REPORT ON THE NEED FOR AND PROVISION OF AN 'IDEAL' INFORMATION RETRIEVAL TEST COLLECTION

K. Sparck-Jones
C J Van Rijshergen

University Computer Laboratory
Cambridge

December 1975

Preface

The preparation of this Report was supported by a grant from the British Library Research and Development Department. The Report was used as a discussion document for a Workshop held on December 11-12, 1975. The authors are very grateful to the Workshop participants for their stimulating reactions to the Report, and for their favourable response to the suggestion that an 'ideal' test collection of the kind indicated should be constructed as a material aid to retrieval research over a wide area.

We have not attempted to incorporate the many comments made into the final version of the Report, as this would have effectively required a wholly new document. We are instead issuing the Report as prepared for the Workshop with only minor clerical corrections, in the hope of receiving further comments from other potential users of the 'ideal' collection, which could be input to a more detailed design study for the collection construction.

K.S.J.

C.v R.

December 1975

0. SUMMARY OF REPORT, AND RECOMMENDATIONS

Summary

This study

a) investigates the need for an ideal test collection(s) for information retrieval research;

b) discusses the requirements it should meet;

c) outlines the characteristics it should have; and

d) considers the administrative implications of setting it up and maintaining it.

The study is in three parts:

1. deals with the need for, and properties of, the ideal collection(s);

2. deals with organisational aspects;

3. deals with (roughly) estimated costs.

The Appendices provide details of existing collections.

Our conclusion is that there is a genuine need for a well-designed multi-purpose test collection. The least collection satisfying these needs would consist of a large document set with core properties, having several small, enriched, subsets, and a number of associated collections comparable with the subsets in size and having other properties. Higher-grade ideal collections would provide more alternatives, large and small. At least some of the basic material could probably be obtained from existing services or projects, but this would certainly have to be supplemented.

The ideal collection(s) could be set up by a one-off project, but it must be maintained and made available to research workers, and some person or organisation is required to do this. The collection itself should hopefully allow a large range of uses, and while the primary intention is that it should be made available to different projects, it could also benefit research through in-house exploitation by the holding organisation.

Very rough cost estimates suggest that the collection could be set up for between £25K and £40K, depending on how much use could be made of existing data and the grade of collection chosen; and it could be maintained for between £25K and £10K p.a., giving a total of between £25K and £50k for a minimum useful maintenance period of five years.

Recommendations

This document is primarily for discussion, and so presents alternatives at many points. We nevertheless feel that it is useful to provide specific recommendations based on our survey. These are

1. that an ideal test collection be set up to facilitate and promote research;

2. that the collection be of sufficient size to constitute an adequate test bed for experiments relevant to modern information retrieval systems; and that it be characterised in a large number of ways suited to different needs. Specifically, that it consist of one or more large document sets with core properties, each with several subsets enriched in different ways, and each accompanied by other collections comparable in size and richness with the subsets;

3. that the collection(s) be set up by a special purpose project carried out be an experienced worker, called the Builder;

4. that the collection(s) be maintained in a well-designed and documented machine form and distributed to users, by a Curator;

5. that the curating project be encouraged to promote research via the ideal collection(s), and also via the common use of other collections acquired from independent projects.

1. THE NEED FOR, AND CHARACTERISTICS OF, AN IDEAL TEST COLLECTION

Introduction

At the recent Workshop on Automatic Indexing (Sparck Jones 1974, Wilson 1974), it became apparent that there was a need for a well-founded information retrieval test collection. This is not to say that this need was not recognised before; it certainly was, but retrieval research has fortunately not been significantly hampered by the lack of such a collection. It has nevertheless now become clear that future research in information retrieval will require better test data than is currently available. The purpose of this report is to say why an 'ideal' test collection or set of collections* is needed, to characterise it, and to consider how it might be provided.

We believe that research in information retrieval is entering a new phase. Perhaps the best way to describe this new phase is to characterise the phase now ending as one of pilot studies. The last fifteen years have seen the publication of many preliminary and isolated results. These results were frequently obtained with data-bases which were not quite right for the kinds of experiment concerned. But since these were the only available data-bases, the tests were the best that could be done; and the experiments were therefore performed and reported, sometimes, though not often enough, with suitable caveats. Other research workers might want to use the same data, to compare their own results with the earlier ones. For this alternative descriptions of the data might be required, which were not readily available. The later workers would thus be driven to other collections to satisfy their particular requirements, making proper comparisons with the previous work impossible.

It is now time that many of the 'mini' results reported so far are incontrovertibly established (or rejected). The major stumbling block seems to be the lack of suitable test data. There is a widespread feeling among research workers that existing test collections are inadequate because they are small and/or careless and/or inappropriate. They may also not be fully machine-readable, or may be in an esoteric machine format.

At present, too many projects are working with different collections. This leads to unnecessary data preparation effort by each project. It also accounts for much of the project disconnection which exists. This makes it difficult to correlate the results obtained by different projects. Further, since the use of a collection is limited by the time span of a single project, data tends to get lost, particularly where it has been temporarily extracted from an operational system.

Further, the recent growth of on-line retrieval services based on large data-bases has changed the conditions and character of information retrieval in many ways, and there is a new requirement for test collections suited to experiments relevant to these services.

* abbreviated to "ideal test collection(s)".

This state of affairs suggests a need for an 'ideal' test collection(s), namely one satisfying requirements for

commonality between projects,

hospitality to projects,

adequacy for projects, and

convenience in projects.

The rest of this Section of the report considers the detailed implications of these requirements.

Levels of collection description

An information retrieval collection is a set of documents, requests, and relevance judgements. Sometimes it may be more convenient to talk of test data than a test collection, but these are to be taken as more or less equivalent.

Past test collections referred to in this report are ones which have been used mainly for experiments in computer-based information retrieval. The ideal collection(s) is also designed primarily for research in mechanised indexing and searching, and some low-level requirements and preparation activities are associated with the provision of an easily-used computer-based collection. The collection itself could, however, in principle be used for manual experiments. We have not pursued this possibility here, though it should be noted that listings for manual use are a simple by-product of machine processing. The distinction between manual and automatic collection of data is quite different, and the extent to which the collection will be characterised automatically appears in our detailed specification.

We discuss collections at four levels:

1) real : this refers to the entire documents, requests and relevance judgements, in their full detail; however in many cases the entire document texts have never been explicitly considered, all processing like indexing, the making of relevance judgements etc. being based on e.g. abstracts.

2) material : this refers to the form of the documents, requests and relevance judgements actually input for considering or processing (i.e. directly used for tests).

3) keyed : this refers to the machine readable form of the material data.

4) formatted : this refers to the keyed data after standardisation and clerical manipulation for easy use. It does not ordinarily refer to collections specially formatted e.g. for operational IR systems, but is represented by KSJ's "standard collections" where standardisation is represented by stemming and clerical manipulation by the use of numbers for words and the systematic provision of a set of transformed data files in a consistent format. (For details see Appendix D.)

Past test collections

Since both research experience and the justification of demands for the ideal collection(s) have been materially influenced by past and existing test collections and their properties, it is useful to summarise salient facts about the more important of the collections used for evaluation tests, primarily of mechanised indexing and searching. The British collections are:

1. Cranfield 2
2. Inspec
3. ISILT
4. UKCIS
5. Medusa
6. NPL
7. Olive, Terry and Datta's

Details of the collections and the projects exploiting them are given in Appendix A, under headings including

collection size and subject
project objective
mode and source of indexing
form of relevance judgement, etc.

The striking feature of this set of collections is their incomparability: there is no one form of index description common to all; there was great variation in the environmental conditions; the projects had quite distinct objectives; and they presented their results in very different ways.

We have confined our detailed analysis to British collections because these collections are sufficiently representative, and because information about them is more readily available. For reference, non-British collections of any status include those used by the Smart Project, by Lancaster, and by Jahoda. Summary information about these collections is given in Appendix B. Several of the British collections have been used by more than one project, and some have been processed for easy machine handling. Some American data has also been shipped about, and the SMART Project collections in particular are presumably available to interested parties. However most collections set up to date have not been easy to use or widely exploited.

