CHAPTER 9

COMMENTS ON THE RESULTS

This test has produced a large mass of data, and even for those who have been actively engaged in the project, it is not always easy to separate that which is significant from that which is of little value. In this report and appendices, the attempt has been made to present at least examples of everything that might be of interest to serious readers so that they may have the opportunity of making their own assessments, and it has been the endeavour to keep to a minimum personal expressions of opinion. In the course of the work, the staff obtained a number of impressions concerning the various systems being used. These cannot be expressed as figures or percentages, but will, to a certain extent, influence the interpretation of the results as given in this chapter.

At first glance, the results as given in Table 3.1 show an unexpected level of performance, particularly if allowance is made for the possible standard error. Including this in the percentage results would produce the following figures:-

U.D.C.	73.1% - 78.1%
ALPHABETICAL	79% - 84%
FACET	71.3% - 76.3%
UNITERM	79.5% - 84.5%

Taking these at their highest or lowest ranges shows that all four systems could be within the range of 3.2%, and it is doubtful if a claim that this is significant in an investigation of this nature could be substantiated. However, the figures at the other extremes would give a difference of 13.2% and this would be a worthwhile figure.

More logical is to consider the figure for the result of the searches which were based on documents in the final sub-programme of documents P12001 - P18000 (Table 3.5), for these represent the work of the indexers when they were operating at their highest efficiency. The figures here give Uniterm with a superiority of 3.8% over Alphabetical which is in turn 5.3% above U.D.C., with Facet a further 3.5% behind.

The breakdown according to system and time, as shown in Table 3.2, is subject to a larger standard error of between 5 and 6%. This can be reduced by ignoring the system and grouping the indexing by time only, and results in the figures as shown in Table 9.1.

Indexing Time Minutes	Percentage Retrieval
2	72.9%
4	80.2%
8	76.2%
12	82.7%
16	84.3%

TABLE 9.1

EFFICIENCY BY TIME FOR ALL SYSTEMS

These figures show an expected increase from the short time of two minutes to the longest time of sixteen minutes, with the exception of eight minutes. In the view of the indexers, there is a possible explanation as to why the efficiency showed a drop at this middle time. When working at the short times of two minutes and four minutes, they knew that there was little time to spare, and concentrated on making entries for the essential concepts. When working at the eight minute level, they felt free to study the whole document more closely before making any indexing decisions. In fact the time allowance was not really sufficient for this to be done thoroughly, so the end result was that the standard of the indexing decreased. This is not a statement that can be substantiated and might appear to be contradicted by the figures given in Table 1.2. This shows that there was an increase in the number of entries in the catalogue according to indexing time, and it could be assumed that, when recall was the only factor being measured, extra entries in the indexes would result in an improvement. On the other hand there is slight support for the argument, in that analysis of the failures showed a greater proportion of errors in indexing at eight minutes than at the other time allowances.

While the figures indicated that two minutes was probably too short an indexing time, four minutes gave results which by all systems are only slightly inferior to the indexing times of twelve minutes and sixteen minutes. By Uniterm the

four minute time was within 1% of the figure for sixteen minutes, and with a collection of documents such as used by the project it appears that there would be no justification in spending more than four minutes indexing. With sixteen minutes indexing, not only would the indexing costs be four times as great, but also there would be the increased posting cost occasioned by having an average of 12.4 Uniterms per document as against 7.7 at four minutes.

The indexing times given in the project were governed by the use of a stop watch, and represent real time spent on actual work. As discussed in Ref. 1, pages 24-25, to translate them into a real life situation would require adding 60% to the basic times, so that the four minute indexing time would thus become about $6\frac{1}{2}$ to 7 minutes. This brings it within the range of indexing time that was most generally used by the supplementary indexers, as shown in Table 8.7, and is also very close to the average time found necessary by Mrs. Aitchison when indexing documents in connection with the W.R.U. test.

