Service to an interdisciplinary need-group from computerised secondary services.

submitted for the Degree of

Master of Philosophy

by

John Eustace O'Brien Martyn

June, 1974.

SERVICE TO AN INTERDISCIPLINARY NEED GROUP FROM COMPUTERISED SECONDARY SERVICES.

A Thesis submitted for the Degree of Master of Philosophy,

by

John Martyn

SUMMARY

The extent to which a specialised interdisciplinary need-group can assemble its reference base from computerised secondary services, and associated problems have been explored. The reference file of journal literature published in 1969, collected by the Biodeterioration Information Centre, was taken as typifying the reference requirement of a specialised need-group. Its coverage by Chemical Abstracts Condensates, Chemical Biological Activities, Chemical Titles, BA-Previews, MEDLARS, Food Science and Technology Abstracts and the services provided by the Institute for Scientific Information (ISI) was determined by item-by-item check. It was found that the best service covered 60% of the reference file, and that all services together covered almost 77%. 23% of the reference file was not covered by the secondary services examined, and inspection of this non-covered subset indicated that non-coverage could not be attributed to triviality or irrelevance of content. Profiles for machine search of the secondary services were compiled, based on comparison of the frequencies of occurrence of terms in the references composing the reference file with frequencies of occurrence of the same terms in the files to be searched, in order hopefully to produce the greatest number of relevant references at least cost per relevant reference retrieved. Searches were carried out on the ISI Base, CA-Condensates, Chemical Titles, BA-Previews and MEDLARS, and the results were assessed for relevance of retrieved references. There are indications that as terms are used in descending order of specificity, both the overall cost per relevant citation retrieved and the proportion of irrelevant items retrieved tend to increase.

contd/.....

The non-availability of complete coverage from existing computerised secondary services and the cost and difficulty of extracting relevant references from them suggest that reliance on these services for the servicing of the reference requirements of interdisciplinary need-groups is not yet practicable.

Acknowledgements.

I wish particularly to thank Mrs Lynda Torrisi, who spent many hours searching abstracting and indexing serials to provide some of the data for the coverage aspects of this study: her help was invaluable. Mrs Torrisi was seconded to me from the Polytechnic of North London, and I thank Mr Jack Mills for arranging her secondment. Thanks are also due to Dr Dennis Allsopp, Dr Howard Eggins and Mr Michael Willsher, of the Biodeterioration Information Centre, for their constant cooperation and ecouragement, even when required to make several thousand relevance judgements. Professor B.C. Vickery, formerly Head of Aslib Research and Development Department, allowed and encouraged me to carry out the research, and he and Mr E.H. Driver, Librarian of the University of Aston in Birmingham, have materially assisted with discussion, comment and guidance. Others who have assisted with advice and criticism include Mr N.W. Briggs, Mr A.E. Cawkell, Dr Alistair Cochran, Dr Eugene Garfield, Mr Peter Lunn, Mr Tony Myatt, and Dr Douglas Veal.

Particular thanks are due to my colleagues in the Aslib Research and Development Department, who have been very patient with me while I have been engaged on this study, and to the members of the Office for Scientific and Technical Information (now British Library Research and Development Department), who were constantly helpful throughout.

My wife kept my children occupied while I was writing up the work, and her contribution is therefore only slightly less than that of Mrs Torrisi. Finally, Miss Maria O'Toole typed the whole, with never a word of complaint.

Services to an interdisciplinary need-group from computerised secondary services.

CONTENTS

List of Tables

List of Figures

Appendices.

Satisfying the reference requirements of an interdisciplinary need group from computerised secondary services.

INTRODUCTION

Information services are created for users. This obvious-seeming statement can be realised in two ways. One realisation takes the form of a number of services, covering between them all or most of human knowledge, which may allow specialised user communities to satisfy their needs by accessing an appropriate selection of the services. The other form is that of a service organised to satisfy from a single source the full requirements for information of a definable and, in terms of the information required, relatively homogeneous group of users. The first interpretation has led to the creation of the major disciplinary-oriented abstracting and indexing services, the second to the development of specialised information centres. The first begins with the information, the second with the user. In practice, both interpretations suffer from defects which act to reduce the efficiency or the effectiveness, or both, of the services developed.

Originally, perhaps, the two viewpoints were believed to coincide. There was a discipline called chemistry, there were people called chemists, many of whom belonged to the same learned society, and there was a substantial and growing body of literature about chemistry. It was natural to assume what in many cases was and is true, that the bulk of the information needs of chemists could be met by organising and indexing the literature of chemistry. Unfortunately, a number of documents which can reasonably be classed as 'chemistry' can also be classed as 'physics' or 'biology', and these tend to be indexed within both the chemical and the physical or biological systems; in other words, discipline-based services tend to overlap. There are also documents which do not appear to fit easily into an existing disciplinary structure, and these may not be collected until a new disciplinary service is organised to cover them. Also, because of the scholarly origin or archival intent of the major services, there is sometimes a

tendency to select by level as well as by subject, which results in an apparent bias in favour of pure science as opposed to technology. All these factors tend in their various ways to erode the effectiveness of the major services, but the most important factor is that with the passage of time the degree of correspondence between the needs of the chemist and the literature of chemistry has been progressively reduced, so that the chemist of today now needs and uses the literature of many other disciplines besides his own.

The specialised information centre concept is more recent in origin, although much of the work of such a centre has been done in the past by the internal information services operating within some industrial organisations. The benefits afforded to a community of specialised users by an effectively operating centre are obvious, including on the document provision side a reduction of user effort in the collection of material and a greater completeness of coverage than can usually be achieved by an individual, coupled with a higher degree of relevance in the material assembled than is usually the case with a larger discipline-oriented service. A major disadvantage is that in order to serve a very large number of specialised need groups, there would inevitably be a high degree of duplication of acquisition of references, and a greater need for information-processing staffs. In addition, there are undoubtedly a number of need-groups of insufficient size to support specialised information centres appropriate to their needs, and consequently information provision based on the specialised centre concept would leave a large number of small areas of under-provision. More or less by definition a special centre is a small organisation, and the difficulties of reference collection in fields where references spread over large numbers of journals are more keenly felt than would be the case with a larger organisation collecting over proportionately more manageable a field.

The core of the problem is that the pattern of demand for

information in the middle of the twentieth century no longer fits the disciplinary classification of the nineteenth, by which the major abstracting and indexing services are structured. As areas of study become increasingly multidisciplinary, so a growing number of groups and individuals are finding it necessary to use not one but a number of secondary services in order to collect information on their topics of interest.

There is no simple answer to the problem. Many of the major services have turned to computerisation in order to improve both their production and their ease of manipulation, and this has improved the speed and ease of access to and search of their files in many respects. It has also assisted the process of subdividing the files in order to provide new services aimed specifically at smaller potential user groups. Unfortunately it still remains true that although a number of the newer fields of interest can be catered for by specialised subsets of the major services, an increasing number cannot. One solution currently being operated at a number of locations is to acquire the tapes of the major services and either to meld them to produce one massive base which contains information relating to a wide range of disciplines, or to process incoming requests against a selected number of the bases sequentially, the appropriate selection being done either by the users or the system operators. Merging tapes to provide a unified base has proved to be technically very difficult, because of the lack of consistency of format or of data elements included among the major services, and both merging and the search of bases selectively have encountered the problem of overlap between services, so that in addition to recovering novel references from each base searched, the user also retrieves a number of which he has already discovered, and, as it is sometimes impossible to structure a search so that this is avoided, the unfortunate user finds himself paying several times for each reference he finally retains.

Considering the problem from the user's point of view, there is a limited number of ways of acquiring references on a specialised topic. They can be discovered by scanning the primary literature,

either directly or via collections of current journals' contents pages, such as Current Contents ®, by scanning abstracting and indexing journals in their printed versions, by arranging that other specialists in the field interchange references to material appearing in their own countries, or by searching several computerised secondary services. Journal scanning is usually not a practical proposition if comprehensive cover is desired, because of the tendency of references on any subject to be distributed through a wide range of sources, many being found in a few locations but many more being scattered thinly throughout virtually the whole journal field. If all that is required is a number of references of high relevance but not all the references published in the field, then these can often be assembled by scanning a relatively small number of core journals (that is, journals whose subject-matter is very largely relevant to the topic of search). This also has the effect of reducing the amount of irrelevant material which has to be scanned. Effectively, scanning core journals produces high precision but low recall, and, because of the relatively small number of core journals in most fields, the cost of each pertinent reference found is low. Additionally, there is no retrieval of duplicate items. If, on the other hand, some attempt is made to approach comprehensiveness of cover, then a large number of journals must be scanned, resulting in increased cost, both of journals and scanning effort, and a much higher proportion of irrelevant material. In other words, as the recall is increased, the precision declines and the costs of each pertinent reference found increases, at a faster rate than recall. Scanning contents pages of current journals where available reduces the cost of scanning core journals to a large extent (with the penalty of some sacrifice of recall, because of the difference in information content of the reference as presented in the title and the full paper as given in the primary journal), but once the fringe journal area is reached, the same effect, of rising recall, reduced precision and increasing cost per pertinent reference found begins to operate.

Scanning secondary services in the printed form suffers from some of the same effects, although providing not too many services need to be scanned, the initial cost, at a fairly low recall level, of each pertinent reference found may be reduced. This is offset by recall failures, due partly to the relevant material not being included in the coverage of the services selected, and partly, if the services are used via the indexes, of recall failures caused by these indexes themselves. There is also a cost penalty incurred by the discovery of duplicate references caused by overlaps in coverage of services. Naturally, since the file content is essentially the same, the overall picture is the same as regards searching machine-readable versions of secondary services.

The use of cooperating specialists equates to the scan of primary publications as described above, with the difference that the scanning is distributed among a number of individuals. The problem of the identification of sources to scan is eased to the extent that each scanner can restrict himself to a sub-set of journals demarcated by language or by sub-subject bias, but there is still no escape from the difficulties, indicated above, inherent in primary journal scanning. There is, too, the likelihood of duplicate notifications.

Given that no perfect method of acquiring all the references to a particular topic exists, it becomes necessary, when a need for an attempt at complete coverage exists, to examine all the reference-acquisition modes and to assess their relative effectiveness, in relation to their costs, in order to decide which mode is likely to afford the best return for the least cost in money or effort. The criteria which require quantification (so far as is possible) in this context are

1. Coverage. The amount or proportion of the published material relating to the topic which is acquired by a particular mode.

2. Precision. The ratio between the amount of relevant material acquired by a mode and the total amount of material acquired.

3. Unit cost. The ratio between the number of relevant items acquired and the total cost of operating the mode.

4. Timeliness. The timelag between the first publication of a piece of information and its acquisition by the mode.

The first three criteria are of obvious importance, and are all related to relevance; timeliness is not relevance-related, and is of varying importance, depending on the degree of urgency of need-to-know prevailing among the community of users of the information. It becomes more important, largely because of its nuisance-value, when considering modes of reference location which are subject to the effects of duplicate notification.

