PART A : DESIGN

1. Objective, scope and conduct of the Design Study

The Design Study project was funded by British Library Research and Development Department, following an earlier investigative project on the need for a well-founded, multi-purpose, machine-readable information retrieval test collection. The investigative project was carried out by Dr. C.J. van Rijsbergen in 1975. The project Report (1) discussed the inadequacies of existing test data, showed a need for superior material, considered the requirements to be met, provided a rough specification for an 'ideal' collection to meet these requirements, and sketched methods and costs of building and curating the collection. The Report was used as a basis for discussion by a Working Party to provide a more detailed Outline Specification (2) of the collection.

Documents (1) and (2) were taken as inputs to the Design Study, which was a six month project managed by Dr. K. Sparck Jones with R.G. Bates as full-time assistant, and supported by an Advisory Panel consisting of Miss E. Barraclough (Chairman), T. Aitchison, E.M. Keen, J. Leigh, Dr. C.J. van Rijsbergen and Dr. S.E. Robertson. The Design Study was intended to see whether, and how, the specification of the 'ideal' collection contained in (1) and (2) could be met. A subsidiary part of the work was the gathering of information from those involved in teaching and research about their possible uses of the collection. This information is not analysed in exhaustive detail in this Report, as it is intended primarily for use by the Panel and by BLR&DD in assessing our own analyses of the form and cost of the collection, and more importantly in assessing the real need for the collection. Its bearing on the collection design is considered below, and it is summarised in Part B.

2. Summary of the collection specification input to the Design Study

The Outline Specification (2) of the collection with which we started may be summarised as follows.

The collection should consist of

1) documents

size: a main set of 30,000 documents broadly representative of service data bases in size and subject composition:
one or more other sets of 3000 documents complementing the main set in subject etc. Thus, for example, if the main set was in a scientific area, one other set would be in social science.
These sets would cover short time periods and English language material; they would have core characterisations (see below).
A random subset of the main set, containing 3000 documents, would be established, with enriched characterisations (see below).

properties: the main set, and complementary other sets, would be heterogeneous on identified collection variables such as subject solidity, document type, author type, etc.

The size of the main set should permit the identification of subsets, say containing 3000 documents, which would be homogeneous on such variables.

In addition, one or more other sets would be required for time and language contrasts, and possibly for gross contrasts on other variables, for example covering monographs as opposed to articles. These would have core or enriched characterisations as appropriate or available.

2) requests

size: a primary set of 700-1000 requests would accompany the main set of documents.
secondary sets of 150-250 requests would accompany other sets of documents.

As the primary sets would be of one form, envisaged as retrospective off-line queries, alternative sets representing different forms, e.g. SDI, and containing 150-250 requests were proposed for the main set. (At least some overlap of primary and alternative sets through derivation from common need statements was suggested; this overlap would define a base set of 30 requests.) These sets would have core characterisations. A random request subset of the primary set, containing 150-250 requests, would be established, with enriched characterisations.

For document other sets representing time and language contrasts, subsets of the primary set would be appropriate as requests.

properties: the primary and alternative sets would be heterogeneous on such variables as topic type, user type etc.

The size of the primary set should permit the selection of homogeneous subsets of perhaps 150 requests.

The request sets should represent many users, as well as many requests.

3) relevance judgements

The proposals here were not fully worked out, and are further developed in the present Report. For reference, the main points were:

default judgements by the users of their own search output;
exhaustive judgements of the random subset of documents;
pooled judgements on variant strategy search output;
these would all use abstracts;
checking judgements for the base set of requests e.g. against the texts of the random subset, against another random subset etc.

The data item characterisations are

a) core: for documents, all regular bibliographic information, abstracts, citations, natural language indexing, controlled language indexing (using thesaurus terms or subject headings) and high level subject class codes, and an about sentence;
for requests, a verbal need statement, lists of free and controlled terms, and a Boolean specification.

b) enriched: for documents, more exhaustive indexing, indexing from different sources, indexing by different people, PRECIS, etc; for requests, term weights, indexing by different people, etc.
Note that the variants for poooled judgements would constitute further request formulations.

Queries in the form of source documents should also be obtained.

It was proposed the full 'sociological' background information relative to requests should be obtained.

c) relevance judgements: essential information would consist of two relevance grades and also a novelty indication; known relevant documents would be recorded. Judgements by different people would be covered by the basic design.

The essential philosophy of the Outline Specification was that the collection should consist of sets of documents and/or requests related in different but regulated ways, so that a whole range of experiments could be carried out in a controlled manner. The different sets would represent variation in some major collection variable, but one only, so that tests with one data set could be systematically related to those with others.

For convenience, we will continue to refer to the 'ideal' collection as a single entity even though we are dealing with several, perhaps disjoint, document and request sets, any pairing of which might at a lower level be deemed to constitute a collection.

The Design Study was thus looking for ways in which the collection just specified could be provided. In particular, since a large volume of real material is needed, it was evident that this would have to be sought in operational services. The Study was thus concerned with identifying suitable sources of data, and with methods and costs not only of obtaining it initially, but with those of processing the collected material to set up the collection in a 'standard', i.e. convenient and portable form with the basic material supplemented by statistical information, a variety of formality defined data sets, etc.

3. Conduct of the Study

The project work was as follows. We made a comprehensive literature search for sources of detailed information about potential data bases and search services, and sent questionnaires to appropriate services. Questionnaires were also sent to potential users of the collection, one to research workers and one to teaching establishments; and following a suggestion by J. Leigh, a questionnaire specifically about the possible use of the collection for on-line education was sent to participants in a BLR&DD -sponsored Workshop. At the same time Dr. van Rijsbergen and Dr. Robertson as informal consultants to the project carried out a statistical analysis of the formal requirements for adequate relevance assessment. Replies to the data base questionnaire were analysed to identify the most attractive sources of documents and requests. It was evident that, contrary to earlier expectations, suppliers of data bases are not also suppliers of search services, i.e. data base suppliers are not search service owners, and we therefore investigated sources of documents and of requests separately. The analysis of the questionnaires was therefore followed by discussions with representatives of data base and search service suppliers. We are especially grateful to the following for their friendly and helpful response to our questions:

T. Aitchison, J. Pache and K. Mayne of Inspec, Dr. J. Newton and R.P. Healey of CAB, G. Pratt of DIALOG, Dr. A. Kabi of UKCIS, and Miss J. Bowron of BLAISE. We are also grateful to members of the Panel, which met twice during the earlier part of the Study, for valuable comments on the questionnaire design and relevance assessment analysis; and to Professor Cleverdon, who attended the meetings in his capacity as director of a complementary project.

In parallel we carried out preliminary analysis of the research project and teaching questionnaire, though we were hampered in using relevant information by a rather slow response.

We finally constructed the detailed specification and rough cost estimates which are given below.

The remainder of this Part is in nine sections.The next five sections deal with obtaining the document data sets, the request sets, and the relevance judgements, with document and request characterisation, and with citations. The various sets cannot really be treated independently, and the conclusions reached under each head are therefore tentative: we have put the points made together in a separate summary section. This is followed by a note on the replies to our research and teaching questionnaires, and their implications for the collection and its obvious uses. The next sections deal with the costs of building the collection and outline a tentative work programme for the building. In conclusion we summarise our findings.

Note that while we have tried to ensure that our detailed information is as up to date as possible, detailed facts and figures inevitably refer, rather loosely, to the period mid-1976 to mid-1977. Data bases and search services are so rapidly changing that information which is both stable and accurate cannot be obtained. We have indeed not felt able to pursue some details, either because the sources of information are remote, or because we have not wished to pester those responsible for data bases and search services unduly. All figures must therefore be treated as approximate.

4. Document sets

As mentioned above, data base and search service suppliers tend to be independent; we therefore studied likely document sources first, and then considered the question of obtaining request sets for selected data bases.

Our first step in seeking information about possible sources of documents was a literatute survey of the various directories now available. Those consulted are listed in Appendix 1. We found the very exhaustive and relatively up to date (details to mid 1976) Directory of Machine Readable Data Bases compiled by M.E. Williams and S.H. Rouse extremely helpful and used it as the base for further investigation.

As the main set of documents is by far the most important of the collection sets, and also involves most effort and expense, we consider it first.

