PART C : DISCUSSION

A draft of this Report was presented to the Study Project Advisory Panel, and a number of specific points were raised. These fall into two groups: those concerned with the 'ideal' collection design, and those concerned with costs. These are considered in Comments on the Study below. The Panel was also presented with a report on an experiment in operational system evaluation conducted by Professor Cleverdon,supporting an alternative approach to retrieval test material collection (3); and with detailed information about possible uses which is summarised in Part B of this Report. Professor Cleverdon's approach and ours are considered in relation to potential research projects, and to information retrieval in general, in the final section of this Report, Retrieval experiments.

1. Comments on the Study

Collection design

a) It was suggested that working with very technical documents could present problems for research projects, for example in attempting failure analysis or alternative indexing, and this point was perhaps not sufficiently emphasised in Part A. It would be a problem, for instance, with the most plausible source of material for the main document set, the physics section of INSPEC. This point was accepted, but it was also recognised that the difficulty was one which could hardly be avoided with any scientific data base.

b) The second main point raised about the design was that the number of relevance assessments per request required of users was unrealistically high, For the main set of documents the user would be asked to assess an average sample of 250-300 documents from the retrieved pool (and he would also want to assess for his own use the output for his own query from the search of the entire data base). It was strongly argued that non-user are good enough, so the design should be modified to allow for formal user assessments only of that part of the user's output falling in the pool, which would be presented to him in a suitably unbiased manner within a slightly larger sample, and that there should be separate assessment of the pool sample by an assessor. This would incidentally allow some overlap, for checking the assessments. The substitution of an assessor for the user would increase the cost: see below.

c) Some implications of the relevance assessment argument were examined. According to the argument of Part A above, supported by Appendix 7, the percentage of the pool to be assessed decreases with an increase in the number of relevant documents per request, as well as with the number of requests. Thus for 700 requests with 10 relevant documents 70% of the pool has to be assessed, while for 75 documents 10% has to be assessed. Our suggestion that the latter would imply assessing 250-300 documents from a pool of, say, 3000, is derived from our own experience with the UKCIS test material: as noted, strict Boolean searches of some 27,500 document titles retrieve an average of about 36 relevant documents, estimated at perhaps 40% of all the relevant to be found, in a total search output of 81 documents. This implies a total relevant of about 90 which we opine are probably retrieved in an exhaustive coordination type search giving an average output of 2670 documents. Our table in fact considers a lower total of 75, reflecting a slightly higher success rate for title searching. However Professor Cleverdon argues that this number of relevant is not realistic, since if there are 75 relevant documents in a set of 30,000, this implies that in the large data base from which this set is drawn, say containing 1.2M documents as in the case of INSPEC, there will be 3000 relevant documents. (This assumes that the set of 30,000 is a random sample of the data base.) This is in fact what the UKCIS figures do imply, since the document set is more or less of a random sample of the whole. But it has been maintained that such figures are characteristic of SDI profiles, which the UKCIS requests are, and that for regular retrospective requests much smaller numbers of documents are relevant, say 10 in 30,000 or 400 in 1.2M. This would apparently imply the assessment of a very high proportion of a pool of 3000 documents retrieved, but the opinion of Professor Cleverdon and the Panel is that if the number of relevant to be retrieved is small, they can almost certainly be retrieved by a much smaller pool. For example if a pool of 300 is retrieved, assessing 70% of it as required for 10 relevant documents implies inspecting 210 documents, while a pool of 400 implies assessing 280 documents. That is, the number of assessments is much the same as the original number, so the original costings would apply. It is, however, possible that slightly more care would be needed in designing alternative queries to generate the pool. It should also be emphasised that if the number of relevant documents is very small, statistically based assessment is impossible, as the column for 5 relevant documents in Appendix 7 indicates.

d) Related to points b and c are some suggestions about the generation of the pool. In Part A it is proposed that searches for alternative (usually broad) queries should be conducted at the same time as those for the user's own query. Professor Cleverdon suggests that a more useful approach would be to adopt a two-stage approach with an initial search on the user's query, so that its assessments could be exploited to design the alternatives, or indeed to reject requests likely to generate too few relevant documents. This would not appear to present any practical problems, though adequate control would be needed, especially to avoid expensive search sessions.