The research reported here has its origin in a realisation of the need for fuller information to be made available on the operating characteristics and performance of reference-acquisition modes. Within the limits of a single study it is not practicable to attempt a detailed analysis of all possible modes, so the decision was made that attention should be concentrated on one mode only, that based on exploitation of the currently available computer-based secondary services. The decision was influenced by the recent growth in the number and variety of such services (there are currently more than a hundred such services, notifying between them more than three million references annually, with much duplication), which has in its turn led to a more urgent demand for data relating to their performance. It is, in fact, something of a two-sided demand, the users requiring to know how they can use the tapes, and tape suppliers or manipulators wanting to know how they can be used.

At this point, then, the question to be studied has become 'How may the information needs of an interdisciplinary need group be satisfied by the use of computerised secondary services?'.

This problem has been studied here by taking an example of an

interdisciplinary need group and examining the possibilities of assembling its relevant reference base from a number of appropriate computerised services. The selected example was the Biodeterioration Information Centre at the University of Aston in Birmingham, a Specialised Information Centre.

Specialised Information Centres were first described as such in a Report of the President's Science Advisory Council (1) in 1963, and following the recommendations contained in that report, some experimental centres were set up in the United Kingdom with support from the Office for Scientific and Technical Information, an office within the Department of Education and Science. (2)

A Specialised Information Centre is an information centre which attempts to meet as many as possible of the needs for information on a particular specialised topic of the workers interested in that topic, regardless of their location. 'Meeting the needs' may mean collecting all documents relevant to the topic, indexing and storing them, disseminating the information they contain through current-awareness services and other devices, supplying copies of the documents on request and operating a question-answering service. It should also mean evaluating the collected information and producing critical and state of the art reviews, but in present practice this is seldom done. Much of the activity of a Specialised Information Centre is very similar to that of a normal special library or information department. The major difference is that whereas a normal information department is run for the benefit of a group of workers in the same organization or at the same geographical location, a Specialised Information Centre serves a widely-distributed, often international group whose only link is one of common scientific interest.

The Biodeterioration Information Centre was established in 1965, by Dr H.O.W. Eggins, of the Department of Biological Sciences at the University of Aston in Birmingham. Biodeterioration, the study of the deterioration of materials of economic importance by living organisms, brings together a wide range of otherwise

distinct fields of materials science and biology, and relevant references are widely scattered throughout the literature.

As an illustration of the Centre's spread of interest, one of the problems of defining Biodeterioration arises when considering organisms in relation to food crops. An organism attacking a food crop is of no great concern to the Centre when the crop is still growing, but if the organism continues its attack when the crop is harvested and stored, then it becomes of interest.

The effects (and possible uses) of toxins produced by the organism are of little interest, but means of controlling the production of the toxin or the growth of the organism are relevant. The topic of organic attack on products of economic importance includes fungal growth on timber, rodent attack on stored grains, bacterial attack on plastic products, fouling on ships' bottoms, organic erosion of monumental masonry and paint films (including cave paintings), and bird strikes on aircraft. Growing yeasts on petroleum products to provide food is not of interest, but the mechanisms of growth may be.

The Centre collects references to relevant documents, collects and indexes the documents themselves, publishes the International Biodeterioration Bulletin Reference Index Supplement (IBBRIS) quarterly, supplies copies of documents on demand, answers queries, performs literature searches on request, and offers advisory and consultancy services, including, where necessary, test and research programmes. It also houses a number of post-graduate students working on problems in biodeterioration.

It processes between two and three thousand references a year, and collects them by scanning a number of primary journals, Current Contents, and several abstracts journals (22 in 1969) produced by the Commonwealth Agricultural Bureaux and by

industrial and research associations. In addition, there are a number (in 1969 about 140) of cooperating specialists in various countries, who scan journals and the report literatures of their own countries, and submit references to the Centre. At one time, the cooperating specialists provided about half the input references.

This Centre was chosen as the test example because of the similarity of its reference input requirement to that of an industrial information department. The supply and organisation of information to and within industry is an area of activity of considerable economic importance in which any research resulting in improved efficiency or effectiveness of the operations involved in gathering and disseminating information is likely, if sufficiently generalisable, to produce benefits substantially greater than its cost. Unfortunately, study of an actual industrial information department is not always easy to arrange, for a number of reasons, including commercial secrecy, unavailability of budgetary information, sensitivity of topic coverage to market pressures and so on. The Biodeterioration Information Centre does not suffer from these disadvantages, and has the additional advantage that its reference base is organised in such a way as to be hospitable to investigation. Additionally, the close relationship built up between the Centre and the present investigator in the course of previous studies of Centre operation on behalf of OSTI,(3,4) meant that the present research could be mounted with the minimum inconvenience to either side.

2. Methodology

The conventional way of examining how a specialised group's needs can be met by using one or several secondary services is to take an actual specialist group and to collect references from the services under examination by manual or machine search, then present the retrieved references to the group for relevance assessment. This approach has some advantages. It is easy to implement, gives an estimate of the degree of overlap among

services that may be expected using the search techniques employed, and can provide some data very quickly. However, it does have a number of defects. In a study of this kind, what is being tested is really the effectiveness of the search procedure employed as applied to the material contained in the services being examined; this means that there are two sorts of recall failure which are not illuminated, the first being the failure to retrieve relevant material actually covered by a service, because of inadequancies in search strategy, and the second, failures caused by inadequate coverage on the part of the services examined. No clear indication is available of how much relevant material is being missed because it is not within the coverage of the secondary services examined, and it is often tempting to assume that, if a number of services are being studied, then between them they must be picking up virtually all the literature. This is, as will be shown, a fallacy. There are other defects to this methodology, but they are relatively minor, and do not need stating here.

The approach adopted for the present study is an extension of a methodology previously employed (5,6) for testing abstracts services. In these tests, the technique was to obtain a bibliography, which could be assumed to be as nearly comprehensive as possible, on a particular specialised topic, and to extract from it all references to material appearing in a particular period of time, usually a single calendar year. Then a number of secondary services which appeared to be likely to contain material relevant to the topic was selected, and the services were then searched, via the author indexes, to determine how many of the items shown in the bibliography had been notified by each of the services examined. The subject indexes were next searched, using the keywords or other index entries considered to be appropriate to the topic of the bibliography, to estimate how much of the relevant material covered by each service could actually be retrieved via the indexes. The studies showed that, in general, about a fifth of the references sought had not been covered by the services examined, and of that material which had been picked up, about three-quarters could be found by diligent and ingenious

search of the indexes to the services. For this type of test to be completely successful it is necessary to have as comprehensive a bibliography as possible, or failing that, to have one which can safely be assumed to be representative of the spread of the subject through the literature, and not biased in favour of any particular service or services. It is, for example, pointless to examine an abstracting service by this method using a bibliography which has been compiled from that service.

It is fair to point out one deficiency in this methodology as compared with that noted above. The bibliography approach necessarily gives a picture of the situation obtaining at some time in the past. Because complete or adequate bibliographies normally take some time to compile and some further time to be published, and because of the delays inherent in the abstracting process, it is necessary to select a time-period for the references to be sought which antedates the present by some three years, in order to allow the secondary services which are to be examined to have notified all the references from the bibliography, which they are ever likely to pick up. As an illustration, it is quite possible for a reference originally published in 1969 not to be notified in a particular secondary service until 1972 or even later. Most secondary services will not exclude a relevant reference on grounds of age alone, and some references in obscure journals or difficult languages are difficult to collect. Consequently, the data of this study relate to the recent past and not to the present, but the conclusions which may be drawn are not altered.

The major practical difficulty in using the bibliographical approach is the location of a suitable bibliography. In the present study a bibliography was required which could be identified with the needs of an interdisciplinary group of workers, representative of the whole spectrum from pure science to applied technology. The topic covered should be one of some economic importance and immediate applicability. The bibliograpy itself should be as comprehensive as possible, within a defined timespan,

and should have been assembled by a variety of means, without reliance on one major mode of reference acquisition. Access should be available to the compiler or compilers of the bibliography, so that relevance judgements on additional material discovered should be easily obtained. With these requirements in mind, it was decided that a suitable bibliography could be drawn from the material assembled by the Biodeterioration Information Centre and notified in the International Biodeterioration Bulletin Reference Index Supplement (IBBRIS).

It was decided that a bibliography should be drawn from the references notified in IBBRIS in 1969 and 1970 which were to material which had been published in 1969. This bibliography, a full analysis of which appears in the next section, could be confidently supposed to represent the major part of the 1969 literature of biodeterioration, without discernible bias towards pure science, technology, language, nationality or any particular indicator of quality. Although some further 1969 material has emerged in the course of the current study (much of it in more recent issues of IBBRIS), there is no reason to suppose that the initial assumptions about the bibliography were ill-founded.

The bibliography, referred to henceforth as the BIO file, was initially transferred to cards, purged of the few duplicate references it contained, and analysed as described later. This file was then to be divided into a number of overlapping subsets, each subset representing the coverage of the BIO file references by a particular computerised secondary service. This was done by an examination of the printed versions of the appropriate services, taking each reference in the BIO file in turn, and using the author index of the examined service to decide whether or not the BIO file reference had been covered. On identification of the author in the author index, the abstract or reference as shown in the body of the service was examined, to ensure identity with the actual BIO file item concerned. When the whole BIO file had been broken up in this way into overlapping subsets, each subset then represented the coverage of the 1969

biodeterioration literature by a given service. The next problem then was, knowing that an identified body of references was contained within a service, how this could be retrieved by a search strategy applied to a machine file, in such a way as to retrieve as many as possible of the identified references, and as few as possible other (and by definition, irrelevant) references, at the least cost. If this could be done, it should then be possible to develop the set of profiles suitable for interrogating an appropriate mix of computerised secondary services in such a way as to produce the maximum amount of the whole BIO file, with the lowest amount of irrelevant material, at least cost, with as little duplication as possible. This is similar to an approach reported by Lynch and Smith (7)..

