CHAPTER 5

SUPPLEMENTARY PROCEDURES

The project is being deliberately restricted to those matters which are concerned with the intellectual problems of indexing documents and formulating search programmes. The question of which particular physical form the completed index should take is a completely separate matter that is largely irrelevant to the investigations. Whereas certain systems such as the alphabetical subject catalogue or the U.D.C. are conventionally used with catalogue cards, a system such as the Uniterm can take several different forms. Originally it was proposed that lists of document numbers should be written on cards to be compared; many organisations use some form of peek-a-boo cards; others make use of punched card equipment and some are reported to use computers. Irrespective of which method is used, if the indexing and the search programme are constant, so will the result be the same.

However, the opportunity was taken in the course of the investigation to record some results of the work involved in the compilation of the necessary indexes. The form which these should take and the methods used were dictated by practical and economic consideration of the equipment and personnel that were available and no way is it suggested that they would be ideal solutions in all circumstances.

With the alphabetical subject catalogue and the U.D.C., conventional 5" x 3" catalogue cards were used. The same was the case with the classified catalogue and the chain index for the facet system. With Uniterm it was decided to prepare lists of document numbers for visual comparison.

Preparation of catalogue cards

For each document that was indexed we had a known requirement for a minimum of six cards up to an unknown maximum, with the average requirement being estimated at twelve. The minimum of six was made up of at least one card for each of the three systems, (i.e. U.D.C., Alphabetical and Facet), plus one card for each of the three supplementary indexes that were being maintained, these being an index in numerical order of the project document references, an author index and a source reference index.

We considered a number of methods of producing these cards, and rejected many either on account of cost, lack of permancy of the master, or poor quality of the product. At one time we did discuss making use of the services of the equipment used for producing catalogue cards for the British National Bibliography, and the Director, Mr. A.J.Wells, was very helpful in working out a possible satisfactory method. A major difficulty would have been in the transportation of masters and cards between London and Cranfield, and while discussions were still in progress, certain equipment was installed in the Business Systems Laboratory of the College which made possible a most satisfactory method.

The equipment was the Graphotype Embossing Machine manufactured by Addressograph-Multigraph Ltd. The Laboratory already contained an Addressograph printer, so we had available equipment which gave us all the flexibility, combined with economic working, which we required. The Graphotype is an electrically driven keyboard operated model, which embosses characters on metal plates. Since we could use this equipment without charge, the only material costs involved were in the purchase of the metal plates and the holders into which the plates have to be put for

printing. (Fig. 2). 18,000 metal plates were purchased at a cost of 0.9 pence a plate. The holders cost $2\frac{1}{2}$ pence each and we found that 300 holders met our requirements. As each holder was used on an average of 60 times, our basic material costs were just about one penny for each document.

The plates which we used had a capacity of nine lines of type, with a maximum of 40 characters in each line. The layout of the cards is shown in Fig. 3 and consists of author and project document number on the top line, the title of the paper or article and the source reference. Typing on the Graphotype is slower than with a conventional typewriter, due to the time lapse necessary between each letter to allow for the embossing of the metal plate, but it was possible to maintain an average of 35 plates an hour for masters of the kind shown in Fig. 3.

The typing of these plates was done ahead of actual requirements, and the plates were stored in order until required. Documents were indexed in batches of 100, and as each batch was completed, the appropriate plates were put in the holders and placed in the Addressograph machine. We knew from the master indexing card the number of catalogue cards that would be required for the various indexes, and this number of cards was run off before passing on to the next plate.

The cards had to be placed in position by hand, and improved ability to do this consistently and speedily came with experience. We found that it was possible to print an average of 700 cards an hour on this machine.

