual search terms in the 221 most used questions was 7.6, median 7.9, and the mode was 7. These figures were obtained at the stage when

the search programmes included every possible word, and a more conventional library search would be made on fewer terms than this, an average of probably 4 to 5.

It is always difficult to prove that any set of questions is really typical, or average in some way, but since each of these questions is a statement of a real need for information that arose in the course of some 180 research projects, they are probably as typical a set as can be obtained outside a real life situation. Many of the questions may have been put to an information service at some stage.

Without the facility to cross-examine the questioner, interpretation of the meaning gave less trouble than expected. A deep knowledge of the subjects would probably have revealed some facts and connections not appreciated, but many replies to the second questionaire included additional search terms suggested by the authors, and in some cases alternative rephrased questions. An example of the intricacies of the subject is seen in the following comment, made by an author to explain why one of the additionally submitted documents was not relevant to his question:-

> "It might seem strange that the paper by Kuchemann and Kettle would be of no use at all in answering my question. This is due to the fact that the influence of end plates is different for streamlined and unstreamlined bodies. In the first case they modify the vortices shed from the tips whereas in the second case they prevent spanwise flow brought about by the blockage of the body. There is no connection between these two effects."

The test design has produced a set of documents which have been assessed as relevant to a set of questions. Since this has not been done in a real life situation, can it be argued that the questions are artificial and the match with the documents unreal?

Considerable discussion and argument on these points has taken place in connection with the questions used in Cranfield I and the Western Reserve University test. Although the present question-gathering method did involve a base or 'source' document, it has not been used in the same way as in the previous tests. Previously the questions were framed so that the source document would be a complete answer to the question, but in the present test the question is the real need or research problem that gave rise to the 'source' document being written. Although the 'source' documents are included in the collection, it is only the cited documents from each 'source' document that are assessed and counted as relevant, with the addition of the extra relevant documents found. The 'source' document for each question is removed from the collection when that question is being tested and does not appear in any of the results at all. There is therefore, no reason for continuing to argue about the unreality of tests based on source document questions, or to continue to imply that the 'Cranfield test method' necessarily involves the use of such questions. However, we have stated a belief that source document questions 'can still be used satisfactorily in situations where time and cost are important considerations, as might be the case in an evaluation of a small operational information retrieval system'.

This comment was given in a reply to an article by D. R. Swanson, on 'The Evidence Underlying the Cranfield Results' (Ref. 4), in which he emphasised what he called 'the artificial' or 'biased' nature of the relationship of the question to the

source document', in Cranfield I and the W.R.U. Tests. Swanson, in a sample taken from the first project, demonstrated that this biased relationship was shown by an unusually close match between the words of the question and the titles of the relevant documents. In his paper, Swanson gives the result of an analysis of the terms used in a set of 100 questions and the titles of their accompanying source documents. This was done by the Cranfield group and discussed at some length on pages 27-32 of Ref. 2, although Swanson does not comment on this work. Instead he prepared an admittedly more exact method, which would give, according to his view, retrieval of the source document and the number of irrelevant documents also retrieved would be small. To do this, he took the 100 document titles given in Appendix 4B, and made a list of all the terms which did not occur more than once. From this he argued that, if such a term also occurs in the matching question, then the document would be retrieved, with an average of 60 other documents also being retrieved. This statement is incorrect, in that Swanson bases it on the view that there were only 6,000 documents in the index searched, whereas there were 18,000, so a search of the nature proposed might be expected to retrieve an average of 180 documents.

However, using this method, Swanson finds a close correlation between the result of his 100 searches and the actual search results, and goes on to imply that the use of questions based on source documents will give predictable results.

To find whether these results could be repeated, we carried out the same procedure with the 114 questions and source documents of the W.R.U. test, as given in Appendices 2a and 2b of Ref. 3. This procedure gave 232 terms, of which 132 occurred only once. The result of this analysis was to show that 38 documents would have been retrieved by the use of a key term occurring not more than once, this representing a recall ratio of 33%, as against the 85% recall achieved by the Cranfield facet index. On the other hand, assuming that each key term occurring once in 114 documents would occur on an average of nine times in the whole collection, this method would have given a maximum precision ratio of 11% as against 16% achieved by Cranfield. Such a precision ratio of 11% could, of course, only be achieved by the hindsight of selecting the correct term and no other. For instance, Q.107 'Effects of increasing molybdenum content by carburising steels' is counted as a success by the fact that 'carburising' occurs in both question and document title. However, 'molybdenum' meets the single-use requirement, so would have retrieved the source document for Q.21, which would have been completely non-relevant. This effect would probably reduce the relevance ratio to less than 5%, but even so, the performance obtained by this method is vastly inferior to the performance obtained by the Cranfield index, and appears to make untenable the criticisms of Swanson.

There would appear to be three possible reasons for the difference in results of the similar tests done by Swanson and at Cranfield. Firstly, the W.R.U. collection was narrower in subject coverage than the collection of the first Aslib-Cranfield project. For instance, one key word given by Swanson is 'Titanium'. Since only some 300 documents in the whole collection dealt with metallurgical subjects, such a term is clearly unlikely to occur more than once in a hundred documents, whereas in the W.R.U. count it occurred on eight occasions. (This is an aspect of the generality ratio discussed later)

A second reason could be a significant difference in the quality of titles. Many documents in the first Aslib-Cranfield test were research reports, with titles which were fuller than usually occur in commercial journals, from which many documents were taken for the W.R.U. test.

Total documents retrieved by bibliographic coupling at strength of 7 or more		312
Documents which had already been assessed for relevance by being references.	87	
Base documents	12	213
Documents which had been located by students	15	114
Submitted to authors for relevance assessment		198

Table 3.8 BREAKDOWN OF 312 DOCUMENTS RETRIEVED BY BIBLIOGRAPHIC COUPLING AT STRENGTH OF 7 OR MORE.

QUESTION 145

Has anyone investigated the unsteady lift distributions on finite wings in subsonic flow

RELEVANT DOCUMENTS

1698. The unsteady lift of a wing of finite aspect ratio, (STRONG MATCH)

1705. On the kernel function of the integral equation relating the lift and downwash distributions of oscillating finite wings in subsonic flow. (STRONG MATCH)

1704. A systematic kernel function procedure for determining aerodynamic forces on oscillating or steady finite wings at subsonic speeds. (WEAK MATCH, because 'finite wings' and 'subsonic' are commonly used terms in this collection)

1700. Two and three dimensional unsteady lift problems in high speed flight. (WEAK MATCH)

1703. General airfoil theory. (NO MATCH)

1792. Some low speed problems of high speed aircraft. (NO MATCH)

TABLE 3.9 EXAMPLES OF QUESTION/TITLE MATCHES FOR RELEVANT DOCUMENTS

Terms underlined in the document titles are those matching the required terms in the question.

The third, and probably most significant reason was the greater care taken with the questions for the W.R.U. test. There appears to be no reason to apologise for the fact that it was not possible to exercise such close control over the question compilers when we had to obtain some 1,600 questions for Cranfield I, but by the time of the W.R.U. test, the importance of the matter had been accepted, and the question compilers were personally selected and more adequately instructed.

In the W.R.U. test, an analysis was made of all documents in the collection against each question and, as given in Appendix 3C of Ref. 3, 42 other documents were assessed as equally relevant as the source documents. As a further check on source document questions, the titles of these documents have also been matched against the appropriate questions, using the list of terms generated with the original 114 source documents. Fourteen documents had a single term match with the questions, so again the recall ratio was 33%, the same as with the source documents. This appears to show fairly conclusively that, in the W.R.U. test, there was no unnatural relationship between the terminology of questions and source document titles, and lends support to the strongly-held view of the Aslib-Cranfield staff that questions based on source documents can still be considered as being, in the right circumstances, a convenient and economic device for testing I.R. systems.

Some unnatural relationship was clearly present in Cranfield I, but it is wrong to conclude from this that whenever there is a substantial match between question and title, then the relationship is necessarily unnatural. Some proportion of questions in a real life situation are bound to have some relevant documents with a close question title match, and if this is not the case then all Permuted Title or K.W.I.C. indexes are useless. However, although as explained earlier, source-document questions are not used in the present test, Swanson still expresses doubt and comments on the present test method:- 'This is some improvement (since the title-question correlation is probably diminished); but it is still dubious in principle - a 'biased' or 'special' relationship between questions and relevant articles persists' (ref. 4). Although no evidence is presented to justify this statement, an examination of some of the questions and their relevant documents has been made, to find out the extent, if it exists, of the bias of the suggested relationship.

Using 35 of the questions*, and their associated 287 relevant documents, we first examined the correlation between the questions and document titles. The words and phrases of the questions were examined for a 'match' with the words and phrases in the titles, and generally an identical word or phrase only was considered as a match, except that synonymous word ending variants were accepted. In terms of the whole question, two levels of matching were distinguished:-

Level A Strong Match Two or more concepts, or important subject words were demanded. A single concept was only accepted if it was one of the vital ones in the question, and in a few cases a single word was accepted as a vital or 'key' term provided it was used less than twenty times in indexing.

Level B Weak Match These rules accepted any match down to a single word, provided it was a subject content word. The general descriptive words such as Problem, System, Solution, Parameters, High, Large, etc. were not accepted.

---

*These questions are the 7 search-term questions and appear as Question Set 1 in the Appendices.

Strength of match	Relevance grades				
	(1)	(2)	(3)	(4)	Totals, all relevant
Strong match	12	17	40	20	89
Weak match	3	20	39	25	87
No match	4	24	54	29	111
(Total)	19	61	133	74	287
Percent strong match	63.2%	27.9%	30.1%	27.0%	31.0%
Percent strong and weak combined match	78.9%	60.7%	59.4%	60.8%	61.3%

TABLE 3.10. RELEVANCE GRADES OF DOCUMENTS WITH SPECIFIED QUESTION-TITLE MATCH

Strength of match	Cited documents	Additional documents	All documents
Strong match	44	45	89
Weak match	38	49	87
No match	67	44	111
(Total)	149	138	287
Percent strong match	29.5%	32.6%	31.0%
Percent strong and weak match combined	55.0%	68.1%	61.3%

TABLE 3.11. COMPARISON OF THE CITED AND ADDITIONAL DOCUMENTS WITH SPECIFIED QUESTION-TITLE MATCH

Some examples of a strong match are: Chemical, Kinetic (Question 9); Viscous, flat plate (Q.82); and Slip (frequency of 19, Q.87). Some examples of a weak match are: High Speed (Q.2); Aircraft (Q.2); Hypersonic (Q.9); and Structural (Q.49). Further examples can be seen by reference to Table 3.9.

Out of the 287 documents examined against the 35 questions, 89 (31%) showed a strong match; an additional 87 had a weak match, and the total of 176 represents 61.3% matching. 28 of the questions had one or more documents with a strong match, and 32 had one or more with a weak match.

This shows that nearly one-third do have a strong question-title match, but since the assessment of relevance has been done in four grades, we may expect that those documents with a strong match will be graded as more relevant than those with a weak match. Table 3.10 divides the results into the four relevance grades, and shows that the probability of a relevance (1) document being strongly matched is more than twice that of the relevance (2), (3) or (4) documents. That the relevance (2),(3) and (4) documents show the same probability may be accounted for by the difficulty of consistently doing such a refined grading of relevance, but the relevance (1) documents seem to indicate a strong trend.

Whether it is taken that these figures show an unusual question-title match or not, the presence of an unnatural question-document relationship cannot be proved or disproved by this. One would have expected a certain strength of title match in this subject, where titles are usually fairly long and a good indication of the subject of the document. The documents examined were the total of those relevant to each question, and included both the original documents cited in the authors' base document, and also the additional documents discovered in the collection. It is obvious that these additional relevant documents, discovered by the students' examination of the collection and by bibliographic coupling, were discovered and assessed as relevant in a situation equivalent to a real life one, and therefore it would be quite absurd to suggest that an unnatural or biased relationship could possibly exist in their case. So a comparison of the question-title match between the 'cited' relevant documents and the 'additional' relevant documents will provide some evidence of any unnatural differences in the question-document relationships.

The 287 relevant documents comprised 149 cited and 138 additional, and the matching scores were calculated for each group. Table 3.11 presents the results, and it is shown that the additional relevant documents had a slightly stronger question-title match than the cited ones, 32.6% to 29.5% for the strong matches, and 68.1% to 55.0% for the weak matches. Ten of the 35 questions had no additional documents at all, and the cited document for these questions have been included in the results; deleting these ten questions would reduce the matches for cited documents to 27.6% and 51.4%.

These results might alter over the whole set of questions, but there is no reason to expect that they would change significantly. On the basis of the question-title match anyway, no real difference exists between the cited and additional documents. We suggest that this indicates that there is no justification for any implication that there is a biased or unnatural question-document relationship, and that the relevance assessments and relevant documents found are not really different from that which might happen in a real life situation. Further evidence can be obtained from some of the test results themselves, where the retrieval performance in recall of the cited documents can be compared with the additional documents.

Co-ordination Level	Cited Documents		Additional Documents	
	Total Recalled	Recall Ratio	Total Recalled	Recall Ratio
1	99	94.3%	128	92.8%
2	80	76.2%	101	73.2%
3	59	56.2%	75	54.3%
4	40	38.1%	46	33.3%
5	17	16.2%	25	18.1%
6	9	8.6%	10	7.2%
7	2	1.9%	3	2.2%
Total Relevant	105		138	

TABLE 3.12 COMPARISON OF RECALL PERFORMANCE OF RELEVANT 'CITED' AND ADDITIONAL DOCUMENTS IN RELATION TO 25 QUESTIONS.

All questions had seven starting terms; the table shows the effect on recall of increasing the search requirements from any one term to all seven terms.

Using the same set as considered in the previous paragraphs, the 25 questions which had some additional relevant documents were used, comparing 105 cited with 138 additional documents. Here again the difference between the two groups is not significant, (see Table 3.12). For instance, at a coordination level of 2, the recall ratios are 76% and 73% for cited and additional documents; at a coordination level of 5, the figures are 16% and 18%. These results (which are, of course, only a small sample of what will be presented in a later report) should have revealed any unnatural question-document bias, whether conspicuous in the title or not, had any bias been present at all. We are confident that there is no measurable unnatural match between the questions and the documents themselves. Questions obtained from a real life situation and tested on an existing collection might give different results in some way, but until such a test is done, and a comparison is made of different test methodologies, it is not possible to state in what ways, and by how much, the present test method falls short of the ideal in this respect.

CHAPTER 4

Indexing Procedures

The function of indexing in libraries and information retrieval systems is to indicate the whereabouts or absence of items relevant to a request. It is essentially a time-saving mechanism. Theoretically, we can always find the relevant items by an exhaustive search through the whole collection (assuming that we can recognize what is relevant when we see it). Since this is economically impossible, the size of the store to be examined is reduced by classification, using this term in its very broadest sense, i.e., as the recognition of useful similarities between documents and the establishment of useful document groups based on these similarities. So documents, or document surrogates, are assigned to a limited number of classes according to certain criteria, in particular, their subject content (although in machine indexing, utilizing complete text scanning, this 'limited number' can become very large - as large as the number of significant words used in the text). Search for relevant items is made via these classes (which are classes of documents); only those with a probability of containing relevant items are examined, and the rest (hopefully the vast majority) are ignored. Clearly, we need to know as much as possible of the nature of the classes to be recognised, and the degree to which they allow reliable predictions to be made as to the probability of relevant items being included in them.

Most library indexes, other than those to imaginative works (novels, music scores, etc.) are aimed ultimately at the retrieval of subject information. Even the great Author-Title catalogues, on which so much care has been lavished, serve for the most part the function of a diagnostic classification, i.e., an author's works are sought in the first place because they are about a certain subject and his name is a clue to locating it. The popularity of the author-title catalogue rests partly on its precision in retrieval. Classes determined by authorship or title are mutually exclusive; there is almost no overlapping, no ambiguity about them and requests can be met with 100% recall and precision. But they are useless if the author or title is not known and it is this situation with which IR is mainly concerned. So the classes investigated by this project are those designed for searching by subject prescription only.

There is one exception to this. Bibliographic coupling (including Citation indexing) establishes classes for much the same reason as author-title catalogues, as an oblique way of getting at subject content. Papers which have cited item x are assumed to have some connection with the subject of x. This particular device is dealt with separately in Chapter 7.

The terms which are used to express a request or a search prescription rarely coincide exactly with the terms used to describe a particular relevant document; this is likely to happen only at a relatively broad level, when a request may be answered by a treatise or monograph on the subject. For example, in the test Q.93 read 'What investigations have been made on the flow field about a body moving through a rarified, partially ionized gas in the presence of a magnetic field.' Two documents relevant to this question were

1296 'Waves through gases at pressures small compared with magnetic pressure'
1446 'Waves of a satellite traversing the atmosphere'

In both cases the match is very imperfect. It is made only by recognising that the

class prescribed in Q.93 can be adjusted in order to coincide at some points with the index descriptions of the documents. The class 'rarefied, partially ionized gas' must be seen to correspond or relate, after suitable manipulation, to 'gases at small pressure' in the one case and 'ionosphere' in the other. The class 'body' must be seen to relate at some point to the class 'satellite'.

So a subject index must provide facilities for adjusting and manipulating its classes; it must allow the index classes examined to be expanded or contracted, and in different directions, until a match with the search prescription is recognized. Index language devices are the agents of this manipulation. They are devices whereby class definitions may be adjusted to meet the requirements of different searches.

Index language devices

The index description of a document is a condensed (usually a highly condensed) statement of the document's subject content; it seeks to convey succinctly what the document is about. Its main, and sometimes only, constituent is the set of substantive terms (lexical elements) which act as clues to the subject of the document. These terms may be supplemented by some indication of the relations between them (syntactical elements), e.g., by the addition of roles, or facet indicators (explicit or implicit) or by such elementary syntactical devices as those of the Alphabetical subject catalogue. In a post coordinate index they are usually kept to a minimum, enough to remove serious ambiguity but no more.

It seems reasonable, then, to assume, as the simplest possible form of index description, a bare list of words, selected directly from the title and text of a document as being good clues to its content, and presented without any reference whatsoever to a control list for synonyms, related terms, etc.

The simplest way in which such a list of words could be used would be to regard each word as defining one of the classes to which the document belonged, without reference to the other words. Searches would then be made simply within these classes, separately. For example a document indexed as being about Wakes - Satellites - Traversing - Ionosphere would be seen simply as a member of four different classes (the class 'Documents dealing with Wakes', the class 'Documents dealing with Satellites', and so on). So a search on Satellite wakes would be made simply by examining all documents in the class Satellites, and all documents in the class Wakes. This is very similar, of course, to what a Permuted Title or KWIC index does. Recall performance figures for this crudest of all forms of index language were assessed in the first Cranfield project as 97%. Precision figures were not available but it is certain that they were very low. It is assumed that all the keywords constituting the question are examined. If a selection were made, recall would probably drop in so far as the exhaustivity of the searching would have dropped. The question of exhaustivity and specificity of searching and indexing is discussed later.

Now will be considered the ways in which, by the use of various devices, this simplest of all possible forms of indexing can be refined in order to increase its capabilities for meeting all the demands which search prescriptions may make on it. Such devices may be separated conveniently into two groups;
1. recall devices - those which, when applied to any existing class, increase the size of the class in terms of the documents responding to the definition; e.g., if the class Bakelite is expanded by hierarchical linkage to include all Phenolic resins, more documents are retrieved.

2. precision devices - those which, applied to any existing class, decrease the size; e.g., if we coordinate Bakelite with Extrusion and examine only the class defined by this simple relationship of intersection, we exclude those documents on Bakelite which do not refer specifically to this operation.
The operation of these devices on the sort of simple index description described above can be seen by a consideration of the relations such a description displays to the precise subject of the document concerned, and to the wider subject field of the information store from which we may wish to retrieve that document or something like it.

An index description a, b, c, d, e, f (where each letter represents a substantive term or lexical element, e.g., Wing, Drag, Control) embodies two sets of relations: firstly, those internal to it, reflecting the local and temporary conditions peculiar to the subject of the document described; e.g., the fact that d is the product, whereas f is an agent of the process a which produces it, or, the fact that b qualifies a while c qualifies d, but that neither of these qualifiers is applicable to the object of the other; or, more subjectively, that a and b, rather than c d e or f, represent the dominant theme of the document. These are, broadly speaking, the interlocking relations between the substantive terms.

The second set of relations are those external to it, reflecting the more permanent pattern of relations in the wider field or subject area to which the document belongs: e.g., that a is a species of x, or that c is almost synonymous with q, or that a represents one participle of a term (e.g., Cooling) which may be usefully related to another participle (e.g. Cooled) in the subject concerned.

The two sets of relations can be utilized to add precision to the original description (using the first set) or to expand the description by reference to the wider relations (using the second set). In other words, they underlie the two groups of index language devices which will now be outlined briefly.

Devices which increase precision

(i) Coordination - i.e., the conjunction of two or more terms to produce a narrower class defined by the intersection; e.g. Shear and Flow to give Shear flow. This is the most important device in indexing. Whilst it is commonly associated with postcoordinate systems where it is implemented mainly if not entirely at the search stage, it is equally fundamental to precoordinate systems; but in these, only the products of selected coordinations are usually catered for conveniently.

(ii) Weighting - i.e., the assignment to a term of a figure representing the relative significance of that term in the total subject description of the document. So a term which represents the central theme of the document gets a high weighting and one which represents only a marginal element in the subject content of the document gets a low weighting. If now a question is also weighted, i.e., greater significance attaches to one or some of its terms than to others, then the search may be directed only to coordinations with that term or only to the same term when it has been given a similarly high value in indexing. In either case, the class of documents retrieved is made narrower.

(iii) Links - i.e., indicating a particular connection between two or more terms in a description where the lack of such an indication would produce ambiguity; e.g., if the same document deals with the hardness of copper and conductivity of titanium a link between Hardness and Copper on the one hand and between Conductivity and

Titanium on the other would make it clear which property referred to which substance. Clearly, a link must involve at least two terms.

(iv) Roles - i.e., indicating the role or function of a particular term in an indexing description. Sometimes a simple link is insufficient to remove ambiguity; e.g., Production of particle x by bombardment of particle y. To link, say, Production with x and Bombardment with y could still allow the description to be interpreted as Production by x, or Bombardment of y. The addition of a role-indicator makes the relationship more explicit, e.g. by labelling x as Product and y as Patient (and possibly a bombarding particle, z, as Agent).

Devices which increase recall

(v) Confounding synonyms - i.e., accepting items indexed by x when searching for y, and vice-versa, where x and y are regarded as synonymous.

(vi) Confounding word forms - i.e., accepting items indexed by different forms of the search terms, such as its singular and plural, participle and gerund; e.g., Injectant + Injected + Injection + Injectors. The most comprehensive operation of this device is where a stem or root is used to define the class and all words containing it are included in the class.

(vii) Hierarchical linkage - i.e., accepting items indexed by terms which are in some generic hierarchical relation to the search term. By this we mean terms which are either subordinate to, superordinate to, coordinate with, or collateral with the search term; e.g., the class Cooling might be extended hierarchically to include the subordinate term Sweat cooling, the superordinate term Heat transfer, the coordinate term Heating and the collateral term Radiation. This is a stricter interpretation of hierarchical linkage than is often used in the literature of IR, confining it to the relations between a thing and its kinds as distinct from numerous other relations such as those between a thing and its parts, its processes, its properties, the operations performed on it, and so on.

It is perhaps necessary to note that hierarchical linkage is essentially a recall device. This is not to say that it cannot be used in order to refine a question and give it more precision; e.g., an enquirer about to search the class Cooling might be led to realise (by the hierarchical display of related terms) that he really wanted Sweat cooling and would then narrow his search by confining it to this species of cooling. This is almost as though hierarchical linkage were acting as a precision device. But clearly, the question asked was wrongly put and the function of the hierarchy would have been to assist the question- programming. In testing devices, it must be assumed that the question has been accurately stated, otherwise a large and disrupting variable will enter the test. Therefore such adjustments as the one described cannot be considered, and hierarchical linkage must be treated as a recall device only.

(viii) Non-generic hierarchical linkage. The device of hierarchical linkage which has always featured prominently in subject indexing and is the central device in what is generally known as 'classification' is the result of selecting one particular relationship (the generic one, between a thing and its kinds) as the basis of various kinds of class definition. This raises the question: should we similarly regard the use of other particular relations (e.g., between a thing and its parts, a thing and its properties) as constituting separate indexing devices, each to be evaluated separately?

The importance of the genus/species, or thing/kind relation needs little explanation. Since Aristotle, it has been the central relation considered in traditional logic. In information retrieval its use as an indexing device ('hierarchical linkage') is that it provides an automatic relevance network based on the inclusion relation, e.g., if laminar boundary layer is a kind of boundary layer, then documents on the latter are likely to have some degree of relevance (possibly a high one) to questions on the former. It is also the basis of the notion of 'specificity', which is a fundamental parameter determining precision in index performance. It might be argued that the non-generic relations displayed in all library classifications also possess an 'inclusion' quality so far as documentary classes go; e.g., a precoordinate class with the index description Delta wing - Drag could be said to be 'included' in the class Delta wings just as much as the true species Sweptback delta wing is included. But the inclusion relation of the first example is weaker, in that it is a shared one; Drag on a delta wing is as much included in Wings-Drag as in Delta wings, whereas Sweptback delta wing is entirely included in the class Delta wing. That is to say, the non-generic relation is essentially one of conjunction ('coordination') rather than complete inclusion.

Certainly, no other single relation rivals it in ubiquity or importance for retrieval, although collectively the other relations contribute significantly to class definition in indexing. In precoordinate classified indexes it is normal in question-programming to move from the terms of a category or facet, each one constituting a hierarchy, to the particular term which gives rise to the facet; e.g., to move from Conductivity (of titanium) in the Properties facet of the class Titanium, to Titanium. The object of systematically arranging classes in such indexes is to provide a permanent and constantly available mechanism for manipulating the classes in this way, and not only by movement within a strict genus/species hierarchy. Classes are expanded or contracted by moving also from one category to another. For example, in a question on compressor operation, the class examined may be expanded by moving from particular parts of the compressor (Blade, Shroud, etc.) to particular processes (Stage interaction, Stage stall, etc.) or to particular characteristics of these (Inlet blade angle, Stall limit line, etc.)