The reports describing major experiments conducted with the British collections (for references see AppendixA) say very little about the detailed design of the collections. In general it seems as if they were designed by default. It is true that the composition and nature of the test data, for example how relevance was assessed, may be described in some details But there is very little discussion of possible choices, say of index description source, or indication of why one particular choice, for example of abstract, was made. In some cases, of course, the choice was dictated by circumstances; but even then its implications may not be explicitly considered. But perhaps this lack of design is not surprising, since the prime concern of the

experiments has been the testing of some major variable affecting efficiency or effectiveness, in a given environment. The same seems to hold for non-British collections.

Following up our study of the literature, we have discussed specifications for an ideal collection(s) which would meet both needs arising from present lines of research and those likely to arise in the future, particularly in connection with on-line searching, with research workers in the UK as follows: Mr. Aitchison, Miss Barraclough, Dr. Brittain, Miss Horsnell, Mr. Keen, Dr. Leggate, Professor Lynch, Mr. Robertson, Professor Vickery and Dr. Wyatt.* We are very grateful to them and to Dr. Holmes of BLRDD for their help. We also sought suggestions from Professor W.S. Cooper, Professor B. Griffith and from Professor Salton. The discussion of collection requirements which follows is based on the experience and predictions of ourselves and our fellow workers.

We emphasise that we have attempted to be forward looking. It is apparent in particular that it is most important that the ideal collection(s) should be a means of relating valid abstract studies of information retrieval and those of operational systems and user behaviour. These both imply a large test collection, with some properties not manifest in existing collections. The specification of the ideal collection(s) are neverthless necessarily derived mainly from experience with past collections, and we therefore make no apologies for references to these. In general prediction can only be based on previous findings. In information retrieval research in particular, past results have been so fragmentary that some future research must be concerned with validating them. At the same time it is clear that new research topics are arising, for example in connection with on-line searching, the availability of really powerful computing and communications facilities, the development of retrieval networks, and so on, for which suitable test data must hopefully be provided. We have therefore sought to specify a fairly 'open' ideal collection(s).

Areas of interest

General areas of interest likely to be of study interest to users of the ideal collection(s) are:

A. relative to collections and their users:

- text populations
- document populations
- source (i.e. journal) populations
- origin (i.e. author, organisation, country) populations
- citation populations (differentiated under the preceding 3 heads)
- request populations
- user populations
- need populations
- expert populations (e.g. indexer, searcher)
- vocabulary populations, natural and index
- language populations, natural and index
- description populations
- catalogue populations
- input populations
- subject populations

---

* We were unfortunately unable to contact Mr. Cleverdon at the relevant time.

B. relative to computer manipulation:

file populations
network node populations

C. relative to economic management:

It is not easy to say exactly how economic questions could be studied with a test collection which is necessarily abstracted from ordinary use. However we think a large test collection could be used to provide comparative information for specific costing studies, and more importantly, since many facts about it will be known, as a means of validating simulations of some library management operations.

Requirements

We can broadly distinguish two kinds of requirement to be met by an ideal collection(s). The first is to ensure the validity of experimental results. The second allows for the control of variables affecting retrieval performance. The control of one variable may not be compatible with that of another. For the moment we will ignore the implications of this incompatibility for the design of test collections, and simply list all the requirements that an ideal test collection(s) should independently meet.

A. General requirements, concerning the sets of documents, requests and relevance judgements

i) substantive requirements re these sets

The ideal collection(s) should be:

1. large

re documents < 500 documents are of no real value

1-2000 documents are minimally acceptable for some purposes

> 10000 documents are needed for some purposes

re requests: < 75 requests are of no real value

250 requests are minimally acceptable

> 1000 requests are needed for some purposes.

reasons: real collections are large
statistically significant results are desirable
scaling up must be studied

(Note that request and document set sizes are not necessarily correlated).

2.1 various in content

documents and requests should cover a range of subjects of varying content and 'hardness' e.g. science, social science, news

reasons: real collections are heterogeneous
consistency of devices must be tested by comparison

2.2 homogeneous in content

documents and requests should cover one subject intensively

reasons: real collections are homogeneous
discrimination of devices must be tested by exhaustion

2.3 various in type

documents should be of different types e.g. popular, specialised, survey, review, patent; requests e.g. broad, narrow

reasons: similar to 2.1

2.4 similar in type

documents and requests should be of the same type

reasons: similar to 2.2

2.5 various in source

documents should cover a range of journals and journal types

reasons: similar to 2.1

2.6 homogeneous in source

documents should cover one or a few similar journal types in depth

reasons: similar to 2.2

2.7 various in origin

documents should represent different author origins and status; requests should represent different users and needs (link relevance)

reasons: similar to 2.1

Collection specifications

A. To satisfy general requirements for set of documents, requests and relevance judgements

a) substantive

There should be at least 2 large collections for an

arts
science subject area respectively.

These should each cover variations in

content, type, source, origin, time and language.

They should preferably be taken from an operational system, i.e. both documents and requests should be thus taken, and accompanied by genuine user relevance judgements.

It should be possible to extract from each of the large collections one or more small subcollections which are homogeneous with respect to

content, type, source, origin, time and language.

The subcollections should be operational as far as possible, but it is highly probable that some requirements e.g. for alternative indexing etc. can only be met by design.

b) formal

The large collections should be various in formal properties, and it should be possible to extract homogeneous subcollections.

B. To satisfy requirements concerning individual documents, requests and relevance judgements

The detailed specification of core and enriched properties for the ideal collections primarily refers to individual documents, requests, and relevance judgements, rather than sets of documents, etc. Our choice of properties from the full lists of pp./(to which the numbers refer) is as follows: 11-13

C = core; E = enriched

Documents

C. Documents in large collections should be represented by

2 abstracts
3 titles
4 free keywords, from abstracts
6 citations)
7 author and bibliographic elements) abbreveiations to be avoided
8 thesaurus or subject indexing, if available

Indexing should be by one simple indexer, and one expert, for 4
by one expert, for 8.

E. Documents in small collections should be represented core plus

4 free keywords, from text, title, to different exhaustivity
5 free sentence
8 thesaurus if not in core, and other controlled languages as available.

Indexing should be by various simple indexers, and various experts, for 4
by one simple indexer, and one expert, for 5
by various experts, for 9

C. Requests

Requests in large collections should be represented by

1a) verbal text
b) coordinated terms
c) Boolean formulation (which could consist of a cumulative log of an on-line search)

Indexing should be by one user, and one expert, for (b) and (c).

E. Requests in small collections should be represented by core plus

1d) terms with weights
e) edited forms of b, c, d
f) modified forms of b, c, d
2 source documents
3 verbal text from source documents.

The total record of any request eliciting procedure should be preserved. For example if a user is asked to mention appropriate known documents, these should be indicated.

Note that while some experiments relevant to on-line searching could exploit requests formulated during previous on-line searches, in other cases new searches would be required; for the latter the ideal collection(s) would provide an adequate set of documents.

Relevance judgements

C Relevance judgements in large collections should allow

2 grades (highly, partially)
one user, and one expert, as judges.

It is unlikely that exhaustive relevance judgements could be made for large collections; however some attempt must be made, e.g. by additional searches, to estimate recall.

E Relevance judgements in small collections should allow core plus

more grades, types
various users and various experts.

Collection recommendations

The general problem is that these ideal specifications may have to be tempered by realism. The exact way in which the requirements listed above can be met must be determined to some extent by what is available in operational systems. A specific problem is that while some data may be available in an operational system, it may not be in machine readable form. In general one might hope to extract material with most core properties from an operational system, but keying of items like abstracts must be allowed for. Much of the data for enriched collections would have to be specially supplied. Clearly, any project to set up the ideal collection(s) would have to have an initial phase for a detailed study of data sources.

Our specifications suggest the following as useful but realistic collection sizes:

large	10-30000	documents;	500-2000 requests
medium	2-10000	"	
small	500- 1000	"	200- 500 " .