REQUIRED NO. OF CARDS	3	5	10	25	50
Cost of plate and holder	1d	1d	1d	1d	1d
Labour charges (at 5/- an hour):					
Typing master plate	1.7d	1.7d	1.7d	1.7d	1.7d
Fixing plate in holder and removing	.2d	.2d	.2d	.2d	.2d
Printing cards	.25d	.42d	.85d	2.1d	4.2d
Total Costs	3.15d	3.32d	3.75d	5.0d	7.1d
Cost per card	1.05d	.66d	.37d	.2d	.14d
Labour charges (at 10/- an hour):					
Cost per card	1.77d	1.13d	.65d	.36d	.26d
Labour charges(at 15/- an hour):					
Cost per card	2.48d	1.59d	.92d	.52d	.39d

TABLE 1 COSTS FOR PREPARING CATALOGUE CARDS BY GRAPHOTYPE AND ADDRESSOGRAPH MACHINES

As has been said earlier, this method of producing cards was made economic because we had the relatively expensive equipment available without cost. The resulting printed card was perfectly satisfactory in appearance and the only obvious limitation of this method is the amount of information which can be put on the plate. Working from the figures that we achieved over a long period, the cost of this method of producing catalogue cards in varying numbers is given in Table 1. It should be emphasised that this is only the cost of printing, and that the cost of the cards is an additional charge.

Entering headings on catalogue cards

For reasons connected with the availability of clerical labour, it was decided that the headings for the alphabetical subject catalogue and the notation for the U.D.C. should be entered by hand. The time taken to do this by any particular individual is a compound of three points :-

a. Length of symbol to be written
b. Readability of symbol to be transcribed
c. Familiarity with terminology or notation

A frequent criticism of the Universal Decimal Classification is the length of the notation required. In practice we found that the number of separate letters in an alphabetical subject heading was larger than that of a combination of numbers in U.D.C., e.g. "WINGS, SWEPTBACK, SUPERSONIC - Stability, longitudinal, Wind tunnel tests" contains 60 letter characters and five punctuation marks. The corresponding entry in U.D.C. would read :-

533.693.1:533.6.013.412:533.6.011.5:533.6.071

and contains 32 figures and 13 punctuation marks, or a total of less

than 70% of alphabetical. Any such arguments are, however, rather inconclusive for we substituted certain letters in place of commonly used groups of numbers, e.g. A = 533.6, B = 629.13, etc. The notation above would have been written as A93.1:A.013.412:A.011.5: A.071, or 20 figures and 9 punctuation marks. Equally so, the alphabetical heading could have been reduced to :-

W., SWEPT., SUPER. - Stab., long. W.t.t.

and still retain sufficient meaning for sorting purposes.

In transcribing the symbols from the master indexing card on to the catalogue cards, greater difficulty was experienced by the clerical staff in transcribing U.D.C. numbers, which meant nothing to them, than to the words used in the alphabetical subject headings. However, for anyone having considerable experience of the notation of the U.D.C., there was no more difficulty in transcribing U.D.C. numbers than alphabetical headings, and the only difference in the time factor was accounted for by the number of characters.

The notation that was used in the facet classification was a combination of upper and lower case letters, and without doubt it caused us more difficulty than either of the other systems. It would seem that a random grouping of letters is more awkward than a random grouping of numbers, for the mind is presumably accustomed to the latter, while with letters one tends to expect some pronouncable form. The difficulties which we experienced could have been lessened if care had been taken, when compiling the notation, to avoid letters which, particularly when hand-written, were easily confused, e.g. 'u' and 'v', 'q' and 'g' etc. From the viewpoint of the length of the complete entry, the facet notation was an improvement on the others.

The example previously given would read, in facet, Cd(Iibb)Nbk Ocb Vn, and therefore this compensated for the other disadvantages in entering the headings for the classified index. It was felt in this case, however, that it was essential that they should be typed and checked by the indexers, so as to minimise errors in filing due to failure to transcribe correctly. To ease the difficulty of reading a complete notation we found it necessary to interpose a space between the end of one element and the start of the next.