In fact, the term 'hierarchical' is frequently used as a synonym for the process of subordination which is the essence of a precoordinate system. In this view of hierarchical linkage a chain such as Delta wing - Sweptback - Transonic speed - Low angle of attack - Lift is regarded as reflecting a hierarchy every bit as much as a true genus/species chain like Metals-Titanium, or Aircraft-Heavier than air-Monoplane-with Delta wing. It is possible to go even further and refer to the precoordination of an Alphabetical subject catalogue (which, of course, subordinates some terms to others in its subject headings, e.g. Aircraft-Design) as a hierarchical system (Ref. 21).

Even if this last extreme view of the term 'hierarchy' is rejected, we are left with the fact that the identification of hierarchical linkage with the full range of relations displayed by a highly organized library classification system does not give us a basic device which can be measured in the same way as can coordination, weighting, etc., but is a varied mixture of relations capable of defining classes.

In traditional library classification it is well established that any relationship may be used to establish 'subclasses' of a given class. In the sense that library classification deals with classes of documents, and that documents on any aspect whatsoever of a subject x can be regarded as belonging to some subclass of 'Documents on x', then Extrusion of plastics is just as much a subclass of Plastics as

Bakelite. But recognition of this is of little use unless the further step is taken of organizing these subclasses according to their particular relations, in terms of subject content, to the class x., i.e., of recognizing that some are the properties of x, some are its parts, and so on.

In a modern faceted classification these relationships are systematically displayed. All terms standing in the same relation to the original class are marshalled together to constitute a category or facet of that class. So, whatever concept comprises the original class, all its Properties are assembled together, all its Operations, all the Agents of these Operations, and so on. So far as they lend themselves to the process, the members within each category are organized in a hierarchy and their relationship within the category is one of a thing and its kinds - kinds of properties of x, kinds of operations on x, kinds of agents of operations on x. But hierarchical linkage in the strict sense, that is generic hierarchical linkage, is established between the terms within a category, not between individual terms from different categories, or between the terms of a category and the original class.

However, it is undeniable that in indexing and searching, classes are manipulated (i.e. expanded and contracted) by going outside hierarchical linkage in the strict sense defined above. The two main paths pursued are those already indicated; firstly to move from a term in one category of a class to one in another category of the same class, e.g. Separation see also Boundary layer control (where Separation belongs to the Process facet of Boundary layer and Boundary layer control belongs to the Operations facet). Secondly, to move from a term in a category to the original class giving rise to the category, e.g. Blowing see also Boundary layer (where Blowing is an operation designed to accelerate the flow in the boundary layer, and belongs to the Operations facet of Boundary layer).

Both the above types of connection imply a more or less definite subject area in which the terms in question stand in some categorical or facet relation. Another type of connection is sometimes recognized, between terms which come from quite distinct areas and which are therefore not considered to have such a facet relation. The 'phase relations' of Ranganathan are one example, and some of the terms connected in syndesis constitute another. For example, a thesaurus might link Automatic control theory with Aerodynamic stability. Generally speaking, the view that this constitutes a quite distinct type of relation assumes a relatively arbitrary map of the field of knowledge in which subjects are assigned to one conventional class or another, such as those found in a general classification. It does not correspond to any fundamental relation between the terms which, if they have any connection at all, can be fitted into the framework of categorical or facet relations. For example, Control may be viewed as a term in an Operations facet of the subject Aerodynamic stability.

It is not feasible to consider syndesis (i.e., the adjustment of classes via a system of linking references) as a device in itself since it clearly uses a mixture of several quite different relations in indicating its further classes. Neither does it seem particularly profitable to make a rigid distinction between the two types of non-generic relations above since it is likely that they will overlap from time to time; e.g., Blowing may well occur in a general Control Operations facet rather than be subordinated solely to Boundary layer; in which case the linkage between Blowing and Boundary layer exemplifies the first type, not the second.

Also relevant to the question of whether these relations rank as discrete devices

is the fact that they are rarely, if ever, used alone; whilst all the devices given earlier may be thus used (at least in a postcoordinate system), the non-generic relations are invariably associated with hierarchical linkage, whether this is via a classified index or a syndetic network of connective references. In a thesaurus, for example, generic hierarchical relations and synonym relations are often indicated separately, but individual non-generic relations are never recognized separately.

We conclude that there is nothing in the practice of indexing to suggest that a separate evaluation of each non-generic relation is necessary, but the collective contribution to index performance of these relations compared with the contribution of generic hierarchical linkage is a matter of some interest, and it seems reasonable to group them together as a comparable device. It may be noted that generic hierarchical linkage is itself an aggregate of several particular relations, just as this group is. The problem is discussed further in the section on Concept hierarchies in Chapter 5.

(ix) Bibliographic coupling is a device for extending a class x (representing the subject of a particular document q) by accepting all, or some of the documents which have cited q: or, by accepting all documents in a particular universe which have a certain number of citations (6 or 7, say) in common with q.

(x) Associative indexing by machine ('clumps', etc.) The possibilities of automatic indexing now being explored by a number of investigators rest mainly on the assumption that classes useful for retrieval purposes can be established on the basis of the statistical characteristics of the index vocabulary (which may in fact approximate to the complete texts of the documents concerned). By using such features as the frequency of occurrence and co-occurrence of individual words and of particular word-clusters, their position in the text, their relative frequency compared with a standard word-frequency list in the subject area concerned, and so on, associations between terms are established which then form the basis of search programmes. The criteria defining the classes to be examined are thus quite different from any of those listed above and therefore the procedure constitutes an indexing device in its own right.

How far it might be feasible to distinguish particular procedures (e.g., the use of one statistical technique rather than another) is as yet uncertain. In particular, the purely statistical methods are in some cases replaced by methods using linguistic analysis, and insofar as these must overlap the 'semantic' devices already described (confounding of word forms, hierarchical linkage of various kinds) they may not merit the status of a discrete and unique index device.

(xi) "L'Unité" system described by te Nuyl (Ref. 22) is a somewhat exotic device whereby a reduced vocabulary is established in a quite mechanical way by lumping together all the terms in a given sequence of pages in the Concise Oxford Dictionary and treating their aggregate as a single class. As is the case with all drastically reduced vocabularies, it is argued that the theoretical absurdities which might arise (e.g., the appearance of documents on Acne in a search for Aconite, or on Conductivity in a search for Cones) do not arise in fact, since subsequent coordination eliminates them.

Any reduced vocabulary may be regarded as a recall device, in that it implies enlargement (by coalescence) of the classes which are formed initially by the individual index terms assigned to a document. Usually, reduced vocabularies are formed

by a mixture of the devices already described - by hierarchical linkage, by confounding word forms, etc. "L'Unité" uses none of these, although the last-mentioned will in fact normally be a prominent accompaniment of its class definition, since different word forms will usually appear in the same alphabetical cluster. So it must be admitted that it forms a discrete index device in its own right.

The above eleven devices may be compared with the list of techniques for controlling index languages given by B. C. Vickery in his book 'On retrieval system theory' (Ref. 9) (see Chapter 1). Of those listed above, three devices, namely Bibliographic coupling, Associative machine indexing, and L'Unité, were not mentioned by Vickery. The others include all the techniques given by Vickery, since a number of these were variants of the more broadly defined devices above.

We have tried to distinguish the basic device itself, as a method of class definition, from the different ways in which it might be implemented in different index languages. The latter may be regarded as different amalgams of the various devices, with further differences resulting from the various methods of file organization.

The different ways in which coordination is applied in precoordinate and postcoordinate systems, have already been mentioned. The fundamental difference is that in the former a limited number of coordinations are made and the resultant compound headings are then filed in linear order, and rules observed (as to citation order, etc.) to allow determination of the exact position of any particular combination. This difference has repercussions for the other devices. Even in a largely single-entry precoordinate system, for example, 'weighting' of an elementary kind is implicit in the restrictions placed on the number of entries which can be recognized. Links are fundamental to a precoordinate system by 'partitioning', e.g., separate entries would be made for Copper-Hardness and for Titanium-Conductivity.

Roles are indicated in a precoordinate system by citation order or by explicit syntactical devices; in a faceted index an index description such as

Wings - High aspect ratio - Drag - Low angle of attack

conveys by its citation order of Thing (Wing, etc.) - Property (Aspect ratio, etc.) - Process (Load, Drag, etc) - Condition (Aerodynamic parameters, Angle of attack, etc.) that High specifies Aspect ratio and Low specifies Angle of attack. In an Alphabetical Subject Catalogue, a similar function is served by such headings as

Children in art,
Pressure vessels - Heat transfer,
Acceleration - Psychological effects.

Synonyms are treated with varying strictness of interpretation in different systems. Often, it is a reflection of the degree of specificity sought, as when one system distinguishes Potential flow from Irrotational flow and another confounds them. The problems of synonymity occurring at the level of multiple term descriptions raises a particular problem for post coordinate systems, since it implies a degree of precoordination at some point in the system; e.g. a Ground effect machine is a synonym for Air cushion vehicle, although there is no synonymity between the individual constituent terms.

Confounding of word forms may be implemented at the indexing stage, via a controlled vocabulary (e.g., Conducting see Conduction), or by search rules (e.g., accept Conducting + Conduction + Conductor). In a precoordinate system, where the different forms are separated in different categories, confounding may be possible only at the search stage; e.g., by consulting the A/Z index of a classified index and observing the variant forms used.

Hierarchical linkage may be implemented in a postcoordinate system by generic posting, by coding based on a form of semantic factoring as in the WRU system, or by search strategies based on a classification schedule or thesaurus. In a precoordinate system it may be achieved to a large degree by the file arrangement, as in the systematic juxtaposition of related classes in a classified file; it may be achieved fragmentarily by inversion of subject heading in an Alphabetical subject catalogue (e.g., Drag, Base; Drag, Form; Drag, Induced; etc.) or by search strategies based on a syndetic network of see also references.

Measuring the performance of index devices

In order to establish recall and precision performance figures for the different devices, both singly and in various combinations, it was first of all desirable that we established as far as possible figures for indexing in which none of the devices was operating. Then it would be possible to determine the impact on these figures of the introduction of each device in turn. This assumes, of course, a test collection and a set of questions to be put to it, where it is known just what documents are relevant to each question, as described in the previous chapter.

Performance figures for an 'unindexed' collection seemed to imply a situation in which the complete text of each item in the collection was searched for each question. This would have been too tedious an operation (although something like it, except that it was on a small scale, using computer facilities, has been described by Swanson (Ref.18)). The alternative which we decided to take, was to use, as the base situation, one in which the simplest known indexing device was used and to measure the impact on this of all the other devices. This simplest device was taken to be that of condensation of the full text into an index language consisting solely of the 'uniterms' thrown up by the title and text of the document itself, quite uncontrolled by any prior index language.

So the first step was to establish, by the indexing of the test documents, a crude, elemental index language from which all the other languages (each one characterized by the addition of a particular device or aggregate of devices) would be derivable. Before this could be done it was necessary to provide for the control of two major parameters in indexing, exhaustivity and specificity.

Exhaustivity and specificity

Exhaustivity in indexing refers to the degree to which one recognizes (i.e. includes in the index descriptions) the different concepts or notions dealt with in a document. Specificity refers to the generic level at which these concepts or notions are recognized. For example, suppose a report has as its main theme the subject 'Drag on swept wings at high subsonic speeds'. If one neglects, for the time being, the various subsidiary themes which are also dealt with, this report may be said to deal with three concepts - an aerodynamic characteristic, an aerodynamic structure and a flow condition. If these concepts were described in the above fashion in the index description, this latter would be exhaustive but not specific. If the description consisted only of Drag - High subsonic speeds it would be neither exhaustive nor specific; for whilst the terms retained are specific, the absence of any reference to Swept wings implies that the subject deals with aerodynamic structures in general (some structure is implicit, of course) and this is less than specific, since to be this a description must be exactly coextensive with the notion represented. There can be no reduction in exhaustivity which is not a reduction in specificity; but the reverse does not hold.

An exhaustive and specific description (either in indexing or in question formulation) is one which allows no further qualification or refinement - it is a completely precise statement of the class concerned (or classes, if the terms are considered individually). It seems clear that such a description should include not only the substantive terms or lexical elements (which are the essential and often sole constituents of most index descriptions, at least for post coordinate systems) but also the full range of interlocking relations, or syntactical elements, which convey the exact relations between these terms in that particular description. To take a rather far-fetched example, another report might refer to a high wing at subsonic speeds and unless, in the first example, High is interfixed or linked with Subsonic speed the two different subjects are not clearly distinguished. Unless we are to recognize these syntactic elements as a third parameter in the precise description of a document, they must be regarded as elements in exhaustivity and/or specificity. In the great majority of cases they do not refer to the generic level of substantive terms but reflect non-generic relations; they constitute 'relational' terms, analogous to the substantives. They are used as such in a few indexing systems and theoretically at least can themselves display varying generic levels; e.g., Influence could be replaced by the more specific Harmful Influence. It would seem, then, that these terms reflect both exhaustivity and specificity, but more often the former. Only when the relation is explicitly a generic one (as can be the case, for example, with Farradane's appurtenance operator, Ref. 23) can they be said to determine specificity.

Exhaustivity of indexing

Recall devices cannot be fully tested unless the indexing on which they are tested is exhaustive; otherwise, loss of recall at any point might be attributable to a lack of exhaustivity rather than to the device concerned. So maximum exhaustivity in indexing was attempted, at least as far as substantive terms (lexical elements) were concerned. At the same time, since it was clearly desirable to measure the effect of varying exhaustivity on different devices, it was necessary to note during the indexing which terms would in fact have been omitted if any level less than complete exhaustivity had been acceptable.

This problem was very conveniently solved by using the figures assigned to terms as a weighting device as indicators also of which particular terms would have been accepted at different levels of indexing exhaustivity. For the highest level of exhaustivity all terms would be acceptable, whatever their weight. For the lowest level of exhaustivity only those terms given the highest weight (i.e., those terms which would be regarded as essential even in relatively superficial indexing) would be acceptable.

In the result, on average, 31 terms were used per document, with 3 levels of weighting. (6, 8, 10). If only those terms weighted 8 or 10 were counted, it represented an exhaustivity level of 25 terms per document and if only those weighted 10 were counted it represented 13 terms per document.

Of course, 'complete exhaustivity' is a relative term here; strictly speaking only the use of the full text of the document including diagrams, tables and graphs, constitutes completely exhaustive indexing. But whilst the economics of mechanised aids in indexing may eventually make this feasible and its testing desirable, the degree of exhaustivity represented by 31 terms per document was thought to be a reasonable approximation to what would be regarded, for documents in this particular subject area, as extremely thorough indexing. The problem of syntactic elements ('relational terms') as an element in exhaustivity will be dealt with later.

Specificity of indexing

Precision devices cannot be fully tested unless the indexing language on which they are tested is of maximum specificity. For example, a question on Elliptical cylinders can only be matched specifically if, whenever that concept appears during indexing it is represented by the exact description Elliptical cylinders and not by a more general term such as Cylinders alone. If the indexing is not specific in the first place, there is nothing the searcher can do to improve precision by altering his search programme.

At the level of substantives, or lexical elements, specificity was fairly easy to achieve, since, by adhering closely to the language of the document and indexing exhaustively, it was reasonably certain that the specific subject of a theme or concept would be brought out. Even if the author used a more general term in the title or summary, as was often the case, the specific term would nearly always appear somewhere in the text. For example, the title might refer to a 'Laminar boundary layer', the summary to an 'Incompressible boundary layer' and the body of the text to a 'Steady, laminar, incompressible boundary layer'; the indexing would give Steady, laminar, incompressible boundary layer.

Effect on indexing procedures of methods of measurement

Having provided for the control of the major parameters of exhaustivity and specificity, the problem arose of how the different devices might be added, one by one, to the basic natural index language, so as to allow for their measurement.

Several possible methods of proceeding now presented themselves:

(1) To make one index completely devoid of any devices, and concurrently, to make a number of separate indexes, each one embodying this first index modified by a single device, e.g. one index in which the varying word forms of a term were confounded, another in which hierarchical linkages were established, etc.

(2) To make a device-less index and measure the impact of devices entirely by variations in search programming; e.g., the result of confounding synonyms could be measured simply by programming a search for 'Disturbance' as 'Disturbance + Perturbation' whereby the expansion of a class is achieved simply by making a sum of the constituent parts of the expanded class. Similarly, measurement of the effect of confounding word forms could be effected by programmes such as 'Injecting + Injection + Injector ...'

In comparison with (1), this method, obviously, would be much less laborious clerically, even in the case of hierarchical linkage. It might be thought that this could be best measured by constructing a classified index: but the latter is an amalgam of several devices and the measurement of strict hierarchical linkage in isolation is measurable quite effectively by such programmes as: Wave + [(Wave x (N + Standing + Blast + Shock)] to expand an initial class like 'N Wave' to the generic containing class 'Wave'.

Of course, such search programmes required the compilation of code dictionaries - of synonyms, of word-forms, of hierarchies (i.e. of classification schedules). But the indexing itself, so far as these recall devices were concerned, could be done without any regard to these devices whatsoever, since the relations concerned did

B 1590	AUTHOR STONE, A. TITLE Effect of stage characteristics and matching on axial flow compressor performance	Indexer	Date
Base Document A137	REFERENCE Trans. American Soc. of Mechanical Engineers 80, 1958, p1273	GM.	17-6-63

THEMES (partitioning)		CONCEPTS (Interfixing)		CONCEPTS (Interfixing)		TERMS & WEIGHTS		TERMS & WEIGHTS	
A	cd effect of a use of ef	a	Stage characteristics	t	Range of operations	Stage	10	Blade	5
B	——— b ———	b	Stage matching			Characteristic	10	Range	5
C	cdh	c	Axial flow compressor	u	Stage flow coefficient	Matching	10	Operations	5
D	cdi effect of j with k	d	Stage performance			Axial	10	Mass	8
E	cdl effect of g	e	Test data	v	Mass flow	Flow	9	Choking	8
F	cmp effect of n	f	Analysis	w	Choking flow coefficient	Compressor	10	Line	5
G	——— vo	g	Mach number			Performance	10	Slope	6
H	cmdq	h	Velocity distribution	x	Surge line	Test	9	Knee	6
I	cmdr	i	Temperature co-efficient	y	Change in slope	Data	9	Table	6
J	cmdt effect of sg	j	Flow co-efficient	z	Knee double valued performance curve	Analysis	9	Curve	6
K	cmv effect of g	k	Constant flow angle			Mach	7	Unstalling	6
L	——— v	l	Cascade losses	aa	Unstalling hysterisis	Velocity	7	Hysterisis	6
M	cbv effect of w	m	Idealised compressor	bb	Inlet guide vane stagger	Distribution	7	Inlet	6
N	cbx effect of o	n	Total pressure ratio	cc	Uprating stage one	Temperature	6	Guide	6
O	cbxy effect of o	o	Percentage of design speed	dd	Uprating stage two	Coefficient	6	Vane	6
P	cbz effect of aa	P	Performance	ee	Blade stagger	Constant	6	Stagger	6
Q	cv effect of bb	q	Stalling point	ff	Stage loading	Angle	6	Uprating	6
R	——— cc	r	Compressor surges	gg	Annular area	Cascade	6	One	6
S	——— dd	s	Pitch line blade speed			Loss	6	Loading	6
T	cx effect of ee					Idealized	8	Annulus	6
U	——— ff					Total	5	Area	6
V	——— gg					Ratio	5	Number	7
W	——— vq					Percentage	8	Line	6
						Design	8	Valued	6
						Speed	8	Two	6
						Stalling	8	Pressure	7
						Point	8		
						Surge	8		
						Pitch	6		

FIGURE 4.1 INDEXING SHEET FOR DOCUMENT 1590

not depend on the context of the individual report being indexed but on the context of the index language as a whole; in other words, the relations were paradigmatic rather than syntagmatic. It may be noted that all the devices measurable by Method (2) are recall devices - i.e., devices for elaborating or expanding the classes given in the index description of a document.

(3) To incorporate particular devices in the original indexing but in such a way as to make them detachable when required; e.g., to attach weights to terms which could be counted when figures for weighted indexing were required but ignored for unweighted indexing. This method, also much less laborious than Method (1), would be particularly appropriate for those precision devices which depended on the context of the individual report being indexed, and which could not therefore be measured by Method (2).

It was finally decided that the indexing proper should be done on the basis of Method (3); that is to say, the indexing would be basically postcoordinate, and take into account only the precision devices of weighting, links and roles (whilst observing a high degree of exhaustivity and specificity). Method (2) was to be used in the measurement of the recall devices of Synonyms, Word-forms and Hierarchical linkage (generic and non-generic). Associative indexing could not be measured within the conditions above, but it was hoped that the indexing would be sufficiently exhaustive to allow some tests of associative techniques to be made by other investigators. Te Nuyl's device was also ignored at this point, since it was clear that our indexing language could always be translated into dictionary-based clusters when necessary for measurement by Method (2). Bibliographical coupling, since its classes are not defined by subject descriptions, required quite separate measurement, and is discussed in Chapter 7.

The major precision device of coordination is, in a postcoordinate system, purely a search device and its measurement does not fit exactly into either Method (2) or (3). It is perhaps necessary to mention at this point that post coordination in itself does not constitute an indexing device. It is essentially a method of recording subject descriptions in a physical form which allows equally free access to whatever combinations of terms are requested. A precoordinate system, on the other hand, allows direct access only to certain selected combinations of terms. Other combinations constitute distributed relatives and access to them is to this extent made less convenient (although it is by no means forbidden, or made impossible, as is sometimes suggested). But the class defined by coordinating two or more terms is exactly the same, whether the operation is performed at the indexing stage (precoordination) or at the search stage (post coordination). The relative convenience with which access to such a class is gained was not something with which this investigation was concerned.

The form in which the indexing was recorded is best shown by Fig. 4.1 which shows the index sheet for document 1590. Author and title details were printed on the sheet. The indexer then analysed the document in four stages: firstly, 'concepts' were distinguished as a first-stage interfixing device ('link'); these are not easily defined (and this must be recognized as a theoretical weakness) but their practical function was reasonably clear. This was to remove the first level of vagueness and ambiguity inherent in words taken singly, by not accepting adjectival forms alone but only in conjunction with the terms they qualified. So terms which in isolation are weak and virtually useless as retrieval handles were given the necessary context; such terms as High, Number, Coefficient, Main, Trailing, Angle, Aspect which in practice do not form classes for which requests are made, appeared in conjunction

with other terms, to produce meaningful class terms - e.g., High subsonic speeds, Mach number, Pitching moment coefficient, Main wing, Trailing edge, Low angle of attack, High aspect ratio. Not only 'weak' terms were combined, however, for the interfixing function of concepts often produces phrases whose constituent terms are quite potent, index-wise even in isolation. For example, Cruciform wing, Circular body, Tail fins, Span loading theory, Force divergence Mach number, Wing vortex field, Rectangular wing surface, Wind tunnel wall. Such combinations make it clear, for example, that in the one document Surface relates to Rectangular wing, not Cruciform wing; that Mach number relates to Force divergence rather than to some other phenomenon; that Low relates to Angle of attack and High to Aspect ratio (and not vice-versa). Or, as in Document 1590 that Distribution relates to Velocity and Ratio relates to Total Pressure, and not vice-versa.

Secondly, the concepts were now grouped into a second-stage link device in order to display the distinct 'themes' into which the document could be partitioned. 'Partitioning' of a document is a well-established procedure in traditional precoordinate indexing and is often referred to as analytical cataloguing . Owing to the exigencies of space in the precoordinate index, such analysis is usually confined to items in which the constituent chapters, sections, etc. can stand alone; examples are symposia of various kinds, festschriften, and collections of plays. In such cases, 'standing alone' could be interpreted almost literally in the sense that each theme is dealt with in a distinct, self-contained physical section of the document. In such circumstances, partitioning could allow greater recall (resulting from greater exhaustivity of indexing) with almost no loss of precision. This was rarely the case in the aerodynamics reports constituting the test collection. In these, a particular concept might run as a thread throughout the document, appearing at different times in different contexts. So themes were not necessarily, or usually, mutually exclusive. This diffusion of various concepts throughout a document seems to be an important cause of many problems in retrieval and particularly that of the inverse relation between recall and precision; for a document whose index description contains all or most of the terms of the question prescription may yet feature those terms in an unacceptable pattern.

The example in Fig. 4.1 (Doc. 1590) demonstrates the salient features of the two stages of linking embodied in the indexing. To economise in the writing down of themes, the concepts were labelled with lower case letters and only these appeared in the themes. Where the relationship between concepts appeared to be potentially ambiguous, it was indicated in an elementary fashion by verbal quasi-role devices such as 'effect of', 'by means of' or 'use of'. So the first theme of document 1590 is to be read as Axial flow compressor - Stage performance - Effect of Stage characteristics - Use of Test Data - Analysis.

Normal practice was to give as the first theme (or first few themes where necessary), the general subject of the document considered as an integrated whole. Subsequent themes would then bring out the particular subjects which made up the whole. In document 1590, for example, themes A and B jointly represent a formal statement of the title, with the addition of Test Data and Analysis. It also demonstrates a fairly common situation whereby the title provides a reliable and succinct statement of the document's general theme.

The third step in indexing was to give weights to the concepts (and subsequently to the terms) - i.e., to allocate to each concept a value indicative of its relative importance in the document. Such a value can be regarded as an assessment of the probability that, should the concept concerned happen to be the subject of a question,

the document indexed would be of some relevance to it. It might be argued that all indexing which involves a selection of terms to describe a document (and that is all indexing - even the use of full text by a computer, with deletion of articles, prepositions, etc.) implies weighting in that the use of a subject term indicates a reasonable probability that the document is of some relevance to questions on that subject whereas the rejection of a term indicates that the probability is low or non-existent. Weighted indexing extends the range of values from two (worth using, not worth using) to whatever number of different weights are recognized. For example, if an index description contains the terms a b c d e, and weighting is assigned to each term in the scale

3. Most important terms
2. Less important terms
1. Least important terms

If a, b and c are now each weighted 3, while d and e are each weighted 1, the implication is that the probability of the document being relevant to a question a b c is that much greater than the probability of its being relevant to a question on d e.