The main problems are clearly those of satisfying the core requirements for the large document sets which are needed for some purposes; and of ensuring that small collections are experimentally valid while not making them too large for the capacities of independent projects which might contribute them.

Similar recommendations are needed for numbers of collections. In particular, since choices of property requirement can be combined in different ways, it is convenient to distinguish three grades of ideal collection(s) which might be built. They are

1 best,
2 acceptable, and
3 least.

Which grade is achieved is determined largely by the ease with which enriched property descriptions can be supplied by operational systems. It is likely that many will have to be supplied by design. In detail, the grades are as follows:

1 best	2 large collections, each with 3-5 subsets having substanial property overlaps, of which 3 are designed and 2 are selected subsets. 3-5 other small collections to complement these.
2 acceptable	2 large collections, each with 3 subsets having some overlap, of which 1 is a designed and 2 are selected subsets. 2 other small collections.
3 least	1 large collection, with 5 subsets having some overlap of which 3 are designed and 2 are selected subsets. 2 other small collections.

These subset specifications do not include ones which could be selected by purely clerical operations, e.g. ones representing all the articles from a specified journal or requests with the same number of terms. Some such selections could easily be made initially, others to order.

Even the least collection would be of great value to research, particularly if it was supplemented by collections from other projects, especially if these were of the good quality which might be achieved by 'bulge' funding. In addition, if the collection was primarily extracted from an operational system, it might be encouraged to grow through the operational system. This would clearly be a very satisfactory way of meeting many ideal collection needs.

Collection form

As noted earlier, it is intended that the ideal collection be machine held. This means that the main collection data is machine held, and in a convenient form. Referring to the categorisation of collection levels on p. 4 , it is clear that some real information , like full document texts, could hardly be keyed; but it should preferably be held in microform. The material collection must, however, be keyed. It must be supplemented by adequate backup information and documentation a) characterising the content of the collection and how it was set up; and b) detailing any processing applied to bring it from level 3 to level 4, and its format at level 4.

It is perhaps not reasonable to require that other projects, even when funded by BLRDD, should provide total information about deposited collections. But these collections should meet some minimum standards of content and format, to make them sufficiently comparable to the core-characterised ideal collection(s); and they should be suitably documented. In principle collections set up by projects not funded by BLRDD might be of value to supplement the ideal collection(s); it might of course not be possible to obtain such material in the desired form, but even raw magnetic tapes and primary documentation should be sought.

2. ORGANISATIONAL ASPECTS

There are two questions here:

1) the ideal collection(s) must be set up, by someone whom we will call the Builder;

2) the collection(s) and possibly other generally useful ones must be kept so that they can be made available to research workers, by someone we will call the Curator.

These are distinct activities, so Builder and Curator need not be the same person. The division between their concerns comes with the provision of the ideal collection(s) at formatted level 4; this could be either the final stage of the building project, or the first stage of the curating one. We emphasise that there is little point in setting up the ideal collection(s) unless proper management and maintenance is provided for. Some organisation is required even to provide tape copies of the level 4 formatted collection. But we believe that the Curator could have the more positive function of stimulating research through the use of the collection(s).

It will be clear that both setting up and maintaining the collection(s) are non-negligible enterprises. The implications of the specifications outlined in the previous section, and possible ways of setting up and maintaining the collection(s) are discussed below. We necessarily assume that funding sufficient for the least ideal collection(s), is available. The higher grade collection(s) of p19are clearly more attractive to the research community, but we should not claim that they are necessary for the well-being of the community. In particular, since the cost of setting up collections involving different primary document sets must be largely additive, the grade to be chosen depends primarily on BLRDD's willingness to provide funds. We think that a very good case can be made for BLRDD's supplying the least collection, both to reduce the cost of individual projects and to promote the research that information retrieval needs; and since managing this collection and supplying it to users is a not wholly trivial task, some committment to the future maintenance of the collection from BLRDD is also required.

The Builder

We do not think that this is the place for nit-grit recommendations as to exactly how the ideal collection(s) are to be set up. This will depend in part on the level of funding, and in part on how far suitable input material exists in current operational systems or has been assembled by research or development projects. However we feel that the general approach to setting up the ideal collection(s) is independent of such specific considerations. The important points are as follows.

Even if ideal collection building is funded only to achieve the least output, the degree of control required, and effort involved, are considerable.

2.8 homogeneous in origin

documents and requests should represent one kind of author and user

reasons: similar to 2.2

2.9 range over time

documents should be of different dates; requests should be of different dates both for different users and the same user

reasons: similar to 2.1

2.10 coincide in time

documents and requests should be contemporaneous

reasons: similar to 2.2

2.11 various in natural language

documents should be in different languages (or at least their titles should, in which case translations should be provided)

reasons: similar to 2.1

2.12 homogeneous in natural language

documents should be in one language

reasons: similar to 2.2

Globally, it should be possible to use the ideal collection(s) to investigate or simulate

retrospective searching	i.e. one request against all documents;
SDI searching	i.e. a repeated request against successive document sets;
iterative searching	i.e. a modified request against some or all documents;
multifarious searching	i.e. a request, modified request or set of requests against multiple document sets.

It should also be possible to use the collection(s) in studying the interfaces between components in a mixed system incorporating, for example, data retrieval, fact retrieval, document retrieval and computer-aided instruction. This may be called hybrid searching.

ii) formal requirements of document, request and relevance judgement sets.

1 documents and requests should be variable in

real length
material length (i.e. index source length)
index length

reason: to test consistency

2.2 documents and requests should be homogeneous in

real length
material length
index length

reason: to test discrimination

It is assumed that appropriate parallel substantive and formal properties of relevance judgements will follow naturally if the above specifications for document and request sets are met.

B. General requirements re individual documents, requests and relevance judgements.

i) substantive, re documents

Document representation

It should be possible to use or study

1. full text (this should be preserved even if not keyed to allow for future new indexing, linguistic studies and question-answering experiments)

2. abstract a) as is
   b) all non-stop keywords, stemmed

3. title a) as is
   b) all non-stop keywords, stemmed

4. free extracted keyword or keyword string indexing

   1) from full text)
   2) from abstract) a) as words b) stemmed
   3) from title)

   where in general if 1 has exhaustivity x
   2 has exhaustivity x and also y>x
   3 has exhaustivity x and also y and also z>y

5. free quasi-extracted sentence, i.e. a single unit sentence incorporating extracted keywords

6. citations a) in full detail
   b) in short code

7. author and other standard bibliographic details

8. controlled indexing, including broad subject codes

   1) using any standard existing thesaurus for the field

   (and or classification, as many as readily to hand)

   1 from abstract
   2 from title

9. probabilistic indexing (using keywords)

10. usage statistics

Document indexing

Indexing should be carried out

re 4

a) by a simple indexer; by one indexer at different times;by different indexers
b) by an expert " " " " " " " ;
"

(also perhaps re 5) by expert consensus

re 8 by an expert " " " " ;

by expert consensus

Request representation

It should be possible to use or study

1 verbal (given the same source text request)

a) running text
b) simple coordination formulation
c) full Boolean formulation
d) terms with user weights
e) edited after consultation forms of the above
i.e. pre search, with librarian
f) modified forms of the above at end search, with
recorded history of subsearches, changes etc.

i) off-line
ii) on-line

2 source document as request

3 verbal (as above) from source document

a) where source document is relevant
b) where source document indicates area of interest but is not
necessarily specifically relevant

Request indexing

Indexing should be carried out

1 by user; by user at different times

2 by expert; by expert at different times; by different experts,
by expert consensus

This indexing may be done with a specific relevance need in mind; if so, this should be indicated with the query. Any other germane background information should be recorded.

Index language

Ensure available, i.e. preserved, even if not used, if relevant language exists at time collection(s) is set up.