While the compilation of the alphabetical index to the U.D.C. and the list of headings for the Alphabetical Subject Catalogue is an important task, yet they do not compare with the clerical effect involved in preparing the chain index. As has been stated earlier, the chain index involved typing a separate card for each entry in the chain and the example given a few lines above would require the following cards :-

Wind tunnel tests: longitudinal stability: supersonic flow: sweptback: wings

Longitudinal stability: supersonic flow: sweptback: wings

Supersonic flow: sweptback: wings

Sweptback: wings

Wings

Whereas once any such card was in the index it was unnecessary to add a further card, yet there is a time loss in checking whether such a card is already there. The chances are that cards already in the index will be those with only a few elements to the notation (see Chapter 6 and Table 7) and therefore the cards involving most typing work all have to be typed. In spite of the time cost of checking previous typing, it is desirable that this should be done since inclusion of unnecessary cards not only wastes typing time but also results in

extra sorting and filing time. The method we used was that the typist made a personal list of each single and two element notation that occurred at the commencement of each notation, e.g. where the notation was

Ca Cd Ea Nbk Ocb Vn

the typist entered Ca and Ca Cd, in a sequential list and if the next complete notation was

Ca Cn Ea Nbk Ocb Vn

she would enter Ca Cn, but would know that it was unnecessary to type a card for Ca. The element combinations were so varied that to have included all these in the list would have made it so long that the effect would have been to make the time-loss greater than the saving. As a compromise, we also entered in the list any combination of three or more elements that occurred more than once, these occasions being revealed when the cards were sorted into the chain index. Certain combinations of elements, e.g. Ca Cd Cn Ea, occurred frequently, and this method is believed to have been of value in cutting out unnecessary work.

Filing cards in the catalogues

With approximately 200,000 cards to file it was desirable to make certain that we were using the most efficient method of sorting and filing cards. This is, of course, largely a matter of common sense but it also appeared possible that we might be able to do some useful investigations on the comparative "filability" of the three systems.

When taking any pack of cards that have to be sorted into a certain order, the first requirement is to divide the cards into smaller packs and maybe subdivide these, until all the packs are small enough to be

rapidly hand-sorted into their final order. The size of this final pack will largely depend on the ability of the individual to hand-sort a batch of cards, and this may well vary from five to twenty or more.

A work study investigation was done to determine the most efficient method of reducing the complete pack of cards down to this final batch. The basic problem resolves itself into striking the best balance between (a) making the fewest possible sorts, since the cards have to be re-handled for each sort, and (b) the time which is taken in placing each card on the correct pack. The fewer the packs, the quicker can be the mental decision as to the correct pack for each particular card; the greater the number of packs, the longer will it take to make such a decision.

Another factor to be considered is the size of the original pack in relation to the maximum number of cards in each of the final packs. Presuming there are 1,000 cards to be sorted and that the objective is final batches of ten cards for hand-sorting, there are various possible approaches. The cards can be put into a comparatively small number of groups, such that there is the certainty that each group will have to be resorted, but which will permit rapid placing of the cards. Alternatively an attempt can be made to sort straight away into 100 groups in the hope that most of the resulting batches of cards will be small enough to hand-sort without further subdivision.

We found that there was a definite limit as to the number of groups that were desirable for the first sort and that this number depended to some extent on the system, but also on the individual doing the sorting.

With alphabetical, or with the facet chain index it was possible to sort into twenty groups, based on initial letters, with a few obvious doubles, such as I and J, K and L, N and O, P and Q, U and V, and X, Y and Z. With the U.D.C. cards, a maximum of 15 groupings was as much as could be managed economically by an experienced sorter. To obtain reasonably level distribution between the resulting batches, the limits of the U.D.C. numbers were as follows :-

0 to 52	621-
53 to 532	621.1 to 621.4
533 to 533.6.011	621.5 to 628
533.6.013	629
533.6.015 to 533.68	63 to 65
533.69	66
534 to 539	67 to 99
54 to 620	

Presuming that the original pack was 1,000 cards, the alphabetical sort would produce twenty sub-packs which would range in size between 20 and 100 cards. This meant that with one further sort of each sub-pack, the stage would be reached where final hand-sorting could be done. Assuming an average time of one second was taken to place a card on the correct pack, and an average of eight seconds to pick up each sub-pack or prepare it for resorting, by this method the time taken to break down the original pack of 1,000 cards into batches small enough to hand-sort would be as follows :-

First sort into 20 packs	1,000 seconds
Picking up 20 packs	160 seconds
Second sort into final packs	1,000 seconds
Total	2,160 seconds or 36 minutes

When we attempted to combine the two sorts into a single sort by having

approximately 100 sub-packs it was found that the complications were such that it would take an average of three seconds to place each card on the correct pack. Therefore the total time for this one sort was 50 minutes, and we still had a number of packs which were too large for hand-sorting.