Main document set

Requirements for the document data sets may be divided into two kinds, intrinsic and extrinsic. The former refers to those properties of the documents required by the Outline Specification, the latter to organisational/administrative properties likely to make them more or less convenient for our purpose. Thus considering some extrinsic properties, we decided that though many data bases are available, we would confine ourselves, for obvious economic reasons, to machine-readable data bases. We further decided to consider initially data bases generated or having local management representatives in the U.K. We would widen our investigations only if none of these seemed satisfactory on intrinsic grounds. It was clear to us that suitable arrangements for obtaining the data for the ideal collection, and for using it for searching to obtain the collection requests, could probably be much more easily made with a U.K. data base supplier.

The information provided in Williams and Rouse about the data bases covered by the directory is generally sufficiently detailed (see the example in Appendix 2) for use to be able to get some idea of whether a given data base might meet our intrinsic requirements of size and content i.e. type of document, type of document information, and type of indexing. However there are gaps in the directory we thought it desirable to try to fill in. We therefore obtained the following list of data bases for further investigation.

In the list, data bases associated with a single supplier and linked in some way are referred to as subbases.

NAME	ORGANISATION
(CA : Chemical Abstracts) 9 subbases	United Kingdom Chemical Information Service
CAB System : Commonwealth Agricultural Bureaux System 18 subbases	Commonwealth Agricultural Bureaux
CANCERLINE : CANCER Information on-LINE	British Library Lending Division
Culham Lab : Culham Laboratory Library Mechanised Information Services	United Kingdom Atomic Energy Authority Culham Laboratory
INSPEC : International Information Services in Physics, Electrotechnology, Computers and Control	Institution of Electrical Engineers
ISMEC : Information Services in Mechanical Engineering	Institution of Electrical Engineers
Mass Spectrometry Bulletin	Mass Spectrometry Data Centre
MEDLINE	British Library (BLAISE)
Research and Development Abstracts	Department of Industry Technology Reports Centre
Rock Mechanics	Rock Mechanics Information Service
SCI : Science Citation Index	Institute for Scientific Information
STI : Specialised Textile Information Service	The Shirley Institute

The list includes the Science Citation Index since it was evident that it might be difficult to obtain the citation information required by the Outline Specification from any other source, though the Index itself would not satisfy many of the requirements for a collection document set.

We sent a detailed questionnaire to those responsible for these data bases in the U.K. (treating the subbases separately). This was based on Williams and Rouse' directory, and we supplied information from the directory for checking to reduce effort. An illustrative questionnaire as sent out is given in Appendix 3. We are very grateful to the data base suppliers for their cooperative responses. As mentioned, since it appeared that data base suppliers may not own a search service, the section on searching the questionnaire was designed to collect information from those who do searching or to lead to independent search services using the data base, which could then be approached separately.

We did not receive replies for Research and Development Abstracts, or a proper one on MEDLINE. It also turned out that some data bases were not in fact suitable for our purpose: thus Culham Lab is not current, CANCERLINE and ISMEC are now run from the U.S., Mass Spectrometry is related to a specialised data base, and Rock Mechanics does not supply data base magnetic tapes. In addition, the CAB System is undergoing rapid development, and the characterisation of and relations between the various subbases is rather complicated. This therefore gave us the following list of prima facie candidate data bases:

CA 9 subbases, listed in Appendix 4
CAB 18 subbases, listed in Appendix 4
INSPEC
MEDLINE
STI

As noted, the SCI is not a suitable source for a regular document data set; it is considered separately below as the prime source of citation information for the ideal collection.

Information for those of these data bases for which questionnaire replies were received is crudely tabulated in Appendix 5, to obtain some idea of the type of information we have obtained. (Apicultural Abstracts is listed with CAB, though whether it is actually in the merged CAB data base is unclear). It should be noted that we did not get replies for many of the CAB subbases, so detailed information on most of the CAB subbases is missing. We have had to make assumptions about these, and hence about some properties of the merged data base derived from the subbases.

Even when the subbases associated with CA and CAB are treated separately, the range of possible sources of the main ideal collection document set is not large. We hoped, however, that it would be large enough to give us one or more sources of the main set, and we therefore concentrated on investigating the candidate data bases in detail for their suitability as a source of the main set, rather than spreading out enquiries further. It will be noted that none of the candidates is non-scientific. We have therefore considered the question of supplying a non-scientific data set as the most important other document set below.

In considering the candidate data bases in detail, the subbases of the CA and CAB data bases present some problems. The CAB subbases are merged in the combined CAB system, but the result is not homogeneous. The CA subbases are not merged but are linked through the presence of common documents in the different subbases and a single CA document identifying numbers. In what follows the individual subbases will first be taken separately, and then together.

Apart from the extrinsic criteria already mentioned, a variety of extrinsic and intrinsic selection criteria can be applied to choose one or more data bases as front runners for the main document set.

Extrinsic selection criteria

1. An important extrinsic criterion is whether enough searches for the large set required by the Outline Specification can be gathered for a given data base, and whether the associated relevance information can be satisfactorily gathered. This question is considered in detail below, but it can be said here that the various subbases, and the CAB ones in particular, would be very unlikely to generate enough searches. We decided, however, that as obtaining detailed information about question sets might be complicated, we would attempt an initial selection on other grounds and then see whether the selected data bases would support large query sets.

2. The other major intrinsic criterion is whether a document set selected from the data base would satisfy the relevance sampling requirement: i.e. whether a small subset of a large data base would actually contain any of the relevant documents for a query addressed to the data base, given that it appears that the only practical method of obtaining queries is to take those addressed to the data base as a whole. This point is considered in more detail in connection with the relevance judgements.

Intrinsic selection criteria

These selection criteria cannot strictly be applied in order for progressive elimination. However the first two are so important that, as some data bases pass them, we have felt able to use them as eliminators of others.

1. One major intrinsic criterion is size. The data base sizes are given below, except that for CAB only the merged file size is given: the sizes of the CAB subbases are listed in Appendix 6. Some of these subbases are currently not large enough, but all but Weed Abstracts would be by the time an 'Ideal' collection could be, so only this one is eliminated on grounds of size forthwith.

Data base sizes

	by 1976/7	1977 increase
CA		
CACon	2.6 M	470 K
CBAC	275 K	47.5 K
CIN		50 K
CT	1.8 M	148 K
ECOL. & ENV.		47 K
ENERGY		35.5 K
FOOD & AGR.		32.5 K
MATERIALS		85.5 K
POST	300 K	45 K
CAB System	400 K	130 K
INSPEC	1 M	150 K
MEDLINE	800 K	300 K
STI	60 K	8 K
SCI	6.4 M	530 K

2. A second criterion is whether machine-readable abstracts are available, or would be in sufficient quantities by the time the collection was being assembled (MEDLINE has taken to abstracts comparatively recently). The cost of keypunching abstracts is so great that it cannot be contemplated certainly for the main set and probably for other sets. The availability of abstracts is given below. This criterion eliminates the STI data base, and also some CA data bases like CT and also CIN, which has extracts rather than abstracts. These data bases have not been considered further. As only about half the MEDLINE documents have abstracts, a selection would have to be made from the file. This would give a large enough document set, but the selection might be rather awkward to correlate with searching, as described later. We may, however, for the moment assume that MEDLINE passes this selection test.

Availability of abstracts

CA		
CACon	no	
CBAC	yes	
CIN	no	
CT	no	
ECOL. & ENV.	yes	
ENERGY	yes	
FOOD & AGR.	yes	
MATERIALS	yes	
POST	yes	
CAB System	yes	(but Index Vet. titles only)
INSPEC	yes	
MEDLINE	yes, for between 1/3 and 1/2 of the documents since 1975	
STI	no	
SCI	no	

3. The third intrinsic criterion is that the data base should be 'solid'. This criterion can only be characterised informally but is felt to be important. The requirement is that the data base should be neither a mere aggregate of disjoint subject fields, nor offensively specialised, the assumption being that as large operational data bases are neither, the ideal collection should resemble them in this respect. The candidates all appear acceptable from this point of view. This criterion is also important from an experimental point of view. In particular, it is desirable that a collection which consists of a sample drawn from a very large data base should not be too miscellaneous for useful experiments. The INSPEC data base, which falls into three sections, is perhaps unsatisfactory as a whole from this point of view, and the merged CAB data base is also to some extent an aggregate. The table below gives a crude analysis for the data bases selected by the previous criterion.