The breakdown by indexer (Table 3.3, page 23) shows no significant differences between the three members of the staff, but worthy of comment is the ability shown by Mr. Hadlow to index with an efficiency equal to that of the more experienced members.

The assumption is commonly made and frequently asserted that scientific or technical documents can only be satisfactorily indexed by a person having detailed subject knowledge of the field. This may be true in certain subject fields, but all the results of the investigation show that it cannot be substantiated for such subjects as physics, engineering or metallurgy. In the main test analysis (Chapter 5) it is shown that only with two of the 1200 source documents could lack of subject knowledge be considered a contributory factor in failure to retrieve. The librarian indexers certainly had difficulties with many of the documents, and frequenctly had recourse to scientific and technical dictionaries, but they were very successful in dealing with a wide range of research papers, particularly considering that Mr. Hadlow came direct from a public library, with no experience either of the subject or of doing indexing of this type. In addition it has to be remembered that, during the two years, neither he nor the other indexers received any feedback as would be the case in a more normal true life situation, where the probability would be that they would not only be indexing documents but would also be helping to answer enquiries, and thus acquiring knowledge both of the type of questions put to the system and of the ability of the indexing to meet the requirements.

Table 3.4 considers the results of searches based on the subject matter of the question. There is a possibly significant difference in favour of general questions. This was to be expected, since the major concentration on documents in a narrow subject field resulted in a multiplicity of alternative locations, whereas in the wider, more general field, there would be less difficulties in selecting the appropriate headings. It was interesting, and in view of other results, possibly significant that the difference does not appear with Uniterm.

Table 3.5 (page 24) indicates that there was an improvement of 17.5% for Uniterm in the last sub-programme as compared with the first, 13.3% for U.D.C. and only 7.8% for Alphabetical. It was surprising to find such a large improvement in Uniterm. It might appear to be accounted for by the fact that the number of Uniterms posted in the final sub-programme was 32% higher than in the first sub-programme (see Ref. 1, Table 5A). This difference in posting, however, is lower than the difference of 60% when indexing at the different times of four minutes and sixteen minutes. Since in this latter case there was an insignificant increase in the efficiency, it would seem that there must have been a greater definite improvement in the standard of indexing by Uniterm than by the other systems. Table 5.5 (page 49) gives some information on this matter. It is a rearrangement of the figures as set out in Table 5.4 (page 49), being the reasons for failure of all searches covering documents in the final sub-programme, with emphasis on the human error involved. There is shown to be an average for all systems of 35% of failures due to bad indexing, ranging from 42% for U.D.C., to 32% for Facet. The number of searches made using questions based on the two first sub-programmes was limited, so the failures are so few that there is doubt as to whether they can be considered sufficient in number to permit of the detailed breakdown of reasons for failure as has been done for the main programme. With this reservation, Tables 9.2 and 9.3 have been compiled in an attempt to show the particular point concerning learning rate for the different systems.

	U.D.C.	ALPHA.	FACET	UNITERM	TOTAL
Personal Errors					
Indexing	15 (43%)	10 (40%)	-	13 (52%)	38 (41%)
Searching	4 (11%)	2 (8%)	-	1 (4%)	7 (8%)
Time Allowance	6 (16%)	4 (16%)	-	3 (12%)	13 (17%)
Question	5 (15%)	5 (20%)	-	4 (16%)	14 (17%)
All Other Reasons	5 (15%)	4 (16%)	-	4 (16%)	13 (17%)

TABLE 9.2

BREAKDOWN OF REASONS FOR FAILURE ON DOCUMENTS INDEXED IN FIRST SUB-PROGRAMME (DOCUMENTS 1 - 6000)