If, however, the original size of the pack was only 500 cards, there was some justification for altering the strategy. Taking two sorts, the time would be 500 seconds plus 160 seconds plus 500 seconds, a total of $19\frac{1}{3}$ minutes. By making fifteen extra packs for some headings that were known to occur frequently (e.g. Aerofoils, Aeroplanes, Bodies, Flow, etc.), the result was that there were 35 packs of which 25 might be small enough for hand-sorting. The time for placing each card in the first sort went up from 1 second to 1.3 seconds, so with this method the timing would be :-

Original sort into 35 packs	650 seconds
Picking up 10 packs	80 seconds
Second sort of 310 cards into final pack	310 seconds
Total	1,040 seconds or $17\frac{1}{3}$ minutes

This shows a small saving of 2 minutes over the other method.

If the original pack was several times larger than given in the first example, a combination of these strategies might be most effective. With five thousand cards to be sorted, the sub-packs might be expected to range in size from 100 to 500 cards. In some cases one further sub-sort would be sufficient, but often two further sub-sorts would be required.

The situation with the project was that there was no particular urgency

for cards to be sorted into the catalogues, since the catalogues were not being actively used until the completion of the indexing. As a result, we could allow the cards to collect until a large number were available, but in normal library practice, it would be more desirable for new cards to be sorted into the catalogues as soon as possible.

It is obvious that the larger the number of cards to be sorted into the catalogue in a single sequence, the less time that will be taken in the second stages of the total operation of sorting and filing cards. On the other hand, the larger the pack of cards to be sorted the longer the pro-rata time taken to sort them.

One pack can be interfiled with another pack most efficiently when the packs are equal in size. This condition would obviously only prevail in the very early stages of a new catalogue, and our tests showed that the time to sort a given number of cards into a catalogue increases regularly until the ratio is reached, of 1 to 20 in regard to cards to be filed as against cards in the catalogue. After this stage down to a ratio of 1 to 100, there is no significant increase in the time. That is to say, if a catalogue already contains twenty-thousand cards, unless the pack to be filed contains more than 1,000 cards, there will be little significant difference in filing time per card if the size of the pack to be filed is 1,000 cards, 500 cards or 200 cards. Such increase as there is, is not due to the time taken in locating the correct position and actual insertion of the card but is due to the time taken in opening and shutting the drawers of the catalogue, particularly when a card retaining rod has to be taken out and put back. The time to do this adds significantly to the time for filing each card if less than ten cards are to be filed in one drawer

The example given above was generalised by making certain assumptions concerning time and ability to hand-sort. Actual figures are shown for three members of the staff who did a number of timed tests in sorting and filing catalogue cards. Figures for this are given in Table 2, while Table 3 gives a detailed breakdown of the sorting of a pack of 2,031 cards for the facet classified catalogue.

In all cases the alphabetical subject catalogue give the quickest filing times, and if the figure for this is taken as unity, the percentage times for the other systems are as follows :-

	Alpha.	U.D.C.	Facet Chain Index	Facet Classified
Sorter A	1	1.3	1.2	1.1
B	1	1.1	1.1	1.1
C	1	1.5	1.3	1.3

Against this there is the fact that we did not file more specifically than the notation or subject heading demanded, so that with the alphabetical subject catalogue, if the heading was one such as "HEATING, AERODYNAMIC", for which there are over fifty cards in the catalogue, a new card could be placed anywhere in this group of fifty cards. For the chain index, however, each card had to be placed in an exact position, and with the U.D.C. and Facet Classified catalogues the placing would have to be more exact than with Alphabetical. The result of these qualifications to the figures given is that there appears to be little significant difference in filing by any of the systems.