Data base subject 'solidity'

CA	
CBAC	yes
ECOL. & ENV.	yes
ENERGY	yes
FOOD & AGR.	yes
MATERIALS	yes
POST	yes
CAB System	subbases yes, whole slightly aggregated
INSPEC	main sections yes, whole slightly aggregated
MEDLINE	?

4. Another important criterion is whether the basic range of forms of indexing information is available, i.e. whether natural language keywords, controlled thesaurus terms or subject headings, and high level subject or class codes are supplied. Specific forms of indexing could of course be provided as part of the collection building operation, but it is clearly more convenient if they are already given. Unfortunately the questions about controlled language indexing in Williams and Rouse are not wholly satisfactory, and our slight revisions of them were not satisfactory either. In neither case is there a clear distinction between the use of an authority list or none, between pre- and postcoordinate indexing, and between a set of descriptors with no or few levels and one with many. It is an area in which it is difficult to get accurate and consistent information from questionnaires. However since the Outline Specification calls, rather crudely, only for controlled indexing of some kind, and for high level classes, we have generally been able to get sufficient information for present purposes. The table below characterises the indexing of the candidate data bases. Some of the CAB subbases do not have all three forms of indexing, so combining them to form the merged CAB System will give 'patchy' indexing for the latter. A further further problem with CAB is that even where the same type of thesaurus/ subject indexing is supplied for the different subbases, the terms are now drawn from a common authority list, but are simply the products of different indexing policies. A post hoc authority list is, however, now being formed. MEDLINE lacks natural language indexing and would be eliminated by this criterion.

Data base indexing

	natural language keywords	controlled terms/ subject headings	high level subject classes	controlled term authority list
CA				
CBAC	yes	yes	yes	yes
ECOL. & ENV.	yes	yes	yes	yes
ENERGY	yes	yes	yes	yes
FOOD & AGR.	yes	yes	yes	yes
MATERIALS	yes	yes	yes	yes
POST	yes	yes	yes	yes
CAB System	sometimes	yes	yes	not really
INSPEC	yes	yes	yes	yes
MEDLINE	no	yes	yes	yes

5. The fifth intrinsic criterion is whether the data base has a suitable mix of document types, since the Outline Specification calls for some heterogeneity. The typical data base appears to consist predominantly of journal articles, with some other items like conference proceedings, reports etc.; the candidate data bases are all of this kind, as the table below shows.

Data base document composition

	journal articles	reports	patents	conf.proc. monographs	misc.
	%	%	%	%	%
CA					
CBAC	85	3	4	8	
ECOL. & ENV.	75	4	10	11	
ENERGY	73	4	13	10	
FOOD & AGR.	74	1	15	10	
MATERIALS	55	2	35	8	
POST	48	1	50	1	
CAB System	?80	? 5		?10	? 5
INSPEC	80	7	6	6	
MEDLINE	mostly	some		some	some

A subsidiary requirement was for documents varying in treatment, author origin etc. We have not been able to collect detailed information about this, but it must be concluded that the range of material covered by data bases as large as those we are concerned with must be varied in all these respects.

Some other points relevant to the choice of collection source, though they are less important than those discussed so far, are as follows.

In setting up or using the ideal collection for the kinds of purpose envisaged there could be difficulties or substantial extra work connected with handling chemical nomenclatures. This is an argument against using CA material. It has also been suggested that

a hopefully forward-looking test collection should not be hampered by depending on antiquated indexing philosophies or products. This argument may apply to MEDLINE, and perhaps also to INSPEC, though the indexing languages of both services are updated. Finally, we may take into account a very informal requirement for a 'friendly' subject. It is not necessary for much information retrieval experimental work that the subject matter of the documents should be well-understood by retrieval research workers; but some inconvenience could arise if the material is so technical as to be totally opaque. This is perhaps a problem chiefly with the chemical data bases but it must be recognised to be a problem with at least some documents in all the data bases.

The application of the intrinsic selection criteria just discussed to the candidate data bases is summarised in the table below. It must be emphasised that this characterisation of data bases in terms of requirements concerns only their suitability as inputs to the ideal collection, and is in no way a comment on their general status. In particular, the requirements for a multi-purpose test data base for information retrieval work are rather different from those for a data base associated with a retrieval service, and no direct inferences should be drawn from utility in the one context about utility in the other.

Summary of intrinsic selection criteria applied to data bases

	Criterion 1		2	3	4			misc.
	large: all	1 yr	abstracts	solid	indexing: 3 forms	auth: list	+ other	
CA								
CACon	yes	yes	no	?	no	no	yes	
CBAC	yes	yes	yes	yes	yes	yes	yes	
CIN	yes	yes	not really	yes	no	no	yes	
CT	yes	yes	no	?	no	no	no	
ECOL.& ENV.	yes	yes	yes	yes	yes	yes	yes	formulae
ENERGY	yes	yes	yes	yes	yes	yes	yes	unfriendly
FOOD & AGR.	yes	yes	yes	yes	yes	yes	yes	
MATERIALS	yes	yes	yes	yes	yes	yes	yes	
POST	yes	yes	yes	yes	yes	yes	yes	
CAB System	yes	yes	yes	?	no	post-hoc	yes	
subbases	most	no	yes	yes	not nec.	not nec.	yes	
INSPEC	yes	yes	yes	?	yes	yes	yes	
Physics	yes	yes	yes	yes	yes	yes	yes	
Elec.	yes	yes	yes	yes	yes	yes	yes	
MEDLINE	yes	yes	partly	?	no	yes	yes	old
STI	yes	no	no	yes	yes	yes	yes	
SCI	yes	yes	no	no	no	no	no	citations

From this table it is clear that the front runners for the data base source of the main document set are the merged CAB file, the CBAC, ECOLOGY and ENVIRONMENT, ENERGY, MATERIALS and POST subbases of CA, INSPEC and MEDLINE. Of these, the CAB System suffers from heterogeneity and a limited number of forms of indexing, the CA subbases

from formulae, and MEDLINE from spasmodic abstracts and limited indexing. INSPEC is thus somewhat in the lead at this stage.

As far as the mechanism of selecting the main document set from one of the large data bases is concerned, the following procedure would seem appropriate. The Outline Specification proposed a short time coverage for the main set; and as for these data bases the number of documents added annually exceeds 30,000, a natural way of obtaining the set would be to take a suitably sized block of the most recent material. Note that, as will be discussed in more detail under the formation of request sets, the document set should be selected from the data base after searching, since it would not be acceptable for a users' searching to be confined to it.

Other document sets

The major requirements of the test collection not met by any of the data bases considered so far are for social science/arts (i.e. non hard science) material*, for more varied types of document, or at any rate less emphasis on journal articles, and also specifically for monographs; for material in a language other than English; and for a document set with a different time characterisation. The intention of the Outline Specification was that other document sets differing from the main set only in one of these respects (and in size) should be set up. Thus a social science collection would differ in subject character but not in document type makeup. However this strong degree of control might be difficult to achieve in practice and it was recognised that, for example, an other set differing in document type makeup might be not merely in a different science, but not in a scientific area at all. However it was intended that the other set with a different time characterisation should be a specific contrast with the main set.

In general, the provision of other data bases would involve much less effort than that of the main set as both fewer documents, and more importantly, fewer requests, are required. We believe that deriving other sets according to different criteria data bases independent of that used for the main set involves much the same effort, and we have therefore considered providing only the social science one in some detail. The remaining other sets proposed are treated rather briefly. There is a general problem of whether enough requests could be obtained for other sets, and also one of relevance sampling. These are discussed below.

1) Other set in social science

Social science is interpreted rather broadly here, and indeed the old fashioned term arts might be more appropriate: the intention is to provide a collection of non-scientific material.

As it has several times been suggested that a large social science document set is desirable, we have considered the question of obtaining one other set, in social science, comparable in size with the main set. Our initial list of data bases, considered in the previous section, shows no current native social science data base. However LISA, Library and Information Science Abstracts, will be available for searching on

*'social science' and 'arts' are not deemed equivalent: they are both contrasted with natural science.

DIALOG shortly, and would provide a large set of social science documents. There are two objections to the use of LISA. One is the difficulty of obtaining a large genuine request set. The other is that it would be a somewhat incestuous data set for information retrieval workers, who might find it difficult to maintain objectivity in their experiments with it. We have therefore considered other sources of social science material. It should however be borne in mind that one recommendation of the earlier ideal collection reports was that the ideal collection could be extended by the addition of document and query sets generated, to suitable standards, by independent projects. Such a set derive from LISA in the course of a library school project, for example, would be welcome.