Collection costing

a) It was noted that printing costs for searches might be slightly higher than those given, since the user would require printing for at least some documents retrieved from the overall data base, which would not be of interest for collection assessment.

b) It was suggested by Professor Cleverdon that searcher effort in Part A is under-costed, though reliable cost estimates are extremely difficult to obtain and the question is open to argument. The original assumption was for staff to conduct searches at the rate of 3 requests per day, taking the set of queries for each request together, for the main document set, and social science other set. A less favourable assumption would be 2 searches per day: there was some resistance to the suggestion that only 1 could be performed. It is likely that the costing for the remaining other sets was also too low, though it may be noted that the formulation of alternative queries was separately treated.

Upward revision to double the specific searching costs, adding £6000 for the main set and social science set, and £2000 for the remaining others, would increase the collection version costs as follows:

	A	B	C	D
£ K	123.4	109.7	94.0	85.3
	8.0	8.0	6.0	6.0
	131.4	117.7	100.0	91.3 .

c) The proposed substitution of assessors for users as suggested above adds the cost of one set of assessments to each request; but the number of assessments might be slightly smaller than those considered originally, and assessors are deemed slightly more efficient than users. On the basis of assessments for 3 requests a day we get an extra year's work for the main set, costing £4000, an extra three months' for social science, costing £1000, and for the remaining others we allow an arbitrary sum, say again £1000. The overall increase in relevance costs for the main and social science sets is therefore £5000, with £1000 for the others. Allowing for these changes the original estimates for the different versions of the collection become:

	A	B	C	D
£ K	23.4	109.7	94.0	85.3
	6.0	6.0	5.0	5.0
	129.4	115.7	99.0	90.3 .

Taking the suggested modifications with increased costings under b and c together therefore, we have the following picture:

	A	B	C	D
	123.4	109.7	94.0	85.3
extra search cost	8.0	8.0	6.0	6.0
extra assessments	6.0	6.0	5.0	5.0
	137.4	123.7	105.0	96.3 .

However, as noted, there was some disagreement among Panel members on costing such large scale searching and assessment, as there is very little hard information available about them; and it may reasonably be argued that these figures are rather pessimistic.

d) A much more important point about both the Part A costings and these is that they are all unofficial ones based on commercial rates for the use of data base tapes and search services. It may be presumed that were the 'ideal' collection to be built as an officially sponsored enterprise of BLR&DD, special arrangements with data base and search service suppliers might be looked for which could reduce some of the costs. The Panel agreed that this was a possibility that should be borne in mind in evaluating the proposed design. For example a reduction of 15% in the rates through such arrangements would approximately wipe out the cost increases just discussed, and if applied to the original costings of the collection versions would reduce them as follows:

	A	B	C	D
	123.4	109.7	94.0	85.3
minus 15%	18.5	16.4	14.1	12.7
	104.9	93.3	79.9	72.6 .

An economy collection

The relatively favourable figures just given are still very high. An obvious response to them is to consider a 'cut-price' collection, i.e. one which though below the specification would not thow the baby out with the bathwater.

Our view is that the essential component of the 'ideal' specification is the large document set with its large request set and well-founded relevance assessments. The requests and assessments in particular would provide much better test data for a range of experiments , and well- organised comparisons, than is currently available.

If we considered providing version D, the least ambitious of those examined without any indexing enrichment, this would only reduce the cost slightly. A more economically effective approach would be to work with only one main set of documents with its request and assessment set. Using the original costing scheme this gives us an economy collection version E, as follows:

Documents				
	main		3000	
	thesaurus	say	300	
				3300
Searches and assessments				
a) staff	searchers		4500	
	formulators		1000	
	assessors		2000	
b) computing			24500	
c) keypunching,		say	800	
				32800
Data processing		say	1500	
				1500
Staff (2 = director, programmer)			13000	
				13000
Miscellaneous		say	4500	
				4500
	TOTAL			55100

As discussed above, increasing the cost of searching and assessment would bring the cost up to over £60K, but this could probably be offset by official negotiation, to reduce some commercial prices. We thus conclude that a reasonable price for the economy version E 'ideal' collection would be £55 K.