It has already been noted that the weights given to terms could be used as the basis for measuring exhaustivity i.e., when a figure for a high level of exhaustive indexing was required, weights could be ignored and all terms regarded; when a figure for less exhaustive indexing was required, those terms which had low weights could be ignored and treated as though they were not indexed. It should be noted, however, that the use of weighting as a measure of exhaustivity is purely an evaluation technique and plays no part in normal indexing, for in the latter, weighting only comes in as a device when it is applied also to the question. Then a question term with a given weight will accept as a match only those index terms which have the same (or a higher) weight. This procedure inevitably alters the boundaries of the classes defined in searching and proves weighting to be an independent index language device and not just a reflection of exhaustivity.

The rejection of an index term as irrelevant is now performed at the search stage, whereas exhaustivity of indexing is decided, of course, at the indexing stage. For example, suppose a question containing terms a b c d e, and a relevant document which has been indexed with weights as a^3 b^3 d^1 e^2 g^2 etc. If we were simply measuring the effect of exhaustivity, we would say there was a match of four terms (a b d e) when indexing was fully exhaustive (all weights accepted). If terms with the lowest weight (1) are now ignored, the match is reduced to three terms (a, b, e); if only the highest weight of terms is accepted (i.e. the lowest level of exhaustive indexing), then the match is reduced to two terms (a b). Here we have been using weighting purely as a measure of exhaustivity of indexing, but to consider its effect as a precision device, assume that each term in the question is weighted, and that the search specification is now a^3 b^3 c^2 d^1 e^3. The term c is rejected because it does not appear in the indexing; e is rejected because the search requirement is for a term with a minimum weighting of 3, whereas in the document e has been indexed with a weight of 2. This now means the class accepted has altered to a b d, whereas with variations of exhaustivity it was respectively a b d e, a b e or a b.

As to the problem of how to weight, two general approaches seemed possible and both were made. Firstly, for three hundred documents, weights were assigned on a quasi-statistical basis, dependent on the sections of a document in which a term appeared. If a term appeared in the title, the summary or abstract, the introduction, the body of the text, the conclusion, and the list of cited references, it received a

weight of ten. If it failed to appear in any one of these positions its weight was reduced by one point for each failure to appear. Since the indexing was strictly according to the natural language of the document this meant that all terms received some weight insofar as no terms were used which did not appear somewhere in the document.

Secondly, for the rest of the collection, weights were assigned subjectively. In most of the literature on weights, notably the pioneering articles by Maron and others (Ref. 24), weights were assigned subjectively to individual terms. This proved unsatisfactory; e.g. in a document in which 'Low aspect ratio wings' constitutes a central theme, can the single terms Low or Aspect or Ratio possibly be regarded as crucial in themselves? The significance of a term in a document is very often lost if the term is robbed of its context. So weights were assigned to concepts and the individual terms within a concept received the weight given to that concept.

A range of six different weights was again adopted. This time it attempted to combine a measure of the importance of the concept in relation to the total message of the document (which is another way of referring to the document's probable relevance to a question entailing the concept) with an assessment of its significance in retrieval terms, i.e., its potency as a retrieval handle in the particular collection indexed. The subject significance was measured by reference to a trio of values: these assume that a document can normally be regarded as consisting of its integrated subject as a whole (its main theme), together with one or more subsidiary themes, which may vary in importance from quite major component themes to quite minor, marginal themes. This assumption is a simple extension of the analysis into themes already described as 'partitioning'. According to the status of the theme in this scheme it received weights between 9/10, 7/8 or 5/6:

Weights

9/10 For concepts in the main general theme of the document
7/8 For concepts in a major subsidiary theme
5/6 For concepts in a minor subsidiary theme.

When assigning the weights to the individual terms, the higher weight of the pair assigned to the concept concerned was used if the term was considered to be a very potent one; potency here was regarded as a mixture of word frequency in the total collection (indicating roughly the generality or otherwise of the request in the context of the particular collection), and 'concreteness', whether the term was likely to be requested as the focal point of a subject or whether it was too vague to be the object of a direct and separate request. For example, in Document 1590 (Figure 4.1) the concepts of Themes A and B were each allocated the top weight; the individual terms which were considered potent (e.g. Stage, Matching, Compressor, etc.) received a weight of 10; Flow (on the grounds that it was a very common term) Test, Data, Analysis (on the grounds of vagueness) received the lower top weight of 9.

One small point to be noticed is that an individual term was always given the weighting of the more heavily weighted concept if it appeared in more than one concept. In the example, concept m is Idealised compressor, and the concept is given a weight of 8. However the term Compressor has previously occurred in concept c, Axial flow compressor, for which it received a weight of 10, and this it retains.

The fourth step in the indexing was to write out the terms individually and attach the weights to them as explained above.

The fifth step was to assign roles to the terms. At this stage a temporary failure of plan has to be recorded. If roles were to be assigned, it was obviously desirable to assign them at the time of indexing. Preliminary analyses were made of a number of complete indexing descriptions in order to establish the major variations of role occurring amongst the terms and the degree to which this might lead to ambiguity.

It was stated earlier that the main function of roles would seem to be the removal of a certain type of ambiguity beyond the power of simple links to remove. Links, it should be remembered, simply assert that some relation exists between two or more terms - between a and b, say, rather than a and c. A role states what the relation is.

From the preliminary analyses, it became apparent that the roles developed by Western Reserve University, American Institute of Chemical Engineers and by DuPont were not appropriate, designed as they were for subject areas (chemistry and applied chemistry) in which the appearance of the same term (e.g. a material) in significantly different roles was a fairly frequent experience. The Cranfield investigation of the W.R.U. index had already shown some of the drawbacks associated with roles, even in those fields for which they seemed particularly appropriate, and our analyses confirmed these. A major problem was the fact that when a term plays one particular role in an indexed document, it does not necessarily make that document irrelevant to a question in which the same term features in a different role: for example, a document on the 'Use of mufflers to control the sound of jets' suggests the roles: Muffler (Agent of operation), Control (Operation), Sound (Product), Jet (Cause). If a question were now asked on the 'Effect of mufflers on jets' the role of Muffler might well be designated as Influencing factor (cause) and that of Jet as Thing affected; the relevance of the document to this question would be obscured by the different roles assigned to the otherwise matching key terms. Such a situation implies that several quite different roles might be acceptable in searching; but this comes dangerously near to negating the whole idea of roles. Such, in fact, was the situation in the W.R.U. test, when the roles were frequently ignored in W.R.U. search programmes, presumably because of awareness of the danger of demanding too close a match.

An alternative plan which seemed to be suggested by the kind of example given above was that roles should not be assigned purely according to the relations between the terms in the document or question concerned (their syntagmatic relations), but also according to a wider picture of the relations between the terms in a subject area (their paradigmatic relations). It is generally recognized that the organization of terms into facets is closely related to the provision of role indicators. In special faceted classifications it is not uncommon to find a term appearing in more than one facet, the difference being due to the difference in role played; e.g., in a classification for pharmaceuticals, the same substance might appear as a Product, a Substance Extracted, an Agent of a reaction, or as an Agent of an operation. In such a system, where prior analysis of the terms of vocabulary has been undertaken, the more enduring relations (that a problem or by-product in the propagation of jets is the production of noise, which demands control, and that mufflers are one agent of control) would be recognized and the fortuitous alteration of roles suggested in the question would not have been allowed to obscure the situation.

Consequently, a set of roles was developed along these lines so that they closely reflected the categories which would be distinguished in facet analysis of the field. But trials of these (i.e., examination of a number of indexing descriptions in order to see whether ambiguities would be removed by the roles) showed that they left untouched what was probably the commonest problem of ambiguity in the vocabulary

being developed. This was the problem of knowing what terms were qualified by what adjectives. The example of 'High aspect ratio wings at low angles of attack' has already been cited. Without some interlocking device (links or roles) this document description would respond to a question on 'Low aspect ratio wings', or 'High angles of attack'. Similarly, an index description 'Transient aerodynamic heating of external loads' might respond to a question on 'Transient loads' or 'External heating', and 'Compressible flow over an adiabatic wall' might respond to a question on 'Adiabatic flow'. However, the situation is typically one which is solved by links rather than roles, since the type of relation is not in dispute, only which of the two (or more) terms are to be linked.

If such modifiers were to be regarded as roles the only solution which would remove all ambiguity would be to give them the same role as the term they modified, and then it would work only on the assumption that links were used as well. This was therefore tried (using such roles as Specifier of agent, Specifier of structure), but only at the cost of duplicating the linking function already performed by partitioning and interfixing.

One of the few situations, in the literature indexed, in which a concept (rarely a single term) might feature in a role which appears to differ significantly from its usual one is that in which a body of data is used as an agent in the investigation of some other phenomenon (to which it therefore takes second place in the document concerned). For example, in B1204, Two-dimensional airfoil section data are used as a means of studying high subsonic speed characteristics of swept wings. But closer examination throws doubt on the necessity for using a role even here; had the section data been referred to in some other way (as a parameter influencing aerodynamic behaviour, say) its essential relation to the primary subject (swept wings) would not appear to have been altered.

It was stated earlier that in transferring concepts to themes (the second step in indexing) the exact relationship was sometimes indicated for clarity. The relative infrequency with which the need to do this arose (and Document 1590 was not typical here) was itself a warning that the field of high speed aerodynamics was not likely to be very fruitful in the evaluation of roles as a device. This impression was reinforced by our preliminary analyses. Since, then, any measure of roles was likely to reflect an uncongenial and unresponsive test environment, the very considerable effort involved in developing a set of appropriate roles, applying them and measuring the impact, seemed of doubtful worth. For this reason the preparations for the evaluation of roles as a device were temporarily abandoned.

CHAPTER 5

Formation of Index Languages

The indexing described in the last chapter provided a number of different index languages: first, one consisting of single terms in the natural language of the documents indexed; second, this initial language made more precise by the recognition of 'concepts', reflecting a first level of interfixed relations; third, a yet more precise language recognizing a further level of interfixed relations in the form of 'themes'; fourth a language in which the relative importance of the terms was recognized (in the form of weights). Combinations of these provided still more precise languages; e.g., combination of the third and fourth.

Insofar as the indexing recognized substantive or lexical elements primarily and lacked the relational, or syntactic device of role indicators, it was something less than completely exhaustive. But apart from this, all the precision devices had been accommodated and the next step was to establish facilities for expanding the elementary classes and forming further index languages - i.e., to construct recall devices. This chapter deals with this activity in relation to

1. Single terms
2. Simple concepts
3. A pre-established thesaurus

1. Single-term classes

A preliminary task was to prune the natural language indexing of certain minor inconsistencies and variants which had inevitably crept in and which were not in themselves regarded as sufficiently serious methods of defining classes to warrant separate measurement. These initial controls involved the following:
(1) Singular and plural forms were confounded;
(2) American and English and other variant spellings were confounded; e.g. gage and gauge, fiber and fibre, Von Karman and Karman.
(3) Certain qualifiers of terms (affixes, hyphenated-forms which were sometimes separated, etc.) were disregarded; e.g., built-up, pitch-up, rolled-up, etc. were treated as built, pitch, rolled; ellipse-like, jetlike, etc. were treated as ellipse, jet.
(4) Numbers as qualifiers were separated and treated as separate terms; e.g. Mach 6 became 'Mach' and '6', N.P.L. 18 x 4 (a wind tunnel) became 'N.P.L.' and '18 x 4'.

Table 5.1 gives the basic data regarding the number of single terms and their frequency of use after the above preliminary controls had been imposed. The full set of indexing terms is given in Appendix 5.1

Salient points are: for a collection of 1,400 documents the total vocabulary was 3,094 terms, with reductions to 2,668 and 1,816 for the less exhaustive vocabularies, (the reduction being based on the weights assigned to each term). The average number of terms used to index a document was 31.3, reduced to 25.2 and 12.9 respectively for the less exhaustive vocabularies. (A discussion of the problem of reduced vocabularies appears below).

As to the use of different terms, whilst the average number of times a term was used was 14.2 this is not a very significant figure in view of the wide scatter. Of the 3,094 terms, 1,169 were used only once; one term (Flow) was used 942 times, another (Pressure) 720. The distribution curve for word-use is shown in Table 5.2 where it is compared with three other indexes, with larger vocabularies. It can be seen that the distribution behaves as expected in view of the fact that it reflects a smaller vocabulary than the other three. In fact, the frequency of use proved to be remarkably consistent with the well-known Zipf distribution of words according to their frequency of use in natural language texts. It will be seen that some 10% of the terms (the most

Collection size	1400 documents
Total postings of terms	43,857
Average postings per document	31.3
Total unique terms	3094

Variations in exhaustivity

	Total terms in vocabulary	Average Postings per document
Maximum exhaustivity (all weights)	3094	31.3
Medium exhaustivity (Weights 7/10)	2668	25.2
Minimum exhaustivity (Weights 9/10)	1816	12.9

Use of terms

Average usage per term	14.2
Terms used once only	1169
Terms used more than once	1925
The first ten terms, ranked by usage:	Flow (942)
	Pressure (720)
	Boundary (512)
	Layer (512)
	Distribution (442)
	Theory (400)
	Velocity (360)
	Supersonic (352)
	Mach (344)
	Equation (312)

Variations in vocabulary size (according to different index languages)

Language 1 (Natural language, single terms only)	3094
Language 2 (Lang. 1 with synonyms confounded)	2988
Language 3 (Lang. 1 with word forms confounded)	2541
Language 4 (Lang. 1 with synonyms and word forms confounded)	2444
Language 7 (Lang. 1 with minimum hierarchical reduction)	1217
Language 8 (Lang. 1 with medium hierarchical reduction)	796
Language 9 (Lang. 1 with maximum hierarchical reduction)	306

(383 Proper names are not included in the counts for languages 7, 8 & 9)

FIGURE 5.1 NATURAL LANGUAGE SINGLE TERM DATA

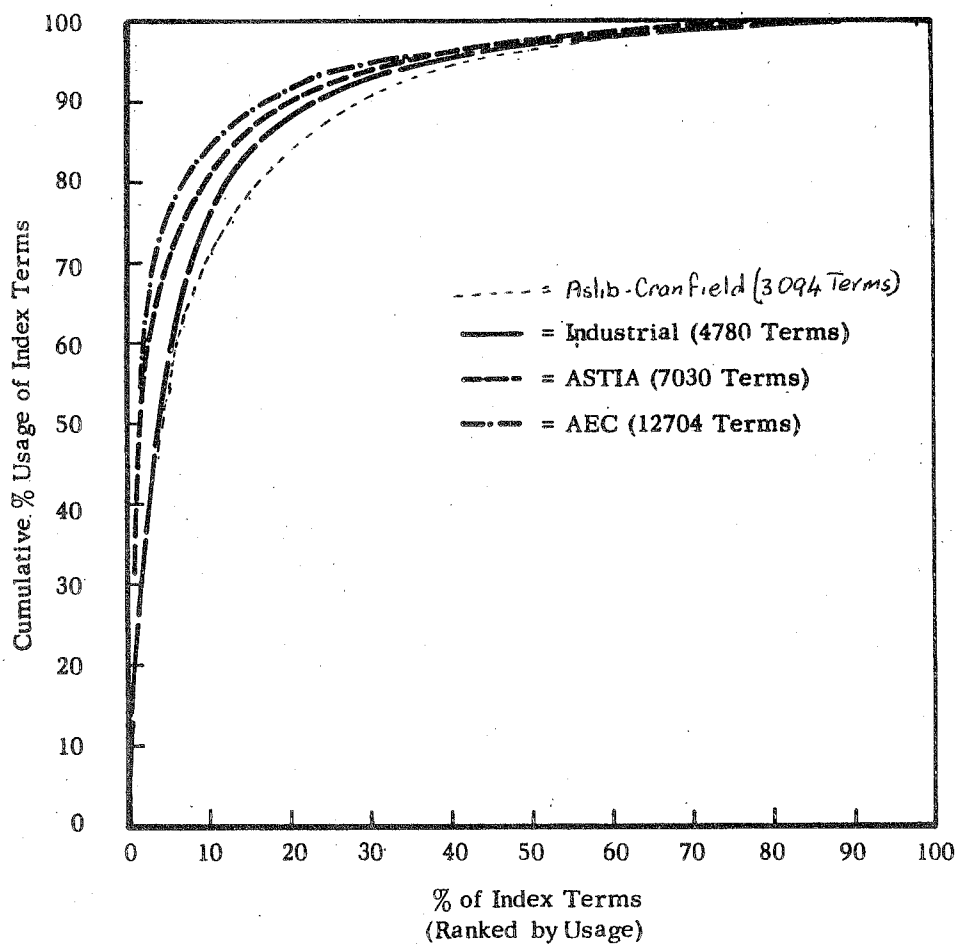


FIGURE 5.2 PATTERNS OF TERM USAGE

(original
(Original figure taken from 'Centralisation and Documentation", Arthur D. Little Inc)

used ones) accounted for 68% of the indexing postings, and 30% accounted for 92% of the postings, after which the curve flattens out.

Reduced vocabularies

Some explanation of the problem of vocabulary reductions referred to above seems desirable. Generally speaking, all recall devices imply a smaller vocabulary (with bigger classes), and precision devices imply a larger vocabulary (with smaller classes). A class is enlarged by confounding two or more classes which previously had a separate existence; contraction is the reverse process. By 'vocabulary', we mean the total number of discrete indexing elements, lexical and syntactic (i.e., substantives and relational terms) provided in an index language. It may seem surprising that links are included in a statement of vocabulary size, since they are not discrete devices in the sense that they are countable in the way lexical terms and roles are, but vary with the number of documents indexed. However, by the fundamental criterion of whether they define particular classes which would not be distinguished without them, they must be regarded as part of vocabulary size.

It should be noted that vocabulary size, under normal indexing conditions, is not necessarily a determinant of the specificity possible in an index language. This is because increased specificity is always obtainable by coordination; e.g., if the vocabulary contains the terms Flow and Supersonic, class Supersonic flow is specifiable by coordinating these two terms. Theoretically it is possible to specify almost anything in this way; e.g., Air x Cushion x Vehicle is a simple conjunction of the separate terms normally used to name this thing; but even where a name in no way defines the nature of the thing it represents, it may be specified uniquely by contrived analytical 'definition' e.g., in the W.R.U. Semantic Code, Tempering is represented by Process x Metal x Heat x (number) where the number is an arbitrary code symbol distinguishing this particular heat process on metal from any other. Perhaps the extreme example of the use of reduced vocabularies, with precise description resting on the various conjunctions of a few fundamental terms was the Malvern experiment (Ref. 25).

In the case of single-term classes without coordination, however, a reduced vocabulary can be an absolute bar on the specificity possible. If no coordination is used, a single-term vocabulary of 1,500 specifies only half the classes specified by a 3,000 term vocabulary. So far as testing devices is concerned, there are two different ways of effecting the expansion of classes. One is by an absolute reduction of vocabulary whereby the reduction is obligatory for all searches; the other is by selective search programmes, whereby the effective reduction is permissive and may or may not be utilized in a particular search. In the first case the reduction is measurable (i.e., in terms of the number of discrete classes distinguishable) and in the other it is not.

Obligatory reduction of vocabulary

Here, there is an absolute 'block reduction' (a block of classes being condensed into one) in the number of classes recognized, and the indexer and searcher has no option but to accept the confounding of more specific classes which is implied. This was the case with reduction by synonym-control and by confounding of word forms. It was also the case with the single-term hierarchies, although reduction by hierarchy may be achieved permissively and was in fact done this way in the testing of 'concept' hierarchies. This point is explained later on.

The use of quasi-synonyms* to enlarge classes is a permissive device; since the final, expanded classes do not totally exclude the continued separate use of the terms confounded (as is the case with real synonyms) no figure showing the exact degree of vocabulary reduction is possible. For example, the expanded class for Bow is Bow + Bowing + Ahead + Front + Forward + Forebody; the expanded class for Ahead is Ahead + Forward + Upstream. Clearly, the expansion of Bow does not result in the obliteration of the separate class 'Ahead', which not only continues to exist but is in turn expandable by the addition of other quasi-synonyms. In an index language which confounds true synonyms only, the reduction is once and for all and the terms no longer have a separate identity.

At this stage of our thinking about the function of vocabulary size as the main determinant of recall and precision, it seemed desirable to have as exact a measure of this parameter as possible. So the first testing of hierarchy took the form of a fixed (and therefore accurately measureable) reduction in vocabulary size.

Hierarchical reduction

The two methods of measuring hierarchy as a recall device (i.e., by obligatory, block reductions in vocabulary size and by selective searching through different hierarchical paths) are demonstrated below. The first example is one of a 'concept' hierarchy - i.e., one not restricted to single terms and one place per term. This is in order to show more clearly the two methods, and also to emphasize the distortion which results from restriction to a 'one-place' hierarchy of single terms. The latter results in the exclusion of some terms which are located in more general categories and the result is seen in the second part of the example. A hierarchical notation is attached to this example in order to make the permissive search clearer (in the schedules actually used, notation was purely ordinal).

'Concept' hierarchy demonstrating hierarchical reduction

a	Experimental wind tunnel methods for investigating flow
ab	Visualization methods
aba	Using smoke, vapours, etc. (3)
abaa	Vapour screen (1)
abab	Fog (1)
abac	Wood smoke (1)
abad	Oil smoke (1)
abb	Using coatings, flows, etc. (3)
abba	Oil flow (1)
abbb	Oil film (1)
abbc	China clay (1)
abbd	Phosphorescent lacquer (1)
abbe	Ink flow (1)
abc	Using spectrum (3)
abca	X-ray spectrography (1)
abd	Using Stroboscope (3)

*'Quasi-synonyms' are terms which can be used synonymously in certain contexts, but which are not true synonyms. (see page 68)

	[Experimental wind tunnel methods]
	[Visualisation methods]
abe	Using shadowgraph (3)
abf	Using photography. Photorecording (3)
abfa	Drum camera (1)
abfb	Photomultiplier (1)
abfc	Television (1)
abfd	Motion picture (2)
abfda	High speed (1)
abfe	Schlieren (2)
abfea	Spark (1)
abfeb	Photomultiplier (1)
abff	Interferometry (2)
abffa	Fringe shift (1)
abffb	Interferential strioscopy (1)
	etc.

The full hierarchy provides twenty-nine classes.

The first reduction, replacing the terms marked (1) by a see reference to their immediate containing head (e.g. Drum camera see Photography), leaves eleven classes, namely

a	Experimental wind tunnel methods
ab	Visualisation methods
aba	Using smoke, vapours etc.
abb	Using coatings, flows etc.
abc	Using spectrum
abd	Using stroboscope
abe	Using shadowgraph
abf	Using photography
abfd	Motion picture
abfe	Schlieren
abff	Interferometry

The second reduction, similarly replacing terms marked (2), which now include those originally marked (1), by see references to their containing heads, leaves eight classes, namely

a	Experimental wind tunnel methods
ab	Visualisation methods
aba	Using smoke, vapours etc.
abb	Using coatings, flows etc.
abc	Using spectrum
abd	Using stroboscope
abe	Using shadowgraph
abf	Using photography

The third reduction similarly replacing terms marked (3), leaves two classes, namely

a	Experimental wind tunnel methods
ab	Visualisation methods

In this way figures are obtainable to show the exact effect of moving, say, from

the quite specific class Interferential strioscopy to Interferometry in general, then to Photographic methods in general and then to Visualization methods in general, and so on.

It will be noted that by the second reduction there is no separate class left for Schlieren photography, whilst a distinct class is retained for Stroboscope. Yet in this subject field Schlieren photography is a decidedly more important class than Stroboscope This suggests that reduction by purely hierarchical criteria may be unsatisfactory. When we reflect that the choice of terms within categories and the choice of the categories themselves is ultimately a matter of literary warrant, it is reasonable to assume that reduction of classes hierarchically should not be a rigid process, but should take note of the weight of literature in the different classes, so that Schlieren photography, for example, might be retained although all other subclasses at that level were removed and incorporated in the containing class. For the single term hierarchies, this line of reasoning led to the abandonment of 'pure' hierarchical reduction and the incorporation of judgements as to the relative importance of particular classes, and the noting of word frequencies in determining which classes should be retained intact at a particular level of reduction.

The above example stresses the primary function of hierarchy as a recall device, whereby the index vocabulary is systematically reduced and the scope of each remaining class is consequently widened (hence the greater recall). In practice, however, by varying search programmes, hierarchical linkage allows movement in both directions - to greater precision by refining class definition or to greater recall by coarsening class definition. If specific indexing is assumed (i.e., each document, or document-theme, is assigned to its most specific class) a search may be made in a number of different directions through the hierarchy. For example, a searcher commencing at Photorecording abf may find the amount of material there unexpectedly excessive and so decide to search a narrower class. This may be done, of course, independently of hierarchy, by adding a qualifier or two (e.g., moving from Photorecording to Photorecording in high speed wind tunnel). But it may also be done by moving down the hierarchy (e.g., to Schlieren tests).

Such a decision implies that the links established by the hierarchy are permissive, not obligatory, and that the searcher selects from a comprehensive hierarchy just those routes he considers likely to be fruitful. It is not certain whether figures produced for a number of such searches would be useful in the sense of allowing firm generalizations to be drawn since much will depend on the subject field, and on the choice of pathways followed in searches (4) and (5) as discussed on the next page. Assume that, as shown below a, aa, etc. are terms in hierarchical relation:

a
 aa
  aaa
  aab
  etc.
 ab
  aba
  abb
  abc
  etc.
 ac
 ad
 ae
 etc.

Then, if a request is made for ab, there are the following alternative basic programmes which can be used.

(1) Term and species - i.e. ab + aba, abb ... abz.

(2) Superordinate - i.e. ab + a (but excluding other subclasses of a; this is how 'generic' search is popularly interpreted in general library practice where a might represent a general treatise on the subject of which ab is a subclass).

(3) Generic - i.e. a + aa + aaa, aab ... aaz + ab + aba, abb ..., abz + ac + aca, acb ... acz + az + aza, azb ... azz (this is how 'generic' search is normally interpreted in machine systems and is analogous to search (1) where the content of class ab is taken to include the individually specified subclasses aba, abb, abc ...).