1 thesaurus

2 classification

3 switching language

Relevance judgements

Ideally these should be exhaustive. But if not some attempt should be made to carry out independent searches using any available information and device, to obtain a pooled output for more broadly based relevance judgements than may be obtained only with simple user evaluation of standard search output. In this case some estimate of the recall sample should be attempted.

It should be possible to separate

1) grades e.g. highly, fairly

2) types e.g. novel, stimulating

of relevance judgement.

Judging should be done by

1 one user; one user at different times; one user specifically sequentially

2 one expert; one expert at different times; several experts; expert consensus

Exclusions

The following do not seem to be called for:

1 books as documents
2 'non-literary' items for documents e.g. technical record specifications, simple data records (e.g. stock, personnel)
3 verification-type requests e.g. for publication dates
4 material in esoteric character sets
5 legal data

Other collections

The provision of a new test collection(s), even if ideal, will not make existing collections redundant. This is in part because a good deal is known about some existing collections, so they may be useful test beds for new ideas. Some may also be of value for making comparisons with the ideal collection. It must also be recognised that the ideal collection(s) is unlikely to meet every research need, and that future collections associated with specific projects may be created. Thus in the future we should allow for

a) some further comparison between existing collections;
b) some comparisons between existing and new collections;
c) some comparisons between existing collections and the ideal collection;
and d) comparisons between new collections and the ideal collection.

These projections imply that steps should be taken to relate new collections, in particular, to the ideal collection(s). They should be regarded as a means of extending the ideal collection(s).

The ideal collection(s)

It is obvious that the listed requirements for the ideal collection(s) are considerable.

In some sense they cannot be provided within a single collection, unless this is no more than a mere aggregate.

The following pairs of collection requirements are particularly important:

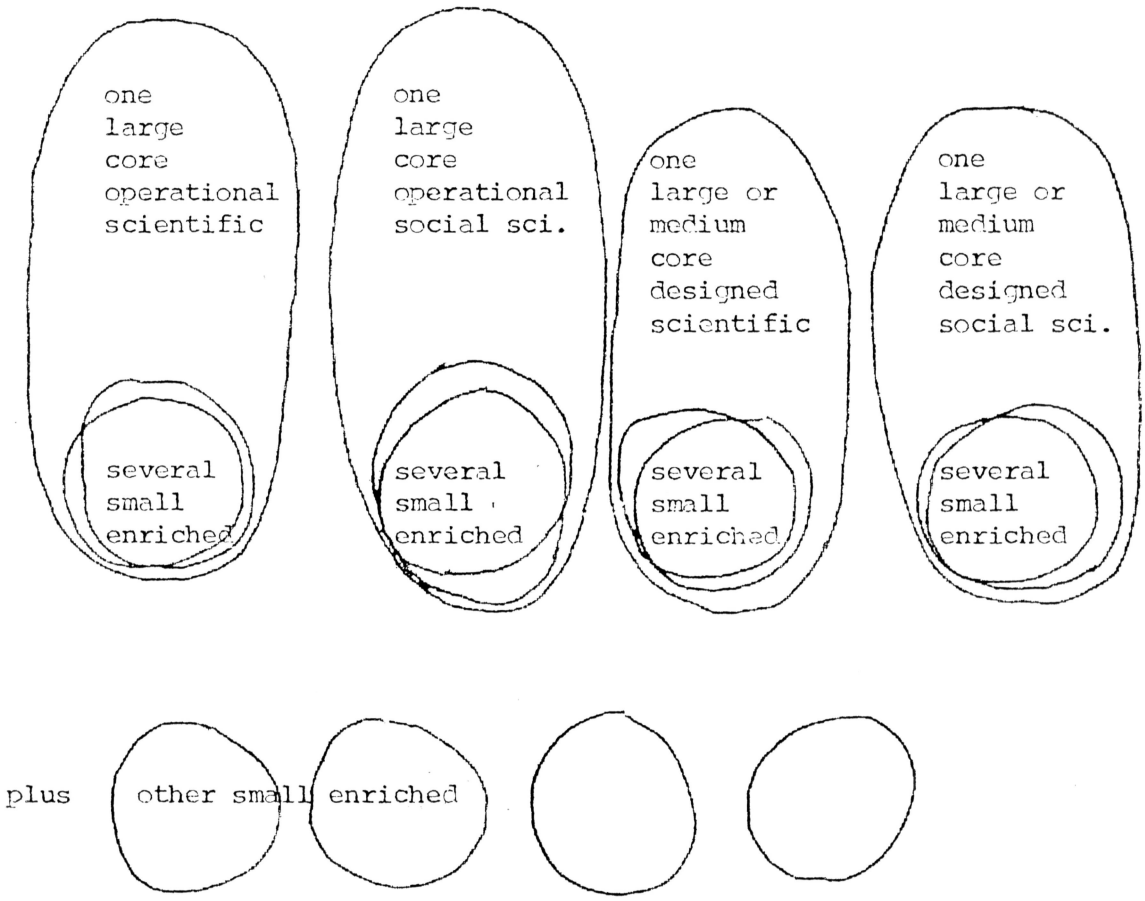
1. The need for sub and super collections;

2. The need for one and several collections;

3. The need for operational and designed collections.

Thus experimental needs are in fact for different collections which can be related to one another, and which have specific properties. Realism suggests that it may be impractical to seek to ensure that each such collection has the maximum set of (compatible) properties (e.g. all variations on the relevance judgement theme), and further that it is unlikely that such collections with all the requisite properties can just be pulled out of operational retrieval systems. It appears more practical to think in terms of large, not necessarily completely characterised collections, with richer small subsets, selected as far as possible from operational systems, but supplemented where necessary by deliberately designed information (e.g. further sets of relevance judgements, index descriptions etc.). The former have 'core' properties while the latter are 'enriched'.

This suggests something like the following will turn out to be needed:



all large, medium, small — collections respectively to be of comparable size in numbers of documents and requests

The following sections work this scheme out in detail

Core and enriched forms of collections

'Core' refers to essential properties possessed by all ideal collections and subcollections; 'enriched' refers to additional properties. Some core property requirements are readily satisfied even for large collections: the problem is to specify a set of core requirements which are both useful for retrieval experimentation and realistic for large collections. Some enriched property requirements are very exigent: it is perhaps unrealistic to suppose that all compatible ones can be satisfied for every subcollection; on the other hand it would be nice if different subcollections of a large collection had more in common than their all being subsets of the same set, with core properties. If possible, some overlap in enriched properties should be provided, to allow for valid comparisons and extrapolations.

This suggests

an experienced head
a project of 1½ - 2 years
a cost of £25-30k (ball-park figure).

We see the project as having three phases:

1) Design study, to by carried out by the Builder as an initial short investigation.
This would survey existing operational or experimental services, and also test collections, to see how they might be exploited to provide input; and it would discuss the mechanisms for collecting the detailed data, with specific cost estimates.

2) Data assembly

This would involve the extraction and bringing together of data from services, and the provision of new data, e.g. alternative indexing, relevance judgements, etc.

3) Machine input

This would include keying the raw input material and applying any appropriate basic transformations to material already in machine readable form. The boring but non-trivial job of raising this level 3 material to formatted level 4 would be done either by Builder or Curator, according to the resources available.

4) Documentation

This would cover a full account of the source material and the way it was collected, with notes on the keying conventions. Level 4 processing if done would require documentation to match.

The most important requirement of the Builder is that he should be experienced in setting up and using test collections. It would clearly be ideal if Builder and Curator were one, but this is perhaps too much to hope for. If they are not, it is most important that there should be adequate liaison between Builder and Curator, perhaps in phase 2, and certainly in phases 3 and 4. A suitable mechanism might be to have the Curator as a consultant on the building project. Since bringing the collection up to level 4 could be done under the maintenance project, it is not necessary that the Builder have direct access to powerful computing facilities. Keying of raw data could be done by a bureau, and basic transformations of material selected from machine-based systems could be done either by the supplier under contract, or a bureau, or by the Curator.

Since willingness to do the job, and the necessary experience, are more important than anything else, we do not feel obliged to specify the Builder's locale. He could be

an independent research worker;
a member of an existing retrieval service organisation;
a member of a consulting establishment like Aslib;
on the staff of, or associated with, BL.