It is possible that the foregoing arguments, which essentially maintain the original cost level, may not be regarded as acceptable: thus Professor Cleverdon might maintain that the increased search and assessment costs considered were not large enough; and it might be that an 'official' reduction in the estimated costs to counterbalance these increases could not be obtained. For the purpose of further discussion, therefore we will make a simple assumption that the original costings should be raised to cover more realistic search and assessment costs by 20%, with no offset. This will give us what we may call pessimistic costings for the collection versions as follows:

	A	A	B	C	D	E
	123.4		109.7	94.0	85.3	55.1
add 20%	24.6		21.9	18.8	1 .0	11.0
	148.0		131.6	112.8	102.3	66.1 .

2. Retrieval experiments

The Outline Specification and subsequent Design Study were responses to a widespread feeling of dissatisfaction with currently available retrieval test data. More and better quality data were required for proper experiments of the kind conducted in the past ten years. Four questions therefore arise:

1) whether an 'ideal' collection as designed would in fact meet such needs;

2) whether there is sufficient experimental work to justify the non-negligible cost of at least £55K;

3) whether the data requirements of potential experiments could be met more cheaply by alternative data gathering procedures; and

4) whether this type of experiment is appropriate to modern retrieval systems and in particular to on-line systems characterised by large size, heavy technology and strong cost constraints.

These questions are best discussed on the basis of some financial considerations. Our experience in our own research is that in the last three years we have spent at least 1.5 man years costing of the order of £10K on putting a variety of collections into standard form, when they have already been supplied with search and assessment information of which the cost has fallen elsewhere, on different projects. The real cost of the limited, currently available collections must therefore have been of the order, depending on size and quality, of £10-20K, and certainly not less than £5K. Research in information retrieval increasingly calls for large scale experiments, and the cost to individual projects of attempting to provide substantial data for their specific purposes is likely to be high. It would either be so high as to act as a deterrent to research, or constitute a gross waste of resources since specific projects put substantial effort in collection generation but tend to produce material of low general utility which is restricted in content and specialised in format. A not unreasonable rule of thumb would be to cost individual project preparation of a non-trivial collection of the traditional kind at £1 a document covering both data assembly and processing to obtain some sort of standardised, portable product; though it must be emphasised that such collections are still likely to be limited in content. The cost per document is higher for small document sets, and lower for really large ones, and the cost of a collection depends substantially on the number of requests. But our experience suggests that a set of 10,000 documents and 200 requests would cost £10K.

1) collection needs

The first question, whether the 'ideal' collection would meet likely data needs, is effectively answered in Part B above. On the whole, the 'ideal' collection in its most ambitious version A would cover virtually all the expressed project needs, and also the teaching ones. Only 4 of the 28 projects submitted were doubtful. Of course there could be other projects not submitted, but the questionnaire coverage was wide, and the range of projects returned is broad. Versions B and D would not satisfy the small number of projects requiring citation data, and version C those explicitly requiring several document sets. Even with version D, the least expensive, between 2/3 and 3/4 of the projects could be done either fully or at least, in our view, fairly effectively. The economy model E presented above is more limited but would probably prove adequate for about 1/2 of the projects. For these 14 projects version E, at the original costing, would work out at about £3.9K per project; for 18 version D would cost about £4.7K per project, and for 24 projects version A would cost £5.1K per project. As noted, it is doubtful whether £5K would buy much of a collection on its own. If the pessimistic costings are adopted the figures are £4.7K, £5.6K and £6.1K respectively.