Unfortunately, this ideal methodology was not fully capable of practical implementation. The services selected as being relevant to biodeterioration, with a probability, that is, of containing a substantial part of the BIO file were the Institute for Scientific Information services, BA- Previews, Chemical Abstracts Condensates (CAC), Chemical Titles (CT), Chemical-Biological Activities, (CBAC) MEDLARS and Food Science and Technology Abstracts (FSTA). Only 1969 and 1970 issues of these services were searched for coverage, because coverage search is immensely time-consuming and arduous and it was argued that as one of the needs information services are intended to serve is current awareness, notification of a reference between one and two years after it has appeared is hardly 'current'. Some sample checks were carried out of later issues of some services, but the amounts of additional material recovered were not sufficiently large to justify what would have been about an extra four months' search effort. Coverage, then, with this limitation, was quite possible to determine. However, when it came to the tape-searching stage, very few of the services had tape available in the United Kingdom covering the appropriate period of time, and of those that had, it was not possible to retrieve the original indexing applied to the BIO file items known to be contained in order to use the assigned indexing as an aid to profile construction. In the event,

the only service which could be fully examined in the way outlined above was that provided by ISI, where the Permutern Index and the Source Index together allow perfect simulation of machine search. The Chemical Abstracts services were not available for the appropriate period, nor were BA-Previews tapes, and the other services were not able to provide the information required for profiling. The services are described in greater detail later, and the difficulties encountered in their use are also discussed.

In the event, the coverage data are available and are presented here. Therefore maximum recall estimates for the services are also available. In order to estimate precision, searches were run on current files of the services studied, and the search products evaluated by the Biodeterioration Information Centre.

The other departure from the outlined methodology was in the makeup of the BIO file itself. As will be seen, not all the references to 1969 material were to journal articles. There were also references to a variety of other forms of publication, including theses, specifications, annual reports and Government handouts or training leaflets. It was decided to confine the study to journal papers, largely on grounds of effort involved if the non-journal material were included, but additionally because it is known that the policies of coverage of non-journal material differ widely between services, some being confined to journal items only, others including one or two specific other forms, but virtually none including all forms. This means that in terms of absolute cover, the results of this study tend to be biased somewhat in favour of the secondary services.

3. The BIO file

The BIO file consists of all the references to material published in 1969 which was notified in the 1969 and 1970 issues of IBBRIS. A number of duplicate entries were discarded, as were six references to translations published in 1969 of material first

published in an earlier year, and two unverifiable references. The 2154 remaining references were made up as shown in Table 1.

TABLE 1.

Composition by medium of publication of BIO file.			
Medium of publication		Number	%
journal (including annual review		1874	87.00
*reports	95	149	6.92
*technical notes, leaflets etc	39		
*newsletters	15		
conference proceedings, symposia		58	2.69
books, monographs (whole or in part)		23	1.07
@newspaper reports		23	1.07
patents		14	0.65
non-serial review (Fermentation Advances)		10	0.46
standards (DIN)		1	
specifications		1	
theses		1	
TOTAL		2154	

* The distinction between the categories is unclear, and for most practical purposes they can be considered together. The material includes annual reports of various national laboratories, other report material, ministry notes for agriculturalists and so on.

@ The newspapers represented are restricted to The Times (16), Sunday Times (5), Financial Times (1) and Sunday Telegraph (1). They are therefore more probably the product of normal recreational reading by the Centre staff than of an organised newspaper scan. The archival value of the references is unlikely to be high, although the current awareness value might be so.

The 1874 journal references are drawn from 517 journal titles, and a list of these, showing the number of references derived from each is given in Appendix B. 43 journal titles (8.32% of the total) produced 945 of the journal references (50.53% of the total journal references) and these titles are listed in Appendix A. The number of journals contributing a given number of references is shown in Table 2 below.

TABLE 2.

Number of journals contributing a given number of references											
No. of refs.	1	2	3	4	5	6	7	8	9	10	11
No. of jnls.	286	89	33	26	13	9	9	3	4	2	1
No. of refs.	12	13	14	15	16	17	18	19	20	21	22
No. of jnls.	6	3	7	1	4	1	1	1	2	1	3
No. of refs.	26	27	28	30	33	35	37	38	40	48	83
No. of jnls.	1	1	1	1	1	1	1	1	2	1	1

A graph showing the percent of journals required to produce a given percent of references is shown in Figure 1. For purposes of comparison, this is shown with a similar curve derived from Physics Abstracts journal coverage for 1965. A brief table of comparisons is shown below.

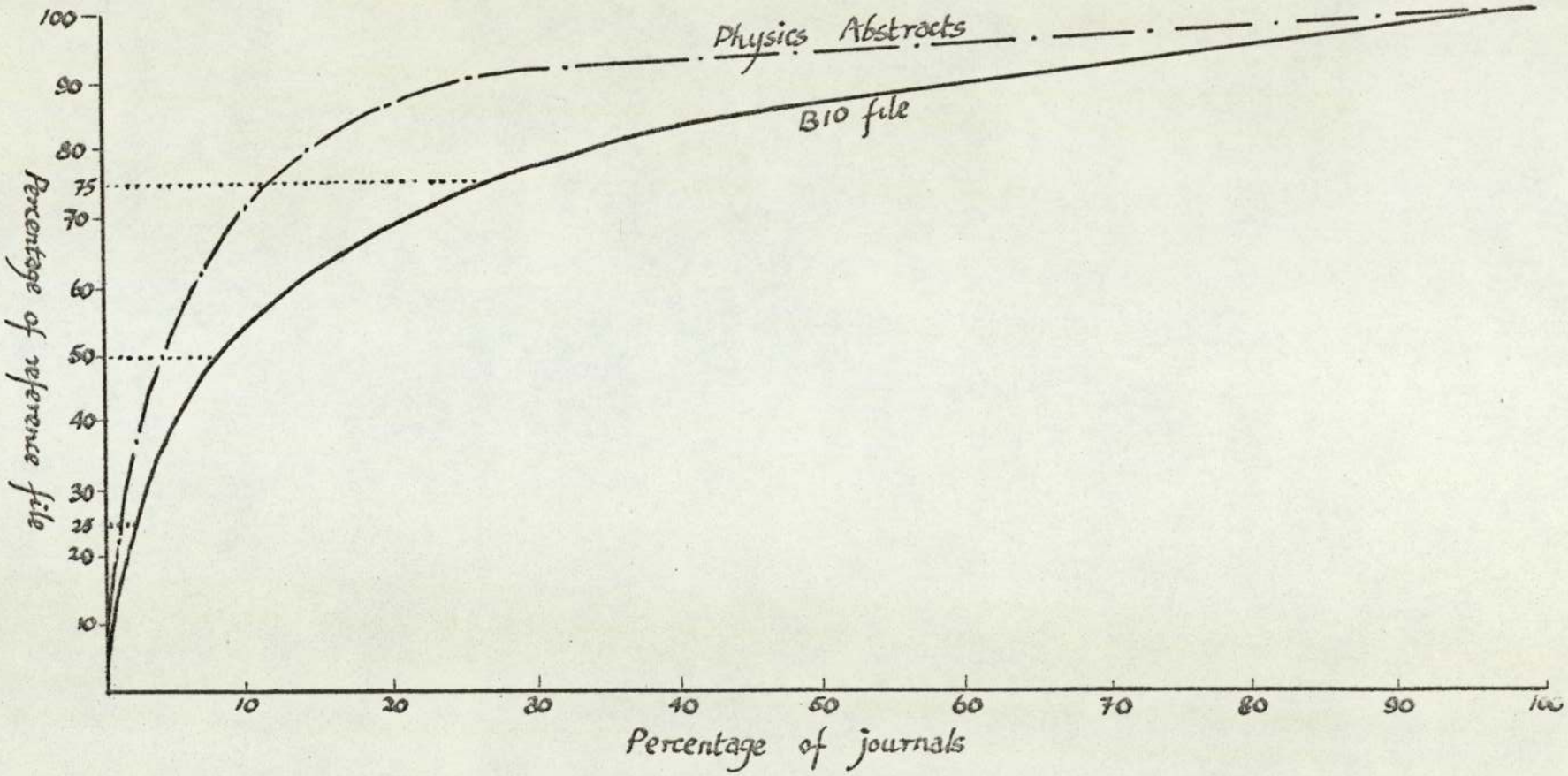


Figure 1. Comparison between Physics Abstracts and Bio file.

TABLE 3.

Comparison between Physics Abstracts and the BIO file.				
	Physics Abstracts		BIO file	
% cover	No. of jnls	% of jnls	No. of jnls.	% of jnls.
50%	23	3.6%	43	8.3%
75%	57	11.5%	119	23.0%
90%	126	25.5%	327	63.2%

The Physics Abstracts 100% coverage was contained in 495 journal titles. 16.2% of the journals contributing references to Physics Abstracts produced one reference each, as against 55.3% of the journals contributing to the BIO file. Physics Abstracts produced 32,279 abstracts drawn from 495 journals; the BIO file has 1874 references, from 517 journals.

This comparison is interesting, because it tells us something about the nature of Biodeterioration as a subject. It is, of course, a very much smaller subject in terms of current literature than physics, about one-seventeenth the size, and the core journals in biodeterioration. However, the scatter of the biodeterioration literature through the journals is much more extensive than is the case in physics. This suggests either that the Biodeterioration Centre is more effective in collecting fringe and peripheral material in its area of interest than is the Institute of Electrical Engineers, or that the criteria for inclusion in Physics Abstracts are more stringent than those applied to IBBRIS, or that the literature of biodeterioration is of its nature more diffusely spread. As the coverage of the BIO file extends downwards from pure science into technology, unlike that of Physics Abstracts, there is almost certainly a difference in the criteria of acceptance of candidate material, but despite this there is still clear evidence that Biodeterioration is, as has been claimed, truly an interdisciplinary subject, and that it may reasonably be assumed that the amount of effort required

to be expended per unique relevant reference acquired is greater in Biodeterioration than in Physics.

The language distribution of items in the BIO file has not been examined, because this would have meant item-by-item scrutiny of every document referred to, but it may be helpful to note that the 1971 issues of IBBRIS carried references to a further 254 items (from all media, 179 being journal items) drawn from the 1969 literature, and that these items covered a range of 18 different languages, including Serbo-Croat, Turkish and Korean. English is the greatest contributor, with Germany the next greatest, and Polish literature is well represented, particularly, as will be seen, in the items not covered by any of the secondary services studied. A large contribution from one language-group might be assumed to be indicative of a relatively high degree of interest and work in the topic-area within the country or countries of the group, but it might also indicate the presence of a particularly industrious cooperating specialist. In the particular case of Poland, there is both strong interest and high activity, and good cooperation. The high proportion of German influence reflects the strength of the German wood industry, and the German meat processing industry.

The examination of a predetermined range of journal titles is one way of collecting references. It has been shown that for Biodeterioration this is not a very easy process, particularly when the problem of selection of titles for scan is considered. There were 286 journals in 1969 contributing one reference each, and there might well be 286 in 1970, but there is no reason at all to suppose that they might be the same 286. As an illustration of the difficulty of forecasting which journals are potential contributors, it is interesting to compare the list of journals from which the 1969 references were drawn with a list of 'Master Journals' likely to contain relevant material, which was produced by O.E.C.D. in October 1966. The O.E.C.D. list was compiled, it is believed, by a committee of experts. The comparison is shown in Figure 2, below.