The Curator

As mentioned above, a minimal view of the Curator's activities would imply that he did no more than hold the established ideal collection(s) and distribute magnetic tapes and descriptive documentation. However other activities for the Curator are implied by the suggestion in Section 1 that the ideal collection(s) might be supplemented by other project collections. The Curator's brief could therefore include the following:

1. Maintaining and distributing the ideal collection(s).

2a. Obtaining existing reasonably solid test collections, if necessary vamping up at level 3 and processing at level 4.

b. Acquiring new collections from individial projects, particularly if BLRDD requires or encourages deposition; if necessary vamping and processing.

3. Carrying out (documented) benchmark retrieval runs; gathering basic, e.g. statistical information about collections.

In terms of day to day operations these activities would imply:

1. holding, over a long period;
2. obtaining, and vamping/formatting;
3. clerical processing e.g. of magnetic tapes;
4. providing and distributing documentation and advice;
5. carrying out simple experiments and counting.

Effective curating over this range of activities would require a Curator experienced in both retrieval work and computing, and fairly powerful machine facilities, and would depend on relatively long-term support at the appropriate level. But it must be emphasised that the ideal collection would probably have a long life, so a long-term commitment of funds, even if maintenance is only on a low level, is needed.

Again, it is not for us to recommend a specific organisational setup for the Curator. The following are alternative possibilities for BL:

1. entering into a non-personal contract with a computing service (commercial or university), for the provision of tape copies, etc; or a similar contract with a retrieval service;

2. establishing a personal Curatorship attached to a Library School, Computing Department or Retrieval Service;

3. establishing a curating project with specified Curator, attached as in 3, with the intention that this should act as a focus of research;

4. setting up an institution with Curator, with intention as 3.

The first of these would almost certainly not promote the fullest use of the ideal collection(s), aid the assembly of other collection material, or ensure benchmark testing. The fourth is objectionable as very expensive and liable to be a white elephant. The second and third

alternatives seem the best bets. In particular they would promote the use of the ideal collection(s) as a focus of research, and hopefully prevent the mere accumulation of dead material. Assuming a more than minimal tape-copying service, these alternatives would imply something like a half-time Curator and half- or probably full-time programmer of some calibre, with suitable support, i.e. an annual cost of between £5K and £10K = £50K over five years (not including machine time). A deliberate attempt to encourage extended research using the collection(s) in association with curating would imply higher costs.

Advisory panel

Since building and curating the ideal collection(s) are significant projects, we advocate a panel or steering committee with the following functions:

a) advising the Builder and Curator on project operations;
b) maintaining technically acceptable standards of data management and distribution;
c) encouraging collection use by advertisement;
d) vetting proposed uses;
e) ensuring general continuity.

3. COST ESTIMATES

The cost estimates given in the following pages should only be taken as ball-park figures. The difficulties in giving accurate estimates now, are

1. Insufficient data on which to base estimates,
2. Data out of date (1971),
3. For some parts commercial rates will not apply.

The figures used are mainly based on a report by Peter Vickers (1974). There is an additional difficulty in allowing for inflation. Although in general a 30% increase may be applicable to the costs quoted we have not adjusted them for the simple reason that in some cases the cost (e.g. computer processing) has gone down. Rather than try and estimate the trend of the cost associated with each item we have stuck with 1971 prices. We also give the raw data on which our estimates are based (taken from Vickers, 1974) in Appendix C. We only give detailed costs of the Building phase of the operation since the costing of the Curating phase depends heavily on what is actually implemented. We do however list some of the major factors determining the cost of the Curating phase.

We ignore the cost of housing the projects and the fact that some of the costs may be borne by separate small projects.

Building (all figures in US dollars; halve for pounds)

The reason for giving most of the costs in US dollars is that we wish to maintain comparability with the figures in Appendix C.

a. Documents

cost of buying a data-base of some 50000 items from an operational system.

Tape with citations	750
+ descriptors	1500
+ abstract	2500
Low-level reformatting	1000

Data prepration at .05 cents/char.

If we have to keyboard per item	50000 docs.	30000 docs.	10000 docs.
500 chars (e.g. index terms)	12500	7500	2500
1000 chars (e.g. abstracts)	25000	15000	5000
2000 chars (e.g. everything)	50000	30000	10000
Proof-reading is about half the keyboarding cost			
500 chars	6250	3750	1250
1000 chars	12500	7500	2500
2000 chars	25000	15000	5000
Equipment is about half the proof-reading cost			
500 chars	3125	1875	625
1000 chars	6250	3750	1250
2000 chars	12500	7500	2500
Total			
500 chars	21875	13125	4375
1000 chars	43750	26250	8750
2000 chars	87500	52500	17500
Computer processing of input at about 0.33 dollars per item	16666	10000	3333

We estimate some of the costs associated with generating small enriched collections

	2000 docs.
Cost of indexing 2.5 - 5.00 per item	5000 - 10000
Cost of abstracting 1.5 - 6.5 " "	3000 - 13000
Cost of acquisition of full text	?

b. Requests

We assume that the requests will be collected during a bona fide use of an operational system. Therefore the cost per query will be mainly that charged by the system. One could estimate 5 - 10 dollars for this. Note however that corresponding to every information need we may have to run 5 - 10 formulations to estimate the relevance set.

c. Relevance judgements

In general one will have to assume that by providing a service free of charge to a user he will in return provide relevance assessments. The exhaustive assessments of small subsets will have to be costed separately

e.g.

Acquisition of full text	?
Mailing	?
Clerical	?

d. Cited references

One of the core requirements is that each document should have as part of its representation the references it cites. Unless this comes with the representation extracted from the operational system the cost of obtaining this further information will have to be estimated separately. The most likely source of the cited references are the ISI tapes for the appropriate period.

e. Generating a collection at level 4

At this stage it is not possible to say whether level 4 should be created by the Builder or the Curator. However this decision will mainly affect the apportioning of costs between the two phases. If for the moment we assume that creating at level 4 is done by the Builder then we will have to allow for

extra machine time,
file storage,
documentation costs.

f. Personnel

Builder)
Programmer) up to £8000 p.a.

Library and clerical support £2000 p.a.
Travel (particularly in the early stages of the design study, see p. 21)
Adminstration (e.g. mailing, xeroxing)

The reason the cost of the Builder and programmer have been lumped together is that to some extent there exists a trade-off between them. If the Builder is experienced computationally he would not need a very experienced programmer. On the other hand if the Builder is not acquainted with the computer technology his programmer will have to be of a higher standard. Also, it may be that if creating at level 4 is left to the Curator the building project would only require a half-time programmer.

Maintenance and Distribution

To some extent the costs of this operation will depend on the demand for the data. The operation should be costed over 5 years following the building phase. The main cost factors are

Curator (half-time?)

Programming support (¼ programmer?)

Equipment (e.g. tapes, terminals, etc)

File storage

Computing time

Travel

Advertising

Clerical

If the ideal collection(s) is to be added to over a period of time then the costs of the building operation will be applicable here pro rata.

Appendices

A. Details of British test collections

B. Summary of non-British test collections

C. Costs table

D. Standard collection format

References

A1 CRANFIELD 2

Project name
Factors determining the performance of indexing systems.

Objectives
'to deal with index language devices..(with).. precise measurement of recall and precision ratios'. To carry out a laboratory test, following up and improving on Cranfield 1.

Chief person/Reference
Cleverdon et al, 1966

Size
221 queries
1400 documents ; several subsets, especially 42 queries
200 documents.

Subject
Aeronautics.

Indexing source
Full texts; also abstracts and titles/titles.

Index languages
3 types, in 30 forms, all applied manually:
single terms; with synonym grouping; with hierarchical reduction
simple concepts; with hierarchical reduction
controlled terms; with related terms.
Abstracts and titles/ titles indexed automatically.

Requests
Authors of selected recent papers asked to state reason (in form of a question) for undertaking research leading to paper, and to provide other questions related to this research.