	U.D.C.	ALPHA.	FACET	UNITERM	TOTAL
Personal Errors					
Indexing	20 (43%)	13 (38%)	12 (28%)	14 (35%)	60 (37%)
Searching	5 (11%)	4 (13%)	8 (19%)	3 (8%)	20 (13%)
Time Allowance	8 (18%)	6 (17%)	6 (14%)	10 (25%)	30 (18%)
Question	6 (13%)	5 (15%)	5 (12%)	6 (15%)	22 (14%)
All Other Reasons	7 (15%)	6 (17%)	11 (27%)	7 (17%)	30 (18%)

TABLE 9.3

BREAKDOWN OF REASONS FOR FAILURE ON DOCUMENTS INDEXED IN SECOND SUB-PROGRAMME (DOCUMENTS 6001 - 12000)

In comparing the tables with Table 5.5, the only significant change in percentage occurs with Uniterm, where from 52% of the personal errors for indexing in Table 9.2, there is a change to 34% in Table 5.5. This, in so far as anything is likely to do in the present investigation, appears to substantiate the view that it was not any inherent difficulties which made Uniterm originally difficult to operate that were the cause of the increase in efficiency, so much as the chance of human error.

Considering the searching, it will be seen from Table 3.7 that the three project staff concerned with making the searches in the first two rounds of testing did not have any significant variation in their efficiency. A check was also made on the results for Warburton and Hadlow when searching for documents which they had

themselves indexed, but this revealed no difference from their general figures. Some more interesting figures are given in Tables 3.8, 3.9, 3.10 and 3.11, as these deal with the searches made by technical staff. These repeat the results obtained by the project staff with the exception that U.D.C. had a superiority of 5.7% over Alphabetical as compared to an inferiority of 5.9% with the searching by project staff. The figure for U.D.C. was also the only system where technical staff returned a higher figure than the project staff, and the result would, to say the least, appear to cast considerable doubt on the oft repeated argument that the notation of U.D.C. is too cumbersome to be mastered by technical staff.

A problem that is attracting much interest at present is the formulation of search programmes. This is particularly the case in organisations which are attempting to use some form of mechanised retrieval, for an incorrect search programme will result in severe time and cost penalties due to the comparative inflexibility of computer searching. Allied to this, a criticism of the project programme as originally proposed was that it would not be dealing with the physical act of retrieval from the store. The reason for our not attempting to cover this point was that the investigation was concerned with what was thought to be the intellectual aspects of information retrieval rather than the clerical. The formulation of the search programme, the decision as to which concepts or which combination of concepts to search first, is the intellectual part of the work, whereas the physical act of locating these items which have been marked with the appropriate tags is a clerical routine. In spite of the fact that there was no controlled test on this latter point, the experience of the project staff may be of some interest. It is clear that we were not prepared to spend the length of time in physical searching which some organisations appear willing to do. This came up in particular with the original Facet catalogue, as can be seen from the comments in the analysis of failures, where several times the conclusion is reached that the searcher considered that the number of different places to be searched in the course of a single programme was more than could be tolerated. It was this point, more than anything else, which caused the comparatively low retrieval figure for Facet, and led us to the decision that fixed order and chain indexing were not suitable for the type of indexing done in the project.

The most reliable figures which can be produced are those given with the W.R.U. test (Table 7.3), where the average search time is shown to be 7.3 minutes. This compares with the general opinion of staff searchers that they would

expect to make from seven to ten searches an hour in the U.D.C. and Alphabetical card catalogue. Concerning Uniterm, we knew that posting document numbers on cards was a relatively archaic method that was inferior to other techniques such as peek-a-boo cards, and the only comment on this point is that our experience proved what we already knew.

The decision as to what is a reasonable search time is a matter for each organisation to consider in the light of its own circumstances, but it has been a matter of surprise to find the time delay which many organisations appear willing to tolerate for the doubtful benefit of using some form of mechanical retrieval. In the field of applied science and technology, the position in England seems to be that most librarians expect to be able to supply some references within five to ten minutes of receiving the enquiry. No-one would suggest that in such a time they would be able to obtain every relevant reference. However, experience again indicates that frequently the requester is satisfied with some documents relevant to his enquiry, and does not, in the first place, require a complete collection of all relevant documents. Naturally, there are occasions when this is required, but an I.R. system which cannot meet the necessity for supplying some relevant documents within a few minutes would fall short of the ideal for many organisations.