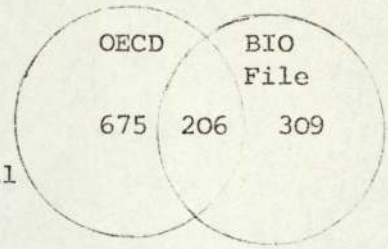


FIGURE 2. Comparison of OECD Master journal list and BIO file journal list.

Another common device for locating possibly relevant material is to search for work by known authors. An analysis of the authorship of the BIO file items indicated that 65 were anonymous, and that there were a total of 1490 distinct first authors. There were 221 authors who were responsible for two or more papers, and the 221 produced 540 papers between them. The distribution is shown in Table 4 below.

TABLE 4.

First-author distribution by number of references written.									
No. of refs.	1	2	3	4	5	6	7	8	TOTAL
No. of authors	1269	162	36	17	2	0	2	2	1490'

It can be seen that using author names alone as a retrieval device is not likely to give high recall, because so many of the authors are producers of only one reference; if one assumed that authors of two or more papers were likely to continue to produce papers of interest at the same rate, then this would only account for about a quarter of the file, if the assumption were entirely true. The precision of an author search would also be poor, because the interdisciplinary nature of the subject enforces interest in material originating in fields the greater part of which are not of relevance to biodeterioration. For example, a chemist may be interested in, and publish much about, a particular series of chemical compounds, one minor member of which may have a fungicidal action; this single item would be relevant to biodeterioration, but the greater part of his work,

dealing with the structure, preparation and so on of the other members of the group would not.

It was anticipated that considerable use would be made of key-words occurring in the titles of journal papers, as retrieval devices in searching computerised secondary services. Therefore a study was made of the vocabulary of the items comprising the BIO file. The whole file was keyboarded and run through the INDACS suite by the London University Computing Services Ltd., which produced an author listing, a journal listing, a KWIC index and a 'Double-KWIC'. The 'Double-KWIC' gave, alphabetically by word, the frequency of occurrence of all words (with the exception of a few non-significant articles, conjunctions and prepositions), together with the frequency of occurrence of pairs of words. In the titles of the journal references which comprise the BIO file, 525 words occur five or more times, and a list of these is given in Appendix C. These words were reduced to stems by taking all the forms of a word present in the file and extracting the longest stem common to all variants. (For example, the words 'fungus, fungi, fungal, fungicidal, fungitoxin, fungous, fungoid' share the common stem 'fung-'.) There are 534 stems occurring five or more times, and these are listed in Appendix D. Because words can occur twice within a single title, the frequencies shown should not be interpreted as being the number of titles containing a given word. The most frequent word, FUNGI, occurs 119 times out of a total of 6802 occurrences of words occurring five or more times. The most frequently occurring stem, FUNG-, occurs 207 times out of a total of 9010 occurrences of stems occurring five or more times. The total number of words (not of unique words) is approximately 12,500, but this figure includes a few prepositions which were not on the stop list, and also includes a number of other words such as 'studies' and 'effect' which would normally be placed on stop lists.

Some stem-pair frequencies were calculated for the most frequently-occurring stems, and these are displayed in Table 12, on page 49.

4. Coverage by secondary services.

The computerised secondary services which contain substantial amounts of material of relevance to Biodeterioration are the services offered by the Institute for Scientific Information (Science Citation Index®, Permuterm®, ASCA IV®, and SCI tapes), Biological Abstracts (BA-Previews), Chemical Abstracts (CA-Condensates, Chemical Titles (CT) and Chemical-Biological Activities (CBAC)), Index Medicus (MEDLARS) and Food Science and Technology Abstracts. PANDEX, which would have been suitable, was not available for consultation. Some other services were considered for inclusion in the study, but sample checks indicated their coverage of biodeterioration material to be so small that their use would not be practical.

ISI Services.

The Institute for Scientific Information's services are largely based on processing a large number of journals (2180 in 1969) on a cover-to-cover basis. The journals selected for coverage are effectively the core journals of science, chosen because by various indicators they appear to be the most heavily used or the most seriously regarded journals, and cover a high proportion of the literature of science and technology. The coverage is multidisciplinary. The Source Index (the printed version of the Source Tape available on-line as Scisearch) contains details of the authors, titles, and journal references of all the papers published in the journals covered by ISI, and is arranged alphabetically by first author, 'See' references being provided for other authors. The Citation Index orders, alphabetically by author, every citation found in the papers which are included in the Source Index, and gives details of the citing papers, sufficient to allow the citing papers to be looked up in the Source Index. The Permuterm index lists every significant word in the titles of the papers included in the Source Index, alphabetically, with every other word with which they occur, and a guide to the Source Index entry relating to them. Examples of the Source Index and Permuterm are given in Appendix E. ASCA IV is a weekly current awareness service

based on the current week's entries to the ISI systems, and access is made via a profile which can be composed of words in titles, authors names, journal titles, cited papers and a number of other elements, which can be associated with Boolean logic.

Coverage of the BIO file material by ISI was checked via the 1969 and 1970 Sources Indexes, using the first author's name as entry. Anonymous items were checked using the journal title as entry. Of the 517 BIO file journals, 225 were included in ISI coverage in 1969 (several more have since been added) and the consequent item coverage should have been 1147, or 61.17% of the total items. Ideally, all these items would be found in the 1969 Source Index, because the annual cumulations appear in the spring following the year to which they relate, and this time-lag is sufficient to allow of catching the majority of items published in the preceeding year. However, some journals, particularly those originating outside the United States, are received too late by ISI to be entered into the year-volume which properly should house them, and these are then processed into the following year. Actual coverage by ISI was 1119 items, representing 39.71% of the BIO file. Of the items covered, 42 were notified in the 1970 Source Index, the remainder being found in the 1969 volume. Thus 96.25% of the relevant material covered by ISI was notified within the year of primary publication.

Of the items not covered, although in accordance with the policy of complete coverage of specific journal titles they should have been covered, some were missing because a particular journal was not added to the ISI list until half-way through the processing year; an example of this, ironically enough, is the International Biodeterioration Bulletin. Some others are missing because a particular issue of a journal was not available, perhaps being lost in delivery or grossly late in publication. A few items, not surprisingly, are undoubtedly missed through human error of one sort or another, but these are extremely few; it is a remarkably error-free system.

Coverage by ISI Source Index is particularly easy to check, because against each author's name appears the title and full reference of the appropriate source item, so that checking a given reference for inclusion in a particular annual volume is a single action. This is in marked contrast to some other services; in Biological Abstracts, for example, authors' names are listed with no more than abstracts numbers against them, and it is necessary to turn from the author index to the abstracts whose numbers are shown, in order to check whether any one of the numbers refers to the abstract of the item which is being checked. This is particularly trying when, as happened once, over fifty abstracts have to be checked for a possible match with a sought item, and none of them does in fact match. In practical terms, this meant that check of the Source Index was completed within a week, whereas most services required at least a month's work.

While searching the Source Index under authors' names, papers not in the BIO file but by the authors of items that were in the BIO file, were noted, and passed to the Biodeterioration Information Centre for assessment of relevance. 89 of these papers were adjudged 'relevant', and a further 53 'possibly relevant', relevance judgements were made solely on the references as given, that is, on the basis of author, title and journal. These references were evaluated by one member of the Centre staff, whereas the BIO file references had been selected by another member, so that, as all the references, being presented in Current Contents, had been scanned (or had a probability of being scanned equal to that of the references included in the BIO file) by this member, some of the 'relevant but not in BIO file' items might not have been judged relevant had they been evaluated by the centre reference selector. Similarly, experience has shown that one evaluator will on one occasion judge a given reference to be relevant and on another occasion will consider it either of possible relevance or irrelevant. Consequently, it is not correct to say that the 89 papers judged relevant as described above represent actual failures to include relevant material; all that can be said

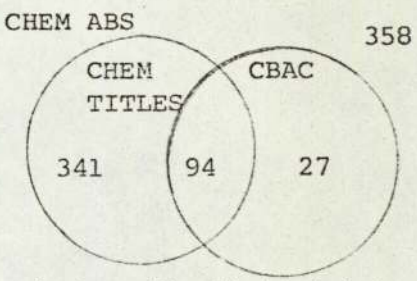
with any confidence is that the Biodeterioration Centre's coverage of the literature is not perfect (and the same applies to any system for collecting references), and that its 'undercover' might perhaps be of the order of four or five per cent, on the basis of search by authors' names. Put in a more practical way it can be shown that searching by authors' names does increase recall, but at the penalty of retrieving also very considerable amounts of irrelevant material.

Chemical Abstracts Services.

The Chemical Abstracts services which were examined for possible content of BIO file items were Chemical Abstracts itself (CA-Condensates being the tape equivalent), Chemical Titles (CT) and Chemical-Biological Activities (CBAC). CBAC has been replaced by an equivalent subset of Chemical Abstracts and CT is being discontinued as tape service in the United Kingdom with effect from the end of 1973, so that for these services the results of this study are largely of academic interest. CBAC and CT are subsets of Chemical Abstracts, and therefore any items covered by either of these two services is also to be found in Chemical Abstracts, but they were included in this study because they provide alternative access points to the chemical literature which, in certain circumstances, might be cheaper than use of the full base. In all three services, coverage was checked via the author indexes. Chemical Abstracts covered 820 items from the BIO file (43.76%) in 1969 and 1970, Chemical Titles covered 435 (23.2%) and Chemical-Biological Activities covered 121 (6.5%). The overlap is shown in Figure 3 below.

Figure 3.

Relationship between Chemical Abstracts, Chemical Titles and Chemical-Biological Activities cover of the BIO file.



CA-Condensates is issued weekly, but is broadly divided into two sections, odd and even, and each section appears fortnightly.

The odd sections cover biochemistry and organic chemistry, and the even cover macromolecular chemistry, applied chemistry, chemical engineering and physical and analytical chemistry. The odd sections covered 693 BIO file items (85%) and the even covered 127 (15%). BIO file items were found in 18 out of 20 Biochemistry sections, 8 of 14 Organic Chemistry sections, 7 of 12 Macromolecular chemistry sections, 10 of 18 Applied Chemical and Chemical Engineering sections and 4 of 16 Physical and Analytical Chemistry sections. The distribution among sections in detail is shown in Table 5 below.

TABLE 5.