	First Example			Second Example		
	Sorting (500 cards in pack)	Filing (Ratio 1 to 100)	Total	Sorting (5000 cards in pack)	Filing (Ratio 1 to 10)	Total
ALPHABETICAL						
Indexer A	2.8 sec.	8.9	11.7	5.5	4.5	10.0
Indexer B	3.3 sec.	8.7	12.0	5.3	5.7	11.0
Indexer C	3.6 sec.	11.1	14.7	6.2	7.7	13.9
U.D.C.						
Indexer A	4.5 sec.	9.5	14.0	6.5	7.0	13.5
Indexer B	5.6 sec.	7.5	13.1	7.0	5.4	12.4
Indexer C	7.8 sec.	12.3	20.1	12.0	9.5	21.5
CHAIN INDEX						
Indexer A	4.3 sec.	8.1	12.4	6.1	5.8	11.9
Indexer B	4.5 sec.	8.6	13.1	6.2	6.0	12.2
Indexer C	6.2 sec.	12.2	18.4	8.0	9.6	17.6
FACET CLASSIFIED*						
Indexer A	4.8 sec.	6.8	11.6	5.9	5.3	11.2
Indexer B	4.5 sec.	7.6	12.1	5.1	6.5	11.6
Indexer C	7.2 sec.	10.2	17.4	8.4	9	17.4

TABLE 2 EXAMPLES OF SORTING AND FILING CARDS IN CATALOGUES

In the filing column, the ratio denotes the number of cards to be filed against the number already in the catalogue.

*With the facet classified, the size of the packs was 200 and 2,000.

FIRST SORT into 17 groups. Time 34 minutes 45 seconds

	No. of cards	2nd Sort Time	2nd Sort No. of cards	3rd Sort Time	Final Sort Time	Total time in minutes
A	18				1.05	1.05
B	254	7.00	8		0.15	
			34	0.50	1.40	
			46	1.15	1.45	
			5		0.15	
			44	1.10	1.45	
			5		0.08	
			17		1.00	
			2		0.03	
			2		0.03	
			9		0.20	
			32	1.00	1.05	
			39	1.10	0.50	
			4		0.08	
			7		0.15	22.07
C	318	8.00	40	1.00	1.15	
			41	1.30	2.30	
			59	1.45	3.20	
			71	1.20	4.05	
			79	1.20	3.40	
			42	1.00	3.00	
			16		0.50	
			8		0.25	
			4		0.06	
			2		0.03	
			16		0.40	35.49
D	33	0.35			1.30	2.05
E	106	2.10	15		0.41	
			8		0.15	
			4		0.06	
			14		0.36	
			10		0.20	
			9		0.18	
			20		1.10	
			9		0.20	
			12		0.28	
			15		0.42	6.56

TABLE 3 DETAILED TIMING FOR SORTING OF CARDS FOR FACET CLASSIFIED CATALOGUE.
(Further breakdowns for G and P not shown)

TABLE 3 (Continued)

	No. of cards	2nd Sort Time	2nd Sort No. of cards	3rd Sort Time	Final Sort Time	Total time in minutes
F	321	7.20	5		0.08	
			81	2.00	5.10	
			54	1.00	2.20	
			4		0.05	
			31	0.40	1.15	
			15		0.40	
			4		0.05	
			8		0.15	
			20		1.20	
			20		1.00	
			41	1.10	2.15	
			14		0.26	
			24		0.58	28.02
G	222	5.05		1.20	12.09	18.34
H	135	3.00			8.44	11.44
I-L	33	0.35			1.15	1.50
M	60	1.20			2.30	3.50
N	138	3.50			9.05	12.55
O	22				0.55	0.55
P	149	4.10		1.05	7.37	12.52
Q-R	37	0.45			1.25	2.10
S-T	35	0.45			1.30	2.15
U	24				1.15	1.15
V-Z	66	1.20			2.20	3.40

Total of all sorts 202 minutes 49 secs.