(4) Coordinate - i.e. a selection of the more likely classes coordinate with ab, e.g. ab + aa + ad; e.g., in a category of three-dimensional shapes a search for Spheroid can be extended generically by searching under Body of Revolution, or by examining all the different kinds of Body of Revolution (Sphere, Hemisphere, Ogive, Cone, etc.). But some of the latter will be more closely connected to Spheroid than others (e.g. Sphere) and an intermediate search, stopping well short of examining every species, can be made. It is true, of course, that a 'closer connection' between several subclasses implies the possibility of an intermediate step of division being inserted. But we have to stop somewhere.

(5) Subordinate - i.e. a selection of the more likely subclasses of ab, e.g. abb + abn.

It will be noted that programmes (1), (2) and (3) are obligatory; no freedom of choice is given to the searcher, but (4) and (5) are permissive, the decision as to the formation of the classes being at the discretion of each searcher.

To return to the matter of variations between single-term and concept hierarchies, the shrinking of a concept hierarchy by restricting it to a one-place hierarchy of single terms is seen by the following, which is the schedule given on pages 62-3 and reduced in this way.

V/2	Experiment + Experimental
V9a/10	Visual + Visualization
V11	Spectrography
V12	Stroboscopic
V13	Shadowgraph
V14,16	Photography + Photorecording
V17	Schlieren
V18	Spark
V19	Interferometry
V21	Interferogram
V23	(Fringe) Shift
V24	Strioscopy
V25	Interferential
V25a	Clay
V25b	China

Of the 29 classes in the concept hierarchies only 13 appear in the one-place hierarchy plus two (Interferogram and Interferential) which appear for the reasons

explained on p. 70. The others are now distributed under the more general categories as explained: i.e., terms like Smoke, Vapour, Screen, Fog, etc. appear in other contexts as well and are therefore placed in more general categories. Under these conditions, as soon as reduction of the original full vocabulary begins, it becomes very difficult to maintain the sensible boundaries of a class like Visualization tests. For if a question on this were now programmed to included such terms as Fog, etc., it is very likely that these in turn have been swallowed up in the reduction of the general categories and that their inclusion in the search programme can only be had at the cost of bringing in a number of other terms, such as Cloud, Snow, etc. which are quite irrelevant to the context of Visualization tests.

Another drawback, related to the foregoing, is the loss of connection suffered by terms treated in isolation and not in coordination. For example, a search in response to a question on 'Flow in channels' would fail to draw in documents indexed by 'Couette flow' or 'Poiseuille flow'. Although there is a clear connection between these at the 'concept level' of types of flow, at the level of single terms there is no connection between Channel (treated as a Structure affording a passage) and the personal names Couette and Poiseuille. This situation reflects a practical difficulty in post-coordinate systems which rely on single terms - that of indicating connections (in a thesaurus, say) when these connections are dependent on particular conjunctions; e.g. this would imply a reference of the kind:

Channel: when coordinated with Flow

see also: Couette Flow
Poiseuille Flow.

It is important, therefore, to remember that the performance results of the single-term hierarchies reflect the use of one particular application of hierarchy as a recall device - i.e., its expansion of classes by fixed reduction of vocabulary size. Also, that this was a procedure determined largely by considerations of measurement rather than regard for the normal use of hierarchy as a recall device in practical indexing. There seems little doubt now that it is a mistake to regard hierarchy as an obligatory recall device. Its essential function is to act as a permissive device, allowing flexible choice of class adjustment according to the demands of the question context in a way which is not feasible within the artificial conditions of single-term hierarchies. From this viewpoint, the performance figures for the concept hierarchies described in the next section are a better guide to the value of generic hierarchy as an indexing device.

Languages based on single-terms and embodying recall devices

Before describing these in detail it may be noted that a certain artificiality inevitably accompanies the application of recall devices to single terms in isolation, simply because, in many cases, words make little sense when stripped of accompanying qualifiers, etc. For example, the problem of synonymity in index languages frequently demands recognition of phrases, as when 'Ground effect machine' is equated with 'Air cushion vehicle' although at the single term level there is no synonymity between the constituent terms; and a term like 'effect' on its own is practically valueless as a retrieval handle (which is what any class, in indexing, aims to be).

Traditional, pre coordinate indexing has always begun with some degree of coordination. Even in analytico-synthetic classifications, where 'elementary constituent terms' are separated out as far as possible, there is no rigid adherence to the single term as the basis of the language; for example, 'Ground effect machine' would be comfortably accommodated in a Vehicles facet. But coordination of terms is an extremely

potent precision device and whilst the measurement of its impact, alone and in conjunction with other devices (including all the recall devices) was of course essential, it could not be included as a variable when measuring the impact of the other devices on single terms. Completely free manipulation of classes is only feasible if we begin with single terms; this is a basic assumption of post-coordinate systems. It was clearly desirable to obtain performance figures for the impact of single devices on single classes before attempting to measure the joint impact of several devices - and even a slight degree of pre-coordination would have compromised such figures.

Confounding of synonyms

This is perhaps the most obvious of all indexing devices and the one least likely to be neglected even in the crudest of indexes. Much of this work was straightforward: e.g., recognition of synonymity between such terms as Acoustics and Sound, Amount and Quantity, Calculation and Computation, Axisymmetric and Axisymmetrical, Vertex and Apex, Viscid and Viscous. However, exact synonymity is relatively rare (there might even be argument about some of the examples above). The commoner situation is a partial synonymity, where terms are interchangeable only in particular contexts. The evident richness of the English language, even in the literature of high-speed aerodynamics, led to quite different terms being used on different occasions (but often in the same document) to represent the same thing; e.g., the notion of Proximity might be conveyed by that term or by Near, Nearest, Nearly, Close, Closely, Off, Adjacent, Contact, etc. Two terms which might be used synonymously on most occasions would occasionally diverge seriously; e.g., Interplanetary flight is equated with Interplanetary voyage; Hypersonic flight with Hypersonic flow, Free flight with Free falling. But Voyage, Flow and Falling cannot be regarded as synonyms.

The establishment of a synonym-list suffered one unfortunate drawback in that it preceded the construction of classification schedules. Ideally, a synonym-list in any given area should be extracted from a detailed classification; only by a systematic organization of all used terms according to their meanings can the ramifications of complete and partial synonymity be exposed. For administrative reasons, however, it was desirable to proceed with the measurement of relatively straightforward devices like synonyms, word-forms, weights, etc., whilst the preparations for the more difficult devices like hierarchical linkage went on.

The truth of the assertions just made was borne out when the classified hierarchies were completed, in that a number of further synonyms, unrecognized in the synonym programme, were disclosed. However, these cases were relatively few and we are satisfied that the synonym-list on which the tests were made was reasonable on the whole.

One difficult decision necessary in establishing the synonym list was whether we should recognize variant word forms as synonyms. Whilst the usual view of synonymity excludes variant word forms as being examples of a grammatical rather than a semantic relationship, the practice of many subject heading lists, thesauri, etc. which fail to recognize variant word forms at all is an implicit acceptance of the view that such variants are virtually synonymous. Certainly, in the process of indexing by natural language terms extracted from the documents, the fact that one word form rather than another was selected was often almost fortuitous and this is shown, with examples, in the section on hierarchical linkage. However, this argument was not regarded as acceptable; a thesaurus, etc. may fail to recognize variant generic levels as well as variant word forms, and so implicitly confound a genus and its species.

Should a synonym list accept this also as a type of synonymity? In view of the fact that separate measurement was being made of these other devices it was decided to interpret synonymity as strictly as possible, but the joint use of synonyms and word-forms (for example) was also measured.

Another weakness, showing itself as a result of the single term basis, has already been referred to - the inability to cater for synonyms which appear only at the 'concept' level of phrases; e.g., Flexural centre and Shear centre, Surface friction drag and Skin friction, Uniform surface temperature and Constant wall temperature.

These considerations led inevitably to the recognition of 'quasi-synonyms' as a variant of 'pure' synonymity.

Confounding of quasi-synonyms

In this device, those terms are confounded which on some occasions, but not all, are used synonymously; e.g. Subsonic and Subcritical; Compressor, Impeller and Pump; Blunt, Blunted, Bluff and Rounded; Medium, Environment and Surrounding; Region, Atmosphere and Material. A certain overlap appears here with the device of confounding word-forms; on many occasions, different word-forms would be used in a report indiscriminately to convey the same notion. The same overlap would appear, of course, with 'true' synonyms if 'conveying the same meaning' were the sole criterion. But in a single-term index language, the extra element of context is lacking; although the phrases 'Seasonal density variation' and 'Variation of density with season' were virtually synonymous in the reports indexed, if the single terms 'Seasonal' and 'Season' are taken alone they cannot be regarded as synonymous.

With quasi-synonyms this restrictive rule did not apply, since acknowledgement of differences conveyed by variations in context is the basis of the device. So variant word-forms were accepted, where applicable, as one type of quasi-synonym, e.g. Flexural and Flexible, Flow and Flowing.

The establishment of synonyms and quasi-synonyms was done as the indexing progressed, with the aid of glossaries, classification schedules, etc. We have already noted that theoretically, the only sure way of tracing synonyms is by a close classification of the field, utilizing the defining functions of classification to expose synonymity between terms used. However, the compilation of classification schedules was a much longer task and, for clerical reasons, the list of synonyms was compiled first, in the manner indicated above. Again, for clerical reasons, the synonyms were worked out fully only in the case of those which were demanded by the search programmes - i.e., the starting terms from the questions.

Confounding of word-forms

There is little to be said of this device, which was the simplest of all to establish. The expanded classes consisted of a comprehensive aggregation of all the various forms a given word-root could take, whether with prefixes, suffixes, participles or gerunds, etc. Examples are: Angle, Angled, Angular and Angularity; Asymptote, Asymptotic and Asymptotically; Axial, Axially and Axis; Blunt, Blunted, Blunting and Bluntness; Bound, Boundary, Bounded and Bounding. Their relations to synonyms and quasi-synonyms have been mentioned and their place in one-place single term hierarchies will be considered in the next section.

The terms used in searching, together with their synonyms, word endings and quasi-synonyms, are given in Appendix 5.2.

One-place, single-term hierarchies

By far the most difficult device to establish was that involving hierarchical linkage. Two major (and connected) problems arose. Firstly, the arbitrary and somewhat artificial restriction implicit in the need to place each term in one hierarchy only. This arose inevitably from dealing with single terms and meant that the assistance normally given to definition by context was absent. Secondly, the problem of interpreting the prolixities and ambiguities of the natural language index vocabulary in terms of this particular type of controlled vocabulary.

Problems of hierarchy

The term hierarchy as normally used in indexing can mean one of three different things:

1) A generic hierarchy; i.e., a system of subordinating some terms to others whereby only terms which reflect the relationship of being kinds of a thing are subordinated to that thing. Other relations are excluded. But the basis for the formation of the species may or may not be a 'fundamental' characteristic.

2) A strict genus/species hierarchy, differing from (1) in that it is confined to the use of 'fundamental' characteristics; e.g., Methane could not be subordinated to Fuel (as it is in the test schedules) since a fundamental definition of Methane does not require characterization by this attribute. A parallel has been drawn by Gardin (Ref.26) with the distinction between paradigmatic and syntagmatic relations, the former reflecting permanent or necessary relations and the latter temporary or contingent ones. A modern faceted classification uses both types of hierarchy in that the same term might appear in two or more different facets according to its status (as a product, an agent, etc.) and not be confined to the facet where it 'fundamentally' belongs.

3) A hierarchy which includes generic and non-generic elements; i.e., one which subordinates some terms to others regardless of the relation involved, so long as the subordinated term can be seen to belong to some category or facet of the 'containing' class, e.g., the subordination under a term of its properties, parts, processes, etc., as well as its kinds. This situation is typical of nearly all existing library classifications.

Reasons why (3) should be treated as a separate device ('non-generic hierarchical linkage') have already been given and are not considered here. In choosing between (1) and (2) for single-term hierarchies, logic seemed to suggest that (2) be chosen; for if each term may go in one hierarchy only it is arguable that that one place should at least reflect the most essential characteristics of the thing represented. On the other hand, the practical purposes of hierarchy in indexing would sometimes be ill-served by such an arrangement. This purpose is to provide for each term a set of class-mates standing in the same relation (of Thing/Kind) to the containing class and thus facilitate the expansion, or contraction, of any given class by the inclusion or exclusion, of some or all of these helpfully related neighbouring terms; and 'helpfully' here depends on the subject context.

If the terms are relegated to a 'fundamental' or 'common' category, these helpful relations tend to become tenuous; e.g., if the term Upper is located in a highly

generalized category of Spatial Phenomena, its class mates will eventually include such terms as Underwater or Buried. In a question on the Upper atmosphere, if expansion of the first term brings in such classes it is clearly unhelpful. Many other examples could be given; e.g., Boosted and Reinforced could reasonably be assigned to a basic category of activities affecting the dimensions of a thing. But in the collection indexed these terms appeared in quite different contexts - Boosted under Rocketry and Reinforced under Structures.

The procedure finally adopted was a compromise. Where, in the test collection, an index term had appeared only in one particular context it was placed in a hierarchy reflecting that context, without regard to whether it was a necessary or contingent relation. For example, Gun would be regarded as a method of propulsion in any fundamental hierarchy; but in the test collection it appears only as a designation of a kind of aeronautical testing device (a special kind of wind tunnel) and so it was located with the latter. An extreme example would be Gamma; as a single term, this could hardly appear in any 'fundamental' category other than Letters; in the test collection it appeared only as the designation of a kind of steel and was located as such.

If, however, a term appeared in several different contexts suffering a significant qualification of meaning, it was placed in a 'fundamental' category; e.g. Integral appeared in its mathematical and structural sense and was therefore placed in a Common properties category. Similarly, the term Working appeared in two main guises: to designate a particular section of a wind tunnel and to designate a fluid (e.g., a test gas). The sense of the term Working is significantly different in the two contexts and it was therefore relegated to a common properties category.

Problems of terminology

Closely related to the above problem was that of interpreting the intended meaning (from the point of view of the test collection) of the terms used in the natural language indexing. The organization of terms into hierarchies constitutes a form of controlled vocabulary, of course; in this case, it was a control being exercised retrospectively, after the indexing stage. The object was to place each term in the hierarchy to which it would have been assigned had the indexing been done using the controlled vocabulary. So where the same essential notion was conveyed in various grammatical styles, these variants would have been ignored and one form done service for all; that is to say, the particular grammatical form of a term might have to be disregarded since its semantic content in the index was the only point of interest now. For example, a writer might refer indifferently to 'reduction of x by compression', or 'reduction of x by compressing' or 'reduction of x compressively' without wishing to convey a significantly different idea. Again, any one of the phrases 'plate with curvature', 'curved plate', 'curve of the plate', and 'curving the plate' might be used in a report without any intention of conveying different nuances of meaning (i.e., without meaning to refer to the structure or the property or the operation in particular). Other examples were Test and Testing; Calculation, Calculating and Calculated; Asymptote, Asymptotic and Asymptotically. All these variations in expression were ignored and the different forms juxtaposed in the hierarchies.

Where different word forms reflected significantly different emphases in meaning they were assigned to their formal categories. So Buckle and Buckling appeared as processes and Buckled as a property; Cantilever was used to characterize a kind of beam, but Cantilevered designated a type of structure. Scooped appeared as a property and Scooping as a process.

In the same way, there were numerous examples of terms which appeared to represent operations or processes (if one regarded only the single terms in isolation) but which represented an integral part of the specification of a particular kind of thing; e.g. Settling chamber, Driving gas, Non-lifting wing, Geared elevator. Wherever such a term had appeared only in that particular context and its function as a class determinant had been to characterize the entity and not the operation, property, etc., as such, it was subordinated in the hierarchy to the entity which it specified.

The exact status of these variants on insertion into the hierarchies created a slight, theoretical problem. The confounding of synonyms in an earlier programme had already established what terms were exactly synonymous and it would have been inconsistent now to add these variants as synonyms (the weakness of a synonym programme derived before the establishment of a classification has already been noted). So they were simply clustered together as though coordinate in relation to each other. Had the measurements of single-term hierarchical linkage taken the same form as in the later 'concept hierarchies', whereby various hierarchical trails were followed in order to distinguish sharply between different relations (subordinate, superordinate, coordinate, etc.): this might have produced a very slight distortion of the performance figures. However, the measurement of single-term hierarchies only took the form of block-reductions in vocabulary size (in the manner discussed earlier in this chaper), so no harm was done.

It must be admitted that a few errors crept in, when unjustified violence was done to a category by the subordination of one of its members to another category. For example, in the overwhelming majority of cases, the term Revolution occurred in indexing as part of 'Body of Revolution'; so, according to the reasoning above, it was located in the category of Shape, since its function was to designate a particular kind of shape. However, its synonym, Rotation, occurred once or twice in its fundamental guise of a process; it is therefore misplaced under Shape. It is not thought that these occasional lapses were serious. We have already seen that in making single term hierarchies, if a term is relegated to a fundamental category this results in classes sometimes being drawn in which are unhelpfully associated; this is also what happens in the case of a lapse like the above.

Construction of single term hierarchies

Having settled on the various solutions to the problems described above, the formidable task of organizing the 3094 terms of the natural language proceeded. The basic operation was one of facet analysis (a facet being a hierarchy). A useful framework for the initial sorting was the Facet Classification compiled for the first Aslib-Cranfield Project by J. Farradane and B. C. Vickery, although high speed aerodynamics (the subject of this test collection) tended to concentrate itself in only a few of the areas covered by the scheme, and was in far greater detail than had been handled before. Particularly large categories were those relating to Bodies, to Shapes, and various Spatial and general relations, to Fluid dynamics proper, with particular clusters of detail under such topics as Compressors, Upper atmosphere studies, and Astronautics. The speed with which the last subject has developed in recent years was reflected in the fact that whereas the Facet Classification barely mentioned it, in this test collection it was a major theme.

Because no attempt was made to establish 'fundamental' categories as such, the common categories which were formed tended to be residual ones in that they contained only those terms which had not found a place in a more limited context.

A1	Particles
A3	Electron + Beta
A4	Proton
A5	Atom
A6	Isotope
A7	Ion
A8	Molecule
A8a	Mol
	(Structure)
A8b	Atomic
A8c	Molecular
A8d	Bimolecular
A9	Homonuclear
A9a	Nuclear
A10	Monatomic
A11	Diatomic
A12	Polyatomic
A13	Polymer
A14	Polycrystalline
A15	Matter + Material
	(By use)
A16	Pigment
A17	Lacquer
A18	Phosphorescent
A19	Ink
A19a	Injectant
A20	Fuel
A21	Methane
A22	Ethylene
A23	Hydrocarbon
A24	Methanol
A26	Propellant
A27	Explosive
A28	TNT
A30	Coolant
A31	Lubricant
A32	Refractory
A32a	Oxidizer
	(By origin)
A34	Electrodeposit
A35	Electroformed
	(By constitution)
A38	Metal
A39	Alloy
A39a	Bimetallic

FIGURE 5.3 SAMPLE SHEET FROM SCHEDULES OF SINGLE TERMS

Nevertheless, some of them were still exceedingly large and detailed - e.g., those reflecting spatial and shape characteristics. For these, and for the common categories of Properties, Processes, Operations, etc. the Thesaurus and Code Dictionary FROLIC produced at the David Taylor Model Basin (Ref.28) proved very useful.

Interpretation of the various word forms, etc. referred to above was assisted by the file, compiled during indexing, of synonyms, definitions, decisions, etc., by the word frequency list, and by reference to the indexing sheets of individual documents where necessary.

Sample excerpts from the single-term hierarchies are given in Fig. 5.3. (The complete schedules appear as Appendix 5.3.) It must be emphasised that only those terms appear which were used in the indexing of the test collection. Whilst this resulted in very detailed schedules in some areas, these still cannot be regarded as exhaustive of the terms in the particular area. Sometimes, if they did not happen to occur in the test collection, quite important terms will be missing.

2. Simple concepts

In the previous section we described the establishment of index languages based entirely on single words, and indicated the limitations on the performance of synonyms and hierarchies imposed by this restriction. These limitations were accepted in order to allow the examination of the performance of the different devices applied to single terms, in the absence of any element of precoordination. The next step was to accept a degree of precoordination from the outset.

Examples have already been given of the sort of simple linking necessary if the meanings of some expressions in the natural language are not to be quite lost; e.g., 'Ground effect machine' must be retained as a single concept if loss of meaning is not to be suffered. The original indexing had, of course, included a statement of the 'concepts' in each document - it was in fact the first step taken in the actual procedure of indexing a document. These concepts were now taken as the basis for the production of new synonym and hierarchy languages.

'Concept' languages

In order to reduce the task of preparing these to reasonable proportions it was decided to take a substantial subset of the full collection of 1400 documents and to make a detailed classification schedule for all the terms appearing in it. The subset consisted of some 200 documents, containing all the documents relevant to some 40 questions. In order to make the new collection reasonably homogeneous, only aerodynamics documents were included.

The performance of the index languages in this same subset was subsequently measured separately for a controlled language (based on a thesaurus) and for the 'options' investigated by G. Salton and his colleagues at the Harvard Computation Laboratory (the SMART system). Figures for the single term languages for the subset had already been obtained - they had simply to be extracted from the figures for the full collection.

No reindexing was attempted, of course, since this would have invalidated comparisons with the single-term tests. One adjustment was made, however; the

original concepts, based closely on the natural languages of the documents indexed, reflected a degree of precoordination which was excessive for our purpose, so over-elaborate phrases were now broken up into smaller units, e.g., Biconvex circular arc cross section was broken up into Biconvex cross section and Circular arc cross section; Dissociated frozen hypersonic laminar boundary layer became Dissociated boundary layer, Frozen boundary layer, Hypersonic boundary layer and Laminar boundary layer. If search were subsequently necessary for the original concept, it would still be possible by postcoordination. Meanwhile, this splitting up allowed maximum freedom in distinguishing facets and subfacets (arrays). In other words, the rigidity attending the excessive precoordination typical of the older classification systems (resulting in the obscuring of the multiple relations between facets and sub-facets) was avoided.

Formation of concept hierarchies

This task proceeded in the normal way, by the now well-established process of facet analysis. However, some of the problems which occur when making a special classification were absent or greatly reduced; at the same time, the unusual basis of the schedules (the 'natural language' concepts, already embodying a certain degree of precoordination) raised some new problems of presentation. These points are discussed later.

The procedure was as follows: the concepts (mostly short phrases like Tumbling entry, Centre of rotation, Crossed flexure pivot, but with some single words, e.g., Strips, Trajectory, Pivot, Inclination) were first grouped into the following major subject areas:

Aircraft types and parts
Bodies (Aerodynamic)
Non-aerodynamic structures
Flight: flying operations
Fluids, gases, atmosphere
Fluid flow: Kinds, Elements (vortices, jets, etc.)
Aerodynamic forces and loads, processes and properties
Aeroelasticity, flutter
Aerodynamic reference parameters (angle of attack, planform, etc.)
Mechanics, dynamics
Heat
Research: Experiment, Theory
General properties and processes.

No particular significance attached to this order; for convenience of reference it approximated to the order of terms in the original Cranfield Facet Classification. Generally speaking it reflected the citation order used in locating concepts; a concept containing notions from more than one area was located under the one appearing first in the above sequence; e.g., Wing-body interference went under Wing-body, not Interference; Spherical segment nose went under Nose (Aircraft parts) not Spherical segment (Bodies); Leading edge stall went under Leading edge. But where a clear relation, explicit or implicit, existed between two elements of a concept, and reflected a clear precoordinate indexing principle (e.g., subordination of agents to the operations or processes they serve) this was observed, even if it ran counter to the broad rule above; e.g., a Shielding mechanism is a structure (non-aerodynamic)

but it was subordinated to Ablation cooling since its functions in the literature indexed was that of an agent of the cooling process.

The fact that Heatshield was subordinated to Ablation cooling devices did not mean, of course, that it was unavailable as a member of the class Structure if this latter subordination had also been required. It was placed under Ablation solely on the score that in the test collection, or the subset, this was its most probably useful hierarchy. The concept schedules were essentially 'one-place' schedules in linear sequence, in the sense that no attempt was made to repeat one concept in several different positions should it happen to belong usefully to several different hierarchies. This last event was provided for by the rotated A/Z index, and by references within the schedules described later. It must be emphasised that the function of these schedules was simply to show as clearly and as comprehensively as possible the hierarchical relations (generic and non-generic) between the terms (concepts) so that searches could be programmed from them. The major relations were most economically displayed by physical juxtaposition. Other hierarchical relations were established via the A/Z index and by internal references (linking, for example, Heat transfer, subordinated to Thermodynamic processes, to Transport properties in general).

The index in the physical sense (the matrix of index descriptions and document numbers) consisted of the separately entered concepts designed to be searched postcoordinately (the strip, or the scan-column, method used is described in Chapter 6). So a compound like Fully developed laminar channel flow could be sought equally in any of the various arrays concerned, or combinations from them - Fully developed flow, or Laminar flow, or Laminar channel flow, etc.

It follows from the above that the problem of citation order was very much reduced, compared with a real life schedule, since it was confined entirely to the choice between two (and sometimes, but rarely, three) elements; e.g. Jet noise, Interference rocket, Laminar boundary layer heating, Surface stress, Slot blowing. But whilst these particular examples offered a serious choice between two or three equally important elements, the great majority did not even demand this; they consisted of combinations such as High pressure ratio compressor, or Hinged flap, where the major element was obvious and the other elements trivial; no hierarchy of Ratios, or things High, or things Hinged was necessary. Since the concepts represented the limits of precoordination, the problem of providing for the accurate prediction of the exact location of all potential synthesized combinations (a major function of citation order) did not arise. The problem of 'distributed relatives' was solved by postcoordination of the concepts.

It also follows from the above that problems of notation were virtually nonexistent. A purely ordinal notation to identify quickly the location of particular simple concepts was the only requirement. Problems of hospitality and expressiveness did not arise; no additions to the schedules were contemplated and no aids to display were necessary in schedules which were relatively homogeneous and designed for internal test-programming entirely.