Relevance
by authors, for own cited papers: exhaustively by experts to obtain additional papers for author vetting. There were four relevance grades.* Relevance judgements were based on full text.

Document collection
Document set consisted of some recent papers, and their cited references, with some others.

Present state of test collection
Queries and single term index descriptions, abstracts and titles, with relevance judgements, available at level 4.

Other users of test collection
Sparck Jones, van Rijsbergen, Salton and SMART Project workers; also Minker, Svenonius.
(Some SMART tests with 24 or 155 queries and 424 documents).

*n relevance grades does not include non-relevance as a grade.

A2 INSPEC

Project name
Comparative evaluation of index languages.

Objectives
A comparative assessment of the retrieval performance, in the INSPEC system, of a number of index languages which might be used as the sole or main means of subject manipulation.

Chief person / Reference
Aitchison et al, 1970

Size
97 queries
542 documents.

Subject
Physics, electrotechnology, and control.

Indexing Source
Abstracts and titles/titles.

Index Languages
1. Titles
2. Abstracts and titles) not normally regarded as an index language
3. Printed subject index to Science Abstracts + free language modifier line
4. Controlled language using a thesaurus
5. Free language terms (applied by the SDI investigation staff to indicate 'subject content of document' before translation into 4).

3-5 applied manually.

Requests
Questioners were asked to ensure that the questions were within the scope of their SDI profiles ('it will need to be answerable by some of the documents already notified to you by the SDI service'). Only questions with at least one document at the higher level of relevance were used in the evaluation. Queries screened by research team 'if detailed study of the profile showed that its scope had been changed in the course of the four SDI services' query would be discarded).

Relevance
Each questioner to be sent for assessment only those documents which he had previously assessed as relevant to his profile. Relevance assessments made by the user on the basis of document texts. There were two relevance grades.

Document collection
2/3 of documents relevant to some query.

Preconditions
1. SDI investigation was in progress
2. Queries were solicited from users who received all four services and had assessed at least 12 documents as relevant to their profile.

Present state of test collection
Queries and free language index descriptions, with relevance judgements, available at level 4.

A3 ISILT

Project Name
Information science index languages test.

Objectives
'to compare the effectiveness and efficiency of different index languages as used in subject retrieval systems'.

Chief person / Reference
Keen et al, 1972

Size
63 queries
800 documents

Subject
Documentation

Indexing source
Abstracts and titles/full texts

Index languages
5 kinds all applied manually
1. Compressed term language - 300 terms from ASLIB + related terms added.
2. Uncontrolled - natural language text words underlying hierarchical index terms of 3. Specific indexing was followed by redundant indexing.
3. Hierarchically structured language - post-coordinate.
4. Same as 3 but pre-coordinate
5. Relational indexing
Exhaustivity and specificity of indexing were controlled.

Requests
These were miscellaneous real requests, formulated considerably later than the dates of the documents.

Relevance
Exhaustive relevance judgements were made based on abstract and title for 408 of the documents; for the rest full text was used. There was a scale of relevance. The assessment was 'non-user relevance by simple subject experts who were not requesters, indexers, or searchers in the test'.

Document collection
set I...408 documents from the Smart project. Abstracts from the period 1961-63 were available in machine readable form. These abstracts claimed to be bad.
set II..392 good abstracts dated up to 1968, some of these were quoted as known relevant ones by requesters.
2/3 of the collection in fact relevant to some request.

Present state of test collection
Queries and uncontrolled index descriptions for whole collection, abstracts and titles for Subset I, with relevance judgements, available at level 4.

Other users of test collection
Sparck Jones, Van Rijsbergen, Horsnell
(Some tests with Subsets I or II with automatic indexing from abstracts and titles/titles for Subset I)

A4 UKCIS

Project name
Retrieval experiments based on Chemical Abstracts Condensates.

Objectives
1. To gain experience using CA-Condensates tapes.
2. To compare the relative effectiveness of searching titles only, titles-plus-keywords, and titles-plus-digests.
3. To measure the variation in performance between profiles covering different subject areas.
4. Investigate automatic profile construction.

Chief person / Reference
Veal / Barker, et al, 1974

Size and Subject
193 requests (subset 48)

documents	size	subject
CAC-1	11518	Biochemistry, organic chemistry
CAC-2	15629	Macromolecular, applied and physical chemistry
CBAC	1568	Biochemistry
POST-J	1412	Polymer science
POST-P	1442	Polymer science

Indexing source
Full texts/titles

Index languages
Keywords applied manually; titles and digests effectively indexed automatically.

Requests
Formulations were from SDI service users and were current for the documents searched. Different versions were written for different data bases.

Relevance
Assessment of output (pooled if appropriate) by users, usually from titles and digests, sometimes titles only. There were 2 relevance grades, and sometimes 2 non-relevance grades.

Document collection
Documents were taken from Chemical Abstracts Service files.

Present state of test collection
CAC-1 and CAC-2 available at level 3 or 4.

Other users of test collection
Sparck Jones, van Rijsbergen

A5 MEDUSA

Project name
Medlars on-line search formulation and indexing.

Objectives
To compare the standard method of search formulation by a trained search editor with a physician's using an on-line terminal.

Chief person / Reference
Barraclough/Barber et al., 1972

Size
58 queries
51000 documents

Subject
Medicine

Indexing source
Full texts

Index language
MeSH, i.e. controlled language, applied manually.

Requests
Requests from real on-line systems users, with two formulations, one by the user and one by a trained search editor.

Relevance
2 grades of relevance, also 2 grades of non-relevance.
Assessment of output based on citation and indexing, by user for both search formulations.

Document collection
Documents taken from monthly files of regular Medlars service.

Present state of test collection
Available at level 3

Other users of test collection

A6 NPL

Project name
The National Physical Laboratory experiments in statistical word associations and their use in document indexing and retrieval.

Objectives
1. To develop methods of clustering words on the basis of especially computed measures of association between word pairs.
2. To explore and evaluate ways of employing these clusters and associations to improve performance especially in the ability to recall relevant material.

Chief person / Reference
Vaswani, 1970

Size
93 queries
11571 documents

Subject
Electronics, computers, physics, and geophysics.

Indexing source
Abstracts and titles

Index languages
A dictionary of 1000 index terms (stems) was constructed based on a sample of 1648 abstracts by semi-automatic means. These were used to index the documents.
1. Weighted / Unweighted terms
2. Clusters derived from associations
3. Expansion through a connection network

Requests
20 people formulated requests based on source abstracts; but these only specified subject and abstract not necessarily relevant.

Relevance
17000 relevance decisions made by the people who formulated the requests. Report claims that 80% relevant documents uncovered by various strategies.

Document collection
Set from published abstract journal.

Present state of test collection

Other users of test collection

A7 UKAEA/NSA

Project name
SDI from Nuclear Science Abstracts

Objectives
A study of the relative performance of two computer matching techniques: (a) of Euratom indexing terms and (b) words in titles.

Chief person / Reference
Olive et al, 1973

Size
60 queries
12765 documents

Subject
Nuclear science

Indexing source
Abstracts and titles?

Index languages
1. Natural language
2. Euratom index terms
3. NSA subject categories
2 and 3 applied manually.

Requests
Formulations were based on SDI service users' interests;

Relevance
User assessment of search output based on title, bibliographic elements and assigned index terms. There were two relevance grades and an option to state that abstract was required to make relevance decision.

Document collection
Documents taken from successive issues of NSA used in a regular SDI service

Present state of test collection
Available at level 3

Other users of test collection

B1

Test collections used in SMART Project tests published by Salton and others from 1968 are:

	Requests	Documents	
ADI	35	82	Documentation
IRE	17	375	Computing
	24	375	
	34	780	
Cranfield	42	200	Aeronautics
	36	200	
	22	200	
	22	424	
	24	424	
	30	424	
	155	424	
	36	1400	
	50	1400	
	225	1400	
Ispra	48	1268	Documentation
	48	468,1095	
Medlars	18	273	Medicine
	24	450	
	29	450	
	35	1033	
(Ophthalmology)	29	852	
Time	83	425	World Affairs
	24	425	

These collections are automatically indexed from abstract and title (but ADI from short full texts); some have indexing derived from a manual thesaurus; the Medlars collections MeSH indexing is not held. The collections are presumably available at something like level 4.