2) projects supply

The second question, whether there are enough projects to justify the 'ideal' collection costs can also be given a positive answer, but here necessarily only a superficial one. As noted in Part B., the number of projects submitted in relation to the current level of research and number of research workers in the U.K. is high. Thus there are at least a good many potential projects. Of course this does not imply that the 28 projects returned would all be submitted formally for funding, or that they would be funded. This is not a point we can properly consider. We can, however, consider the consequences of funding smaller numbers of projects within the range of research topics of which those described may be deemed representative. Suppose that 10 of the projects returned, or 10 like them, were funded, and further that these were spread over the whole range of requirements, implying the use of version A. This would cost £12.3K at the original costings and £14.8K at the pessimistic ones, which is not so much in relation to the total cost of the typical research project. If the 10 were similar in their requirements and could be satisfied with version D this would cost £8.5K and £10.2K respectively. If only 5 projects were funded, this would make version

A rather costly at £24.6K or £29.6K, while version D would cost £17.0K or £20.4K. But it should be noted that all these figures presume no benefits in terms of economies elsewhere in other research or teaching activities, which is hardly reasonable: if such collection data existed it would be used, as previous collection data has been used, for a range of worthwhile purposes not originally envisaged. In particular, its potential value for on-line education is suggested by the relevant section in Part B.

3) alternative data

The third and fourth questions are more fundamental than the first two. The third is whether the data requirements of the projects considered could be met more cheaply.

Professor Cleverdon argues in his Report (3) in favour of operational system-oriented research and, more specifically, for tests working within modern on-line systems. He implies that if retrieval data is wanted for independent study, it may be derived from such systems and will be more useful in research because it comes from them. This is in fact just what the 'ideal' collection as designed above would be. The implication must therefore be that such data may be got more cheaply. This must either be because it may be obtained strictly as a byproduct of other investigations, or because less ambitious data is adequate.

In the first case we have either data which is a byproduct of some research project, which will have to be funded to collect it; or which, as has sometimes been the case in the past, is the incidental product of a commercial service's own investigation. But unless the data to be obtained is less ambitious than that we propose the first option will not produce data more cheaply than we would. The second may well produce data which the courtesy of those producing it may make available at low cost to other research workers, but there is no particular reason to suppose that those topics of particular interest to specific data base or search service suppliers, and the subject of their own, typically restricted investigations, should be suited to general research.

The more serious proposal must therefore be that scientifically adequate and/or practically useful experiments can be conducted with much less ambitious data than that advocated in Part A of this Report. The point to be made here is that individual experiments can (sometimes) be so conducted: these would typically be the classic closed comparison between specific available choices (of indexing, search field, search strategy, or whatever), with each evaluated against the others. Such an experiment is illustrated by the pilot study of Cleverdon's Report. Professor Cleverdon has in particular argued that for such experiments relatively few requests, and more limited relevance assessments, are quite adequate. But even if this argument is accepted, it may be argued that the data produced is of little or no general value in part because it is associated with specific system options, and in part just because the request and relevance information is limited: this may make assessments of the performance of new options difficult or impossible because there is little output overlap, and the initial relevance information may be too restricted to allow reliable statistical extrapolation. That bane of past research, the difficulty of making comparisons across projects, is therefor liable to be perpetuated, though individual projects generating the data are likely to be cheap enough. Thus searching an operational system with limited output printing and assessment might cost, say, £20 a request, so a project working with 100 requests would cost a mere £2K for essential data. But as soon as more search variants are allowed, or more output is generated for study, or document sets are extracted for independent investigation, the cost per request goes up. The limited size of request set may make it difficult to pinpoint request features affecting performance through the study of request subsets, and that of assessment set may inhibit experiments with new search variants, particularly over a period of time when the original user may be lost or the data base changed. It is worth pointing out that working primarily through operational services, as it were live, may be an administrative hassle for individual projects and costly for organisational reasons.