This being the case, the formulation of such programmes must be a reasonably straightforward matter, and this was the position with the project searches. The majority of source documents were retrieved by the first or second search programme and it can be seen from Tables 5.4 and 5.5 that searching was only responsible for 17% of the failures, and of this 15% was adjudged to be due to human errors. The analysis of failures which has been made in Chapter 5shows most decisively that the failures were, for more than all other reasons together, due to mistakes by the indexers or the searchers, and that a third of the failures could have been avoided if the project staff had indexed consistently as well as they were capable of doing. Put another way, this means that in every hundred documents, the indexers failed to index adequately five documents, the failure usually consisting of the omission of some particular concept. They might legitimately plead that the conditions under which they worked for two years, such as the monotony of the work or the continuous re-adjustments of indexing times, were not conducive to that concentration which is required to obtain perfection. In fact, they have no reason to plead, for the analysis of the supplementary indexing and of other systems has shown tnat their performance was above average.

As has been emphasised in Chapter 6, the emphasis in the main test was on recall, and other tests had to be developed in order to put the figures in perspective. To do this with any real precision turned out to be impractical, but these further tests, together with the tests on other systems, have shown that the general working level of I.R. systems appears to be in the general area of 60% - 90% recall and 10% - 25% of relevance, the shaded area of Table 9.4. This is a considerable distance away from the oft-made assertion that systems are operating in the general area of the top right-hand corner.

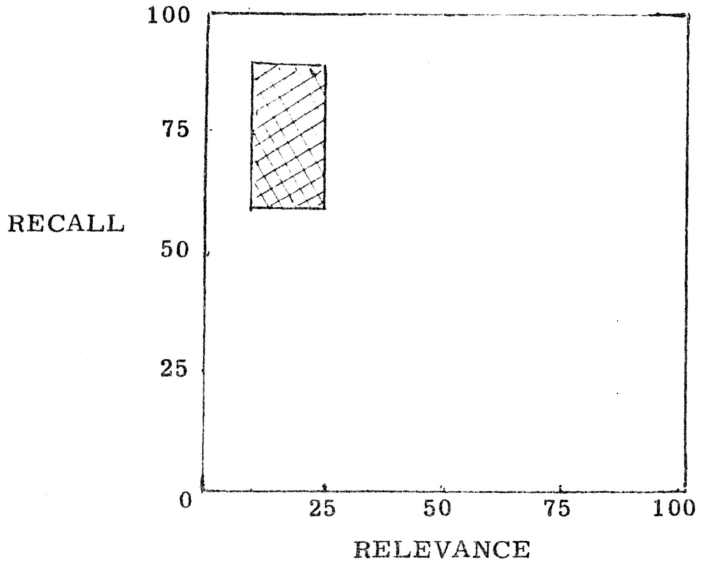

TABLE 9.4

PROBABLE OPERATING AREA OF I.R. SYSTEMS

Further, it can now be said that the inverse relationship between recall and relevance has been conclusively shown, and it should be now possible to design and operate systems that will satisfy, in the most economic way, stated requirements. There will be situations where the emphasis must be on the highest possible recall level, and the resulting penalty of the low relevance figure will be accepted. In other cases recall is less important, and greater emphasis will be placed on improved relevance. It must be stressed that the relevance figures depend entirely on the references which have been retrieved by the programmes used in the search, and no attempt has been made to weed out the irrelevant documents. It must not, therefore, be imagined that there is necessarily a severe time penalty by the librarian or the searcher having to look through a number of useless documents. It would be dangerous to be too definite on this point, but it would not be unreasonable to suggest that, in the conditions prevailing in the systems tested, a rapid visual check of the titles and in some cases of the abstract would enable the large majority of the useless references to be eliminated straight away. To express this in figures, one might say that of every hundred documents retrieved in the W.R.U. test, fifty could be quickly and safely eliminated as being of no possible interest.