Distribution of BIO file items among Chemical Abstracts sections

Section heading	No of items
EVEN issues	
Macromolecular chemistry. (secs 35 - 46)	
Cellulose, lignin, paper and other wood products	37
Coatings, inks and related products	21
Fats and waxes	2
Industrial carbohydrates	1
Plastics manufacture and processing	1
Surface-active agents and detergents	8
Textiles	7
Applied chemistry and chemical engineering (secs 47 - 64)	
Cement and concrete products	2
Essential oils and cosmetics	2
Extractive metallurgy	2
Ferrous metals and alloys	2
Mineralogical and geological chemistry	2
Petroleum, petroleum derivatives & related products.	2
Pharmaceuticals	6
Pharmaceutical analysis	1
Sewage and wastes	16
Water	7

Table 5. (continued)

Physical and analytical chemistry (secs 65-80)	
Electrochemistry	5
Phase equilibriums, chemical equilibriums and solutions	1
Organic analytical chemistry	1
Surface chemistry and colloids	1
ODD issues	
Biochemistry (sections 1 - 20)	
Animal nutrition	6
Biochemical methods	13
Enzymes	46
Fermentations	59
Fertilizers, soils and plant nutrition	10
Food	84
General biochemistry	5
History, Education and documentation	1
Immunochemistry	1
Mammalian biochemistry	3
Microbial biochemistry	238
Non-mammalian biochemistry	10
Pesticides	97
Pharmacodynamics	12
Plant biochemistry	16
Plant growth regulators	24
Radiation biochemistry	5
Toxicology	51
Organic chemistry (secs 21 - 34)	
Aliphatic compounds	1
Alkaloids	1
General organic chemistry	1
Heterocyclic compounds	3
Non-condensed aromatic compounds	3
Steroids	1
Synthesis of amino acids and proteins	1
Terpenes	1

Biological Abstracts

Biological Abstracts is produced by the BioScience Information Service of Biological Abstracts (BIOSIS), and notifies roughly 18,000 references a month drawn from about 8,000 journals, plus reports, monographs and conference proceedings. BA Previews is the magnetic tape version of all items appearing in Biological

Abstracts and the BioResearch Index, BA Tapes being available fortnightly and BIO I tapes monthly. Biological Abstracts printed versions for 1969 and 1970 were searched for inclusion of BIO file items, via the author index. This stage of the project was most arduous, because the author indexes give only an abstract number against an author's name, and this abstract has to be checked for possible candidacy; in the case of some prolific authors (one of whom had 119 abstracts numbers against his name, the 105th being the BIO file item sought) a great many abstracts had to be examined to determine which, if any, referred to the particular BIO file item being sought. The coverage of BIO file items within the two years of Biological Abstracts examined was 746 (39.81%) items.

Index Medicus.

MEDLARS (Medical Literature Analysis and Retrieval System) is one of the oldest computer-based systems, the magnetic tapes used being first produced for the photo-composition of Index Medicus. Index Medicus still remains as the printed version, with the difference that more indexing terms per document are available on tape than are shown in Index Medicus. The coverage of BIO file items in Index Medicus 1969 and 1970 was 463 (24.71%) items.

International Food Information Service.

Food Science and Technology Abstracts were first published in 1969, and became available on tape in 1971, sponsored partly by the Commonwealth Bureau of Dairy Science and Technology. Although the tape service was not, strictly speaking, available for use in 1969, the target year of the BIO file, it was thought justifiable to include it in the study because of its obvious relevance and the possibility that it would be available on tape in the future. In fact tapes are now available which cover the whole of F.S.T.A. from its inception. Its coverage of BIO file items was 301 (16.06%) items.

The overall coverage of the BIO file.

The overall coverage by the services studied of the BIO file material is illustrated in the Venn diagram in Figure 4 below. Because of the complexity of this diagram, it is worth extracting some of the data and presenting it separately in Table 6, and Table 7 below.

TABLE 6.

Some aspects of coverage by secondary services of BIO file items		
No. of items covered by 1 service only	354	18.89%
No. of items covered by exactly 2 services	432	23.05%
No. of items covered by exactly 3 services	420	22.41%
No. of items covered by exactly 4 services	234	12.49%
No. of items covered by 5 services	19	1.01%
No. of items covered by 2 or more services	1105	58.96%
No. of items covered by 3 or more services	773	35.91%
No. of items covered by 4 or more services	253	13.50%
No. of items not covered by any service	435	23.21%

TABLE 7.

Unique coverage of BIO file items by secondary services.	
Service	Unique items covered
I.S.I.	144
C.A.	88
F.S.T.A.	55
B.A.	47
I.M.	20

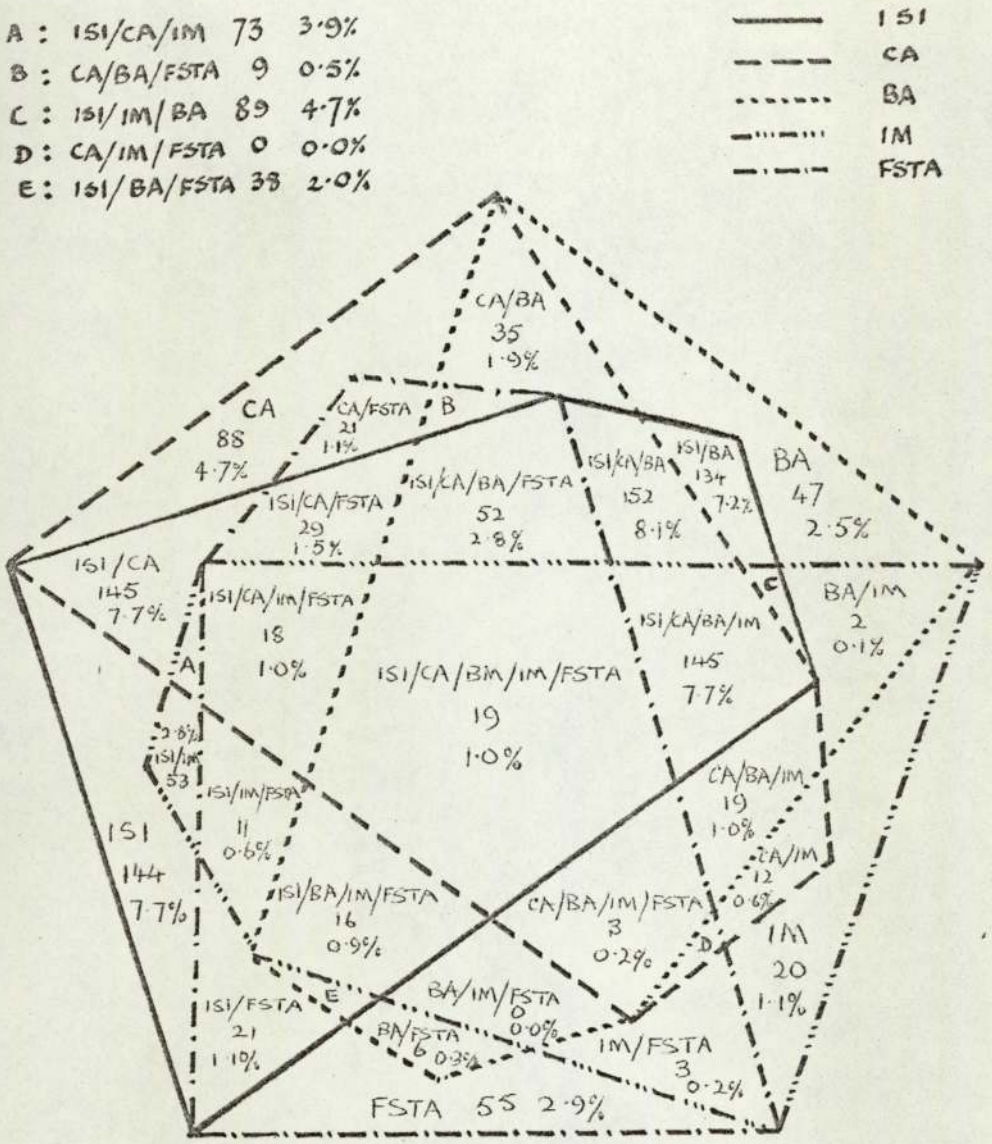
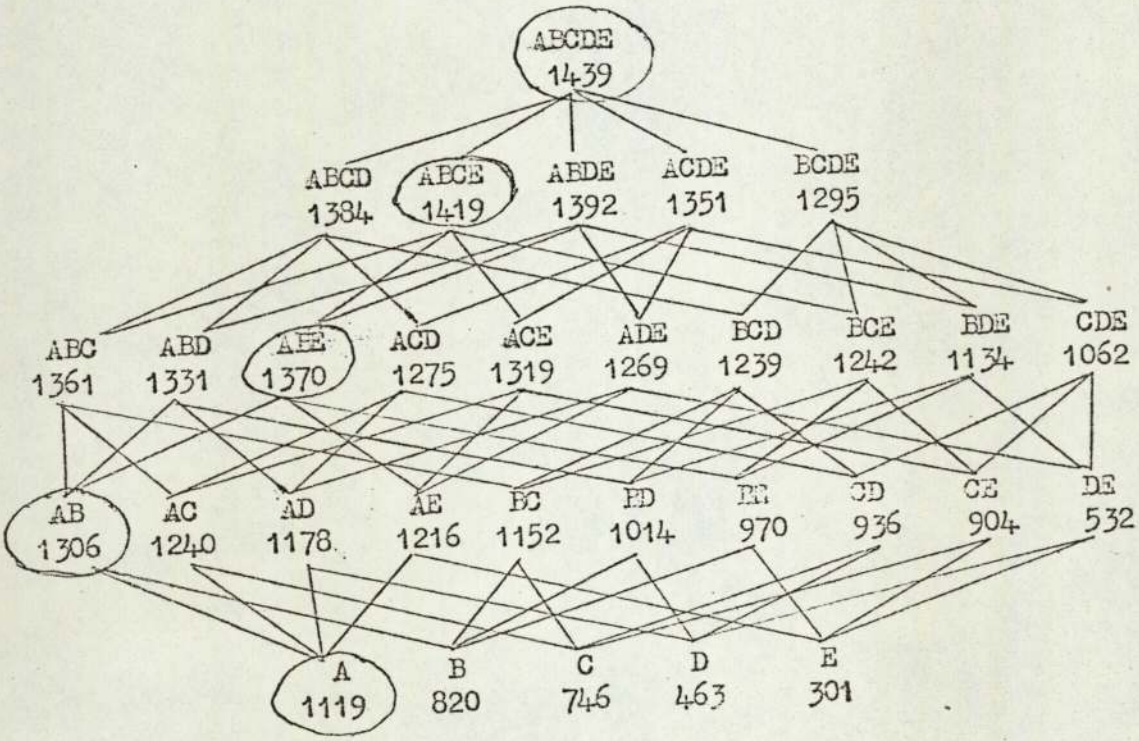


Figure 4. Coverage of the BIO file by secondary services.

ISI ——— CA ---- BA ..... IM +++++ FSTA -·-·

The same information shown in figure 4 can be presented in the alternative way shown in Figure 5 below, which illustrates the coverage of each individual secondary service, the coverage by any pair, any triple, any quadruple, and all five. The usefulness of this presentation is that the choice of best combination of services to provide maximal coverage of the topic is simplified.