The other sets were defined in the Outline Specification as of 3000 documents and 150-250 requests, and those discussed below are of this size. We think, however, that some of the requirements of those interested in social science material could be met in a relatively painless way as follows. The major cost of setting up a document and correlated request set is that of obtaining queries and their relevance judgements; the cost of the document set itself is relatively small. We therefore suggest that where a complete other set as defined is provided, the documents involved should be part of a larger set comparable in size with the main document set. If the 3000 documents of the other set were a ransom subset of the larger one, the latter, even if not involved in searching and assessment, could constitute an extension to the former of value for some purposes; and it could be provided as small cost.

Since non-native data bases would have to be sources of the social science other set, we considered those available to searchers in this country through the major on-line search services. Applying the criteria used for the main set, the following data bases appear likely sources of material: that is, they would probably support enough searches, are of adequate size, have machine readable abstracts, are solid in character, and have different forms of indexing and types of document. Note that our information about these data bases is derived only from Williams and Rouse, and so is not completely adequate.

NAME	ORGANISATION
CIJE : Current Index to Journals in Education	Educational Resources Information Center
HA : Historical Abstracts	American Bibliographical Center - Clio Press
PATELL : Psychological Abstracts Tape Editions Lease License	American Psychological Association, Inc.
Sociological Abstracts	Sociological Abstracts, Enc.

Strictly, HA and CIJE are not satisfactory in including a range of document types, since the data bases consists of 100% Journal articles. But HA has a good range of indexing forms and CIJE is included since one possibility might be to exploit the existence of the existing ERIC test collection provided by DIALOG as their ON-TAP teaching file. The ON-TAP file would not satisfy the ideal collection requirements on the size of request set, as it has 30 as opposed to the desired minimum of 150, and we feel it would be difficult to enlarge

the ON-TAP file request set in a sufficiently controlled way. An objection to PATELL is that the material is too scientific in character to provide a proper contrast with the main document set. The data base may also not be very cohered, though a solid subject could doubtless be obtained. Sociological Abstracts may also be open, though less strongly, to the same objection. The characteristics of these data bases are summarised in the following table.

Social science data bases

	size		m-r	solid	indexing				Composition	
	all	1 yr	abs		nat.	cont.	high.	auth. list	arti-cle %	other %
CIJE	120K	20K	yes	yes	yes	yes		yes	100	
HA	37K	7.5K	yes	yes	yes	yes	yes	yes	100	
PATELL	230	25K	yes	yes	yes	yes	yes	yes	95	5
Sociological Abstracts	77	6	yes	?	yes	yes	yes	yes	80	20

2) Other set for varied document types

The readily available data bases containing very different material are those like the FUNK and SCOTT INDEXES of Predicasts Inc. covering U.S. and international business information. Unfortunately many of the Predicasts data bases are strongly U.S. oriented, and it is therefore unlikely that sufficient requests for them could be obtained in this country. The most suitable Predicasts data base would seem to be INTERNATIONAL STATISTICS. We have not, however, pursued the provision of another set derived from this data base, partly through lack of information about potential sources of queries, but mainly because there is no evidence in our survey of possible uses of the ideal collection of a real demand for this type of material.

Research in Education subbase might also be a suitable source but we doubt whether enough queries could be obtained. A more modest approach to the provision of a mixed document other set would be to make a deliberate selection of non journal articles from one of the regular data bases containing them, or at any rate a selection of documents with a different distribution by type from that of the data bases as a whole. A matter for decision would be whether it is more appropriate to take this set from a data base not used for any other purpose, or from a data base supplying the main or some other document set: either course would have advantages and disadvantages in terms of tests on experimental variables. The CAB System and Sociological Abstracts appear to be good sources for this other set, or some CA subbases if a high proportion of patents is desired.

2a) Monographs

Some of the data bases already considered contain some monographs, but rather haphazardly; we feel they would therefore not provide suitable test material. The obvious way to obtain an adequate monograph other set would be to select from British Library Marc tapes in some broadly defined subject area: thus it appears that the available tags would permit such selections as of English books on history.

3) Other set in another language

A number of the data bases considered cover foreign language material, CAB in particular. However though this material is represented in the data bases, it does not in fact appear in its original language, or at any rate so appears only partially. Thus the title may be translated into English, and the abstract translated into or supplied in English, accompanied by an indication of the original language. Indexing is in English. The only substantial foreign language element likely to occur is a title. To obtain a foreign language other set would therefore involve reference back to the sources of documents which would themselves be quite easily identified as foreign language in the data bases, and additional keypunching. The CAB and MEDLINE data bases at least contain foreign language titles.

4) Other set for different time

As noted earlier, this other set is intended to represent a different time slice from the same material as the main set. Unfortunately, none of the machine readable data bases goes back far, so it would be difficult to complement a main set representing, say, the current year, by an other set with a longer time span, or one representing an earlier year. However, some of the abstract publications underlying the data bases are old (though MEDLINE does not have one), so an early time slice could be obtained from them, for which keypunching would be required. A problem would be that this data set would not be easily incorporated in the data bases used for searching, so it would presumably be necessary to obtain special requests for the document set.

5. Request Sets

As indicated in the previous section, the suppliers of data bases, and particularly those we have been considering, are not generally sole or prime suppliers, i.e. owners, of a search service for these data bases. Some of the data base suppliers have a substantial in-house search operation, especially UKCIS, CAB and INSPEC, for files which are also available on independent search services. In other cases the files are available in this country only through a search service like DIALOG or ORBIT. Some suppliers may indeed exploit such services as part of their own search service.

The Outline Specification characterised requests in terms natural to older test collections: i.e. they were regarded as fixed search formulations derived from the user's original informal need statement and used for one-off searches. Search formulations would be provided for each indexing language available, and would themselves, along with the original statement, form part of the test collection. Some at least of the relevance assessments would also be quite strongly tied to the output of searches with these formulations.

The development of on-line searching suggests that a different view is needed. While the initial need statement is fixed, and relevance assessments are related to it, the latter may be obtained by a succession of complete or incomplete formulations. These formulations, which may be referred to as queries, cannot be treated either

individually or collectively as fixed requests in the old fashioned way, though they should be recorded. We have therefore extended the use of the word request to cover the whole combination of initial need statement and one or more queries in the set used for searching, whether by the user himself or independently to obtain additional documents for assessment. A scan of a document set for a particular query may be referred to as a search, so a request in our sense may have a set of searches.

Request sets have to be considered from two points of view: first in their own right, and subsequently for their implications for the choice of document set source.

Primary request set

The primary request set is for the main document set. It is essential for adequate data collection, both about requests and about relevance assessment, that search information should be gathered in this country. For the data bases considered as main set sources, the types of search available here are tabulated below.

Search modes

	off-line	SDI	on-line				
			DIALOG	SDC	NLM	BLAISE	INFOLINE
CA							
CACon	yes	yes	yes				
CBAC		yes			yes		
others		yes					
CAB System	yes	yes	yes				
INSPEC		yes	yes				
MEDLINE						yes	
SCI			yes				

As this table shows, most of the CA subbases are not available for on-line searching. This does not necessarily mean that they could not be used as sources for the ideal collection, since searches of CACon could in principle be linked to subbases involving abstracts through the CA abstract numbers, but this would be a major administrative hassle to be avoided if possible. The only subbase allowing straightforward use in CBAC.

The most important intrinsic requirement of the primary request set is that it should be large enough. This is not merely because a large set of requests is required for study of request properties, but because as the discussion of relevance judgements below indicates, as the size of the request set increases the number of documents to be assessed per request decreases. This is important from a practical point of view. We will therefore assume that of the order of 700 requests are needed, and consider how they may be obtained.

As indicated in the previous section, the probability that an adequate number of requests can be obtained, say over a period of up to a year, is an important selection criterion for the document set.