The problems arising in indexing by the four systems have been dealt with in detail in Chapter 4 of Ref. 1., and the following comments should be regarded as supplementary to the earlier statements.

Universal Decimal Classification

In that most of the professional staff searchers were reasonably familiar with this system, it is difficult to record fresh opinions. Most striking, perhaps, was the great value and importance of the alphabetical index for the schedules, which was of a higher standard than any of us had previously used. It completely justified the considerable amount of effort which was entailed in its compilation, and was clearly responsible for the high level of success achieved by the technical searchers.

It soon became obvious that the indexing was more specific than the searching could possibly be. More than half the entries in the catalogue consisted of notations having three or four coloned elements, e.g. 533.6.013.412: 533.6.057: 533.6.011.35, but it was only in approximately one search out of ten that the searcher felt justified in using more than two elements.

The main advantage in the classified arrangement appeared to be in being able to combine in alternate searches specific and broad concepts. The question, for instance, might be "Buckling stresses of circular fuselages". The specific heading would be 629.13.012.213.1:531.224.4, which represents a literal translation of the two concepts in the question. If this failed, each concept could be broadened in turn, searching first under 629.13.012.213.1:531.2+ (Stresses plus any subdivision of 531.2) and then 531.224.4:629.13.012.2 (Fuselages) or even 531.224.4:629.13.012+ (Aircraft structures plus any subdivision of 629.13.012). Such searches frequently resulted in success without paying too high a penalty in retrieval of irrelevant information. It was not always possible to do this type of search in Alphabetical, due to the restriction on combined headings that is discussed later, and even when practical, it involved far more search effort. It would have been possible to do it with Facet, but the effort would have been prodigous. With Uniterm it presented few difficulties.

Alphabetical

This system was far more effective than had been expected, and met with the general approval of the technical searchers. The main frustration came in searches where it was desired to combine two main headings, which is another way of saying that the headings were not sufficiently specific. This point received critical comment from the technical searchers, who also felt that there was inconsistency in certain headings, in particular the terms that could be associated with 'Boundary Layer'. This was because of the basic rules which we had compiled to cover the types of terms that could be used as main headings or sub-headings (see Appendix 8A), and can be considered more a defect of the authority list than of the method of indexing.

As explained in Ref. 1, no "see also" references were used in the main testing, and it was surprising to find that such a small percentage of failures were considered to be due to this. The technical staff had, as might be expected, more difficulty in this respect, and found that they tended to "go all round the subject" before alighting on the heading used by the indexer.

Facet

Most of the comments concerning Facet have been made in the earlier chapters. Its weakness undoubtedly lay in the insistence on the use of a fixed order combined with chain indexing, and from our experience it appears unlikely that this technique

could have been successfully applied in this investigation. There are indications throughout the test results that the more elements making up the complete notation, the greater the difficulty in locating the source document, but even with less specific indexing, the use of the chain index would have resulted in difficult and tedious searches.

As a classification covering the subject field of the investigation, the schedules were considered to be very satisfactory by the indexers, a view which was not changed by the sear ching. This was particularly shown in the subsequent test (see page 59), when the fixed order was abandoned and entries made by logical combinations of the elements.

The notation of the Facet scheme, consisting of upper and lower case letters, was most unsatisfactory, for there were frequent cases of confusion caused by insufficient difference between letters. This could have been largely avoided if certain letters and combinations of letters had been omitted. As it was, an eight letter notation was more awkward to handle than a numerical notation twice its length.

Uniterm

Uniterm, as a descriptor language, can be given a high rating on many counts. It achieved the best overall figures in the test, it presented no serious difficulties for the technical searchers, it was the highest scoring system in the supplementary indexing, both for indexers in the United States and in England, and was notably successful with short indexing times. It appears to have as good a relevance figure as any other system, and Table 6.6 (page 58) indicates that it did not compare unfavourably in the recall of non-source relevant documents.