A: I.S.I. B: C.A. C: B.A. D: I.M. E: F.S.T.A.

Figure 5. Coverage of the BIO file by secondary services: choice of service.

It can be seen that choice of the single best service, ISI, provides 59.71% of BIO file coverage. Addition of the next best service, CA, raises cover to 69.70%, an increment of 12.86%. Addition of FSTA raises cover to 73.11%, an increment of 3.42%. Addition of BA gives an increment of 2.61%, to 75.72%. and finally IM, with an increment of 1.07%, raises cover to the total of 76.79%. These increments carry the penalties of discarding, for two services, 633 duplicate references, for three 870 duplicates, for four 1567 duplicates and for all five 2010 duplicates.

The information presented in the preceding figures is, while factually accurate, possibly misleading. The identification of secondary sources of references and the choice of combinations of sources suggested in Figure 5 takes account only of the possible recall from the given services. That is to say, if it were possible to recover all the relevant references contained in a service, then the effects of using single services and combinations would be as shown, and the choices made would be as indicated. However, simple recall alone (assuming perfect recall to be possible) is never the sole criterion. Cost of recall must also be taken into account. For instance, referring to figure 5, service A is shown as containing 1119 references, while B and C together contain 1152. A, being a single service, gives each reference once only, whereas use of B and C together give, in addition to 1152 unique references, 414 duplicates, which would in an actual operation have to be identified and discarded, a process entailing a certain cost. If A, B and C each cost the same amount of money to access, P, then the cost per unique citation (usable) offered by A would be P/1119, whereas the cost per unique citation offered by B and C would be $(2P$ + cost of eliminating 414 duplicates$) \div 1152$, so that the cost for each usable citation obtainable from A would be something less than half that for each obtainable from B and C together. But if it happened that the cost of accessing A was high, while the cost of B and C were very low, then the cost advantage, and consequently the more economic choice, might lie with the BC combination.

This situation could be explored by means of a model, which would take account of different degrees of overlap among services, and different pricing structures, in order to arrive at an algorithmic decision rule. The situation via-a-vis the BIO file is relatively simple, but it is conceivable that in some situations, where for instance substantial non-overlapping coverage is provided by the services in the B - E positions, that at the level of pair combinations, the pair providing highest recall would not necessarily include the single service which at the single level provided the best.

A further complication which will be exemplified subsequently is that although a service may well include a given number of relevant items, the retrieval of all of the items by any strategy which will not also retrieve the whole of the non-relevant items included in the file is seldom if ever possible, and that in practice, retrieval of more than about 60% is seldom economically feasible. Additionally, retrieval of relevant items is invariably accompanied by retrieval of a number of irrelevant ones which must, at some cost, be screened out, so that a realistic model of the situation would have to take account, not of contained items, but of economically retrievable items. This is, unfortunately, impossible to test experimentally, except at enormous cost. In order exactly to compare the 'recall coverage' of a group of services, searches would have to be carried out which would be retrieving elements of a single identifiable set (typified in the present instance by the BIO file), but the differing timelags inherent in different secondary services means that 'recall coverage' for each service would relate to a different universe in each case. The only way in which the universes peculicar to each service could be harmonised would be to take a series of issue of each service - for example, Chemical Abstracts for 1969, 1970 and 1971 - and take from the series all items other than those relating to a single year; thus one would have a synthetic CA 1969, BA 1969 and so on. This having been done, then the synthetic services could be treated as exemplars of the real, and comparable recall coverage figures obtained.

Implications of the coverage study.

The three salient points of the coverage study are that no single service covers more than 60% of the BIO file items (the coverage figures for the services examined are extracted in Table 8 below), that there is a high degree of overlap of coverage between the services examined, and that 435 BIO file items were not covered by any of the services within the time-period of the study.

TABLE 8.

BIO file coverage by secondary services.		
Service	No items covered	% coverage
ISI	1119	59.71
CA	820	43.76
CT	435	23.22
CBAC	121	6.55
BA	746	39.81
IM	463	24.71
FSTA	301	16.06

Because timeliness of coverage bears to some extent on the problem of deciding which service(s) to use in collecting a reference base; Table 9 below gives, for each service examined, the number of BIO file items notified by each in 1969 and 1970 issues, and the percentage of total items notified by each service which appeared in 1969 issues.

TABLE 9.

Timeliness of secondary services			
Service	Items in 1969	Items in 1970	1969/Total x 100
ISI	1072	47	95.80
CA	539	281	65.73
CT	404	31	92.87
CBAC	99	22	81.82
FSTA	174	127	57.81
IM	232	231	50.11
BA	294	452	39.41
(BIO/IBBRIS)	501	1373	26.73

The implications of the three points mentioned earlier as emerging from the coverage study are that because no single service

provides more than partial cover, it is necessary to look in a number of services in order to attempt reasonably complete coverage of the topic. This inevitably entails finding the same reference in a large number of cases in several different services, and it also must be realised that complete coverage by the use of existing secondary services is, to all practical purposes, impossible.

It is not really surprising that no single secondary service provides complete coverage. If complete coverage had been available from a sole service, then there would have been little justification for the existence of IBBRIS as a separate publication. The Biodeterioration Information Centre came into being under OSTI sponsorship, and with OECD interest, because it was felt that Biodeterioration was a definable topic of economic importance, with an associated group of potential users of information on the topic whose literature access requirements could could not adequately be met by any previously existing secondary information service. That the Centre's services are now regularly used by several hundred subscribers, a sufficient number to make the Centre economically viable, is adequate demonstration of the existence of the need group; the present study shows that the assumption that their needs could not be met by any single existing service was correct. However, it must be remembered that we are here dealing with a need group, and not the needs of an individual; it is conceivable (although not not very likely) that any single individual member of the Biodeterioration need group might, if his particular interests were sufficiently narrowly defined and fell entirely within one of the formal disciplinary areas, be able to service his needs from one service alone. Whether he could extract his own interest-slice as conveniently or as cheaply as by the use of the Centre services is doubtful, but this point is outside the scope of the present study.

Being obliged to use more than one information service for references on the topic means finding the same reference, in many cases, more than once. One consequence of this is, that

because a cost is entailed for the acquisition of each reference, whether one already has it or not, there are many occasions on which one pays several times over for the same reference. This means that if the cost of acquiring a reference were standard for all systems, then the unit cost of acquiring one reference would be greater if using two systems than if using one, and the cost per unit reference acquired increases (non-linearly) as more services are used. Another consequence is that where there is danger of acquiring the same reference more than once, then some routines must be set up to avoid re-entering already held references into the system. The simplest way is by setting up a card file with one entry for each reference acquired, and matching each new candidate item against the file to determine its novelty or otherwise. However, this is a combrous process, and entails an extra cost. Rather than comparing candidate items with the entire store, it would be a considerable saving of effort were it possible to compare the incoming material with only a part of the store, that which had entered the system within a certain fixed period of time. Unfortunately, this does not effect a very considerable reduction, because as can be seen from Table 9, the timeliness of services varies considerably. In effect, a reference may be obtained several times during a three-year period amd the file which must be held for identification of non-novel items by comparison should cover the three preceeding years. So the costs incurred by use of more than one system are, increased costs per unique reference added to the store, and costs of identification and elimination of duplicate material.

Perhaps the most disturbing point arising from the coverage study is that a substantial number of the BIO file items were not covered by any of the secondary services examined. It has been credibly estimated that about three and a half million references are being put onto tape annually, (8) and that about a million plus items are appearing annually in the journal literature; it seems remarkable that with the high redundancy implied by these estimates, something more than a fifth of the literature of a particular scientific topic can have slipped through the net.

It does not, of course, necessarily follow that because not all the literature of the topic is covered, there is any consequent loss of information, because the same piece of information may appear in a number of locations and languages, written up for the consumption of a variety of different audiences. But secondary services are concerned primarily with documents, and only incidentally with the information content of documents, and it is unrealistic to suppose that any major service can afford the time and intellectual effort necessary to determine the novelty of the information content of every document offered for entry into the service; the primary criterion for inclusion is subject, not novelty. It might be thought reasonable to suppose that the missing items had been excluded by virtue of being trivial, ephemeral or in some other recognisable sense 'rubbish'. With this thought in mind, each non-covered item was therefore extracted from the collection held at the Biodeterioration Information Centre, and the document inspected. Each document was rated by three persons, a subject specialist on the Centre staff, the Centre staff member who selects references for inclusion in IBBRIS and the present investigator. Out of the 435 documents examined, 54 (12.41%) were noted as being trivial or ephemeral by at least one examiner; although not all judgements coincided, there was a high degree of correspondence. The remaining 381 documents were considered by all three examiners to be worthy of inclusion in IBBRIS, and therefore worthy of permanent retention in the Centre collection.

A possible explanation for the omission of these items was that they appeared in obscure languages. The distribution by language of the journals carrying these items and of the items themselves is shown below, in Table 10.

TABLE 10.

Language distribution of non-covered items and of journals carrying non-covered items.

Language	No of jnls	No of items	Language	No of jnls	No of items
English	93	241	Sebro-Croat	2	3
German	41	88	Spanish	2	2
Polish	12		Roumanian	2	2
French	10	13	Hungarian	2	2
Russian	9	20	Portugueses	1	1
Japanese	7	10	Ukrainian	1	1
Italian	4	5	Danish	1	1
Dutch	3	3			

One German journal (Material und Organismen, mit Beihefte) accounted for 33 of the non-covered German-language items, and one Polish journal (Biblioteka Muzealnictwa I Ochrony Zabytkow Ser. B, a museum journal which devoted one issue in 1969 to problems of material preservation) accounted for 19 of the non-covered Polish items.

It is evident that the failure to collect the bulk of the non-covered material was not attributable to its appearance in obscure languages. 52 of the journals containing non-covered items had carried other items which had been covered by one or more of the secondary services studied, so that it may be assumed that not all the missed items were missed because of their appearance in 'obscure journals'.

Careful examination of the non-covered documents suggests an alternative hypothesis. Few of the documents carry 'new' information or appear to constitute the first publication of a new discovery, theory or other increment to the corpus of science, although there are a number of instances of reports of the occurrence of an organism, or a biodeterioration problem, in a new location. The bulk of the missed documents appears to be written by scientists for technologists, passing on and sometimes

summarising the results of recent research for the benefit of practitioners; they constitute, that is to say, examples of the transfer of knowledge from its origins in pure science to its point of use in technological application. This class of material is not perhaps what a scientist wishes to find when he searches the literature for information on his core interest, but it is very much the material a technologist wants when he has an information need. It was the view of members of the Biodeterioration Information Centre that this class of material was of the greatest value in carrying out consultancies, as being more nearly the level of information which the practical user required. This being so, it is unfortunate that it seems to be excluded from the major secondary computerised services. It is suggested that its exclusion arises from the old conception of a secondary service as being a device for simplifying the information-seeking problems of the research scientist; in an age when the economic exploitation of every asset, including information, is of importance, this conception is no longer tenable.