It is not possible to obtain any very accurate estimates here for the following reasons. First, the general pattern of search activity is changing quite rapidly with the spread of on-line search facilities. For example, the advent of BLAISE must affect the use of MEDLINE, and CAB is currently actively promoting the use of its data base through DIALOG and expects an increase in use. Second, it is difficult to tell what the response would be to the conditions of search associated with setting up the ideal collection. As at least some user assessments of relevance are deemed essential, it is evident that the classical strategy would have to be adopted: i.e. the user would be offered a free search (that is free to him, the cost being borne by the collection building project) in return for providing information about his needs, and carrying out assessments. It is difficult to say how many genuine as opposed to frivolous potential users there are who are currently inhibited from searching from lack of money; but some would be needed as the number of actual users of the data bases is not large enough. The third problem is that some actual or potential users might nevertheless not be willing to cooperate as they are especially concerned about confidentiality and maintaining secrecy about their interests. Fourthly, it is clear that if the requests are to be obtained in a well-organised way, along with background information and relevance assessments, this can only be achieved by working preferably through one centre, or at any rate through a relatively small number. Thus while a sufficient number of requests might be achieved for a given data base for the U.K. as a whole, we require an adequate total from, say, six centres. (In this connection it may be noted that there where a data base is available through a service like DIALOG, the data base suppliers may not have any detailed information about the use made of the data base through the service, while for reasons of confidentiality and commercial interest the service operators may be reluctant to provide much information about past searches on their system.)

The final question is whether the type of requests suggested for the primary set in the Outline Specification could be obtained in sufficient numbers. The Specification suggested retrospective off-line search requests for this, with alternative sets of retrospective on-line and SDI requests. The rapid growth of on-line searching suggests that retrospective on-line requests, which in practice differ little from off-line ones, should be adopted for the primary set. Some controls would be needed to maintain a proper relationship between original need, the request or request variants searched, and the relevance assessments. These are considered below.

We have discussed the provision of a request set with those responsible in the U.K. for the candidate sources for the main document set, and are grateful for their help in attempting an assessment in the light of the problems outlined above. It appears that something like 700 requests could be collected for the INSPEC physics section, that 400+ could be relied on for the merged CAB system, probably 700 for MEDLINE and 300 for CACon, but far fewer for CBAC.

With respect to the other intrinsic requirements for requests, i.e. variety of user, type, etc., it appears that these could all be satisfied.

Alternative request sets

As noted, it appears from observation that in practice on-line retrospective searching does not give rise to types of request materially different from those used for off-line searching: it seems that for economic reasons, extensive browsing is not common. We have therefore abandoned the idea of on- and off-line as distinctive types of request. We propose instead that when on-line search information is gathered, the queries used should be vetted to ensure that as they are developed do not diverge so radically from the original need statement as to imply that relevance evaluation is not based on the original need.

The remaining proposed alternative request is the SDI profile. These appear to be genuinely different in character from one-off search requests; but it is evident from our discussions that there would be considerable difficulty in obtaining large numbers of SDI profiles, and especially of individual rather than standard or group ones. Our study work has also emphasised the cost of obtaining the test requests and we have therefore concluded that there is little point in looking for an SDI alternative request set. It should however be noted that if it was thought sufficiently important to collect SDI profiles this would involve much the same methodology as that used for collecting on-line requests, and the cost would be much the same, or at most a little more as a larger output might be required.

Secondary request sets

The problem of obtaining an adequate request sample to some extent applies to the secondary request sets for the other document sets, though it was proposed that they should be smaller, consisting of 150-250 documents. We believe that there would be little difficulty in obtaining 150 on-line requests for whatever social science other document set was chosen. But as mentioned, we have no idea of whether sufficient requests for a business information other set could be obtained. We believe it would not be difficult to obtain enough requests for a monograph set, though possibly of a rather simple kind: for instance BL Marc tapes can be searched in simple postcoordinate mode in Cambridge. The strategy for obtaining foreign language request for the relevant other document set would have to be that of translating search formulations into the appropriate language and searching on the available field, namely the title (assuming this is permitted.) As noted, the other set covering an early time slice of the main set could probably not be searched concurrently; the requests would be essentially the same as those for the main set but variants to deal with historical terminology would have to be constructed, at main search time though presumably for future search when the document set was available.

The discussion so far has been concerned essentially with obtaining need statements for the different document sets. As noted, these would individually be associated with one or more queries actually used in searching by the user or a professional searcher on his behalf. The Outline Specification calls for careful and full recording of the original need, supplemented by any known relevant documents. In addition, the actual searches carried out would be recorded, so specific query formulations would be available as part of the test collection material.

There is not likely to be much consistency about these, as they may be explicitly in different indexing languages, if these are clearly distinguished for the documents, or implicitly so, if the same verbal query form is to be applied to different document indexing fields. Even where one indexing language is concerned, particularly natural language, queries may vary in applying them to different document surrogates, e.g. title or abstract texts. The effects on the retrieval of relevant documents of such variations over requests should be counteracted by the use of additional queries specifically designed to exhaust the relevant document set, for the reasons considered in the next section. We think it important that, according to the resources available for a given document data base, for a given request the users queries and the additional queries taken together should cover a consistent range of language and field search options. Thus for the request set to be gathered for a data base, a schedule of query types would be needed, all of which would be searched to provide documents for assessment, though for a given query, the distribution of types between user queries and searcher queries would vary.

Relevance judgement sets

The provision of relevance judgements was recognised in the previous reports as the major problem in starting the 'ideal' collection. The essential requirements are that individual judgements are reasonably reliable, and that the set of assessed documents for a query should be 'adequate': this means not only that enough documents should be assessed for specific studies of individual document status - which may be called the qualitative requirement, but to provide a basis for evaluating the the output of future experiments - the quantitative requirement.

Assessments in the primary requests of the main document set

In pure principle, these relevance requirements are satisfied by exhaustive assessment of a document set in relation to the request need statement. But this is quite unrealistic for a collection of 30,000 documents: even if it is feasable with relays of assessors, it is too expensive. Some suggestions for providing good enough assessments were sketched in the Outline Specification, but these were very crude, and as mentioned above, early in the Design study further work was done on improving the specification of assesssment strategies. The Outline schemes in particular were rather complicated in seeking to supplement the relatively limited assessments that could be expected from the original user with further assessments designed to determine the real recall of the user assessed output, while ensuring that judgements involving several assessors would be reasonably consistent.

These schemes were subjected to a rigorous analysis by Dr. van Rijsbergen and Dr. Robertson, who were able to provide a proper statistical rather than intuitive basis for the ideal collection assessments, and in particular for the main set assessments. Their arguments lead to simpler procedures than those originally envisaged, and, provided that certain assumptions are met, require fewer assessments than might be expected. This is useful since it means that all or a large proportion of the assessments could be made by the user, and so would be guaranteed authentic. In some cases non-user assessments would be needed, since the calculations are for average numbers of assessments and some requests could require more. However it appears to be accepted that reliable enough assessments can be obtained from information officers experienced in the given subject field, given a good need statement to work from; and as a check some overlap in assessment between user and assessor could be introduced. As even in the worst cases, very large numbers of assessments would probably not be needed, ensuring consistency over several assessors would not be a major problem.

Van Rijsbergen's and Robertson's arguments concern primarily the quantitative rather than the qualitative requirement. We see no way of fully meeting the qualitative requirement, and simply hope that in practice enough individual documents would be assessed for likely qualitative investigations. The arguments are set out in detail in Appendix 7, accompanied by figures showing how many assessments per request are required, on average, for request sets of different sizes. They may be summarised, in a relatively non-technical way, as follows.

Van Rijsbergen and Robertson's basic argument as so far developed, but it needs further checking and refinement, is that, given search output for a request which may be presumed to contain all (or virtually all) of the relevant documents for that request in the entire data base, and also to contain the output of future searches based on the request need statement, it is sufficient to assess a random sample of this output. Clearly these requirements would not be met only by taking the output of the user's own search, or that conducted on his behalf by a searcher: searches by a range of strategies, i.e. the additional query searches mentioned in the previous section, would be needed: and the random sample would be drawn from the pooled output of all the searches for the queries associated with a need statement. The formal argument does not require any specific size of pool; and, as noted, its nature is such that as the size of the request set rises, that of the random sample to be drawn from the pool and assessed falls.