Recents tests have mainly exploited the Cranfield 24x424, Medlars 24x450 and Time 24x425 collections. For relevant information see Salton 1975a and b.

Reference	Re - quests	Documents	Source	Experiment	Description
Litofsky 69	-	46821	NSA	Document clustering	Manually assigned index terms from EURATOM thesaurus
"	-	46942	NSA	" "	Keywords extracted from titles
Chan 73	-	179	CACM	Document classification and indexing	Titles
Lo 72	-	7000	Comp. Sci.	Iterative feedback	Titles
Akiyama 72	-	1572	Accoustics	Document classification	Titles
Jahoda 69	55	3204	Chemistry	Comparison of title index with subject index	Titles
Augustson 70	-	2267	Science	Term clusters by graph theory	Index terms manually assigned
Virgo 70	23	1276	Ophthalmology	Comparison of Index Medicus and Medlars	MeSH
Cagan 70	31	250	Medicine	Co-occurence patterns for terms	Terms derived from abstracts
Svevonius 72	42	200	Cranfield	Index term frequency weighting	Keywords
Jacquesson73	-	40000	Economics	Term association analysis	Terms from controlled vocabulary
Minker 73	35	82	ADI	Keyword clustering	Indexing based on full text
"	34	780	IRE	" "	" " " abstract
"	18	273	Medlars	" "	" " " "
Feinman 73	-	261	Physics	Document classification	Abstracts and titles. Manual indexing
O'Connor 73	18	82	ADI	Search strategies	Full text
Crouch 72	34	780	IRE-3	Document clustering	Smart indexing
"	98	424	Cranfield	" "	" "
Lancaster68	302	nK	Medlars	Medlars evaluation	Mesh
Lancaster72	47	8000	EARS	Effectiveness of natural language	Abstract, full bibliographic citation, and index terms
Hansen 73	20	nK	CAC	Retrospective searching	Titles and keywords
Tell 71	7	nK	Inspec	Effectiveness of titles and subject headings	Titles and subject headings
"	53	nK	POST	Same plus abstracts	Titles, subject headings and abstracts

APPENDIX IV

INPUT COSTS

All costs shown in U.S. dollars

INPUT COSTS

	S.1	S.2	S.3	S.4	S.5	S.6	S.12	S.13	S.14	S.15A	S.15B	S.16	S.17
(a) No. of items per year	26,000	102,000	60,296	149,000	17,000	40,000	13,400	22,000	10,000	47,000	33,000	220,000	240,000
(b) No. of characters per record	800	1,200	663[1]	1,700	1,200	2500 - 3000	405	2,000	1,300	1,600[6]	1,600[6]	385	100[17]
ANNUAL COSTS													
(c) Acquisitions	15,000	-	45,000	49,000	-	130,101	39,200	14,580	36,750[13]	-	-	-	117,000[18]
Intellectual Processing													
(d) Scanning & selection	incl.in (g)		118,000	-	4,905			-	8,085	28,000	12,000	-	-
(e) Cataloguing	-		91,000	-	21,989	38,076		5,968	4,900	110,500	83,000	inc.in (1)	48,600
(f) Indexing	92,309	330,000	156,500	-	3,905	43,292	13,475	54,313	4,900		147,000	-	449,250
(g) Abstracting	171,431		96,000	-	52,829				17,885	282,000	225,000	-	
(h) Translation	-		20,000	-	613			-	4,900		-	-	97,335
(j) Total	263,740	330,000	481,500	166,600	84,241	81,368	13,475	60,281	40,670	420,500	467,000	696,362	595,185
(k) Thesaurus maintenance	8,594	20,500	21,500	12,250	122		341	-	735	7,050	(12,000) 4,850	146,300	-
Mechanical Processing													
(l) Keyboarding (staff) & correction	17,860[5]	50,000	23,188[2]	220,501	36,088		2,197	8,942	3,655	57,200	37,180	20,370	24,000[13]
(m) Proof-reading	8,640	20,000	8,575	41,650	1,838		see note (12)	1,549	5,104	57,200[7]	37,180[7]	30,556 [10]	24,000[13]
(n) Equipment	4,272[4]	11,352[14]	3,600	inc.in (1)	inc.in (1)		1,558	44,545	inc.in (1)	14,400	9,360	7,179	4,320[19]
(o) Total	30,772	81,352	35,363	262,151	37,926	64,174	3,756	55,036	8,759	128,800	83,720	215,656[11]	52,320
(p) Computer Processing	28,600	24,000[15]	58,786[3]	49,000	38,367		4,792	7,786	355	18,716	(31,857) 13,141	130,000	-
(q) Supervision	46,154	-	120,724	51,450	1,838		-	-	2,460	-	-	55,000	-
(r) Total = j + k + o + p + q	377,860	433,852	717,873	541,451	162,493	-	22,364	123,103	52,979	757,066	568,811	1,251,318	-
UNIT COSTS													
Total cost per item added to data base = $\frac{r}{a}$	14.533	4.469	11.906	3.634	9.559	-	1.669	5.596	5.298	12.102	17.237	5.688	-
Cost of int. processing per item = $\frac{j}{a}$	10.144	3.056[16]	7.986	1.118	4.955	-	1.005	2.740	4.067	8.947	14.152	3.165	2.480
" " selection " " = $\frac{d}{a}$	-	-	1.957	-	0.288	-	-	-	0.808	0.596	0.364	-	-
" " cataloguing " " = $\frac{e}{a}$	-	-	1.509	-	1.294	2.47[8]	-	0.271	0.490	2.351	2.515	-	0.203
" " indexing " " = $\frac{f}{a}$	3.550	-	2.596	-	0.230	4.99[9]	-	2.469	0.490	6.000	4.455	-	1.872
" " abstracting " " = $\frac{g}{a}$	6.594	-	1.592	-	3.108	-	-		1.788		6.818	-	
" " thesaurus maint. " " = $\frac{k}{a}$	0.331	0.190[16]	0.357	0.083	0.007	-	0.025	-	0.073	0.150	0.150	0.665	-
" " data prep. per character = $\frac{o}{a \times b}$	0.0015	0.0007	0.0009	0.0010	0.0019	-	0.0007	0.0013	0.0007	0.0017	0.0016	0.0025	0.0022
" " keyboarding " " = $\frac{l}{a \times b}$	0.0009	0.0004	0.0006	0.0008	0.0018	-	0.0004	0.0002	0.0003	0.0008	0.0007	0.0012[10]	0.0010
" " proof-reading " " = $\frac{m}{a \times b}$	0.0004	0.0002	0.0002	0.0001	0.0001	-	-	-	0.0004	0.0008	0.0007	0.0018[10]	0.0010
Apparent rate of keyboarding (char/hr) = $\frac{a \times b}{\text{no. of operators} \times 2080}$	2,500	5,350	7,700	-	-	-	2,600	7,050	6,250	1,800	1,950	2,700[10]	-
Cost of computer processing per item = $\frac{p}{a}$	1.100	-	0.975[3]	0.329	2.257	-	0.358	0.354	0.036	0.398	0.398	0.627	-

Notes

1. Based on estimate of 40M char. on tape.
2. Indexing input only.
3. Cost of subcontract computer work only; computer costs incurred in-house not known.
4. Estimate based on rental of $89.00 per month per machine.
5. Includes $2,500 for punched cards.
6. Estimate based on 700 chars. for citation plus 900 for abstract.
7. Verification cost.
8. Based on 15,400 items.
9. Based on 8,666 items.
10. For 44,000 items keyboarded in-house.
11. Total for all data prep., including subcontract work.
12. Typist checks own work; spelling checked by computer.
13. For 5,000 items only.
14. Estimate based on 11 Flexowriters @ $86.00 per month.
15. Probably excludes some computer operations.
16. (a) here is 108,000.
17. Index input only.
18. Staff costs only.
19. Estimates based on observation only.