5. Profiling

A bibliographic record carried on magnetic tape contains a number of elements which together comprise the record, and is recalled by a search specifying one or more of the elements it contains. The number and type of elements varies from system to system. (9) but the basic set of elements making up the record for a particular journal item generally contains entries for the author or authors of the item, the title of the item, the journal from which it originates, the year of publication, and the volume and page numbers of the item (although in some systems only the number of the first page is carried). There is usually also a unique identifying number, equivalent to the accession number of an item entered into a formal library system. Other elements which appear frequently but not invariably include assigned keywords or other indexing elements, indicators of the nature of the item (e.g. letter, bibliography, note etc), language of item, organisational

affiliation of authors, and in the case of some ISI services, citations appearing in the original document.

A profile is a list of terms, of the same type or types as the elements making up the bibliographic records in the system to be searched, which in a specified logical association are equivalent to a question to be put to the system being searched. Not all systems can be searched by all the elements present on the tape (for example, a word-in-title search cannot be done on MEDLARS), nor can all possible logical combinations necessarily be used to search all systems. There are sometimes restrictions on profile length. A list of profiling devices follows.

TABLE 11.

Profiling devices.

Recognition devices		
Single term	Identifies all records which contain the term, exactly as stated.	
Capitalisation	Identifies all records which contain the term, but only if it appears in capital letters.	Identifies acronyms that look like common words (EARS not ears)
File indicator	Identifies the file or subfile to be searched.	
Field indicator	Identifies the field within a record which is to be searched.	
Logical operators		
OR	Identifies all records which contain any one or more of the terms which it links.	
AND	Identifies all records which contain all the terms linked with it, occurring together.	

NOT	Excludes all records which contain the negated term.	
IGNORE	A variety of negation excluding particular words which contain a specified stem but are not required words.	SPOR* IGNORE SPORT*finds spore, spores,excludes sport, sports, sporting.
ABS	Indicates that the term so specified is to produce a match irrespective of any other logic (Overrides NOT).	

Context operators

Universal character	Identifies all records containing the term-fragments given, with any single character replacing the universal.	ON*LINE finds on line and on-line.
Right truncation	Identifies all records containing the term, with any suffix or none.	SPOR*finds spore, spores, sporiferous etc.
Left truncation	Identifies all records containing the term, with any prefix or none.	*MYCIN finds streptomycin, aureomycin etc.
Left and right truncation	Identifies all records containing the term, with any prefix and/or suffix or none.	
Infix truncation, imbedded truncation	Identifies all records containing the words which begin and end with the term-fragments given, with any character or characters (except space) intervening.	SUL*UR finds sulfur, sulphur.

In using truncation, the numbers of characters to be accepted at the truncation site may be specified. If specified, right truncation equates to ITIRC selective masking; if unspecified, right truncation equates to ITIRC unconditional masking. Truncation with one character only to be accepted equates to use of the universal character.

ADJ	Identifies all records containing the specified terms in the order specified, if adjacent.	INFORMATION ADJ RETRIEVAL finds information retrieval, not retrieval of information.
WITHIN	Identifies all records containing the specified terms within a sentence or specified number of words.	

These devices, which vary in form, can be used to retrieve references to multi-word concepts where the meaning of the concept depends on the order of the words, and where the words may have intervening words separating them. An example is retrieving INFORMATION RETRIEVAL and INFORMATION STORAGE AND RETRIEVAL but not RETRIEVAL OF INFORMATION. (This is a subset of INFORMATION AND RETRIEVAL).

Other devices.

Limit or range indicator	Indicates limit of interests, usually in searching numerical data.	e.g. temperatures, 0^{o} - 100^{o}
Term lists or back referencing	The practice of arranging search terms in groups prior to linking the groups with logical operators. The members of individual groups are effectively linked with OR logic.	

Context operators (except ADJ and WITHIN) have the effect of increasing recall or reducing precision, as does logical OR, and are used in broadening a search. Other logical operators except ABS, and ADJ and WITHIN have the effect of improving precision or reducing recall.

Weighting is an auxiliary or alternative device which can be used in profile construction, and its application is best described by the following extract from 'Techniques of Information Retrieval' By B.C. Vickery (10).

The cryogenic or low temperature behaviour of metals, alloys, superalloys or superconductors' can be expressed by the equation (A + B). (C + D + E + F), where

A	=	Cryogenic	5
B	=	Low temperature	5
C	=	Metals	1
D	=	Alloys	1
E	=	Superalloys	1
F	=	Superconductors	1

The numerals on the right are 'weights'. These were introduced first in order to rank the output of a search , which is necessary if a large number of references match a search question. A weight, as shown above, is assigned to each search term of the question. For each matching record, the total weight is calculated by the computer. Thus a record containing all six terms A to F would have a total weight of 14, whereas one containing only A and C (still meeting the search criterion), would have a weight of 6. Before printing out search results, the computer can rank them in order of weight. The 'weightiest' references are most likely to be relevant to the query.

The weights can also be used to simulate search logic. Instead of specifying a logical equation, the search terms can be weighted as shown and searched as a simple logical sum A + B + C + D + E + F, with the stipulation that the least total weight acceptable is 6. This allows any combination of A or B with one or more of C, D, E or F. The combination CDEF is not accepted (total weight is only 4). The combination AB (total weight 10) is accepted, although it does not meet the original search requirement, so the search logic is not perfectly simulated. More complex group weighting is also possible.

(* (A or B) and (C or D or E or F)).

Parameters often used in the performance evaluation of information retrieval systems are recall and precision, recall being defined as the proportion of relevant documents that are retrieved and precision as the proportion of retrieved documents that are relevant. Considering the total number of relevant documents that could be retrieved as being the total number published in a given year, and using the BIO file as being a good approximation to this last, then maximum possible recall for any given system examined equates to its coverage. Cost, both overall cost and cost expressed as cost per pertinent unique citation retrieved, timeliness, and (when more than one system is accessed) the uniqueness of retrieved pertinent references, are other parameters of importance to users in evaluating information systems. Different sorts of user are likely to attach different weights to these parameters. An individual in receipt of regular notifications from an SDI service may perfer to have a few highly relevant references regularly and may accept that a quantity of relevant items may be missed in return for the

assurance that a minimum of irrelevant material is sent to him. He would effectively attach greater weight to precision than to recall. Conversely, an individual requesting a comprehensive search for the purposes of preparing a state-of-the-art review would be forced to accept a large amount of irrelevant material in return for a reasonable assurance of as near an approach to comprehensiveness as the system's coverage would permit. In the present situation, it is reasonable to assume that a Centre which provides a service to an interdisciplinary group, aiming to fill as much as possible of the group's information needs for both current awareness and retrospective search, will try to maximise comprehensiveness. This means that it will give a high weight to recall, and will of necessity accept the consequent reduction in precision. In other words, it must accept a fairly high proportion of noise in order to be sure of having as many relevant references as possible; for example, if one search strategy produces fifty references, of which forty are relevant, and another produces five hundred of which fifty are relevant, then the Centre should prefer the second, because it needs to collect as many relevant references as possible. However, a complication is introduced when considering costs, because what is accepted with low precision is the cost of eliminating the irrelevant material. Cost is usually of extreme importance to a self-supporting information centre, because its funds are limited to what it can recover by selling its services, and it must therefore strive to acquire its reference-base at as low a unit cost per reference as possible. Therefore the type of profile it must use when drawing references from a computerised secondary service must be one which will provide as high a level of recall as possible, with as good precision as can be achieved consistent with maximal recall, and at as low a cost per pertinent citation retrieved as possible. Timeliness, although desirable, is not very important in this context, because for any one service timeliness is strictly a function of the service itself and is not affected by profile manipulation; if timeliness is given a high weight, it affects the decision as to which service or services to use, not the way profiles are written.

Choice of elements for profiling.

The most meaningful bibliographic elements which are present in virtually all computerised secondary service bibliographic records are authors' names, journal titles and titles of documents. The effects of using the first two on the ISI file may be noted. 935 of the BIO file authors are present in the ISI Source Index. Therefore to retrieve their items (1110 items, excluding the anonymous) 935 author names would have had to be used as search terms. On an ASCA profile, these would have cost $5 each, a total charge of $4675, or $4.212 per relevant citation retrieved. However, this assumes that the productive authors are known, which, in view of the high numbers of authors of single papers would not be likely to be the case. If authors who produced two or more papers in the year were used, this would have cost $1125; the number of references retrieved by doing this is not known, but 29.2% of the BIO file is covered by these authors, so if the proportions are preserved, the same percentage of ISI coverage might be guessed as being 29.2% of 1110, or 324 items, at a cost of $3.47 per pertinent citation retrieved. Clearly, authors' names are not very suitable profile terms for retrieval of BIO file items. This is not to say that authors' names are not very suitable in all cases; if the productive authors are known, and their topics of discourse are clearly known also, then they operate to provide high precision, but in a situation where high recall is required, the virtual impossibility of forecasting all authors of relevant material militates against their use.

Journal titles suffer from something of the same handicap, because in an interdisciplinary area, almost by definition the forecasting of all titles in which material may be present is practically impossible. In the present study, and in the majority of cases, the problem is exacerbated by the presence of several major journals which are small net contributors to the required file. The best example of this is 'Nature', which costs $154 on an ASCA profile, but contributed only 18 items to the BIO file from its approximately 3500 items published in 1969. If the top thirty contributing journals covered by ISI were used

as profile terms, this would produce 704 references for $483, giving a cost per pertinent citation retrieved of 68.6 cents, which is acceptable; however 415 references from the ISI file are left uncollected, which using journal titles, would have required listing a further 225 journals. At an average of slightly less than two relevant references per journal, it can be seen that the retrieved noise would be considerable. Again, this is not to decry use of journal titles in all search strategies, solely in this particular one.

In a sense, authors' names and journal titles are themselves something like precision devices in profile construction, and would seldom be used alone. They also tend to be used in answering questions other than 'I want all I can find about Thing'; as ISI says (11), the Source Author question enables ASCA IV to keep you informed of current research by key scientists in your field.

Because of the unsuitability of authors' names and of journal titles as profile terms in compiling the type of search required to give maximised recall, the type of profile used in this study is one based on natural language, that is to say, on the terms naturally occurring in the titles of documents, usually truncated to cover the majority of grammatical and trivial variants of a term. Because of the need to maximise recall (which is assisted by the use of truncated terms), the only logical operator used is the logical OR.