Within each major area the various facets and arrays (subfacets) were now distinguished. At this point the problem of displaying generic and non-generic hierarchies arose; it was met as it usually is in conventional library classification, by subordinating to a thing its various categories - its kinds, parts, properties, processes, etc. Below is a brief extract from the schedules which are given in full in Appendix 5.4, followed by an explanation of some of its features:

E64 Compressor
E65 Centrifugal + Radial flow compressor.
+ Radial flow turbomachine
E66 Axial flow c. + A.c. + A.f. turbomachine
E67 Drum construction
E68 Disk construction

E69 Axial flow compressor blade
E70 Naca 65 (12) 10 Blower blade

E71 Jumo 004

E72 Single stage compressor
E73 Multi stage compressor

(etc. - i.e., other kinds of compressor)
[Parts]
E89 [Stage] Q.
E90 Stage characteristic
E91 S. efficiency + S. performance
E92 Cascade losses
E93 S. matching
(etc. - i.e., other Stage characteristics)
[Blade]
F19 Rotor blade
F20 Stator blade
(etc. - i.e. other kinds of Blade)
(Blade characteristics)
F35 Blade shape
F37 B. curvature
(etc.)
[Flow phenomena]
F91 Irrotational flow
[Rotational flow
F92 Inlet whirl
F93 Prewhirl
(etc.)

The first subclasses under Compressor are Kinds of compressor: Centrifugal and Axial flow (based on direction of flow): Single stage and Multi-stage (based on stage numbers) and so on. The synonyms which appear at the concept level automatically sort themselves out (e.g., the three variants at E65). Any categories (generic and non-generic) which refer to a given subclass follow that subclass immediately. So under Axial compressors is found Kinds (Drum construction, Disk construction, Jumo 004) and Parts (Blade) - and a particular kind of axial compressor blade follows that. (N.b. - a clerical error has resulted in the Kind of a.f.c. 'Jumo 004' following the Part, 'Blade' instead of preceding it; such errors did not affect the programming of searches).

The Kinds facet is followed by the Parts facet (Stage, Blade, etc.). The bracketed term Stage followed by 'Q' indicates a term which appeared in one of the questions but not in the indexing of the subset documents; it has been inserted for programming purposes. The array of 'Stage characteristic' demonstrates a recurrent problem in the subject field analysed, that of maintaining 'generic' relations in a situation where strict definition of terms would result in an uncomfortably large number of tiny subfacets

consisting often of a sole member; e.g., Stage efficiency is evidently a property and Stage performance a process - but they are treated as virtually synonymous; Cascade losses constitute a factor in efficiency or performance, but hardly a 'kind of efficiency'. Stage matching is another concept which lies on the borderline between processes and properties. It is possible to say, however, that all these rather subtly related notions are 'Stage characteristics and in this way the facet structure is maintained without undue complexity. Other examples of a certain amount of violence being done to the strict nature of generic relations may be found, as at M4/33 (see Appendix 5.3) where complexly related terms are grouped as Atmosphere properties and characteristics, or at P7/26 Processes and properties of Vortices. Similar situations inevitably arose in the single-term hierarchies in areas like Mechanics and Dynamics where the conditions of what Ranganathan has called 'canonical' classification tend to hold.

Another minor liberty, not demonstrated in the example above, was taken in the treatment of qualifying terms like Theory, Approximation, Experimental data, when these were found precoordinated as in Hypersonic flow approximation. It could be argued that these do not narrow the extension of the term they qualify and should therefore be disregarded - i.e., treated as synonymous with the term alone. Theoretically, this is why they are usually placed (in the guise of 'Form divisions') at the very beginning of the subdivisions of a term in conventional classification. In the search programmes, however, they were included in the 'Terms and species' sub-programme. This was later seen to be a mistake, but it is not thought that this was serious in view of the very few terms involved.

Although the differently related facets follow and interrupt each other without clear signs of demarcation in the schedules, the different relations were strictly observed, of course, when the search programmes were compiled; i.e., when expanding a class by generic hierarchy, only those terms standing in a true generic relation to that class were counted; e.g., Compressor + Centrifugal c. + Axial flow c. + Drum construction + Disk construction + Jumo 004 + Single stage c. + ... would be given as the full generic expansion of a particular kind of compressor (see the Generic - Broad search below). Any terms not standing in a true generic relations (e.g., Axial flow c. blade, Stage characteristics, Irrotational flow, etc.) would be ignored.

Multiple hierarchical relations

The major weakness of the linear display of classes just described, in which a particular class (concept) is located in one place only (albeit a carefully chosen one) is that it fails to show the further generic relations a class may have. For example, Jet interference is subordinated to a category Jet characteristics, in which its class mates are Jet exit, Jet location, Jet energy, Jet structure, Jet emission, etc. It could equally well be subordinated to a category of Causes of interference with class mates like Wake interference, Forebody interference, Support system interference, Wave reflection interference, Wing-Body interference, etc. But in the schedule described these last terms are 'distributed relatives'.

There are two traditional methods of handling this problem in a real-life classified index: by multiple-entry as with UDC, where the number of entries for a given compound-class-description are multiplied, and filed according to a different classification, so that Jet interference appears in a class Jet (divided into Jet characteristics) and also in a class Interference (divided into Causes of interference). Or, by leaving these other connections to be indicated by an A/Z relative index, in which all the

different contexts in which a term appears are gathered together as qualifiers of that term.

Although in a real-life situation the first method provides easier access to these further relations, this advantage was not significant in the test environment. The test collection subset was relatively small and a fully rotated A/Z index was easily producable by clerical labour. In any case, such an index was necessary for other reasons, too, as will be described. Moreover, although the working out of full alternative hierarchies would have involved a considerable effort, there was no guarantee that more than a small fraction of them would ever be used, since only those hierarchies relevant to the terms occurring in questions would be required.

The assumption above is that such an A/Z index will automatically disclose the existence of other possible hierarchies. Indeed, it is difficult to see how such additional hierarchies could be economically developed unless we are guided by the literary warrant afforded by the actual occurrence of the terms concerned in conjunction with these other contexts, in which case the A/Z index automatically picks them up. Nevertheless, further connections were indicated by references in the schedules wherever it seemed desirable, particularly where it seemed that the A/Z rotation of terms might still fail to show the connection; e.g. Small disturbance theory, subordinated to Disturbance, contains a reference to Boundary layer theory to which it is also relevant. Streamlines, subordinated to Flow elements, has a reference to Relative stream surfaces (in Compressor flow phenomena) to which it is also generic. Or, within a given class, references were added to link concepts occurring in different arrays; e.g., Performance discontinuities in the Performance facet of Compressors contains a reference to Stall and Surge in the Flow facet of the same class.

It may be noted that in a real-life classified index, the A/Z index usually shows even those connections just exemplified, since its entries contain more qualifying material (providing further information regarding the context) than the test index, where the concepts gave the sole element of precoordination. For example, a document dealing with Small disturbance theory in the context of Boundary layer theory would produce rotated A/Z index entries:

Small disturbance theory: Boundary layer theory
Boundary layer theory: Small disturbance theory

and these establish the connection which in the test collection had to be established by references.

The significance of multiple hierarchical linkage as an element in the recall performance of generic hierarchical linkage generally is probably not very great. Most questions impose a particular context of their own and the likelihood of relevant material being found in radically different contexts of the particular terms of the question is probably small. For example, a question on the kink in the surge-line of a multi-stage axial compressor imposes a context on the notion of 'surge'; clearly, documents indexed under Surge as a general concept should be examined, but it is unlikely that extended examination of the classes flanking Surge in the general hierarchy of Aerodynamic processes would be very fruitful.

The A/Z Index (see Appendix 5.5)

In order to provide for multiple generic hierarchical linkage as discussed above, and for other reasons, a rotated A/Z index of the concepts was now produced; e.g.,

	Class No.
Afterbody, Conical Base	
,Cylinder	
,Cylindrical	
Drag	D6
Surface	D5
,Truncated	
Vehicle, Conical	
(etc.)	
Base Afterbody, Conical	
Bleed	T16
,Flat	
Forward attitude	K61
(etc.)	
Cone	H55
Cone, Blunt Nose	
,Blunted	
,Circular	
Cylinder	H76
Cylinder Bodies	H76
(etc.)	
Conical Afterbody Vehicle	A57
Base Afterbody	C99
Camber	W33
(etc.)	

It can be seen that each concept appeared as many times as it had distinct words. So the first concept above appeared in three different contexts - that of Afterbody, of Base and of Conical. The class number appeared after the direct form of the concept.

The index served the obvious purpose of a key to location besides its other major purpose - that of indicating all the different contexts in which a given term had appeared in the schedules. One aspect of this second function, the capacity to reveal other generic hierarchical relationships, was discussed above. But this was only one kind of context revealed. In the example above, Afterbody surface, Base bleed, Base forward attitude, etc. reflect non-generic relations. The index therefore acted as a valuable supplement to the schedules proper in displaying these relations. The major display of these was, of course, by the subordination of a thing's categories to that thing. But these would not necessarily exhaust the non-generic relations, and the A/Z index not merely supplied further relations, but could lead the question programmer back into the systematic order to explore further categories, if necessary. For example, examining the entries adjacent to Heat transfer leads to Heat sustaining leading edge (subordinate to Leading edge), to Heat transfer at the wall (subordinated to Surfaces and Walls, where related concepts such as Constant wall temperature and Wall temperature gradient are found) and to Heated air (subordinated to Air, where related concepts such as High temperature air and Dissociated air are found). Many of these other concepts do not contain the term 'Heat' or its variants and might not have been picked up had purely alphabetical considerations governed the search.

A third major function served by the rotated A/Z index was to provide a recall device based on the 'accidental' alphabetical juxtaposition of concepts enjoying a limited

degree of precoordination. It has already been shown how the cluster of concepts around a given term (which might also be a root term for a number of word forms) such as Heat or Interference reflects a variety of relations; e.g., Interference, Blockage; Interference, Forebody; Interference, Jet; these all reflect kinds of interference according to source. Interference filters reflects Interference as an experimental agent (in temperature measurement): Interference load reflects Interference as a source of another phenomenon. When these assorted relations are added to a certain degree of word-form confounding, (e.g. expanding an initial enquiry for Dissociation by the addition of classes like Dissociated stream or Dissociating fraction) the result is an eclectic recall device which utilizes elements of hierarchy, non-generic hierarchy, confounding of word-forms, and linking (an element of precoordination is essential to the programme). Such a mixture cannot, however, rank as a 'device' in the way this notion was understood in Chapter 4. It is further considered in the next section.

Formation of Classes by Search Programmes

A significant feature of hierarchical linkage as an indexing device is the rich variety of relations it displays, enabling a number of different paths to be pursued in adjusting the size and content of the class or classes with which a search begins. Some of these paths were briefly mentioned in the last section, using the example of Visualization tests.

In exploiting these relations two different policies can be followed; either classes are expanded by bringing in all the terms related in a particular way - e.g., all the terms subordinate to the original one, as when all the different kinds of compressors are added to a search for Compressors. Or, classes are expanded eclectically, choosing just those members of a given relationship which seem most likely to be relevant in the context of the whole question. The latter policy is the one normally followed in the conventional classified index.

The former policy has the merit of simplicity in programming (once the schedules are established) and this is clearly pertinent in the case of machine searching and is, in fact, generally implied by the term 'generic search'. Equally obvious is the fact that it will tend to result in a lower precision ratio than a selective search, but possibly also a higher recall ratio.

In the testing of the concept hierarchies it was decided to attempt both approaches and the following different searches were programmed, each one producing a differently defined class.

(1) The simple natural language concept alone

(2) Confounding of synonyms. It has already been pointed out that a classification should automatically throw up synonyms as a result of its analysis; also, that a number of synonyms only become apparent at the level of concepts. Both these factors operated to produce a programme for synonyms quite different from that using single terms alone. Examples are: Temperature distribution + Temperature profiles + Temperature history; Angle of incidence + Angle of attack + Arbitrary angle of attack + Incidence; Initial expansion region + Prandtl-Meyer region.

(3/8) From this point onwards, the classes formed by (2) were regarded as the basic classes to be expanded. This expansion was achieved by adding further classes to (2) on the basis of the following programmes:

(3) Term and species: If the basic class were Non circular cylinder and its synonyms (H75), this would be expanded by the addition of Cone cylinder + C.c. bodies + Elliptic c. + Elliptic c. of eccentricity $\frac{1}{2}\sqrt{3}$. + Hemispherical c. + C. with h. nose + Ogive c. model + Flat faced c. + C. without corners.

(4) Term and species (selection): a choice was made from (3) based on the context of the question asked. For example, in a question on the kinetic theory of gases, when programming the term Gases, only those kinds of gases which reflected in some way the problem of the question were selected - such as Ideal gas, Real gas, High temperature gas, Dissociating gas, Equilibrium gas.

(5) Superordinate - i.e., adding to the basic class its immediate containing genus and as many more genera beyond that as appeared sensible; the number of steps included would rarely exceed three. To Non-circular cylinder (H75) would be added, for example, Cylinder + Body of revolution + 3-dimensional body. It should be noted that only the superordinate term was taken - not its species as well; the search is the equivalent of the traditional library search under 'more general' heads.

(6) Generic (narrow). - i.e. adding to the basic class its immediate containing class (genus) and all the other species in the same array (subfacet) as the basic class; e.g., to Non-circular cylinder would be added Cylinder and the rest of the array based on circularity of shape, but excluding those kinds of cylinder (Inconel cylinder, Flat faced cylinder, Long cylinder, etc.) reflecting other principles of division (Material, Edge properties, Length, etc.). Similarly, if the basic class were Supersonic flow, this programme would add to it all other kinds of flow designated by speed, but excluding kinds of flow based on other principles, such as viscosity, compressibility, degree of turbulence, etc.

(7) Coordinate (selection): a choice was made from (6) of the most likely terms, but excluding the superordinate term. Since by definition the classes of an array are mutually exclusive this was never a very promising search and in fact was not often productive of any terms. But in those border line situations referred to above, where the concept of generic hierarchy can only be realized practically by accepting a less-than-precise category such as 'characteristics' or 'phenomena', the likelihood was greater; e.g., in a question on Air drag the coordinate class Atmospheric rotation was accepted. Another example is that of opposites, or near-opposites, like Laminar flow and Turbulent flow, where a document frequently refers to the one even when its primary subject is the other.

(8) Generic (broad): this added to (6) as many more superordinate terms as seemed reasonable, together with all their species - i.e., not just those restricted to the immediate array (subfacet) in which the basic term appeared. For example, if the latter were Supersonic flow, this search would now bring in documents indexed by any kind of flow - Laminar and Turbulent, Conical and Parabolic, Equilibrium and Non-equilibrium, etc. This somewhat undiscriminating acceptance of the complete contents of a hierarchy is the equivalent of the 'generic search' as usually understood in machine searching.

(9) Systematic Collateral (selection): this was a selection from (8) analogous to the selection from (6) which produced coordinate classes (7) - again excluding the superordinate terms themselves. This search was more productive than (7) since there is often a close connection between concepts from different arrays of the same genus. This fact underlies the correlation of properties and the principle of definition by

aggregation of attributes, where a term is defined by a number of attributes, each of which reflects a different principle of division of the genus which lies at the heart of the definition, e.g., Poiseuille flow may be defined as Compressible, viscous, laminar flow between closely parallel planes - and each attribute reflects a different characteristic of division for the genus Flow. So where the basic class was Boundary layer flow, for example, related classes brought in by this programme would include Shear flow, Separated flow, Viscous flow, etc.

(10) A/Z collateral: the rotated A/Z index of concepts has already been described. This search was made first by examining the index to find the basic class (question concept) and any other concepts containing it (i.e., consisting of the basic class with further qualifications). Those which seemed likely to be relevant were now added to the basic class. For example, to the basic class Axial compressor was added Axial flow compressor blade since this included the basic class and seemed relevant. Or, to Heat transfer would be added such concepts as Convective heat transfer rate, Surface subjected to heat transfer, Laminar heat transfer distribution, etc. It may be noted that most of these further classes represent non-generic hierarchical relations of the basic class. Also, that most of the question concepts already consisted of two or more words and that in many cases there would not be any more concepts containing the one sought; e.g., this was the case with Multistage compressor, Non-circular cylinder, Dissociated stream.

For those concepts containing more than one word a 'second-level' search was also made, in which each significant word (and any of its adjacent variant word forms) was examined separately and further classes selected from the total body of concepts containing it. For example, to Axial compressor would be added Axial inlet impeller; to Surge line would be added Stall limit line and Surge. It should be stressed that these selections were made in the context of a given complete question and might vary somewhat for the same concept if the context differed. For example, in a question on the Surge line of an axial compressor, the 'second-level' for Axial compressor would reject Compressor surge (although it would be relevant to the question as a whole) because this approach was already covered by the programme for Surge line. Again, it may be noted that many of these further classes represented non-generic hierarchical relations; in addition, the combined first and second level searches generally included those terms selected from generic hierarchies in searches (4), (7) and (9) which also included the actual terms used in the basic concept.

(11) Residual hierarchical linkage. The A/Z collateral searches, although providing a large number of non-generic and generic hierarchical linkages, were restricted to those which included one or more of the terms actually used in the question. This still left a number of possibly helpful classes excluded. They could be divided into two groups: firstly all those from the non-generic hierarchies which appeared in the schedules where the question concept (the basic class) was located, but which failed to include the actual question term or terms (in which case the A/Z collateral would have disclosed them.) It was a simple matter to establish these, by scanning the various facets subordinated to a given concept, or adjacent to it.

Secondly, all related classes not already disclosed by the hierarchical relations of the ten searches described. A number of these were already provided for in the schedules, by references; e.g., Surface combustion (D66) see also Ablation; Vaneless diffuser (in compressors) (F84) see also Ducts; Compressor surge see also Rotating stall; Mass flow fluctuation (U44) see also Sound waves.

It has already been noted that the formidable task of adding to the 'one-place' schedules all other possibly useful hierarchies was not attempted. This was partly because much of the effort would have been wasted (if no questions were asked involving these alternative hierarchies), and partly because the A/Z index was likely to disclose the most important ones. It was also thought that the detailed analysis of reasons for failure (an integral part of the test programme) would disclose any examples of failure due to the absence of such alternative hierarchies.

It should be remembered that the hierarchies actually established were those reflecting the most likely approaches to the material and that for many of the concepts alternative approaches (manifested in different citation orders) were quite obviously unnecessary; to take some at random for example, Mixture of cold gases could not conceivably enter into a search for kinds of mixtures, or kinds of cold things; the same applies to a number of other concepts involving the term Mixture. Similarly, in the case of a number of concepts involving the word 'modes', or 'models', it was unnecessary to contemplate the possibility of having hierarchies based on these (although hierarchies of particular kinds of model, e.g., wind tunnel models, were of potential value, of course).

The references already provided in the schedules and by a file of 'notes and decisions' assembled during the indexing were now supplemented by those in various thesauri and subject heading lists in the field of aeronautics and astronautics, since these are in principle the product of similar observation of connections between classes. Examples of such references are those from Heat transfer to Transport coefficients, Large Peclet number and Prandtl number; from Dissociated stream to Ionized boundary layer; from Kinetic theory to Diffusion and to Transport properties. It has already been noted that all such connections could, if necessary, be incorporated in a hierarchy of the kind being tested, although no distinction was drawn between generic and non-generic relations when utilizing these references.

The combination of search programmes (10) and (11) represents, by and large, the performance of non-generic hierarchical relations largely, combined with a smaller element consisting of those supplementary generic relations not shown directly in the 'one-place' schedules. Although it has already been argued that both these hierarchical relations are generally quite secondary to the main display of generic relations, it must be regarded as a weakness of this joint presentation that separate programmes were not made for the two distinct situations.

3. Control by pre-established thesaurus

A major objective in producing the concept hierarchies described in the last section was to afford a degree of precoordination sufficient to remove the artificialities accompanying the use of single words only in the 'one-place' index language and to provide where suitable, that minimum of syntactical linkage necessary to the clear conveyance of unambiguous meaning in the index descriptions. It was thought that the resulting index language approximated more closely to the usual environment of index devices than did the first language.

By this time, the search methods developed in the course of testing the first languages were producing the first detailed performance figures for the various devices and languages concerned. Although the operation of the large number of variables produced an extremely complicated picture in that many ways of aggregating these variables and their different values presented themselves, the general picture seemed to suggest clearly enough that the performances were not very encouraging.

High recall was obtained only at a low level of precision, and as soon as the latter was improved, a precipitate drop in recall ensued.

A number of contributory causes of this were suspected. The ambiguities and inconsistencies of the language of aerodynamics suggested one. The match between the terms of the questions and the relevant documents which was, in some cases at least, very poor was another. The possibility of defective indexing was not thought to be very serious in the sense that exhaustive selection of keywords and phrases and the organization of these into concept and themes appeared to be reasonable. But a failure to recognize fully the connectivity between the terms of the languages so far established undoubtedly caused some of the failures.

Another possible factor was the unusual route by which the initial concept indexing had been translated into the different languages. In a real-life situation, this translation is done concurrently with the indexing itself, which is channelled into the controlled language as the first stage. The central elements in the test languages had so far been applied almost entirely retrospectively. Although there appeared to be no reason why this should have affected index performance, it seemed that validation of it as a method (by comparing it with a normally produced index) would be useful.

One way in which improvements in performance were thought to be possible was by putting more sophistication into the search programmes (by distinguishing between terms of different potency, between different combinations of these, and so on.) It was thought that maximum discrimination and control in searching implied the need for maximum discrimination and control in the indexing if optimum performances were to result. Again, although it was probable that the controls effected retrospectively were as valid as those imposed concurrently (as in indexing by a recognized, pre-established, control language) the slight element of doubt suggested that it would be wise to demonstrate this.

These considerations led to a decision to set up a conventional index with a different set of connectives based on a predetermined list of terms and to compare its operation with that of the natural language with retrospective controls already tested. For this, the Engineers' Joint Council Thesaurus of engineering terms, (Ref.28) was chosen as providing an up-to-date control language in the field of physical science and engineering, which contained clearly defined connectives grouped in a manner allowing convenient comparison with a number of the hierarchical searches described in the last section. A second subset of 350 documents was selected; this included the 200 documents from the first subset that was used in testing the concept hierarchies, thus allowing direct comparison with all previous programmes.

As in the case of the simple concept languages, no reindexing was contemplated, only another translation of the indexing done originally, since reindexing would have introduced an immeasurable variable; but the production of the new indexing language simulated the normal indexing situation. In this, each document is subjected first to 'concept-analysis' when it is decided what the document is about, what its significant terms are and how these are related in concepts and themes. This is followed by the translation of this information into a particular index language, with pre-established controls as to the level of specificity to be allowed and the recognition of synonyms and of other connectives between terms and between concepts.

Production of controlled index language using E.J.C.

The main problem raised by the use of E.J.C. was due to the fact that a

relatively general thesaurus was being applied to a special field. Although some loss of specificity (compared with the natural language) was regarded as inevitable, a considerable extension of the vocabulary was necessary if the specificity were not to suffer seriously. This extension raised problems of maintaining consistency with the principles with which the existing vocabulary and its syndetic structure of connectives had been developed. To assist this, the Rules for preparing and updating Engineering Thesauri (4th draft) November 1964 were observed as far as possible.

Selection of terms

Generally speaking, the aim was to incorporate the extra detail as unobtrusively as possible, without disturbing the distinctive character of the E.J.C. index language. The various E.J.C. methods for keeping down the size of the vocabulary were observed where feasible:

(i) Outright rejection of highly specific terms (Rule T-1) when the sense of the term could be approximated with reasonable adequacy by a broader term. E.J.C. omitted a number of prominent aeronautical and aerospace terms which did not appear to meet this criterion (e.g., Sonic boom, Tail, Stall, Bodies, Buffeting, Chord) and these were simply added. It also omitted a very large number of more precise terms and phrases occurring in the natural language indexing but which qualified for consideration under this rule. Particularly affected were those terms reflecting spatial, dimensional and temporal characteristics many of which were in adjectival form (which E.J.C. avoids); e.g., Normal, Perpendicular, Vertical, Horizontal, Behind, Outside, Below, Nearly, Large, High, Circular, Rectangular, Octagonal, Radial, Circumferential, Zero, Rate, Without, Free.

In some of these cases, where the notion was obviously dispensable because of its poorness as a retrievable handle; the term was omitted. Examples of this were Behind, Complete, Continuous, Degree, Direct, Coefficients, Effects, Horizontal, Vertical, Near, Nearly, Normal, Outer, Outside (although some of these appeared in phrases, such as Continuous loading). Outright omission was used cautiously since it diminishes the exhaustivity of the indexing. It may be noted that the main reason for holding exhaustivity constant is its effect on recall. However, the absence of a term which is completely 'non-potent' as a retrieval handle will not affect recall except in one circumstance - the use of single term searching. Theoretically, if a question includes the term Degree or Normal and this single term is searched it might retrieve a relevant document which would otherwise not be retrieved. This possibility is removed if the term is totally obliterated from the index vocabulary. However, this situation was regarded as sufficiently remote from reality to allow it to be ignored.

Strictly speaking, the only condition under which exhaustivity is affected by index language (as distinct from the personal decision of the indexer to include or not to include a notion) is when the language completely fails to provide an appropriate term even at the highest level of generality. This sometimes occurred with E.J.C. and the solution was simply to use the name of the category to which a term belonged; e.g., the term Shape was used for a whole cluster of natural language terms - Biconvex, Concave, Circular, Configuration, Diamond, Elliptical, Octagonal, Rectangular, etc. Or, the category term Position (location) was used for terms like Beneath, Outboard, Between. In this way, although specificity suffered, there was no lessening in exhaustivity.

(ii) Confounding of opposites: this was used occasionally, as in Continuum flow, Use Free molecule flow.

(iii) Avoidance of precoordination: consistency here was not assisted by the E.J.C. rules, one of which (T-1) warns against being too specific and another (T-4) warns against not being specific enough (in the matter of bound terms). One arbitrary limitation on the degree of precoordination (to 34 characters) is evidently imposed by the three-column format used in printing the Thesaurus. But only in a few cases did new combinations exceed two words (e.g., Mass transfer cooling, Blunt leading edge, Wing-Body-Tail configurations).