Whether the proposed method of obtaining relevance information is judged satisfactory depends on the way in which it would be used to evaluate retrieval experiments. This is essentially similar to that conventionally used, but is not quite the same in detail. The assessed sample would be used for comparative evaluation i.e. given two retrieval experiments, differing in indexing searching, or whatever for a given request set, we compare their retrieval performance with respect to the assessed set of documents only, and not with respect either to the whole document set, or to their own output, which will typically include unassessed documents. It is claimed by van Rijsbergen and Robertson that if the assessments are made as proposed, and particular confidence limits are set, when experimental results are compared using the assessment information, the performance differences observed will represent genuine performance differences between the methods being tested.

The procedure is of course essentially the same as when the entire collection has been assessed and so constitutes the sample, or when closed comparisons between procedures are made using assessments of their pooled output. The point is that the proposed method avoids the problem which arises in the first case of unrealistic numbers of assessments, and that which arises in the second of precluding new comparisons.

The assumptions that the procedure is based on are rather strong. It is therefore proposed that at an early stage in the collection building project, before any actual searches are carried out, some investigation be made of whether they are likely to be met by the proposed schedule of search strategies, and further statistical work be done to provide amended procedures if needed, say to allow for only 95% of relevant documents in the pool.

A variety of techniques have been adopted for conducting alternative or parallel i.e. additional searches to supplement a user's own, for example using UDC numbers. In general it appears that user searches are relatively restricted. Thus the (carefully constructed) UKCIS Boolean profiles as used in our experiments retrieved an average of 81 documents from a file of some 27,500, searching on titles. The UKCIS workers originally estimated that title searching only retrieved 40% of the relevant documents to be retrieved. On the other hand in our experiments we have found that if the set of profile terms is used for coordination level searching on titles, down to level 1, an average of 2670 documents is retrieved. This experience can be used to determine likely conditions of searching in connection with the ideal collection, and appropriate procedures for obtaining documents for assessment.

Thus it may be suggested that a useful technique for obtaining additional queries to complement the users' own would be to conduct coordination searches on a hospitably constructed word or word fragment list. A second possibility, given the apparent requirement is to increase output, would be to apply the standard technique of using high level class or section identifiers.

As a practical proposition, when 700 requests are used, so that, according to the argument detailed in Appendix 7, a 10% of the pooled output documents would have to be assessed, this implies the user assessing 250-300 documents. Our discussions with those who work with search systems or have conducted tests relating to operational systems suggest that inviting a user to assess this number of documents using abstracts in return for a free search would be acceptable. Where the pool is larger additional assessor(s) would be needed. A study of our own output from the UKCIS coordination searches mentioned suggests that they would be required for about 25% of the requests. Assessment would then be done by taking as many random samples of the percentage sample of the pool with some overlap between them as would be required to exhaust the percentage sample. Thus if a 10% sample of the pool contained 700 documents, three assessors altogether doing 300 documents each would be appropriate. But perhaps two, the user and a professional, would be sufficient since it is reasonable to assume that paid assessors could handle more documents than the user.

The assessment procedure just described refers only to assessments relative to the main set of documents. It must be emphasised that the real users providing requests will ordinarily be interested in searching the whole, or at any rate a substantial part, of the data base from which the main set is drawn. Thus the complete processing of a request for the ideal collection will involve on the one hand a search of the data base as a whole for the query devised by the user or his professional searcher, and on the other searches only of that subset of the data base flagged as the main document set for the additional queries. Relevance assessments will therefore be required for documents drawn from a pool consisting of those documents in the user's output coming specifically from the main document set, and the outputs of the additional searches. Our view is that because of its special information status, the complete set of documents (or as much of it as the user can stand) output by the user's own search and in the pool should be assessed by the user. The UKCIS experiments suggest this would not be a large set. There is of course no need to obtain assessments for documents retrieved by the user's search from the rest of the data base. A possible procedure would be to invite the user to assess the pool sample first, and then the rest of his own output. Fairly careful management of the detailed assessment procedure would be required to avoid biases.

Alternative request assessments for the main document set.

The procedure for providing assessments for alternative request sets, and specifically SDI requests would be the same as for the primary requests. The only point worthy of comment is that for different forms of request drawn from the same initial need statement (the base requests) user search output may be rather large so more reliance may have to be placed on assessors.

Relevance judgements for other document sets

The procedure to be adopted for relevance assessments for the secondary request sets and their other documents sets cannot be that used for the primary set, since sampling would not be reliable enough. It appears that exhaustive assessments are needed. An appropriate strategy would therefore be to take a 10% random sample of the other documents for assessment by the user, the remainder of the other set being exhausted by samples for further assessors in the manner described above for large pools; the user would also assess any of his own search output not falling in his random sample. (But the character of some of the other sets might preclude direct assessment by the user: for example: if the document set for another time is specially processed. Assessors only might have to be used here).

The question which arises in connection with all the document sets, and especially the other sets, is of relevance sampling adequacy. If a request has only five relevant documents in a data base of 300,000 items, the chances of any appearing in a sample of 30,000 are not large, and of any appearing in a 10% sample of the latter consisting of 3000 documents are small. This suggests that the use of really large data bases, which are somewhat heterogeneous, as sources for the ideal collection would perhaps be a mistake. Of course if a request has no relevant documents in the selected sample, it may be rejected for the 'ideal' collection: but this may undesirably reduce the size of the request set; and problems could arise even for requests with one or two relevant documents in the sample. A rule of thumb might be proposed to the effect that a database should not be used as a source if the required document sample size is less than 10% of the whole typically used for searching. But this would be rather stringent; a requirement of 5% would be less restrictive, but implies more relevant documents per query. It is difficult to offer a definitive statement on this point, since the density of relevant documents per query will be partly a function of the concentration of the data base subject matter. We propose some sampling for actual user requests, which could be fairly easily done, and in the meantime suggest that CACon, the full INSPEC, and MEDLINE should be approached with caution for the main set, and CIJE and PATELL for the social science other set.

7. Document and request characterisations

The documents in the main set were required to have core characterisations; and indeed two of the selection criteria for data bases as sources of the set were the availability of machine-readable abstracts, and of three main forms of indexing. These requirements together cover a large part of the core characterisation specification. All the data bases include substantial bibliographic detail, so the outstanding remaining items in the core characterisation are an 'about' sentence and citations, neither of which are available in the candidate sources for the main set. Our view is that it would not be feasible to provide an about sentence for the main set of documents, and we have therefore made it a requirement of the enriched characterisation for the random subset, for which it could be provided without large expense. The question of providing citations is considered separately below. Other document sets have core characterisations, and are therefore essentially provided for by their source data bases. However not all have the three forms of indexing. We do not think it sensible to build a thesaurus for 3000 documents, though it would be reasonable to supply subject headings if lacking.

The enriched specification for the random subset of the main document set covered alternative keyword indexing (e.g. from different sources), and also PRECIS indexing. We feel on investigation that the range of alternatives called for is unrealistically large, and would be rather costly. However each option would cost much the same, so the cost of the total range is a matter of multiplication. We have costed for the alternative which seems most needed, given the indexing generally available in the data bases, namely more exhaustive language description. We have also costed for an about sentence as a component of the enriched characterisations. We consider it essential, given the present status of PRECIS, to provide PRECIS indexing, and have therefore costed for supplying this for the random subset.

Requests

The core characterisation of requests given in the Outline Specification covered an initial need statement, a user keyword lists, a set of controlled language terms, and a full Boolean statement. This specification is inadequate in not indicating whether the Boolean formulation is intended for free or controlled terms, and the presumption must be that two formulations were intended.

This specification is not well geared to the conditions of actual search services, and especially on-line ones. As mentioned above request specifications may not be well-defined units in on-line searching, and individual queries may cover, explicitly or implicitly, more than one form of indexing. A further problem is that it was apparently envisaged in the Specification that one form of of the requests would be used for the regular data base searches, and the others essentially as the additional query forms for retrieving for the pool. But again, as noted, it is unlikely in practice that any one form of request would be generally used.

It is evident, therefore, that the user queries as searched must be simply taken over as a source of information without any attempt to treat them as a well-defined set for controlled comparisons. Such sets must be generated independently. However they may be searched very cheaply, and are welcome as contributors to the pool. Thus following the suggestion made earlier that there should be schedule of query forms for generating the pool, we feel that two separate natural and controlled language fixed request specifications should be provided, derived from the need statement, though whether each should be used for searching in both Boolean and coordination form would depend on the search service facilities available: it appears that both would be practicable.