Average per card 5.9 seconds

TABLE 3 DETAILED TIMING FOR SORTING OF CARDS FOR FACET CLASSIFIED CATALOGUE.
(Further breakdowns for G and P not shown)

Posting of Uniterms

While it was accepted that other methods might be more attractive in practice, it was decided for reasons relevant to local conditions that we would write uniterms on aspect cards for visual comparison. The time taken for posting uniterm numbers on cards has been the subject of earlier critical comment but we hoped that we should manage without too great difficulty. Inevitably the time taken increases as the size of the index grows and after a few thousand documents had been posted this way, we found that there was a growing tendency for clerical errors to be made, and that the time being taken merited an investigation into an alternative method.

We therefore decided to use a method originally proposed by Dr. Sanford at the National Security Agency, who commented that "the bottleneck caused by posting threatened the collapse of our entire system" (Ref. 18). This involved the punching of cards with the document number and a uniterm code number, the sorting of the cards into uniterm and document number and then transferring these to aspect cards. Having the necessary equipment available in the Business Systems Laboratory, this method was adapted for the final 12,000 documents.

We first had to punch approximately 100,000 cards, in that there was an average of $8\frac{1}{2}$ uniterms for each document. We had a four-figure code number for each uniterm, so with the five figure document number, nine holes had to be punched in each card. The document number was gang-punched, and therefore these five holes had only to be punched once for each document. As a result we had to punch a total of 12,000 x 5 + 100,000 x 4 = 460,000 holes. This could be done at an average speed of 4,000 punches per hour, making a total time for this stage of 115 hours work. We next put the cards through an interpreter which printed the

numbers on each card. This worked at a speed of 4,000 cards an hour giving a further 25 hours work. It was not essential to do this but it had certain advantages for other reasons than the immediate objective. The sorter was used for putting the 100,000 cards into uniterm and document order and worked at an average speed, including putting in and removing from the machine, of 25,000 cards an hour. As it was necessary to make nine sorts of the 100,000 cards, this involved a further 36 hours work. The final stage of printing the cards on to rolls of paper took a further 30 hours, making a total time for the whole operation of 216 hours.

In the work described by Dr. Sanford, the punched cards were used as masters to enter the document number on the aspect cards. In our case we were not in the position of requiring this to be done continuously throughout the two years of the project, and therefore all this work, apart from the initial punching of the cards, was done in one operation at the completion of the indexing. As a result we were able to use the printed lists, cutting them up and pasting them straight on to the aspect cards.

We did, however, investigate whether this method might be reckoned to show any saving in time over the random posting of numbers in a more conventional situation. Assuming that 200 documents with an average of $8\frac{1}{2}$ uniterms per document have to be posted, then the time taken to prepare the lists of these documents would be (from our figures given above) $\frac{216}{12,000} \times 200$ hours = $3\frac{1}{2}$ hours. To enter the 1,700 numbers on to the appropriate aspect cards, took 6 hours, making a total of $9\frac{1}{2}$ hours. To enter the number of uniterms without presorting would have taken us 15 hours.

As a postscript, it might be stated that the machine times given are representative of what we were able to obtain when the various pieces of equipment were working satisfactorily. Unfortunately for us the equipment which we used was old and had only been used intermittently for demonstration purposes during the last few years, and we had numerous breakdowns, with the result that the time spent on the operation was probably well over 500 hours. However, we have no reason to think other than that with properly maintained equipment, the figures which are given would be quite practical.

CHAPTER 6

STATISTICAL DETAILS

In Tables 4 - 9 are given various statistical details of the indexing. Some of these may have little value at the present stage, but will become significant in relation to the test results.

The first set of tables gives detailed figures of the postings for each group of 100 documents during the indexing of the final 6,000 documents. As was to be expected, there was a regular falling off in the number of postings with indexing times; in addition we have, as was hoped, a variation between the indexers. While it would obviously be incorrect to suggest that there is a correlation between the standard of indexing and the number of entries required for each document, yet we hope that in the testing we shall be able to ascertain whether over-indexing has an equally bad effect as under-indexing.

Table 5 compares the indexing done during the first two sub-