This policy included the representation of some concepts by an instructed coordination of single terms; e.g., Aerodynamic noise Use Aerodynamics x Noise (sound), Dynamic systems Use Dynamic characteristics x Systems, Sounding probes Use Sounding rockets x Space probes, Radiating body Use Radiation x Aerodynamic configuration, Reflected wave peak overpressure Use Shock wave x Reflection x Pressure. This device is not very clearly described in E.J.C. (using the ⌗ and & devices) and some of the examples of precoordination make the policy no clearer; e.g., under the term Pressure is given a large number of precoordinated phrases (Pressure distribution, Pressure measurement, Pressure gradient, etc.). When a 'new' term Pressure plotting occurred, it was not clear whether to precoordinate or keep separate or confound as a near-synonym of Pressure measurement. Again, a 'new' term Circular wind tunnel might lead to acceptance by analogy with Circular saws, etc. But should Rectangular wind tunnel and Octagonal wind tunnel be similarly distinguished? Sometimes, this sort of economy in precoordination avoiding highly specific new terms, led to strange equivalents such as Root section. Use Foundations x Profile.

The record of these rejected terms and phrases, together with the ones to be used in their place, grew to large dimensions and constructed a massive 'lead-in' vocabulary from the terms and expressions of the natural language to those of the controlled E.J.C. languages. Over 1,500 entries were made for the subset, which totalled 350 documents. It should be noted, however, that a number of these rejects were simply word-form variants, e.g., Oscillatory Use Oscillations; Oscillating Use Oscillations; Oscillatory motion Use Oscillations; Elastic Use Elasticity; Edged Use Edges.

Selection of references

(1) UF (Use for) These have already been considered above as forming a lead-in vocabulary.

(2) BT and NT (Broader terms and Narrower terms) The definition of these two reciprocal relations is reasonably clear in E.J.C.

The BT reflects a true generic (Thing/Kind) relation excluding not only the obviously non-generic ones, like operations, or Properties, but also, explicitly, the Whole/Part relation, which is often loosely associated with the generic. The BT also excludes "generic families constructed on the basis of usage", so Platinum, a member of the class Metal, is not regarded as a member of the class Catalysts since it is only sometimes used as a catalyst. This seems to suggest an even stricter interpretation by the E.J.C. of the notion of 'class' - i.e., one which excludes from membership all but 'true' species in the sense that they possess permanent and fundamental characteristics, uniquely defining them.

However, this is not borne out by an examination of the Thesaurus, which suffers some inconsistency on this point. For example, Wind tunnel nozzles gives, quite correctly, Nozzles as a BT. But the previous term Wind tunnel models fails to give Models as a BT and under Models gives Wind tunnel models as a Related term (RT). i.e., a non-generic relation. Again, the term Materials has eighteen RTs; some of them are kinds of material based on Structure (Composite, Granular, etc.) some based on Properties or Behaviour (Radioactive, Magnetic, etc.) some based on Use (Structural, Molding, etc.). It would appear that the first characteristic, at least, designates true species. Again, under Plastics are listed numerous resins as NTs. The term Resins itself is treated as a synonym of Polymers, however; but no reference of any kind connects Plastic to Polymers or vice-versa, although there is limited duplication between the NTs for each of these terms. Similarly, there is no connection established between Pumps and Compressors (the latter being treated as a synonym for Air compressor) - although Pumps has numerous NTs in the form of pumps of particular application. e.g. Fuel pump.

(3) RT (Related terms) These are designed to show non-generic relations (as defined above) and the rules state that it is undesirable to make RTs to 'more specific' terms. However, there are numerous examples in E.J.C. of RTs which do not observe this. For example, Hydraulic equipment has numerous references to particular types of Hydraulic equipment (Hydraulic brakes, Hydraulic presses, etc.) Apart from the fact that there were cases of true species (e.g. of Hydraulic equipment) being included in the RT framework, problems arose regarding the references from terms like Quartz (as a heat shield); if we assumed that Heat shield is in the relation of RT to Quartz, should we, under Heat shield, have added Quartz as an RT? By analogy with Insulation (say), which gives as RT the material Magnesium Oxide, Quartz should have been added. But if it represents (as it does) Quartz as designating a kind of heat shield according to material, it is a 'more specific' term and such references are not encouraged.

Generally speaking, E.J.C. observes the old rule of Subject Heading lists which avoids references to adjacent headings on the score that their juxtaposition makes further reminders to the searchers (the question programmers) unnecessary. For example, there is no reference between Shock waves and Shock tubes. So this should lead to the avoidance of a reference from Ionization to Ionosphere (which was, nevertheless, made). Also under Molding materials there are references to four RTs which are adjacent entries beginning with 'Molding ...' .

Apart from these efforts to observe consistency and method in making references, the usual variety of relations appeared to be permissible and consequently RTs were added for new terms and for existing E.J.C. terms when these had inadequate connectives. Examples of the latter situation were fairly common since Aerodynamics is not a particularly favoured subject in E.J.C. For example, there are no connections between Vibrations and Elasticity, or between Supersonic flow and Shock waves; in the case of Poiseuille flow, (which may be defined as viscous, laminar flow in pipes or between closely parallel planes), the term is rejected and referred to Laminar flow but without any references linking it to the notion of pipe flow, which in this context is just as important as laminar flow.

References from post coordinated terms.

E.J.C. provides a number of instances in which a term is distributed between two or more wider terms; e.g., Pressure gas welding had BTs which include

'Gas welding &' and 'Pressure welding &' (the ampersands mean that the two broader terms Gas welding and Pressure welding can be jointly substituted for the single and more precise term Pressure gas welding). This procedure led to a number of similar new references; e.g., Prandtl-Meyer flow Use Supersonic flow x Expansion.

However, this device leads to a difficulty, inherent in post coordinate indexing, when these particular coordinations generate new reference structures of their own which are not apparent at the level of single isolated terms. E.J.C. offers no guidance on this point. For example, assume that for Conical flow the instruction is to use Cones x Flow. In this case, some sort of reference seems desirable, either from the rejected phrase Conical flow or from its constituent terms in the form:

Cones: when coordinated with Flow, RT Shock waves.

In this particular case an intermediate connective was established by using a heading Mach cones. Examples of where the reference structure could be fairly elaborate are Flow x Parameters and Vapour x Screens x Procedures. Each of these subjects has its own set of related terms, generated entirely by the conjunction of their constituent terms, e.g. Vapour screen method, which now has its own related terms, such as Carbon tetrachloride vapour, Humidity control, Temperature control, Operational fog density.

The implied need for such connective references if the syndetic structure were to be developed raised problems of complexity in the scan-column search techniques (which were designed to be purely clerical in operation). In view of this, together with the fact that E.J.C. quite ignored such post coordinate reference needs, it was decided to follow the E.J.C. policy and rely on future analysis of searches to show where this weakness contributed to failures in performance.

E.J.C. roles

The E.J.C. system of roles which appear, without explanation, in Table 3 of Ref.28 was used in the indexing. The reasons why roles were not tested in earlier languages have already been discussed, but the availability of a ready-made set of roles seemed a useful opportunity to investigate whether some of the assumptions made there (those relating to the applicability of roles to aerodynamic literature) were justified. Unfortunately, at this stage of the project, the time factor was beginning to limit the amount of new testing which could be undertaken and it was decided that this validation was not possible.

However the decision was not made until the tentative examination of the feasibility of using the roles had been undertaken by adding them to the indexing descriptions of a small sample of reports. These, in fact provided examples of most of the objections and difficulties we had already met in the earlier attempts to use roles. An example of the difficulties inherent in using the roles may be seen if we consider a particular document and some of the questions to which it had been judged relevant. In a document (1014) on the application of piston theory and the study of aeroelastic problems, one of the themes indexed was calculation of panel flutter in supersonic flow by piston theory. Some of the problems immediately raised by the addition of roles, taking particular terms, were as follows:-

Panel is clearly a patient (role 9), but could conceivably be regarded also as a support or host in a process (role 5). Vibration (the E.J.C. term for flutter) is undoubtedly an undesirable component (role 3) in the aerodynamic behaviour of the panel.

But this is a permanent feature of Flutter in the collection context and does not alter from one document description to another. But roles are expressly designed to clarify the local and varying relations between the terms of a particular description and perhaps this rules out role 5 from consideration. Flutter could also be regarded as the primary topic (role 8). However, firstly, role 8 is not strictly speaking a role at all; it does not show semantic relations between the different terms of an index description, only its subjective state as to a hypothetical reader; it is, in fact, a weighting device. Secondly, it could be argued that piston theory is a primary topic (role 8). Perhaps both terms could be labelled 8, but piston theory could also be regarded as an agent (role 10), although this would overlook the fact that it is not an agent of Flutter but of its analysis. Supersonic flow could be considered as an environment (role 5) or as a cause or influencing factor (role 6). If it is treated as (6) however, Flutter, which it affects, would be the factor influenced (role 7).

When the questions to which document 1014 was one of the relevant documents are considered, the further difficulties in ensuring a match become apparent. Question 97 refers to the Prediction of flutter on lifting surfaces. Flutter could conceivably be given role 3 (undesirable component), or 8 (primary topic). It could also be considered as role 9 in the sense that it is the object of analysis or predictive operations. Two other questions, numbers 98 and 276, to which document 1014 is also relevant, are on how flow characteristics or leading edge bluntness affect Flutter. So now Flutter is an influenced factor (role 7). Yet another question, number 3, is about aeroforces acting on high speed aircraft, and if Flutter is regarded as a kind of aeroforce, it may also be given role 6 (cause or influencing factor).

The general import of these considerations seems to be that the term can be forced to play simultaneously a number of different roles in the same document according to what the particular user is seeking, and that attempting to label them too precisely is liable to result sometimes in unjustifiable rejection of indexing descriptions because they do not match exactly in the roles assigned them.

CHAPTER 6

Testing Techniques

The choice of the physical method to be used for searching was important, but difficult to make. Since the work was entirely concerned with index languages, it was essential that the physical form of the index should in no way impede the investigation by introducing any controls or restrictions of its own. Although it was not possible to forecast exactly the many different tests that would be made, it was clear that for each question there would be the necessity of obtaining several hundred sets of performance figures.

It was decided that a small test should be made soon after the project had commenced; this was to be done partly to check the indexing procedures but also to validate the proposed design of the tests and to provide experience that would assist in deciding on the physical form of the index. For this pilot test, 116 documents had been indexed, and fourteen questions were available for searching, for which there were 26 known relevant documents. It was planned to investigate five sets of recall devices and four sets of precision devices, based on the single-term, natural language indexing. These variables alone appeared likely to result in some 80 searches for every question, and when other variables were added in the main test, the potential number of different searches could run into several hundreds. It was unlikely that every combination of the various devices would be required, but the method used had to be flexible enough to provide for all possible variations of searches, since it would only be after some searches had been made that it would be known which were unnecessary.

Co-ordination was certainly the basic precision device, and some form of post-coordinate index was clearly required. For the pilot test, the decision was taken to prepare a peek-a-boo type index. This was done in a conventional way but a complication arose due to the fact that, at this stage of the work, six different indexing weights were being used, and, to investigate the effect of these, it was necessary to have, for every term, six cards each of which represented a different weighting.

The first search for a given question was carried out on the natural language terms. Subsequent searches were made bringing in the various recall devices and precision devices; the nature of these searches is considered in more detail later in this chapter. The results of this test were interesting in themselves, but the main objective had been to obtain information on the techniques being used. In this respect, the test showed that the general test theory was reasonable and that the indexing was satisfactory for the objectives of the test. Quite definitely, however, it showed that a peek-a-boo index would be quite unsatisfactory for the main test.

This was because much of the testing involved use of increasingly large numbers of terms in the search as the recall devices were tested, with the continual need for co-ordination of all the different combinations. For example, if a question had five terms searched on initially, and each of the five terms had one synonym, two word forms and four quasi-synonyms, then in coordination of all five terms using all the recall devices, 32,768 different combinations are possible. After this, it would be necessary to search for any four of the five sets of terms, then any three and so on. It is true that by use of the lowest posted terms first, the number of coordinations to be done can be reduced considerably, but the use of natural language

for the file, together with weights, resulted in serious difficulties. Another problem that loomed large was that of recording the aggregate of the different documents retrieved out of all the possible coordinations at a given ordination level, since many documents would be retrieved several times. One possible solution to these problems was to prepare a new peek-a-boo index for each of the recall devices; that is to say, there would be one index for natural language terms, a second index with the synonyms controlled, a third index with word endings confounded, etc. However the manual re-punching of new indexes would have been a big task, and at that time no equipment could be found to aggregate a set of postings from a number of different cards all on to one card. Other considerations mitigating against a peek-a-boo index were the task of withdrawing and refiling large numbers of cards during a search, and the difficulty of performing more than one search at one time.

As a result other conventional index forms were considered but offered no satisfactory solution. At this point in the project, several people working on associative retrieval expressed interest in the possibility of using the indexing being performed on our collection for their own testing of statistical associative techniques, clumping, etc. With the agreement of the National Science Foundation, arrangements were made to make the indexing available in machine readable form, on magnetic tape. The format used for this is given appendix 6.1, and details of supplementary tests being made are given in Chapter 7. With the indexing available on magnetic tape, the use of this for computer searching for the testing was then considered.

A number of discussions were held with various groups, and we received cost estimates for programming and searching which varied by a factor of ten An effort was made to discover whether any suitable computer programme already existed, which could be used to do the required searches. Discussions were held during a visit of one of the project staff to the U.S.A., but no suitable programme was discovered to do the minimum of what was required. This led to a reconsideration of preparing programmes in this country, but not only were the cost estimates high in relation to the present project, but also the time factor was becoming critical. Particularly discouraging was to learn that the searches which we had requested would result in seven million lines of print out; for these reasons and our own lack of experience in the field, the idea of using computers was abandoned.

The flirtation with computers had not been entirely wasted, for by this time we had a clear idea of exactly what was needed, and this helped in producing a method which met the main requirements. At the time when the solution was first proposed, no similar method was known to exist, for it is quite unconventional and it is difficult to visualise any application in real life circumstances. However, it was later discovered that a somewhat similar suggestion had been made by Dr. John O'Connor, known as the 'Scan-column index' (ref. 31), although no actual example of its use in practice is known. It had the advantages of flexibility to meet changing circumstances, so that it would give results for the many different types of search, and also of permitting quite complex analyses to be done clerically.

The first stage in the preparation of the index was a complete posting of each single term used in indexing on to a set of cards. These cards also contained information regarding the weights assigned to each term. The indexing decisions regarding Document 2076 are shown on the master indexing sheet in Fig. 6.1. From this sheet, the single terms and their weights were posted on to cards, with a separate card for each term. Thus 'Insulated 10', together with the document code number (2076) would be posted on one card, 'Two-dimensional 10' on another card together with a code number and so on to every index term. These cards were then sorted into alphabetical order and sub-sorted into document number within each term.

PARTITIONS & THEMES		INTERFIX & CONCEPTS		INTERFIX & CONCEPTS		TERMS & WEIGHTS		TERMS & WEIGHTS	
A	a d e f g	a	Insulated two-dimensional surface	i	Laminar boundary layer displacement thickness	Insulated	10	Distribution	8
B	b d e f g	b	Insulated flat plate	j	Laminar boundary layer momentum thickness	Two-dimensional	10	Displacement	8
C	c d e f g	c	Insulated curved two-dimensional surface circular arc profile	k	Skin friction	Surface	10	Thickness	8
D	a d e f h effect of i g	d	Hypersonic flow	l	Wind tunnel tests at Mach 6.86	Flat	8	Momentum	8
E	a d e f h effect of j g	e	Compressible boundary layer			Plate	8	Skin	8
F	a d e f h effect of i g compared with l	f	Linear velocity profile			Curved	8	Friction	8
G	a d e f h effect of j g compared with l	g	Approximate calculation			Circular	8	Wind	7
H	a d e f k g compared with l	h	Pressure distribution			Arc	8	Tunnel	7
						Profile	8	Test	7
						Hypersonic	9	Mach	7
						Flow	9	6.86	7
						Compressible	9		
						Laminar	9		
						Boundary	9		
						Layer	9		
						Linear	9		
						Velocity	9		
						Profile	9		
						Approximate	9		
						Calculation	9		
						Pressure	8		

FIGURE 6.1 MASTER INDEXING SHEET FOR DOCUMENT 2076

The 361 questions which it was proposed to use for searching produced a total of 723 different terms, and these became known as 'starting terms'. As such they were terms used in the questions without being subjected to any controls, and were equivalent to the natural language index terms. For each starting term a set of sheets was provided, these sheets bearing the document numbers 1001-2400. As an example, consider the starting term 'Flow'. The pack of cards which had been posted with this term was taken, and the information transferred from the cards to the set of sheets. The code 1 was used to denote that it was the actual search term (i.e. Flow) that was being posted and Figure 6.2, which is an extract from the set of sheets dealing with 'Flow', shows that a large number of documents were indexed by this term. In particular it can be seen that document 1933 was indexed by Flow at a weight of 9, as were documents 1939, 1940 and 1941. Document 1942 was also indexed by Flow, but on this occasion the weighting is 8. After all the indexing by Flow had been entered, additional entries were made for terms related to Flow. The authority sheet for this is shown in Fig. 6.3, from which it can be seen that Flux and Stream are considered as synonyms. The packs of cards posted for these terms would be taken, and entered on the sheets for Flow. Referring to Fig. 6.2, it will be seen that, for example, document 1978 is marked A6. This indicates that Flux, (which is coded A in Fig. 6.3) was indexed in this document at a weight of 6, while document 1974 is one of several that was coded by Stream(B) The variant word ending, Flowing, (coded E) was used in document 1968; of the quasi-synonyms shown in Fig. 6.3, Motion (K) and Moving (M) are examples which both appear in document 1978. It will be noted that multiple posting can occur on one document number; 1978 has, in addition to Motion and Moving, also been posted with Flow and Flux. The reason for doing this will be explained later.

The completion of this meant that there now existed a record of every time the starting term Flow or any of its synonyms, word endings and quasi-synonyms had been used as index terms. Since the codes for these were always kept constant (A-D for synonyms, E-J for word endings and K-Z for quasi-synonyms), the staff always know to which group any particular entry belonged.

The posting had been done on foolscap sheets and these were now cut into narrow strips, $\frac{3}{4}$ in. wide, each strip being serially numbered so as to maintain the document sequence order. These sets of strips were then filed in two specially constructed 'beehive' cabinets (Fig. 6.4).

In effect, a separate index was now compiled for each question by the preparation of a set of search sheets. The production of these in relation to a particular question was controlled by the question starting term card, an example for question 181 being shown in Fig. 6.5. This listed the starting terms for the question and the order of the terms on the search sheets, this order being of importance in relation to some of the searching options. To prepare the search sheets, the sets of strips for each of the starting terms were obtained and assembled one page at a time by being clipped to a set of 23 prepared boards. These boards showed the document numbers at the extreme sides, and the strips were arranged in correct alignment with the numbers. When all 23 boards had been thus prepared, a xerox copy was made of each board; the result is shown in Fig. 6.6, which illustrates one of the 23 sheets for question 181 in relation to documents 1931-1992.

1745	1807 K6	1869 1.7	1931	1993 1.9 B9	2055
1746 B6	1808 1.9	1870	1932	1994 1.9 B9	2056
1747	1809 1.9	1871	1933 1.9	1995 1.9 B10	2057
1748 1.9	1810	1872 1.9	1934	1996 1.9	2058
1749 1.9	1811	1873 1.9 M9	1935	1997 B10	2059
1750	1812	1874	1936	1998 1.9 B5	2060
1751 1.9	1813 K10	1875	1937	1999 1.9	2061 B9
1752 1.8	1814 1.6	1876	1938	2000 1.9 K8	2062
1753 1.9	1815 1.9	1877	1939 1.9	2001	2063
1754 1.8	1816 1.9	1878	1940 1.9	2002 1.9	2064 1.8
1755	1817 1.9	1879	1941 1.9 K9	2003 1.9	2065
1756	1818 1.9	1880 1.7	1942 1.8	2004	2066
1757 1.9 L8	1819	1881	1943 1.9	2005	2067
1758	1820 1.9	1882	1944	2006 1.10	2068
1759	1821	1883	1945 1.9 K9	2007 1.9	2069
1760	1822	1884	1946 1.9 B9	2008	2070
1761	1823	1885	1947	2009 1.9	2071
1762	1824	1886	1948 1.9	2010 1.9	2072 B5 M8
1763	1825 1.9	1887	1949 1.10	2011	2073
1764 K9	1826	1888	1950	2012	2074 1.9
1765	1827	1889	1951	2013	2075 1.9
1766	1828	1890	1952	2014 1.7	2076 1.9
1767	1829	1891	1953	2015	2077
1768	1830	1892 1.7 K9	1954	2016	2078 1.10
1769	1831	1893 1.5	1955	2017	2079
1770 1.9	1832	1894 B9	1956	2018	2080 1.9 B8
1771 1.9	1833	1895 1.9	1957	2019	2081 1.9 K7
1772 1.9	1834	1896	1958 1.7	2020	2082 1.9 B8
1773 1.9	1835	1897	1959 1.9	2021	2083 1.9
1774 1.9	1836	1898	1960 1.9	2022	2084 1.9 B6
1775 1.9 K8	1837	1899 1.7 M9	1961	2023	2085
1776 1.7	1838	1900 1.9	1962 1.7 B9	2024	2086
1777	1839	1901 1.9	1963 1.8	2025	2087
1778	1840	1902 1.9 K9	1964 1.8	2026	2088
1779 1.7	1841	1903 1.9	1965 1.9	2027	2089
1780	1842	1904	1966 1.9	2028	2090
1781 1.7	1843	1905	1967 1.9	2029	2091
1782	1844	1906 1.7	1968	2030	2092
1783	1845	1907 1.7	1969 1.9 B9	2031	2093 1.9 B9
1784 1.9	1846	1908 1.6	1970 1.9 B9	2032	2094
1785 1.9	1847	1909	1971 1.9	2033	2095
1786	1848	1910	1972 1.9 B9	2034	2096 A8
1787 K8 M9	1849	1911	1973 B9 1.9	2035	2097
1788 K9 M8	1850	1912 1.9	1974 B9	2036	2098 A6
1789 B9	1851	1913	1975 1.9	2037	2099 1.6 B8
1790 1.9	1852	1914 1.7	1976	2038	2100 1.6 B8
1791 1.9 K5	1853	1915 1.9	1977 1.9 K9	2039	2101
1792	1854	1916 1.9	1978 1.7 A6 K7 M9	2040 1.9	2102
1793 1.8	1855 1.9	1917 1.9	1979 1.9	2041	2103
1794 1.9 L7	1856 1.9	1918 1.9	1980 1.7	2042	2104 1.6
1795 1.9	1857	1919 1.9	1981 1.9	2043	2105 1.9 B8
1796 1.6	1858 1.9	1920 1.9	1982	2044	2106 1.9
1797	1859	1921 1.7	1983	2045	2107 1.9
1798 1.9	1860	1922 1.9	1984 1.9	2046	2108 1.9
1799 1.7 K7	1861 1.9	1923 1.9	1985 1.9 B9	2047 K8	2109 B7
1800 1.8	1862	1924 1.9	1986 1.9	2048	2110 1.9 K10
1801 L8	1863	1925 1.9 B5	1987 1.9 B9	2049	2111
1802 1.9	1864	1926	1988 1.9 B8	2050	2112 1.9 K6
1803 1.9	1865	1927 1.9	1989 1.9	2051	2113
1804 1.9	1866	1928	1990 1.9	2052	2114
1805	1867	1929	1991 1.9 B5	2053	2115
1806	1868	1930	1992 1.9	2054	2116
(13)	(14)	(15)	(16)	(17)	(18)

FIGURE 6.2 POSTING SHEET FOR 'FLOW' IN RELATION TO DOCUMENTS 1745-2116

Starting term	1	FLOW
Synonyms	A	FLUX
		STREAM
Word endings	E	FLOWING
Quasi-synonyms		MOTION
	L	MOVEMENT
		MOVING
	N	FLOWING

FIGURE 6.3 STARTING TERM AUTHORITY SHEET SHOWING TERMS RELATED TO 'FLOW'

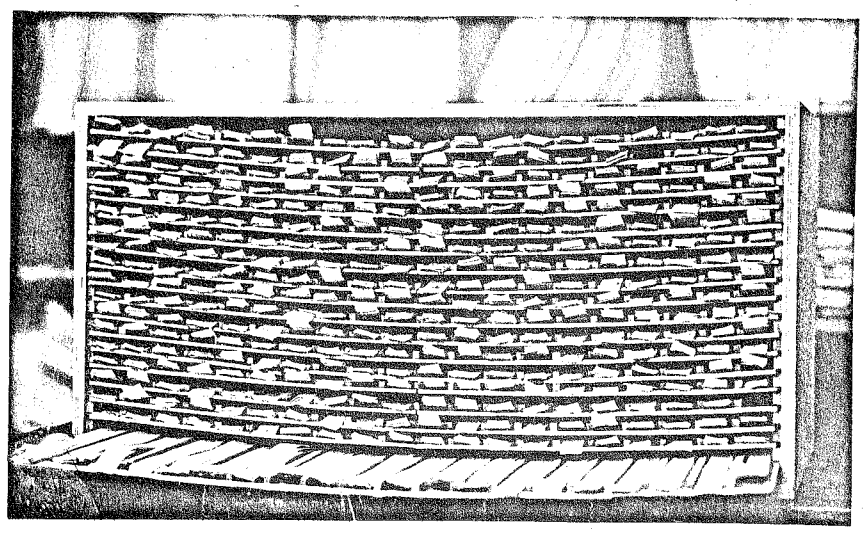


FIGURE 6.4 THE 'BEEHIVE' FILING CABINET

a CHANNEL
b VISCOUS
c COMPRESSIBLE
d FLOW
e STRAIGHT
f DETERMINATION
g NATURE

FIGURE 6.5 SEARCH STARTING TERMS FOR QUESTION 181

	Nature	Flow	Compressible	Viscous	Channel	Straight	Determination
1931							
1932			H·10				L·7
1933		1-9					
1934			H·10				1·7
1935			H·6				
1936							
1937							
1938			H·10				
1939		1·9	1·9		1·10		
1940		1·9					L·9
1941		1·9, K·9	1·10 N·5	1·10 E·8	1·10		
1942		1·8	M·8				
1943		1·9	1·9				
1944							
1945	K·5	1·9, K·9					
1946		1·9, B·9	N·9				
1947	L·10		N·10				L·6
1948		1·9	N·9				
1949	K·10	1·10					
1950						1·9	L·9
1951							L·8
1952							L·9
1953							
1954							
1955							
1956			H·10				
1957			J·6				
1958		1·7	H·8				
1959		1·9					
1960		1·9					
1961			1·10				
1962		1·7, B·9	M·7				
1963		1·8			1·8		
1964		1·8					1·9
1965		1·9					
1966	K·7	1·9	F·10		1·10		
1967		1·9	1·9	1·10			
1968	L·7	E·9					
1969		1·9, B·9	M·9				
1970		1·9, B·9	K·9 M·9				
1971		1·9	M·9 N·9				
1972		1·9, B·9	N·9				
1973		1·9, B·9	M·9				
1974		B·9	M·9				
1975	L·6	1·9					1·6
1976			N·9				
1977		1·9, K·9				1·9	
1978		1·7 A·6 K·7 M·9	N·6				
1979		1·9	M·9				
1980		1·7					
1981	K·9	1·9					
1982	K·7		N·10				
1983			N·10				
1984		1·9	1·10 J·10				
1985	L·9	1·9, B·9	1·10, J·10				
1986		1·9	J·10				
1987		1·9, B·9	B J·10 K·9 M·9				
1988		1·9, B·8					
1989		1·9					
1990		1·9	1·9, J·10				
1991		1·9 B·5	L·10, M·5				
1992		1·9	M·9	1·8			
(16)							

FIGURE 6.6 SEARCH SHEET FOR QUESTION 181 IN RELATION TO DOCUMENTS 1931-1992

It can be seen that for the search term Flow, the appropriate information which was first posted on the sheet shown in Fig. 6.1 for documents 1931-1992 has now been included in the second column of Fig. 6.6. The information relating to the other starting terms would have come from similar strips. As an example, the search sheet reveals that in document 1966 Nature did not appear, but the quasi-synonym Property (coded K) was indexed at a weight of 7. Flow was indexed at a weight of 9. Compressible did not appear, but it was present in the variant word form Compressibility (F) with a weight of 10, while Channel was indexed at a weight of 10. The remaining three starting terms did not appear in any way in this document.