The enriched specifications proposed for the subset of the primary requests associated with the random document subset would not have any practical function in the construction of the 'ideal' collection, in retrieving documents for assessment; and it might no be convenient to use them for searching as the standard search service might not allow, e.g. the use of term weights. They really have a comparative role, and therefore might be more usefully searched on the 'ideal 'collection when set up. However, we feel that it would be useful to have the various versions of the requests set up initially, and have therefore included a fairly modest cost of doing so. The versions of the requests specifically designed to obtain the pool for assessment, e.g. by searching on high level subject classes, would incidentally contribute to the range of request characterisations, in fact for requests provided only with core characterisations.

8. Citations

There are no data bases other than those of the ISI which contain citations in full. The provision of citation information was regarded as very important for the 'ideal' collection, and we have therefore considered ways of providing it. It seems clear that keypunching of the citation data from source documents would be prohibitively expensive so the only obvious way of obtaining it would be by amalgamation of the information about documents in the regular data bases and that in the science (or social science) citation index. This would be a quite demanding operation involving considerable effort in supplying the means of identifying common documents in the two sources, and heavy computation. But we believe it could be done to an adequate level of accuracy. It should however be noted that non-journal articles in a mixed data base could not be supplemented from a citation index and keypunching of citation information for these would therefore be required. Hand punching would probably also be needed for documents missed in the merge operation, which could probably be identified through the citation counts given in many of the regular data bases. Punching might be needed for as much as 15% of the document collection.

Our view is that the merging process is feasible, but it must be recognised as a major collection-building task. Fortunately, some of the document data processing which would be required to set up a file for matching against the citation file would almost certainly be required in any case as part of the general data preparation operations of the collection building.

It should perhaps be noted that citation information supplied for a short time slice of document literature would not support some types of citation study, since references would typically be out from the document set: i.e. the collection would support investigations of some document links through common citations but not those requiring a set of documents with closed or nearly closed citation links.

9. Summary of data base selections

We may now bring together all the points made in the discussion of data sets for the ideal collection, and in particular see how requirements for the document, request, and relevance sets marry. For all the document sets we confine ourselves to data bases with machine-readable abstracts.

Summary of criteria applied to data bases having machine-readable abstracts

	large enough	solid enough	all index-ing	mixed docu-ments	friend-ly	enough requests	relev-vance sample
Main document set - science							
CA							
CBAC	yes	yes	yes	yes	no	no	yes
others	yes	yes	ves	yes	no	no	yes**
CAB System	yes	?	no	yes	yes	yes	yes
subbases	yes	yes	?	yes	yes	no	yes
INSPEC	yes	?	yes	yes	?	yes	?
Physics	yes	yes	yes	yes	?	yes	?
Elec	yes	yes	yes	yes	?	no	?
MEDLINE	yes	?	no	yes	?	yes	?
Other document set - social science							
CIJE	yes	yes	no	no	yes	?	?
HA	yes	yes	yes	no	yes	?	?
PATELL	yes	yes	yes	yes	yes	?	?
Sociological Abstracts	yes	yes	yes	yes	yes	?	yes
Other document set - mixed documents							
CAB System	as above						
subbases	as above						
Sociological Abstracts	as above						
monographs							
BL MARC	yes	yes	no	n/a	yes	?	?
Other document set - foreign language							
CAB System	as above						
subbases	as above						
MEDLINE	as above						
Other document set - different time							
CA							
CBAC	as above						
others	as above						
CAB System	as above						
subbases	as above						
INSPEC	as above						

** but search linking needed

It will be evident that there is no universally satisfactory data base source of the main document set, and of the social science other set. For the former INSPEC seems the best bet, for the latter Sociological Abstracts. The latter is indeed not commercially available, but might be obtainable for 'ideal' collection purposes.

10. Possible uses of the 'ideal' collection

The investigation of possible uses of the 'ideal' collection if it existed was primarily intended as a means of estimating demand for the collection, and as such is discussed in Part B. But it was also useful in checking whether the proposed specification would in fact be adequate for potential uses. We studied the uses of the collection for research purposes and for teaching separately.

a) implications of possible research uses

The questionnaire used to obtain information on these is reproduced in Appendix 8. We sent out 48 questionnaires and received 27 responses. A few of these indicated negative interest, usually because the person concerned was not considering any research work at all, and some people returned more than one project. A total of 28 projects is considered here. These fall basically into two groups: those based on exploiting the proposed file of documents and requests more or less as provided for a variety of investigations; and those which would involve new processing, for example to supply further indexing or conduct new searches. On the whole both classes of project could be sustained by the collection as proposed (ordinarily by the main document set). It may be noted that for projects of the second type it may be assumed that the collection would be available for on-line searching, though not necessarily with all the frills of an actual on-line service. Some of the projects involve requirements which would not be met by the proposed collection: for example in calling for far more documents. However our opinion is that valid projects with the same objectives as most of these could be carried out with the collection. As noted, some types of citation study would not be feasible. However we have only received one project for which the collection would be quite unsuitable: this is on library circulation and management. Such interests could only be satisfied by very different, and effectively disjointed, data.

b) implications of possible teaching uses

We initially circulated a questionnaire on teaching uses to library schools,and subsequently one specifically on the possible uses of the 'ideal' collection for on-line education. The two questionnaires are reproduced in Appendix 9.

1) The first questionnaire was sent to 17 schools, of which 8 responded. Our analysis of the replies indicates that in general the ideal collection would be satisfactory for the kinds of need indicated, except possibly in the range of index language thesauri etc. available for study. In one case two large collections are needed which could only be supplied if the 'enlarged' social science other set was implemented. There is also a requirement for a large file of MARC records which would not be satisfied (though such an extension of the proposed other set of monographs would probably be easily obtained). In comparison with ON-TAP somewhat less relevance information per question would be available, but far more requests.

2) The second questionnaire was sent to the participants in an on-line education workshop sponsored by BLR&DD. This mainly involved representatives of the library schools, the remaining participants being e.g. representatives of BLR&DD. Responses were obtained from 10 out of the 12 schools. On the whole the indicated on-line education needs of the schools could be met by the ideal collection, with the exception of experience of working with very large files. However to meet these needs two conditions of collection use would have to be satisfied: the collection would either have to be put up on a standard on-line search service, for example as a "private file" on DIALOG, or would have to be embedded in a satisfactory local computer environment. The software implications of the latter would have to be investigated: thus needs for a wide range of search procedures, and for on-line thesaurus consultation, would have to be met.

11. Costs

The costs of the whole enterprise cannot be estimated at all accurately. First, some of the cost information e.g. in Williams and Rouse, is not very up to date. Second, the costs of buying data base tapes and of conducting searches on-line are sure to change between now and when the collection is built, if it is built. Third, as it is evident that the collection building operation would depend to a considerable extent on cooperation with data base suppliers, search service suppliers, and establishments where searches are conducted, detailed costs would have to be negotiated to cover collaboration for this one-off job, and cannot really be established other than on an official basis, by or through BLR&DD. Fourth, it is impossible to estimate precisely what the costs of conducting searches will be.

We have therefore sought to identify the cost components of building the 'ideal' collection, and to provide as good guides as we can to the orders of magnitude of each. We are grateful for help given on these matters by representatives of the various services.

The main cost elements are as follows

1. obtaining document data files
2. conducting searches and obtaining relevance judgements
3. setting up material in an appropriately straightforward and portable standard form
4. staff
5. miscellaneous costs.

These items are considered in more detail below. The figures are approximate only, representing costs more or less as of mid 1977 (with dollars at the rate of 1.50 to the pound), and depend on simple views of staff costs and capacities for such activities as keypunching: we assume keypunching 3K strings like basic document specifications would take 2 weeks.