It was deliberately operated in the simplest possible form, and, as discussed in Ref. 1, p 74, it took, compared to the other systems, very little time to prepare.

The project originally began as an investigation into the comparative efficiency of four indexing systems, and it might appear to be shirking the issue if no attempt was made to answer the question that is so often put to us, "What system is recommended ?" This is impossible to answer without qualification, for no system which has been investigated has shown itself to be so markedly superior as to justify its use in all conditions. The size of the collection, the number of users, obviously the subject matter; these are the type of considerations which would influence a decision.

To take a personal case first, an academic organisation where all the students are working on research theses. By discarding older references, the collection remainds fairly constant at 40,000 - 50,000 documents. At certain times of the day there may be a dozen or more people wishing to consult the catalogues. The majority of the search questions will be for background material but many requests are for specific information. A card catalogue is, in these circumstances, the most practical store. Indexing is at a level of an average of five entries per document, with an alphabetical arrangement of the cards in the catalogue. The general use of inverted headings in the alphabetical subject list gives a measure of classification, to a level which appears to be most useful to the searchers. Combined with the list of headings is given a set of classified schedules to enable the searchers to find the appropriate and related headings.

As a different environment, we consider an information service in a research organisation where the staff of the library would normally carry out the physical search. In this case a post-co-ordinate system would appear most effective. Possibly the quickest and cheapest store would be peek-a-boo cards, which are certainly effective for a collection with an annual intake of up to 20,000 documents. The alphabetical subject index to the facet schedules would be perfectly satisfactory for the list of terms; the facet schedules would enable the terms to be used more effectively in indexing and searching.

If, however, the catalogue was to serve a number of different establishments within a large organisation and covered many different subject fields, then there would appear to be reasonable grounds for deciding to use a standard system such as the Universal Decimal Classification. On the other hand, if the subject field is reasonably concentrated, then facet analysis is (to quote from Mr. Sharp's comments in Ref. 1) "probably the most powerful tool ever to be introduced into the science of classification and it undoubtedly provides a most rigorous method for the proper marshalling of terms in a given field". The facet schedules used in the project and in connection with the W.R.U. test were completely effective when the original principle of fixed order was abandoned.

However, our own predilection (a word which we use in its precise meaning, for we can produce little data to substantiate it) would be that, where a large computer and high speed print out was available, printed indexes should be made available to as many individuals as required them. To do this effectively,

simply and relatively cheaply, each document would be given a serial number and a list of these, either titles and references or with the addition of abstracts, would form the basic permanent store, to be supplemented with additional entries as required. The documents would be indexed by a special facet classification designed for the purposes of the organisation, with the separate elements being combined as required, and an average of eight such combinations being allocated for each document. A master classified card catalogue could be maintained if desired, but the entries would be entered on tape, so that print-out would give a copy of the classified card catalogue, except that it would not be necessary to repeat the notation for all the separate entries that might have that notation. To take a sample from the facet catalogue in the W.R.U. Test, it would read:

Nu Jg Dk	11799
Nu Ldr	11247, 11396
Nu Mahf Nbi Ok	11209
Nu Nbe Mahp	11514, 11632, 11895
Nu Nbe Vls	11516, 11732
etc.	

In a small collection, as was used in the W.R.U. Test, the number of different notations will be large in relation to the number of documents, but will decrease as the collection grows. To up-date such a file would be relatively simple, and it would also be possible to subdivide the printed classified index so that individuals received only those sections which covered their main interests. It is estimated that for a collection of 100,000 documents, the classified index would contain about 400 pages, if printed in a double-column layout such as Chemical Titles. With KWIC indexing, a similar collection would require an index five times the size, and would not be amenable to subdivision.

We await with keen interest news of any such application.