When the search sheets had been printed, the 'boards' were dismantled, the strips sorted into order and redistributed into the beehive ready for further use with another search question. The boards finally used were of rigid hardboard, together with 'bulldog' type clips; earlier trials with cardboard sheets and perspex covers had failed because the strips moved out of position too easily. The time taken to mount a question on to the boards varied with the number of starting terms, but usually took between thirty and sixty minutes. The xeroxing and checking took ten to fifteen minutes, and redistribution of the strips a further ten to fifteen minutes. A minority of questions had more than eleven starting terms, and therefore needed two sets of sheets. It was usually possible to pick two questions with quite different sets of starting terms, so that both questions could be prepared at the same time. A system of double checking the search sheets was used to correct any errors which occurred; these were usually due to misfiling of individual strips in the re-distribution stage. While this method might seem cumbersome, it appears to have been justified by results, since it gave the flexibility that was required, and although expensive in man-hours was relatively cheap compared to what would have been the cost for any form of machine searches.

The end result of this exercise was that we had 361 sets of search sheets, 23 sheets in each set, posted with all the occurrences of the terms to be used in searching each question; there were, in fact, 361 question-indexes, and it was now possible to carry out the first series of searches. These were performed on single terms, and investigated three variables.

1. The recall devices of synonyms, word endings and quasi-synonyms, tested in six aggregations (known as 'index languages').

2. The precision device of simple coordination without any linking in the indexing, where the search rules allowed any combination of terms to be accepted, and every level of matching to be recorded.

3. The three levels of indexing exhaustivity, indicated by the weights (5-6, 7-8 and 9-10).

The six index languages investigated in the first series of tests were as follows:

Index Language	
1	Natural language terms (code 1)
2	Natural language terms + synonyms (codes 1 and A-D)
3	Natural language terms + word forms (codes 1 and E-J)
4	Natural language terms + synonyms + word forms (codes 1, A-D and E-J)
5	Natural language terms + synonyms + quasi-synonyms (codes 1, A-D and K-Z)
6	Natural language terms + synonyms + word forms + quasi-synonyms (codes 1, A-D, E-J and K-Z)

These six index languages appeared to cover all reasonable permutations, since it was not logical, for instance, to contemplate the use of quasi-synonyms without the use of synonyms.

The searches were carried out by clerical labour, and the results were recorded on a score sheet as shown in Fig. 6.7. The actual operation of carrying out a search became known as 'putting the ruler down the sheets', since the use of a straight edge to successfully uncover the postings for each document was found to be the best method. The searches were made on the sets of search sheets (as in Fig. 6.6), where each vertical column deals with one of the question starting terms, and shows not only the occurrence of the starting term itself, but also the related terms as described earlier. Often an examination of the postings for a certain question needed some care in working out, since in one operation the search results would be recorded for the six different index languages and for the three weights. However, after a relatively short learning period, the clerical staff had no serious difficulties. The time required to search a single question varied greatly, with this particular set of six index languages, it might be anything from ten minutes to one hour, being dependent on the number of starting terms, the frequencies of postings for each starting term, and the number of terms related to the starting terms.

The score sheets list the document numbers on the left hand side, and across the sheet space is given for recording the coordination level (i.e. the number of search terms that match with the document terms) of each document for each of the six index languages at each of the three levels of exhaustivity. The way this is done may be seen by examining a search sheet (Fig. 6.6) for question 181 'Has any work been done on determination of the nature of compressible viscous flow in a straight channel', in relation particularly to documents 1963, 1966 and 1978.

The search sheet shows that document 1963 has two of the search terms present, and a look at the codes shows that they are coded 1, the natural language terms, which are included in all six languages. Both terms have a weight of 8, and therefore do not come out at the lowest exhaustivity (weights 9 or 10), but do at the medium and high levels. The score sheet (Fig. 6.7) records this, the coordination score of 2 being put in every language at the medium and high levels of exhaustivity. Document 1966 has four of the search terms present; two natural language, (1) one word ending (F) and one quasi-synonym (K). So taking the highest level of exhaustivity (5-10), every index language will have a coordination score of at least 2; Index languages 3 and 4 will score 3, (1, 1 and F); Index language 5 will also score 3, (1, 1 and K), but Index language 6 scores the maximum, 4, (1, 1, F and K) since it accepts both word ending variants and quasi-synonyms. Considering now the various levels of exhaustivity, index languages 1 to 4 have all their terms weighted 9 or 10, and so keep the same coordination score at medium and low exhaustivity, but index languages 5 and 6 have the quasi-synonym weighted 7, so at low exhaustivity the coordination score drops to 2 and 3 respectively.

As a final example, for document 1978, one of the two search terms (Flow) is shown to be present in natural language at a weight of 7, as a synonym (A-6) and also as two quasi-synonyms (K-7 and M-9). All these, of course, only count as a coordinate score of one since they are all separate alternatives to one of the search terms, but the last quasi-synonym (M-9) is important because it is the only term at low exhaustivity. The coordination scores for this document in table 6.3 are 1 for index languages 1 to 4, and 2 for languages 5 and 6, with exhaustivity reducing these scores as shown.

Since the search rules at this stage allowed any combination of terms to be accepted, it was never necessary to note which search terms occurred. Some combinations accepted were obviously nonsense, e.g. document 1982 retrieved by the starting terms Nature and Compressible is not meaningful, and is even worse when the quasi-synonyms

Question 181

Index / Languages / Documents	1 1			2 1, A-D			3 1, E-J			4 1, A-D, E-J			5 1, A-D, K-Z			6 1, A-Z		
	5-10	7-10	9-10	5-10	7-10	9-10	5-10	7-10	9-10	5-10	7-10	9-10	5-10	7-10	9-10	5-10	7-10	9-10
1956	-	-	-	-	-	-	1	1	1	1	1	1	-	-	-	1	1	1
1957	-	-	-	-	-	-	1	-	-	1	-	-	-	-	-	1	-	-
1958	-	-	-	-	-	-	2	2	-	2	2	-	-	-	-	2	2	-
1959	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1960	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1961	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1962	1	1	-	1	1	1	1	1	-	1	1	1	2	2	1	2	2	1
1963	2	2	-	2	2	-	2	2	-	2	2	-	2	2	-	2	2	-
1964	1	1	-	1	1	-	1	1	-	1	1	-	1	1	-	1	1	-
1965	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
1966	2	2	2	2	2	2	3	3	3	3	3	3	3	3	2	4	4	3
1967	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
1968	-	-	-	-	-	-	-	-	-	-	-	-	1	1	-	1	1	-
1969	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2
1970	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2
1971	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2
1972	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2
1973	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2
1974	-	-	-	1	1	1	-	-	-	1	1	1	2	2	2	2	2	2
1975	2	1	1	2	1	1	2	1	1	2	1	1	3	1	1	3	1	1
1976	-	-	-	-	-	-	-	-	-	-	-	-	1	1	1	1	1	1
1977	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
1978	1	1	-	1	1	-	1	1	-	1	1	-	2	1	1	2	1	1
1979	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2
1980	1	1	-	1	1	-	1	1	-	1	1	-	1	1	-	1	1	-
1981	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2
1982	-	-	-	-	-	-	-	-	-	-	-	-	2	2	1	2	2	1
1983	-	-	-	-	-	-	-	-	-	-	-	-	1	1	1	1	1	1
1984	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
1985	2	2	2	2	2	2	2	2	2	2	2	2	3	3	3	3	3	3
1986	1	1	1	1	1	1	2	2	2	2	2	2	1	1	1	2	2	2
1987	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2
1988	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1989	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1990	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
1991	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2
1992	2	2	2	2	2	2	2	2	1	2	2	1	1	3	2	3	3	2

FIGURE 6.7 SCORE SHEET FOR QUESTION 181 IN RELATION TO DOCUMENTS 1956-1992

on which it was retrieved are decoded as Property and Hypersonic. Intellect was put in on later searches, to eliminate such unwanted combinations.

Contrary to the example shown in Fig. 6.7, in practice the score sheets for a question rarely recorded documents with only one search term present, since this would usually have involved recording the large majority of the documents in the collection. The decision as to what coordination score to begin recording documents varied for each question, depending partly on the number of starting terms in the question. The objective was to examine an average of about 100 documents from the collection (involving two or three score sheets), and this decision was fairly easily made by looking at the density of postings on the search sheets. In some cases, when postings were very heavy, a proportion of the collection only was examined (e.g. if half the collection, the odd or even numbered documents only, etc.), and the results scaled up. This was done to reduce the large clerical effort involved in searching so many questions this way (involving looking at nearly 400,000 'documents' on the search sheets in this first series of tests alone), but was only done when the results were statistically valid. An exception to this was that the relevant documents were always fully recorded.

To obtain the final results for a question, the documents which had been assessed as relevant were recorded on a separate score sheet, and deleted from those first produced. The base document for the question being tested was deleted altogether at this stage. Then the actual numbers of relevant and non-relevant documents were totalled up, a separate total being obtained for each index language, at all coordination levels and at each exhaustivity level. The final record is seen on a Results Sheet, (Fig. 6.8). Here, for question 181, it is noted that the Search rule is type A which, as stated previously, allowed any combination of terms to be accepted; the question has 7 starting terms. The search sheets were examined for all documents having a coordination score of 3 or more, and there are two relevant documents sought in this question. Three tables of figures are given, for the three levels of exhaustivity, each table recording the coordination score and language variables. For example, using the highly exhaustive indexing (weights 5-10), a three term coordination score using language 3 retrieves both of the relevant documents, and 60 non-relevant documents. At the next level of exhaustivity (weights 7-10), the non-relevant documents drop to 45; at the lowest level of exhaustivity, the non-relevant documents drop to 10. In this case the recall is maintained throughout, but with index language 6, for instance, at a coordination score of 4, the effect of moving from high exhaustivity to low exhaustivity is to lose the one relevant document retrieved. It will be noticed that no non-relevant figures are given for coordination scores 1+ and 2+, although the relevant documents are shown here. In general, an attempt was made to cut down the clerical effort by ignoring the count of non-relevant documents when the precision ratio was less than 3%, although, as will be recounted in the next volume, some sampling was done at these low precision levels. The figures obtained from this particular question are then ready to be totalled with those from other questions to provide results for a set of questions. This, and the various methods for arriving at these totals, will be considered in the next volume.

There were many additional tests, in which were investigated the effect of such matters as the single term hierarchies, the set of concept languages, again incorporating the various recall devices such as alphabetical and hierarchical grouping, and also the various searches with controlled terms. These other tests meant, of course, that the preparation of the question-indexes had to be commenced from the beginning. For instance, the single-term hierarchies resulted in a group of terms

QUESTION 181

SEARCH RULE A — RELEVANT DOCUMENTS 2

STARTING TERMS 7 — LOWEST COORDINATION FOR NON-RELEVANT 3 — DOCUMENTS TO BE EXAMINED. ALL

INDEX LANGUAGES 1-6

HIGH EXHAUSTIVITY (WEIGHTS 5-10)

INDEX LANGUAGE	COORDINATION LEVELS 1+ R	1+ N-R	2+ R	2+ N-R	3+ R	3+ N-R	4+ R	4+ N-R	5+ R	5+ N-R
1	2		2		1	25	-	-	-	-
2	2		2		1	35	-	-	-	-
3	2		2		2	60	-	2	-	-
4	2		2		2	60	-	6	-	-
5	2		2		2	235	-	32	-	2
6	2		2		2	265	1	42	-	4

MEDIUM EXHAUSTIVITY (WEIGHTS 7-10)

1	2		2		1	20	-	-	-	-
2	2		2		1	30	-	-	-	-
3	2		2		2	45	-	2	-	-
4	2		2		2	45	-	4	-	-
5	2		2		2	175	-	20	-	-
6	2		2		2	195	1	26	-	-

LOW EXHAUSTIVITY (WEIGHTS 9-10)

1	2		2		1	5	-	-		
2	2		2		1	10	-	-		
3	2		2		2	10	-	-		
4	2		2		2	10	-	-		
5	2		2		1	90	-	10		
6	2		2		2	100	-	10		

FIGURE 6.8 RESULTS SHEET FOR QUESTION 181 FOR INDEX LANGUAGES 1 to 6

associated with a starting-term that was different from the group of synonym, word endings and quasi-synonym described earlier. There were some minor modifications in preparing the indexes, but in general the basic procedure described above was used for this further testing.

There was the additional necessity of investigating, on the single terms, the precision devices of interfixing and partitioning, which, as described earlier, are the two stages of links which were recognised, interfixing being concerned with single terms within a concept, while partitioning deals with concepts within a theme. This operation was done by examining the original indexing sheets for the relevant and non-relevant documents that had been retrieved as a result of the searches described above.

To illustrate the procedure adopted, Fig. 6.9 shows the processing of one of the relevant documents (2076) to question 51. This question has eleven starting terms; these are set out at the top of the table, with the double dividing lines indicating the concepts into which the question terms are divided, namely Displacement-Thickness; Plate-Flat; Flow-Compressible; Boundary-Layer-Laminar; Formula-Approximate. These concepts are the pairs and triplets of terms which must be interfixed within concepts. In testing partitioning, all the terms in the search are demanded to occur in one theme of the indexing. Each asterisked term in Fig. 6.9 is the basic term in its concept, and the search rules in operation at this stage of the test demanded that no subsidiary term (i.e. non-asterisked term) would be accepted unless the basic term was present. Thus in the index terms contained in document 2076 listed in the second row, the last term Approximate is not accepted, since Formula is not present. This row shows all the index terms in document 2076 that match with the terms requested in the search prescription, with the weights in brackets, this information resulting from decoding the entries on the search sheet. The index sheet of document 2076 (fig. 6.1) is examined next, the index terms in row 2 are located in the indexing, and the code letters assigned to the concepts in the indexing are recorded in the third row. The first two terms, Displacement and Thickness, both occur in Concept i, and therefore are interfixed; the fourth and fifth terms, Flow and Compressible, occur respectively in concepts d and e, so no interfixing is present. However, an alternative quasi-synonym acceptable in place of Compressible is Hypersonic; this occurs in concept d and thus interfixes with Flow. The fourth row shows the themes from the indexing that contain the greatest number of search terms; theme 02 does not include 'Displacement Thickness'; while theme 04 has this concept, it does not include 'Plate Flat', so both themes give the same results, since both eliminate one concept of two terms. From this data the results can be calculated for interfixing, for partitioning and for partitioning with interfixing, in all of the six index languages and at the three levels of exhaustivity. The results for this single document in regard to these devices are shown on the score sheet (fig. 6.10). This procedure was carried out on all the relevant documents in the questions tested, and also several of the non-relevant documents were examined. The totals of relevant and non-relevant documents for a question are again recorded on a results sheet as before, and from this can be seen the effect on recall and precision of these powerful precision devices.

The testing of the simple concepts involved more index languages than the single terms, since 16 aggregates of recall devices were tested. In this case the code letters used in the columns were each allotted to a single device, rather than a group of letters to a device. (e.g. B was synonyms, C was species, so that even if there were five synonyms or five species, they were all coded with B or C). This was done not only because of the large number of separate results wanted, but because the search

QUESTION 51 DOCUMENT 2076 (Relevance 2)

1. Starting terms	*DISPLACEMENT	THICKNESS	*PLATE	FLAT	*FLOW	COMPRESSIBLE	BOUNDARY	*LAYER	LAMINAR	*FORMULA	APPROXIMATE
2. Search terms in indexing of Doc. 2076, with weights	DISPLACEMENT (8)	THICKNESS (8)	PLATE (8)	FLAT (8)	FLOW (9)	COMPRESSIBLE (9) HYPERSONIC (N9)	BOUNDARY (9)	LAYER (9) FLOW (K9)	LAMINAR (9)		APPROXIMATE (9) but not accepted as FORMULA is not present
3. Concept codes assigned in indexing, showing interfixing	i	i	b	b	d	e d(N9)	e i j	e i j	e i j	-	-
4. Themes assigned in indexing, showing partitioning	Theme B b d e f g - Theme D a d e f h i g i	- i	b -	b -	d d	e d d	e e	e e	e e	- -	- -

FIGURE 6.5 PROCESSING OF DOCUMENT 2076 IN RELATION TO QUESTION 51 FOR ANALYSIS OF INTERFIXING AND PARTITIONING

* These terms must be present for other terms in the concept to be accepted.

QUESTION 51. DOCUMENT 2076

INDEX LANGUAGE Weights	1			2			3			4			5			6		
	5-10	7-10	9-10	5-10	7-10	9-10	5-10	7-10	9-10	5-10	7-10	9-10	5-10	7-10	9-10	5-10	7-10	9-10
NO LINKS	9	9	5	9	9	5	9	9	5	9	9	5	9	9	5	9	9	5
INTERFIXING	8	8	4	8	8	4	8	8	4	8	8	4	9	9	5	9	9	5
PARTITIONING	7	7	5	7	7	5	7	7	5	7	7	5	7	7	5	7	7	5
INTERFIXING & PARTITIONING	6	6	4	6	6	4	6	6	4	6	6	4	7	7	5	7	7	5

FIGURE 6.10 SCORE SHEET FOR LINKS WITH DOCUMENT 2076 FOR QUESTION 51.

STARTING TERMS

a. Compressible Flow

b. Viscous Flow

c. Channels

d. Straightness

SEARCH REQUIREMENTS

Only the following combinations of terms will be accepted.

2-term coordination ac, bc. cd

3-term coordination abc, acd, bcd

4-term coordination abcd

FIGURE 6.11 INSTRUCTION SHEET FOR SEARCH WITH CONTROLLED TERM VOCABULARY

prescriptions contained more related terms than the single term searches did, and would have required more divisions than the 26 in a single letter code. Another answer to the posting problem was not to post any related terms on a document when the natural language term or synonym term (both included in every aggregate of devices) was already there. This could be done provided that a related term did not improve the weights. For example, in document 1978 in fig. 6.6, Flow appears as such as 1-7. Because of this, A6 and K7 are really redundant, but on the other hand the posting of Moving (M) at a weight of 9 is required since this improves the performance in regards to weighting. This superfluous posting was done deliberately on the single terms to enable decoding of all search terms for the interfixing test, but no such requirement existed in the concept searches, and such posting was left off.

As stated, the first series of tests had been done using the minimum of intellect in the search programmes, with the result that many documents were retrieved on nonsensical combinations of terms. At later stages in the test, increasing intelligence was put into the search programmes; this is another way of saying that the requirements were more stringent. This was done in various ways, and each time the attempt was made to identify the particular intellectual decision which had been taken. One example of this is given in Fig. 6.11, where the search was being carried out on the Controlled Term Vocabulary. There are four starting terms, Compressible flow, Viscous flow, Channels and Straightness. Instead of any combination of these being accepted at the various levels of coordination, the search instructions specifically state, for instance, that Compressible flow and Viscous flow are not acceptable on their own. In fact, the definite requirement is that Channels must always be present.

This chapter has only considered the general techniques which were used in carrying out the tests. Quite inapplicable as far as can be seen to any operational situation, they gave, albeit with a large amount of clerical effort, all the flexibility that was required. One point which should be made clear concerns the prior knowledge regarding which documents were relevant to which question. This knowledge was not available to the indexers at the time of indexing, so therefore there is no question of the indexing being slanted towards a particular question. In theory it could have been available to Mills at the time when he was preparing the groups of related terms and the various hierarchies. In fact, Mills was doing this work in London while the indexes were being prepared and the searches were being carried out 50 miles away at Cranfield. Even if he had had access to this data and had attempted to use it in preparing these lists, we do not believe it would have made any significant difference to the results. With regard to the searching, the description given in this chapter of the methods used should make it obvious that its comprehensive nature precluded any possibility of influencing the results.

CHAPTER 7

Additional Tests

The first year of the project coincided with a time when a number of groups, who had been investigating various methods of statistical association, were becoming interested in the possibility of putting their methods to the test, and we received some enquiries regarding the possibility of the project test collection being used as a 'common sample'. All such groups were, of course, working with computers, so with the agreement of the National Science Foundation, it was arranged that a tape should be prepared of the indexing for the 1400 documents in the test collection. This was done by I.B.M. (U.K.) Ltd., and an example of the printout for document 1420 is given in Appendix 6.1.

In the end, for various reasons, none of the groups in America was able to make use of these tapes. However, in England, Drs. Roger and Karen Needham decided to use the Cranfield collection for a test of the 'clumping' technique developed at the Cambridge Language Research Unit. (ref. 29). Since the computer to be used was the Atlas, it was necessary to prepare a set of paper tapes from the punched cards. The problems involved in this are not for us to relate, but the indexing has now been completed, and a copy of the printout for document 1420 is also given in Appendix 6.1.

At a later stage in the project, when the results were coming through, a meeting with Professor Salton made it clear that the research which he had been undertaking at Harvard was basically along similar lines to the work at Cranfield, in that both groups were concerned with comparing the performance of various index language devices, The difference lay in the methods adopted for the clerical processes of the testing, and the SMART programme (ref. 30) gave the flexibility of rapid testing of any set of documents for which the necessary relevance assessments had been made in relation to a set of questions, so long as these were in a subject field for which suitable vocabularies had been prepared. The original testing of the SMART programme had been carried out on a collection of abstracts dealing with computers, and for both groups the prospect of using the programme to test the subset of documents taken from the Cranfield project was very attractive. For Professor Salton, it gave the opportunity of testing his programme in a different subject area; for us it opened up a completely new field. There would be the opportunity for directly comparing the results of the devices being investigated at Cranfield with the similar, but more complexly calculated, devices used at Harvard. Secondly, there was the possibility that it would assist in solving some of the interesting problems involved in the presentation of results. The recall-precision curves, based on a series of cutoffs, were producing at Cranfield quite different figures from the normalised recall and normalised precision based on the ranked output at Harvard. This was only to be expected, since the method of calculation was so different, but it was important to be able to find how to equate the different sets of figures. The final point of interest was that though the Harvard searching was normally done on document abstracts, the flexibility of the SMART programme made it practical for the searches to be carried out on both the abstracts and the indexing which had been done at Cranfield, thus providing for the first time a comparison between searches based on abstracts and on indexing.

A member of the Cranfield group spent a week at Harvard, and as a result of the visit, it was arranged that a subset of the collection, consisting of 200 documents and 42 questions, should be processed at Harvard, and that searches should be made

on both the indexing of these documents and also their abstracts. Subsequently the decision was taken to extend this work so as to cover the whole of the Cranfield collection.

Citation indexing

It was in 1961 that the first major grant was given for a citation index (ref. 32), and the following year we were asked for our views on how a citation index could be evaluated. Citation indexing is basically a method of forming classes of documents which are all related through a common reference to a base document. There will, of course, be many occasions when the class consists of only a single entry; however, in the more numerous cases where the class consists of two or more documents, then citation indexing can be considered equivalent to bibliographic coupling at a strength of one. Bibliographic Coupling has been developed by Dr. M. Kessler at the Massachusetts Institute of Technology (ref. 20), and in relation to citation indexing, can be considered as a precision device, since it progressively narrows the class of documents as the demand for common references increases in number. Citation indexing and bibliographic coupling could therefore be tested in the same way as any other device; it was, however, necessary to prepare an index for this purpose.

The first stage was to prepare xerox copies of the citations in the 1400 documents in the test collection; against each citation was put the code number for the citing documents, after which each citation was cut up so as to appear on a separate slip of paper. This resulted in some 20,000 slips of many various sizes, which had to be sorted into author alphabetical order. This being done, the slips were pasted onto sheets of paper; where two or more slips related to the same cited document, only one example was pasted in; the references to the additional citing documents were entered alongside. This can be seen in Fig. 7.1, which covers a series of references to papers by H.J. Allen, in particular a paper written with A.J. Eggers entitled 'A study of the motion and aerodynamic heating of missiles entering the earth's atmosphere at high supersonic speeds.' (NACA TN 4047). This is shown to have been cited by thirteen papers in the test collection.