1. Document data sets

Obtaining these involves

a) the purchase of raw magnetic tapes from data base suppliers for

   1) the main set
   2) the other sets

b) manual enhancement to fill gaps or provide additional information e.g. extra indexing

c) the purchase of citation information.

a) Some difficulty arises in considering tape costs since suppliers figures may refer to large units of a specific file, like one year's intake, and it is not clear whether a random selection to give a set of the required size could be made for a tape for purchase, as opposed to subsequently. A further point is that the figures for purchases listed are essentially commercial, and it is possible that more favourable ones could be negotiated for a non-commercial and non-competitive use such as the 'ideal' collection. Fairly realistic costs can be arrived at from available figures for the main document set, and the social science other set, but the costs of the remaining other sets can only be very rough estimates. The details are as follows.

main set costs		£	£
CA - CBAC for 30K documents, take 1 year (47.5K)		3900	
CAB System for 30K, if can be selected alternatively 1 year (130K)		1620	6240
INSPEC - Physics for 30K documents take ¼ year (37.5K) alternatively 1 year (150K)		850	3350
MEDLINE			
(note: INSPEC figures from Williams and Rouse, therefore old and should be increased somewhat; the others from recent handouts)			
Suppose, therefore, a cost for the set of		3000	
Other set costs: social science other set			
CIJE for 3K documents, take 1 year (20K) (as cheap as less)		85	
HA for 3K documents, take 1 year (7.5K)		900	
PATELL for 3K documents, take 2 months alternatively 1 year (25K)		350	2000
Socialogical Abstracts not commercially available			
(note: all figures from Williams and Rouse)			
Suppose for the set, a cost of		500	
mixed document other set			
PTS INTERNATIONAL STATISTICS or a subset of the above data bases considered or a subset of the data bases considered above,	say	500	
monographs	say	200	
another language other set Keypunching of abstracts, etc	say	1000	
different time other set Keypunching of abstracts, etc.	say	1000	
Note that for all the documents sets, a copy of any controlled language used for indexing is needed. Machine-readable forms of these apparently are or will be available, but copies would have to be purchased: so allow,	say	500	

			£
b)	Manual enhancements, consisting of three lots of additional indexing (regular, sentence, and PRECIS) for the random subset of the main document set are required. It appears that 30 entries a day is a fair rate, so allowing 6 months each for 3 indexers at £4000 each, we have		6000
	plus keypunching	say	500
c)	The provision of citation information would unfortunately probably be expensive, since a whole year's worth would be required to cover the documents in the main set. The cost of one year's source for the SCI is		6700
	plus miscellaneous keypunching	say	500

2. Searches and relevance judgements

Costs under this head involve

a) recording need statements and background information about the user
b) conducting searches
c) printing material for assessment
d) making assessments
e) recording relevance information

These cost elements are most conveniently allocated to three groups, for staff, for computing, and for keypunching.

1) staff:

Following the discussion in the text we arbitrarily assume that searching would be carried out through six centres, both for the main document set and for the other sets. At each one of these centres a searcher would be required to interview the user and conduct searches on his behalf. The searchers will also be needed to devise and search the additional requests designed to contribute to the pool of matched documents for relevance assessment, for the main set; and he will have to administer the selection of the document sample actually to be assessed. Sufficient staff would be needed for 750 requests for the main set and 250 for the social science other set, and for some arbitrary amount of effort for the remaining other sets, depending on their relationship with these sets. For 1000 requests which we may suppose would represent 3 searches a day, we should allow the equivalent of a full time searcher at £4000 for 1½ years, giving

6000 .*

For the remaining other sets we assume the equivalent of a searcher for 6 months, giving

2000 .

We also require staff for the formulation of the extra versions of the requests not necessarily used for searching but forming part of the request characterisations for the request sets. This would perhaps require the equivalent of 6 months work at

2000 .**

* breaks down as main 4500, social science 1500
** breaks down as main 1000, social science 500, others 500

For relevance assessments for the main set, an assessor is required in addition to the 'free' user, for 25% of the requests, say 200. At a presumed capacity for assessment of 2 requests with 300 assessments per diem we have 6 months work

2000 .

Relevance assessment for the other sets is rather costly as it must be exhaustive.

For the social sciences other set we require 6 assessors for the 150-250 requests, which at the same rate as for the main set would amount to £12000. We must assume that a greater productivity could be achieved, and allow

8000 .

For the remaining other sets we again have the difficulty of making an estimate and simply allocate

6000 .

2) computing:

The machine costs are those of the actual searching on-line, and of printing abstracts for assessment. Printing turns out to be rather expensive, and indeed for the numbers of abstracts required more expensive than the searching. Fair estimates of the cost of searching alone seem to be £10 per search, and with £17.5 for printing 250 abstracts, giving a 'standard' search total of £27.5. For the main set of documents, as noted, extra assessments would be required for 25% of the searches, so further printing would be needed. The cost of 'standard' searches for the 750 main set requests would therefore be

21000 .

with extra printing for 200 requests at 3500 .

In addition, for the social science other set, for which extra printing would not be required, we have 250 standard searches, giving

7000 .

For the remaining other sets in their unspecified relation to those mentioned we allow

10000 .

3) keypunching

By comparison with the previous costs, this is a minor item. In particular, since the identity of documents output for assessment will be recorded, punching of relevance information is required only for those documents deemed relevant, say 50 per request. The initial need statement and perhaps user information as well requires keypunching. We thus have

punching need statements	500
punching relevance information	500

3. Data processing

The computer costs involved in the data processing for the collection may be divided into two elements, 'external' and 'internal' to the Building project. The external costs are in fact those covered by the purchase prices of the document data magnetic tapes, and of conducting searches, and have already been listed.

The internal costs are those involved in processing the acquired data to combine material from different sources, and to set up the collection in some kind of 'standard' form. We may divide these costs into those required for the regular processing, and those of citation data handling. The amount of computing would be fairly heavy. We assume that it could be done in an academic environment, and at academic cost rates. Using the Cambridge University 370/165 charges of £250 per hour, and our experience with the UKCIS material, we may allow

for regular processing, 12 hours	3000 .
for citation processing, 6 hours	1500 .

4. Staff

The 'dispersed' staff handling searches etc. have already been costed. We consider here the essential project staff. The detailed organisation would depend on actual persons available and their location, but we may assume

Senior Project Director at £10000 pa, 1/3 time, for 2 years	7000
Chief Programmer at £6000 pa, full time, for 2 years	12000
Junior Assistant at £4000 pa, full time, for 2 years	8000 .

5. Miscellaneous

The essential items here are

Xerox of at least the random subset document texts

travel

documentation

general overheads

for which we may allow 7000 .

Summary of costs

We summarise the costs as follows:

A = gross cost, all options covered
B = as A, without citations
C = cost for the main set and one social science other set
D = as C, without citations

	A		B		C		D
Documents							
main	3000		3000		3000		3000
soc. sci.	500		500		500		500
others	2700		2700				
thesauri	500		500		500		500
enhancement	6500		6500	say	4000	say	4000
citations	7200				7200		
	20400		13200		15200		8000
Searches and assessments							
a) staff: searchers main, soc. sci.	6000				6000		
others	2000				2000		
extra queries	2000						
assessors main	2000		as A		2000		as C
soc. sci.	8000				8000		
others	6000						
	26000		26000		18000		18000
b) computing: main	24500				24500		
soc. sci.	7000		as A		7000		as C
others	10000						
	41500		41500		31500		31500
c) keypunching	1000	say	800	say	800	say	800
	68500		68500		50300		50300
Data processing							
regular	3000		3000	say	2000	say	2000
citations	1500				500		
	4500		3000		3500		2000
Staff	23000	½jun	19000	½jun	19000	½jun	19000
	23000		19000		19000		19000
Miscellaneous	7000	say	6000	say	6000	say	6000
	7000		6000		6000		6000

The grand totals are therefore:

	A	B	C	D
Documents	20400	13200	15200	8000
searches and assessments	68500	68500	50300	50300
data processing	4500	3000	3500	2000
staff	23000	19000	19000	19000
miscellaneous	7000	6000	6000	6000
	123400	109700	94000	85300

These figures are daunting. It is possible that we have underestimated some computing costs, but overestimated people costs for searching, assessment and keypunching. Further, the figures take no account of any special arrangements which might be made for a project of this kind through BLR&DD, which might well be expected to reduce the costs noticeably.

Work programme

A detailed building programme whould be embodied in a building proposal. However we think it would be useful here to summarise the kind of programme which appears to follow from the Design Study investigations described above.

We assume that the project itself, which would be a two year one, could only start after suitable agreements had been made with the data base and search service suppliers. A natural sequence of work would then seem to be:

1)	set up search operation plan machine processing (conduct statistical analysis through consultant, not costed above)	3 months
2)	start to conduct searches commence data base processing	6 months
3)	start to incorporate search results start to merge citations provide alternative indexing	6 months
4)	finish searching finish citations set up standard collection	6 months
5)	document	3 months
		24 months