The important point, however, is how far research can be conducted by working within operational systems, irrespective of whether the data collected is of use to others. Such systems in fact, as Cleverdon himself recognises, place heavy restrictions on the area of experiment allowed. Working with the given data and search procedures, there can be no change to the inputs (document indexing, index language used, etc.); and there can be none to the search logic permitted. The only type of experiment made easy is that essentially concerned with the exploitation of the system options open to the user, whether of search field, logical structure, level of detail, time taken, etc. Other types of experiment can be simulated, but with major effort, or achieved only through an obliging service operator who permits software modification. For example, an experiment to test the effects of coordinated searches with weighting, not a normal option in existing services, could only be conducted through an operational service by indirection, either through manipulation of the search input, or of its output, both of which are liable to be complicated, and hence costly, and inefficient. Some types of experiment, for example with document clusters, could not be done at all in this way. We may also doubt whether software modification is a practical alternative, except in the most costly form of system duplication.

The foregoing may be placed in context by considering whether the projects listed in Part B of this Report could be conducted through an operational service. This analysis can only be informal, since in some cases insufficient information is supplied, and in others it is possible that a slightly modified project could be conducted though as given the project could not. But our estimate is that, as given, about 1/3 of the projects could be conducted through an operational service, while about 2/3 could not. The number which could not would be reduced if such a facility as DIALOG's "private file" could be exploited for searching a new special data base, though this of course would have to be supplied. But on this basis we may perhaps say that about 1/2 of the projects could be conducted through an operational service. It must, however, be emphasised that this categorisation does not imply that those projects which could not be conducted through a current operational service are not relevant to such services.

4) <u>experiment type</u>

The final point, therefore, is the answer to be given to the fourth question, what kind of information retrieval research is needed. Professor Cleverdon argues that research should be strongly operational system-oriented, and so that research should accept the constraints of such systems: namely that cost is very important, that system efficiency is low, that searching is crude in relation to the resources available; that users are chiefly interested in precision. He would apparently imply that projects which could not be conducted through such services, or at any rate in close connection with them, must be mistaken. In his view the prime need of research is to improve cost/benefits within the existing system framework and in relation to observed user habits and needs, so research should be of a strongly hands-on character.

Our view is that this argument, even if accepted, does not provide much specific direction on the choice of worthy research projects, and much less than Professor Cleverdon apparently believes. It may be conceded that the set of projects listed in Part B omits some which would be of value to operational services, and equally inclydes some of doubtful relevance to such services at any rate in the immediately foreseeable future. But as the table classifying research projects by character in Part B shows, half are relevant in some visible sense to operational services, in terms of their properties and needs, though by the categorisation above some of them could not be conducted through current operational services. An example is the study of term weighting, which is not obviously irrelevant to operational services, but which could only be conducted in relation to a current operational service in an extremely inefficient way. Indeed even quite modest projects would present difficulties, for example that advocated by Professor Cleverdon himself in his Report: he suggests that an experiment in ranking the output of crude Boolean searches be conducted. This would have to be conducted by a mixture of service and independent computation of an organisationally tedious kind, and if evaluated in the style recommended by Professor Cleverdon would also certainly run into difficulties in testing developments in the weighting scheme in response to early results. Experience suggests that such developments would almost certainly be proposed but these would with difficulty be accommodated in an experiment accumulating search and assessment data over a period of time.

We therefore conclude that unless information retrieval experimentation is to be grossly limited, some sort of general-purpose test data is needed; and that this, if set up by the byproduct method favoured by Professor Cleverdon, is liable to be of low utility and hence more expensive than it appears. Thus if BLR&DD can satisfy itself that, say, 7 good projects will be forthcoming, 'ideal' collection version D at £85K or even £100K is a good buy.

References

1 Sparck Jones, K. and van Rijsbergen, C.J. Report on the Need for and Provision of an 'Ideal' Information Retrieval Test Collection, Computer Laboratory, University of Cambridge, 1975.

2 Sparck Jones, K. Outline Specification for the 'Ideal' Information Retrieval Test Collection, Computer Laboratory, University of Cambridge, 1976.

3 Cleverdon, C. An Investigation into the Use of Operational Data for Research into Information Retrieval, Cranfield Institute of Technology, Cranfield, 1977.