This procedure resulted in a normal citation index; to obtain the index for bibliographic coupling required three further stages. First, each cited reference having two or more citations was given a code number, the paper by Allen and Eggers being A25, and a separate card was prepared for each cited reference. On this reference card was written the code for the cited document (i.e. A25) and then, in numerical order the codes for the citing documents. Fig. 7.2 illustrates the reference card prepared in connection with the paper by Allen and Eggers shown in Fig. 7.1.

The reference cards were sorted into numerical order depending on the lowest number on each card. Since these numbers represented the codes for the citing documents, they ranged from 1001-2400. Each card was then taken in turn, and the information from all reference cards having the same starting number was transferred to a master card. As an example, the master card shown in Fig. 7.3 illustrates the position with regard to document 1067, this number being posted in the top left hand corner. In the column headings are entered the code numbers for the documents which have been cited by document 1067, this information being obtained from the reference cards such as Fig. 7.2, and these being, in this particular case, A25, W32, A23, E23, F90, and O24. In the first column of the master card are entered the document numbers of all other citing papers, this information being again obtained from the reference cards. A tick is put against each number in the column under the appropriate heading

Allen, D. J.: The Application of Multhopp's Subsonic Lifting Surface Theory to the Calculation of the Aerodynamic Forces Acting on a Wing of Finite Aspect Ratio Oscillating in Arbitrary Elastic Modes With Control Surface Freedom. Design Dept. Rep. No. 1191, Hawker Aircraft, Ltd. [Kingston-on-Thames, England], June 1953.

D. N. de G. Allen and R. V. Southwell — Relaxation methods applied to determine the motion, in two dimensions, of a viscous fluid past a fixed cylinder. *Quart. J. Mech. App. Math.* III. Part 2. June, 1955.

[15] Allen, F. J., *An Elastic-Plastic Theory of the Response of Cantilevers to Air Blast Loading*, BRL MR No. 880, April 1955.

Allen, H. J.: General Theory of Airfoil Sections having arbitrary Shape or Pressure Distribution. NACA Tech. Rept. No. 833, 1947.

[18] Allen, H. Julian, *Pressure Distribution and Some Effects of Viscosity on Slender Inclined Bodies of Revolution*, N.A.C.A. T.N. No. 2044, 1950.

Allen, H. Julian: Motion of a Ballistic Missile Angularly Misaligned With the Flight Path Upon Entering the Atmosphere and Its Effect Upon Aerodynamic Heating, Aerodynamic Loads, and Miss Distance. NACA TN 4048, 1957.

4. Allen, H. Julian: A Simplified Method for the Calculation of Airfoil Pressure Distribution. NACA TN No. 708, 1939.

3. Allen, H. Julian: Estimation of the Forces and Moments Acting on Inclined Bodies of Revolution of High Fineness Ratio. NACA RM A9I26, 1949.

8. Allen, H. Julian: Calculation of the Chordwise Load Distribution over Airfoil Sections with Plain, Split, or Serially Hinged Trailing-Edge Flaps. NACA Rep. No. 634, 1938.

[38] Allen, H. J., "Hypersonic flight and the re-entry problem," J. Aerospace Sci. 25, 217-227 (1958)

Allen, H. Julian, and Eggers, A. J., Jr.: A Study of the Motion and Aerodynamic Heating of Missiles Entering the Earth's Atmosphere at High Supersonic Speeds. NACA TN 4047, 1957.

[24] Allen, Harrison, Jr., and Fletcher, E. A., *Combustion of Various Highly Reactive Fuels in a 3.84-by-10 inch Mach 2 Wind Tunnel*, NASA MEMO 1-15-59E.

H. J. Allen, M. A. Heaslet and G. E. Nitzberg — The interaction of boundary layer and compression shock and its effect upon airfoil pressure distributions. N.A.C.A. RM A7A02 (1947).

Allen, H. Julian, and Nitzberg, Gerald E.: The Effect of Compressibility on the Growth of the Laminar Boundary Layer on Low-Drag Wings and Bodies. NACA ACR, Jan. 1943.

8. Allen, H. Julian, and Perkins, Edward W.: A Study of the Effects of Viscosity on the Flow Over Slender Inclined Bodies of Revolution. NACA Rep. 1048, 1951. (Supersedes NACA TN 2044.)

Allen, H. Julian, and Perkins, Edward W.: Characteristics of Flow Over Inclined Bodies of Revolution. NACA RM A50L07, 1951.

12. H. J. Allen and W. G. Vincenti, "Wall Interference in a Two-Dimensional Flow Wind Tunnel, with Consideration of the Effect of Compressibility," *N.A.C.A., T.R.* 782 (1944).

14 Allen, R.A., Keck, J.C. and Camm, J.C., "The Recombination of Nitrogen at 6400 K, "Avco-Everett Research Lab., Research Note 243, June 1961.

FIGURE 7.1 CITATION INDEX SHEET

A25			
1067	1077	1163	1164
1715	1719	1815	1816
1978	1982	2002	2346
2379			

FIGURE 7.2 CITATION INDEX REFERENCE CARD

1067

	A25	W32	A-23	E23	F90	0-24
1077	✓					
1163	✓			✓		
1164	✓			✓		
1715	✓	✓			✓	✓
1719	✓					
1815	✓					
1816	✓					
1978	✓					
1962	✓					
2002	✓					
2346	✓					
2379	✓		✓		✓	
1116		✓				
1150		✓				
1428		✓				
1448		✓				
1518		✓				
1531		✓				
1660		✓				
1704		✓				
1767		✓				
2137		✓				
1639			✓		✓	
1716			✓		✓	
1816			✓			
2001			✓			
1225				✓		
2329				✓		
2345				✓		
1499					✓	
1759					✓	
2000					✓	

FIGURE 7.3 CITATION INDEX MASTER CARD

to indicate that it cited this particular reference. However, when a document number appears in connection with another cited reference, the number is not repeated, but a tick is put in the appropriate column. For instance, in Fig. 7.3, it can be seen that this document has two references in common with document 1163, 1164, 1639 and 1716, three references in common with document 2379, and four references in common with document 1715. To return to Fig. 7.2, when document 1067 had been entered, then this number was crossed off and the reference card was re-sorted in the pack under the next number, namely 1077; again this number was crossed off when the master card had been entered for document 1077, and the reference card re-filed on the next number and so on until all the document numbers had been entered.

The final stage was to go through the master cards and prepare the bibliographic coupling card (Fig. 7.4). This showed the master document and all the other documents with which it had two or more references in common.

It is clearly a matter for argument as to how a citation index should be tested operationally, but within the context of these experimental investigations, it was relatively simple to decide on the method to be used. Our concern was how a citation index operated in regard to recall and precision and the procedure adopted was as follows. For a certain question, the relevant documents were known as well as their relevance level. The numbers of the relevant documents were written across the score sheet as shown in Fig. 7.5, referring to question 34. In order to avoid complexity, a fairly simple example has been taken, where there were six documents all of relevance 3.

The numbers in the left hand column indicate the coupling strength going from a maximum of 6+ down to 1, which latter represents citation indexing. The appropriate bibliographic coupling cards were then taken from the pack, the first of these relating to document 1067. As can be seen in Fig. 7.4, document 1715 had a match of 4 with document 1067; there were no other documents at this level of match, and since document 1715 is also a relevant document to question 34 (see Fig. 7.5), this is counted as a success and the score is entered appropriately. The document which matches at a level of 3 is not relevant, so this now makes the score one relevant and one non-relevant. At a match of 2, three of the documents are relevant (1164, 1639 and 1716), so the score here becomes four relevant and two non-relevant. By referring to the cards shown in Fig. 7.3, we can calculate the number of documents involved with a single match. There are no other relevant documents in this set, but many non-relevant, so the score for this is shown as four relevant and thirty-two non-relevant.

This process is repeated for all the other relevant documents, as shown in Fig. 7.5. When this has been done, the scores can be totalled to give a set of figures where obviously the maximum recall and the lowest precision will be obtained at a match of 1, and maximum precision with lowest recall obtained at a match of 6+. However, there are various approaches that can be taken in compiling the score, and these will be considered in the volume of test results. Such analysis was done for documents of all degrees of relevance.

In bibliographic coupling as discussed by Kessler, account is only taken of the actual match rather than what might be called the proportional match. For instance, two review-type articles may each have fifty references, as against two other papers which have only three references. If the former pair of papers have five references in common, this would be considered a stronger coupling strength than the latter pair

1067						
2	3	4				
1163 1164 1639 1716	2379	1715				

FIGURE 7.4 BIBLIOGRAPHIC COUPLING CARD

RELEVANT DOCUMENTS

COUPLING	1067		1164		1639		1715		1716		1717	
STRENGTH	R	N-R	R	N-R	R	N-R	R	N-R	R	N-R	R	N-R
6+												
5+												
4+	1	0	0	1			1	0				
3+	1	1	0	2	1	0	1	0	1	0		
2+	4	2	1	6	2	3	1	5	2	2	0	3
1+	4	32	1	49	2	63	1	47	2	33	0	18

SCORE

	RELEVANT	NON-RELEVANT	RECALL RATIO	PRECISION RATIO
6+	-	-	-	-
5+	-	-	-	-
4+	2	1	33%	66%
3+	4	3	66%	57%
2+	5	19	83%	21%
1+	5	201	83%	2%

FIGURE 7.5 SCORE SHEET FOR BIBLIOGRAPHIC COUPLING FOR QUESTION 34.

1067 (6)	
3.	1715 (9)
7.	1716 (7)
11.	2379 (17)
12.	1163 (9)
16.	1639 (11)
25.	1164 (17)

FIGURE 7.6 RECALCULATED BIBLIOGRAPHIC COUPLING CARD
Figures in brackets represent number of references in documents. Figures in first column give the calculated weighting.

which have, for example, two references in common. However, proportionately, it could be argued that the latter represents a stronger match than the former. To test this, the number of references in each document of a matching pair were multiplied, the resultant figure was then divided by the square of the number of matches and the final figure was considered as the level of coupling strength. For example, document 1067 (see Fig. 7.4) had six references, and document 1163 had eight references, giving a multipland of 48. These had a match of 2, so dividing by 2^2 gives a final weighted figure of 12. Document 1715, however, had nine references which, combined with document 1067, gives a multipland of 54. In this case, since there is a match of four, this figure has to be divided by 4^2, giving a final weighted figure of 3. When the matches for document 1067 had all been worked out, the weighting becomes as in Fig. 7.6. In many cases, the result of this exercise showed significant changes in coupling strength, and therefore the collection was re-tested in the manner described earlier, only this time the scoring was based on these new coupling levels.

CHAPTER 8

Comments

This report has attempted to outline the reasoning and the procedures adopted in the second Cranfield project. It could be argued that comments on these matters should await the publication of the test results, but it is felt more appropriate to conclude this volume by briefly considering some of the short-comings of the design and the techniques used, and showing how the results might possibly be affected.

In Chapter 2, some aspects of the test design were considered from the viewpoint of the decisions which seemed correct in 1961, at the time when the project was prepared. While the test results and the conclusions which can be derived from them will show to what extent the test design is such as to allow the objectives to be achieved, there are certain matters which can be discussed immediately.

The original proposal suggested a collection of 1,200 documents with some 300 questions to be used for searching. For no very good reason, the total of documents in the collection was increased to 1,400; while there would have been no difficulty in finding 300 usable search questions from the 641 that were submitted, only 279 questions were used, and, for most of the tests, this number effectively was reduced to 221. The amount of data which has been obtained from this question-document set is vast, and is more than sufficient for validation of the test results. It can at present only be a matter for discussion as to whether the question-document set was larger than necessary. In many of the tests, sub-sets of the collection were used, sub-sets such as 200 documents and 42 questions. There is a double danger in the use of such comparatively small sets; firstly that they will produce results which are unrepresentative, and secondly that the performance measures will be seriously distorted. To consider the latter point, investigating the effect of generality ratio was a part of the project and although the matter is somewhat complex, it has been possible to work out the relationship between the performance figures for varying generality ratios. This work is reaching the stage where it can be applied in all situations, so this particular problem need no longer create any difficulty in the use of a small collection.

Far more serious is the question of whether the collection size is large enough to give valid results. It has to be remembered that this investigation has been concerned with only one variable, namely index language devices, and this is quite unlike the situation in Cranfield I, in which additional variables were such matters as indexing time, indexers, and type of document. The result is that a much smaller set than the 18,000 documents and 1,200 questions of Cranfield I was required and there does not seem to be any doubt but that the collection of 1,400 documents was large enough for the test. The experience at Cranfield and Harvard of working with a sub-set of 200 documents and 42 questions has produced some useful evidence on the question as to whether the total collection was larger than necessary. With the knowledge that the sub-set produced results very similar to those obtained with the complete collection (when due allowance is made for the generality ratio), it now seems possible that a smaller collection would have served equally well. However, lacking this hind-sight knowledge, it is very likely that the results obtained with a smaller collection would have been subject to criticism which could not have been satisfactorily refuted.

The method of obtaining a document collection and a set of questions turned out to be a perfectly satisfactory way of operating. The response from the authors of research papers was remarkably good, and can be interpreted as showing that the

scientific community - at any rate in the field of aeronautical engineering - is interested in documentation problems, and is willing to co-operate in helping to find an answer to these problems. The selection of the comments from the authors (given in Chapter 3) is only a sample, illustrating various points, of the many interesting and encouraging letters which were received.

Tied in to this method of obtaining the set of documents and questions was also the matter of obtaining relevance assessments, and here some reservations have to be admitted concerning the method adopted. This is not to suggest that there is any experimental evidence of there being any better or more satisfactory technique, but rather to say that the matter of relevance assessment is, without any doubt, the most difficult intellectual problem - in fact, one of the very few remaining problems - in the evaluation of information retrieval systems.

In the evaluation of operational systems, there will be many occasions when the only satisfactory technique will be that of using actual questions for test searches, with the questioners assessing the relevance of the documents retrieved at the time when the information is required. Such would be the case if it was desired, for instance, to investigate the effect of different levels of questioner participation in the search programme. As soon as any deviation is made from this technique of operating in a real-life and real-time situation, a less realistic method is being used, although there will frequently be situations where this could be justified for economic or other practical reasons. This latter point is certainly true of an evaluation of an operational system, and it is equally true of the test of an experimental system, where no real user group can be said to exist. A possible weakness of the method adopted in this test lies in the fact that the subjectivity of the relevance assessments might have been such that it will mask the variation in performance of the various devices which were being tested. There is no experimental evidence of any kind at present available that makes it possible to affirm that this is so, but the possibility is such that it requires investigation.

As stated earlier, the problem of relevance decisions is presently the most serious in the field of evaluation, and is attracting the attention of many groups. There is the very interesting work of Katter (ref. 33) in which a large number of people will be asked to make 'distance' judgements between small sets of documents. In this work the important aspect of the test design is to find which type of document surrogates result in distance judgements which match most closely those judgements made by assessing the complete documents. Then there is the work of Cuadra (ref. 34) where up to one hundred individuals will be asked to assess a set of documents in the field of information disemination, storage and retrieval. Here the attempt will be to identify and investigate the variables which influence an individual's response, and a somewhat similar investigation is being directed by Rees at Weston Reserve University (ref. 35). More empirical is an investigation proposed by Cleverdon which is to be undertaken by ASLIB. This is intended to identify the reasons why individuals reject documents which apparently meet their requirements and alternatively why they accept, as relevant, documents, which to a third person seem no more acceptable than those rejected. This investigation will be carried out on some 600 individuals in twelve different organisations and, unlike the other three projects, the relevance assessments will be made in actual operational conditions.

However, none of these investigations into relevance apply to the problem raised in this test. Here the situation is that a series of tests on various index languages have been carried out, where the scoring for each test is based on the relevance decisions of individuals simulating, as far as possible, a real life situation, with individual

variations hopefully evened out by having nearly two hundred different questioners In an investigation that had very similar objectives, Salton went to the other extreme in his original tests. Using only seventeen questions in the general field of the test collection, these questions were specially prepared for the test and did not represent any actual requirements. The set of 400 documents in the collection were then assessed against these prepared questions by a number of students, this assessment being based only on short abstracts. Since the searching was also done on the abstracts, there was obviously the probability of even more distortion than was the case with the source document questions of Cranfield I. The interesting point, however, is that this seemingly crude technique of question preparation and relevance assessment did, in fact, allow a considerable amount of useful data to be obtained concerning the performance characteristics of the various index language options, and this data appears to be sufficiently valid for certain conclusions to be reached. When this evidence is added to that obtained from Cranfield I, there are some grounds for suggesting the possibility that everyone is over-emphasising the importance of relevance assessments in experimental testing, and that, however relatively unscientific the method used, reliable information can be obtained.
It is intended to investigate this point in future work at Cranfield by having various people make new relevance assessments of the document-question sets used in the present project. The search results can then be re-scored on the basis of these new - and presumably somewhat different - relevance assessments, and analysis will show whether the comparative performance of various index languages is thereby affected.

In experimental testing, the common practice, not unnaturally, is for the groups to work with document collections with which they have some familiarity, and this project was no exception. The language of aerodynamics might be said to fall somewhat to the left of centre in regard to its precision, it is, in fact, mushy rather than firm. As such it presented a number of difficulties; not only could one find the same notion being expressed in different ways by different authors, one often had the situation where the same notion was expressed in different ways in the same paper. Discussing this point with one of the authors, he said that certain people considered it good style if, after expressing a notion in the title in one way, a new phrase could be used for the abstract and another phrase be found for the actual text. Even without this particular complication, the subject matter was full of semantic problems. An illustration of this is provided by a question (not in any way a-typical) which, as originally received, read

'Has anyone investigated relaxation effects on gaseous heat transfer to a suddenly heated wall'

When asked to suggest alternative search terms, the questioner sent back the following comments.

Relaxation effects. Could be replaced by 'excitation of internal molecular energy modes (or states)'. The excitation could result from collisions between gas molecules alone, or gas molecules and molecules in the solid.
Gaseous heat transfer. 'Gaseous' could be omitted, but does help to limit the field. 'Heat' could be replaced by 'energy', 'transfer' by 'conduction' or 'transmission'.
Suddenly heated wall. 'Suddenly' could be replaced by 'rapidly', 'heated' by 'cooled' and 'wall' by 'solid'

Finally, if any of the above permutations are unsuccessful, the question could be rephrased to read 'Has anyone investigated the conditions at the wall behind a plane reflected shock front in a real gas by theoretical analysis'.

The semantic difficulties of papers in aerodynamics provided a very stiff test of the recall devices, and as such it could be considered a suitable subject area for the test. However, the lack of syntactic difficulties caused a change of plan, as considered on pages 56 and 57, in that it was not a practical proposition to use roles. It is an interesting point to consider as to whether another inverse relationship exists, this time between the semantic and syntactic problems involved in the indexing of any particular subject field. Alternatively, and possibly more likely, the position may be that with a mushy subject language, the over-riding necessity of obtaining a reasonable recall ratio inhibits the use of precision devices; in other words, the semantic problems are so difficult to solve that they completely overshadow the syntactic problems. However, in a firm subject language area the semantic problems are more easily solved, so the syntactic problems loom larger, and one can afford to use precision devices, such as roles. If either of these situations exists, it will obviously have consequences in the endeavours to obtain a common sample of documents that can be used to illustrate and evaluate different types of systems, such as the work at Chicago. Here the intention is to have 'an open-ended collection of exemplars of indexing systems applied to a common sample of documents' (ref. 36). The indications are that any given sample of documents would favour certain types of index languages, but handicap other index languages, this being dependent on their strong and weak points in relation to devices intended to overcome the semantic or syntactic problems.

It would seem, that next to the question of relevance assessments, the determination of the effect of subject language precision is the most important problem to be tackled. This is certainly true of experimental situations where it is necessary to compare the performance of tests based on different document collections. For instance, in the comparison of the results obtained by the SMART tests and in Cranfield II, it is now possible (by the methods discussed in a later volume of this report) to normalize the different measures used and the effect of generality ratio. Since it is also theoretically possible to match similar types of index languages and the method of relevance assessment, any remaining difference in performance figures must be due to the firmness level of the language of the two subject areas, namely computers and aerodynamics.

In addition to experimental situations, knowledge of this factor is also important for the design of an operational system which covers a broad subject field, and where there is thereby a wide range in the firmness level. An investigation of this problem could be attempted by a linguistic analysis of the variation of terms in different subject fields - how many different terms or phrases can be used to express the same notion and conversely how many meanings a single term has. The experimental method of investigating the problem to be used at Cranfield will be a procedure that reverses the present project. Instead of testing a large number of index languages against a single document set, it will be necessary to find the different performances achieved when a large number of document sets in different subject fields are tested against a single index language.

No particular fault is at present apparent in regard to the indexing which proceeded according to schedule and was completed during the first year of the project. In the

next stage of the work, there was probably an error of judgement in putting so much effort into the preparation of the single-term hierarchies. It is doubtful if anyone has previously attempted to compile classification schedules consisting entirely of single words, and it proved to be a very difficult, but therefore very interesting task. It was right that, in this project, the attempt should have been made, but an earlier realisation of the limited affect which the schedules would have on the performance of the system would have led to a decision that less time should be expended in their preparation.

The main objective of the test is to ascertain the effort of various index language devices on the performance of information retrieval systems. To conclude this volume, it is reasonable to claim that, although there are some operations which might have been done better another way, nothing has happened seriously to militate against the possibility of achieving this test objective.

REFERENCES

1. CLEVERDON, C.W. — Report on the first stage of an investigation into the comparative efficiency of indexing systems. Cranfield, 1960.

2. CLEVERDON, C.W. — Report on the testing and analysis of an investigation into the comparative efficiency of indexing systems. Cranfield, 1962.

3. AITCHISON, J. and CLEVERDON, C.W. — Report of a test on the index of metallurgical literature of Western Reserve University. Cranfield, 1963.

4. SWANSON, D.R. — The evidence underlying the Cranfield results. Library Quarterly, 35, 1965, pp. 1-20.

4A. CLEVERDON, C.W. — The Cranfield hypotheses. Library Quarterly, 35, 1965, pp. 121-124.

5. LANCASTER, F.W. — Project SHARP. Information storage and retrieval system; evaluation of indexing procedures and retrieval effectiveness. 1964. Bureau of Ships, Washington.

6. MONTAGUE, B.A. — Testing, comparison and evaluation of recall, relevance, and cost of coordinate indexing with links and roles. Proceedings of the American Documentation Institute. Vol. I, 1964, pp. 357-367.

7. BOURNE, C. — Review of the criteria and techniques used or suggested for the evaluation of reference retrieval systems. Unpublished note prepared for N.S.F. Evaluation Meeting, October, 1964.

8. RICHMOND, P. — System evaluation by comparison testing. 1965. (To be published in College and Research Libraries).

9. VICKERY, B.C. — On retrieval system theory. London. Butterworths. 1961.

10. GULL, D.C. — Seven years of work on the organisation of materials in the special library. American Documentation, 7, 1956, pp. 320-329.

11. CLEVERDON, C.W. and THORNE, R.G. — A brief experiment with the Uniterm system of coordinate indexing for cataloguing of structural data. R.A.E. Library Memo 7, 1954.

12. METCALFE, J. — Information indexing and subject cataloguing. New York. Scarecrow Press. 1961.

13. TAUBE, M. — The pseudo-mathematics of relevance. American Documentation, 16, 1965, pp. 69-72.

14. SALTON, G. — The evaluation of computer-based information retrieval systems. F.I.D. Congress. Washington. October, 1965.

15. BRANDHORST, W.T. and ECKERT, P.F. — NASA Search System analysis sheet. American Documentation, 16, 1965, pp. 124-126.

16. GIBB, M. — Keywords to information. New Scientist, 26, 1965, pp. 662-663.

17. ALTMANN, B. — Multiple testing of the ABC method and the development of a second-generation model. Harry Diamond Laboratories. Washington. 1965.

18. SWANSON, D. — Word correlation and automatic indexing. Phase I Final Report. Ramo-Wooldridge. 1960.

19. O'CONNOR, J. — Some suggested mechanized indexing investigations which require no machines. American Documentation 12, 1961, pp. 198-203.

20. KESSLER, M.M. — Bibliographic coupling between scientific papers. Massachusetts Institute of Technology, R-2. 1962.

21. JONKERS, F. — Indexing theory, indexing methods and search devices. New York. Scarecrow Press. 1964.

22. Te NUYL, Th.W. — The L'unité mechanized documentation system. Revue de Documentation, 28, 1962, pp. 140-147.

23. FARRADANE, J.L. — Relational indexing and classification in the light of recent experimental work in psychology. Information Storage and Retrieval 1, 1963, pp. 3-11.

24. MARON, M.E. and KUHNS, J.L. — On relevance, probabilistic indexing and information retrieval. Journal of the Association for Computing Machinery, 7, 1960, pp. 216-244.

25. WHELAN, S. — Library retrieval: the R.R.E. pilot retrieval scheme. R.R.E. Journal. October, 1958, pp. 59-68.

26. GARDIN, J. — SYNTOL. Rutgers Series on Systems for the Intellectual Organisation of Information. Vol. 2, 1965.

27. ENGINEERS JOINT COUNCIL — Thesaurus of engineering terms. New York. 1964.

28. WALTON, T.S. — FROLIC: Thesaurus and code dictionary. Washington. David Taylor Model Basin. 1963. (Unpublished).

29. NEEDHAM, R.M. and SPARCK JONES, K. — Keywords and clumps: recent work on information retrieval at Cambridge Language Research Unit. Journal of Documentation, 20, 1964, pp. 5-15.

30. SALTON, G. — Progress in automatic information retrieval. I.E.E.E. Spectrum. August, 1965, pp. 90-103.

31. O'CONNOR, J. — The Scan-Solumn Index. American Documentation, Vol. 13, 1962, pp. 204-209.

32. Three-year project to study citation index methodology. Scientific Information Notes, 3, 1961, No. 3, p. 10.

33. Investigation into a method for analysing document representation techniques. Scientific Notes, 7, No. 3, 1965.

34. Laboratory approach to the study of relevance assessments in relation to document searching. Scientific Information Notes, 7, No. 4, 1965, p. 14.

35. Field experimental approach to the study of relevance assessments in relation to document searching. Scientific Information Notes, 7, No. 4, 1965, p. 14.

36. Application of selected indexing systems to a common sample of scientific documents. Scientific Information Notes, 5, No. 6, 1963, p. 9.