Sparck Jones' standard formatted level 4 collections are obtained

a) by processing the document descriptions and indexing vocabulary automatically to delete stop words and generate stems with associated term numbers; and

b) by regularising the document, request and relevance judgement sets and deriving basic files from them.

These files all conform to regular layout principles. Thus a standard collection consists of files, or streams,as follows

1/ 0* documents, with original document identifying numbers, and sorted term numbers

1 documents serially numbered, with sorted term numbers

2 requests serially numbered, with sorted term numbers

3 relevance judgements, serially numbered, with sorted original document numbers

4 original document number - serial number equivalence list

5 term dictionary, giving term numbers in serial = alphabetical order and alphabetically first variants in each word group with a common stem

6 term dictionary with words in alphabetical order, if corresponding terms not serial

7 documents, with original numbers, and sorted term names

8 requests, serially numbered, with sorted term names

9 inverted documents, i.e. inverted stream 0

10 inverted requests, i.e. inverted stream 2

11 inverted relevance judgements, i.e. inverted stream 3

12 document frequencies, i.e. a list with the number of terms in each document in stream 0

13 request frequencies, i.e. a list with the number of terms in each request in stream 2

14 relevance judgement frequencies, i.e. a list with the number of relevant documents for each request in stream 3

15 distribution data, giving the numberof items, maximum, minimum and average length, and length distribution of streams 1/ 0,2 and 3 and 2/ 0,1 and 2

2/ 0 document term frequencies, i.e. a list with the number of documents for each term in stream 9

1 request term frequencies, i.e. a list with the number of requests for each term in stream 10

2 document relevance frequencies, i.e. a list with the number of requests for each relevant document in stream 11

3 original request number - serial number equivalence list

* conventional numbering with historical rationale

D2

Collection input data processing normally generates a variety of other streams which, since they all conform to the common layout conventions, constitute a natural extension of the basic standard collection. These streams may include an alternative to

stream 0, with within document frequencies of terms indicated
1 " " request " " " "
3 with serially numbered documents

plus different sets of relevance judgements etc, listings of the full dictionary, indicating truncation and grouping, and so on.

Note that a standard collection refers to a particular set of documents (and requests) indexed in a particular way, i.e. to what may be called a collection version. Thus the Cranfield 1400 documents and requests indexed by manually assigned terms, and by terms automatically extracted from titles, lead to two standard distinct collections. Also, when subsets of documents and requests are selected, all the frequency information is different, so these also generate distinct standard collections. Thus the Cranfield 200 manually indexed document collection is different from the 1400 one.

The particular form of standard collection just given is merely illustrative. It is evident that more complex collections like the ideal one(s), or collections with radically different characteristics, might require more elaborate, or alternative, standard forms. But we feel the principle of standardisation is very important. Data formats, particularly for operational systems, are not necessarily suited to research, so some modification may be needed; but full-blooded standardisation is usually more convenient in the long run. We have certainly found standard collections set up on the lines indicated very helpful. It should also be pointed out that standardisation is a non-trivial operation, so there are clear gains if it is done only once, in a competent way.

References

Aitchison, T.M., Hall, A.M., Lavelle, K.H. and Tracy, J.M. Comparative evaluation of index languages, Part I, Design; Part II, Results, Project INSPEC, Institute of Electrical Engineers, London, 1970

Akiyama, S. Automatic document classification systems, M.Sc. Thesis, Department of Computer Science, University of Alberta, Canada, 1972

Augustson, J.G., and Minker J. Deriving term relations for a corpus by graph theoretical clusters, Journal of the American Society for Information Science, 21, 101-111, 1970

Barker F.H., Veal, D.C. and Wyatt, B.K. Retrieval experiments based on Chemical Abstracts Condensates, Research Report No 2, UKCIS, University of Nottingham, 1974

Barber A.S., Barraclough, E.D. and Gray W.A. MEDLARS on-line search formulation and indexing, Technical Report Series, No. 34, Computing Laboratory, University of Newcastle upon Tyne, 1972

Cagan, C. A highly associative document retrieval system. Journal of the American Society for Information Science, 21, 330-337, 1970

Chan, F.K. Document classification through use of fuzzy relations and determination of significant features, M.Sc. Thesis, Department of Computer Science, University of Alberta, Canada, 1973

Cleverdon, C.W., Mills, J., Keen, M. Factors determining the performance of indexing systems, Vols 1 and 2, College of Aeronautics, Cranfield 1966

Feinman, R.D. and Kwok, K.L. Classification of scientific documents by means of self-generated groups employing free language, Journal of the American Society for Information Science, 24, 382-396, 1973

Hansen, I.B. CA condensates as a retrospective search tool. A commentary, Information Storage and Retrieval, 9, 201-205, 1972/3

Horsnell, V., Intermediate lexicon in information science, School of Librarianship, Polytechnic of North London, 1974

Jacquesson, A. and Schieber, W.D. Term association analysis, Information Storage and Retrieval, 9, 85-94, 1973

Jahoda, G. and Stursa, M.L.A comparison of a keyword from title index with a single access point per document alphabetic subject index, American Documentation 20, 377-380, 1969

Keen, E.M. and Digger, J.A. Report of an information science index languages test, Aberystwyth College of Librarianship, Wales., 1972

Lancaster, F.W. Evaluation of the MEDLARS demand search service, National Library of Medicine, Bethesda, Md., 1968

Lancaster, F.W. Evaluating the effectiveness of an on-line natural language retrieval system, Information Storage and Retrieval, 8 223-245. 1972

References (contd.)

Litofsky, B. Utility of automatic classification systems for information storage and retrieval, Ph.D. Thesis, University of Pennsylvania, 1969

Lo, A.K. An automatic optimum iterative feedback document retrieval system, M. Sc. Thesis, Department of Computer Science, University of Alberta, Canada, 1972

Minker, J., Peltola, E., and Wilson, G.A. Document retrieval experiments using cluster analysis, Journal of the American Society for Information Science, 24, 246-260, 1973

O'Connor, J. Text searching retrieval of answer-sentences and other answer-passages, Journal of the American Society for Information Science, 24, 445-460, 1973

Olive, G., Terry, J.E. and Datta, S. Studies to compare retrieval using titles with that using index terms. SDI from 'Nuclear Science Abstracts', Journal of Documentation, 29, 169-191, 1973

van Rijsbergen, C. Further experiments with hierarchic clustering in document retrieval. Information Storage and Retrieval, 10, 1974, 1-14

Salton, G. A theory of indexing Regional Conference Series in Applied Mathematics No. 18, Society for Industrial and Applied Mathematics, 1975

Salton, G. Dynamic information and library processing, Englewood Cliffs, N.J.: Prentice-Hall, 1975

Sparck Jones, K. Automatic indexing 1974, Computer Laboratory, University of Cambridge, 1974 (OSTI Report 5193)

Svenonius, E. An experiment in index term frequency, Journal of the American Society for Information Science, 23, 109-121, 1972

Tell, B.V. Retrieval efficiency from titles and the cost of indexing, Information Storage and Retrieval, 7, 241-243, 1971

Vaswani, P.K.T. and Cameron, J.B. The National Physical Laboratory experiments in statistical word associations and their use in document indexing and retrieval, Publication 42, Division of Computer Science, National Physical Laboratory, Teddington, 1970

Vickers, P. The costs of mechanized information systems, Directorate for Scientific Affairs, OECD, 1974

Virgo, J.A. An evaluation of Index Medicus and Medlars, Journal of the American Society for Information Science, 21, 254-263, 1970

Wilson, E. Report on Automatic Indexing Workshop, April 29-30, 1974, Computing Laboratory, University of Kent, 1974 (OSTI Report 5194)

Crouch, D.B. A clustering algorithm for large and dynamic document collections, Southern Methodist University, Dallas, 1972.