One major assumption which has been made is that use of computerised secondary services by an interdisciplinary group such as the Biodeterioration Information Centre would be made by the purchase of searches rather than by the purchase of tape for local searching. This means that ISI services would be searched by ASCA IV profile rather than by purchase or other acquisition of ISI tape for searching in a computer under the Centre control, or use of commercially available on-line search facilities. Chemical Abstracts services are searched by submission of profiles to the United Kingdom Chemical Information Service,

chargeable at the normal rates, and other services are searched similarly. This greatly simplifies the costing, in that in some cases fixed tariffs exist for service use, and costs are therefore reasonably consistent, whereas for locally-run searches, costs tend to be unique to location because of hardware and software variations. Also, in most cases purchase of tape by a small centre would be ruled out by its high capital cost, and the ridiculousness of acquiring a whole service in order to take off a few hundred references.

Profiling and ISI search.

The way in which profiles were compiled is best explained by discussing compilation of a profile for searching ISI tape, because the method of profiling adopted here evolved in the course of experiments with ISI services. It is worth commenting that ISI services are uniquely adapted for experiments of this sort. The composition of the Source and Citation tapes corresponds to the printed Source and Citation indexes, the published ASCA IV tariffs permit easy and exact costing of profiles, and the existence of the Permuterm index allows the effects of different profile terms and most forms of profile logic to be tested quickly and easily by hand. Since the printed versions are available for the period from 1961 onwards, there are no problems with the unavailability of early tape. ISI services lend themselves so readily to tests and experiments of the type conducted here that it is as though they had been designed with this sort of application in mind.

The Permuterm index, which was heavily used during this study, is compiled from the titles of all items entered into the Source Index in a given year. It permutes all significant words within each title to form all possible pairs of terms, with the exception of certain frequently-occurring words which are placed on a 'Stop' list. It is therefore possible with any word to see with what other words it co-occurs, and to see also what authors' documents will be retrieved by the use of the word in a search. A portion of the 1969 Permuterm index is reproduced in Appendix E.

Having discarded authors and journal titles as profile terms, words or word-stems were the remaining possible entry to the system (with the notable but untested exception, for ISI services, of citations). It is obvious that the profile which would retrieve the whole of a specialised collection from a file composed solely of that collection, would also retrieve all of the collection contained in a larger file indexed in exactly the same way; in this instance, this means that a profile made up of all the words contained in the titles composing the BIO file (a natural language system), when applied to the ISI file (also a natural language system) would inevitably retrieve all the BIO file references contained in that system. It is true that a number of words might occur in the BIO file which do not occur in the ISI file, because of the absence of the references containing them from the ISI file, and that there might be a number of non-productive words. It is true also that precision in this case is affected by the multidisciplinary nature of the ISI file, so that, in a general case, a physicist retrieving on 'plasma' collects both physical and medical references. (An amusing instance of this occurs using the stem 'fung-'; there is, it appears, a disease known as 'fung disease', on which one paper was written in 1969). In the BIO file, there are 207 occurrences of the stem 'fung-', which derive from 197 documents containing the stem (because of multiple occurrences of the stem in some titles); an inspection of Permuterm produces a list of 568 authors who wrote papers containing the term in 1969, and since some of these authors have written more than one paper containing the stem in the time-period, there are approximately 710 references retrieved by this stem. However, although precision can be seen to be affected by the contents of the major file, recall of the relevant subset contained in it is not.

The first step, then, is to compose a profile which, applied to a notional computerised version of the BIO file, would retrieve the whole of that file by use of the terms contained in the titles of the records. The simplest way would be to write a profile containing every word occurring in any of the titles, but since the average title contains between five and six

meaningful terms, such a profile might retrieve the whole file five or six times, an economically unacceptable degree of over-kill. Therefore a reasonable strategy seems to be to use the most frequently occurring term, then the next, and so on until the whole file is retrieved. Unfortunately, as has already been indicated, the figures given in the table of stem-frequencies cannot be interpreted as indicating the number of documents retrieved from the BIO file by the use of a particular stem, because the frequencies shown are those of the occurrence of stems, and not of the documents containing them. That is to say, if a document title has the stem 'fung-' in it twice, this gives a stem-frequency of two, but a document frequency of one. The table of stem-stem co-occurrences shown below as Table 12 allows correction of the frequencies to document frequencies.

TABLE 12.

Stem-stem co-occurrence frequencies.

		FUNG-	BACTERI-	WOOD-	MICROB-	STOR-	SOIL-	AFLATOXIN-	ACID-	FOOD-	GROW-	PRESERV-	CONTROL-	INSECT-	CELL-	ENZYM-	ASPERGILL-	MICROORGANISM-	TOXI-	ISOCAT-	TEST-
FUNG-	207	20																			
BACTERI-	130	0	6																		
WOOD-	124	36	1	16																	
MICROB-	115	1	2	0	6																
STOR-	103	13	2	4	2	6															
SOIL-	99	22	13	1	9	1	22														
AFLATOXI-	92	2	1	0	0	2	1	10													
ACID-	95	8	9	1	7	2	0	0	48												
FOOD-	92	2	3	0	9	7	0	4	2	10											
GROW-	83	9	13	1	6	1	5	1	7	2	2										
PRESERV-	77	15	2	33	2	0	0	0	3	6	1	4									
CONTROL-	71	8	0	1	5	17	0	1	0	2	1	0	2								
INSECT-	70	4	1	5	0	16	4	0	2	2	0	4	10	14							
CELL-	80	20	4	5	2	0,	3	2	1	0	8	0	0	0	24						
ENZYM-	66	9	6	1	2	0	1	0	12	1	0	0	1	0	16	4					
ASPERGILL-	64	9	0	0	2	0	3	13	4	0	4	0	1	2	6	4	4				
MICROORGANISM-	57	2	3	1	3	1	11	0	3	1	6	1	0	0	2	1	0	4			
TOXI-	59	6	0	4	2	4	0	5	6	5	1	2	2	4	0	0	3	2	6		
ISOCAT-	55	7	16	5	3	1	10	1	1	0	3	1	0	1	2	0	5	8	3	4	
TEST-	53	20	4	17	0	1	1	0	1	3	0	13	0	5	0	0	1	3	1	0	8
RESISTAN-	49	17	1	13	2	1	0	0	0	1	0	6	1	4	0	0	0	1	1	1	4

This table allows correction of the top stem-frequencies to a closer approximation of document frequencies, by subtraction of half the number of times a term occurs with itself from the total frequency of occurrences. This yields the following table (Table 13) of documents containing a given term. Some frequently-occurring but non-significant terms such as 'studies' have been removed.

TABLE 13.

Numbers of documents containing a given term			
FUNG-	197	CONTROL-	70
BACTERI-	127	CELL-	68
WOOD-	116	ENZYM-	64
MICROB-	112	INSECT-	63
STOR-	100	ASPERGILL-	62
SOIL-	88	TOXI-	56
AFLATOXI-	87	MICROORGANISM-	55
FOOD-	87	ISOLAT-	53
GROW-	82	RESISTAN-	49
PRESERV-	75	TEST-	49
ACID-	71		

This table is a close approximation to actuality, but departs from exactitude because of the occasional presence of documents which contain the same term more than twice. Given, for example, a title of the type P A_1 QR A_2 : S T A_3 V, where the As are repetitions of the same term, and the other letters represent other words, the computer program counted the occurrence of A_1 with A_2, and of A_1 with A_3 as being two occurrences; it then counted A_2 with A_1, A_2 with A_3, A_3 with A_1 and A_3 with A_2, yielding a pair-frequency count of six. In practice, this meant that the actual numbers departed from the estimated in two cases. 88 documents contained AFLATOXI-, and 78 contained ACID-. This is one of the minor annoyances inherent in using a standard

computer package; it would be possible to write a word-frequency counting program specifying that only one occurrence of a term in a title should count towards the total number of occurrences, but this feature was not available in the INDACS suite used. It is only a minor annoyance, because the appearance of the same term in a title more than twice is a low-probability event, and the probability may reasonably be expected to decline with the frequency of use of given terms.

It can be seen from Table 13 that the use of the stem 'FUNG-' in searching the notional BIO file would retrieve 197 documents, 10.4% of the BIO file. It does not follow that use of the stem 'WOOD-' would retrieve a further 116, because 'FUNG-' co-occurs with it a number of times, and so some of the documents containing 'WOOD-' are already retrieved. When deciding the sequence of terms to use in retrieving successive maximal portions of the file (that is to say, when constructing the profile which will maximise recall with the most economical use of terms), the only accurate method is to take the most-frequently-occurring term (the term which retrieves the most), remove from the file all documents retrieved by its use, then take the term occurring most frequently in the remainder of the file, remove all documents it would retrieve, and continue the iterative process until all, or as much as is required, of the file is retrieved. This would mean calculating and displaying the frequencies of occurrence of all terms or stems in the residual part of the file after every use of a term or stem, a process which would be very laborious in practice, and most costly whether done manually or by computer. A crude approximation is to operate on the list shown in Table 13, taking the term giving the highest recall, and subtracting from all the other term-frequencies the frequencies of their co-occurrence with the chosen term. Then select the term next appearing as giving highest recall among those remaining, and iterate the process. This inevitably introduces a progressive error, because the co-occurence figures being subtracted represent or are derived from documents in whose titles the terms co-occur, and some of these titles also contain terms which have been eliminated at an earlier stage,

consequently one is continually risking the subtraction of a larger number than in reality should be the case. This is demonstrated in Table 14, which shows the effects of using terms in sequence, derived as suggested above; the forecast number of documents retrieved by using the terms in the sequence shown is compared with the actual number observed to have been retrieved by using them.

TABLE 14.

Use of maximum-retrieval stems in sequence				
Stem	Estimate of does retrieved	Observed no.of does retrieved	Cum total	Cum %
FUNG-	197	197	197	10.51
BACTERI-	127	127	324	17.29
MICROB-	109	109	433	23.11
AFLATOXI-	84	85	518	27.64
STOR-	81	83	601	32.07
WOOD-	75	77	678	36.18
FOOD-	62	63	741	39.54
GROW-	49	53	794	42.37
ENZYM-	45	45	839	44.77
MICROORGANISM-	37	40	879	46.91
INSECT-	35	39	918	48.99
ASPERGILL-	27	31	949	50.64
CONTROL-	23	33	982	52.40
TOXI-	20	33	1015	54.16
SOIL-	17	42	1057	56.40
PRESERV-	9	37	1094	58.38
ACID-	5	39	1133	60.46

(Under 'Cum %' is shown the percentage of the full BIO file retrieved.) It can be seen that instead of an estimated 1002 documents (53.47%) retrieved, the actual figure is 1133 (60.46%), a difference of 6.99%, 131 documents. It can also be seen that the difference between estimate and observation increased as one descends the list.

The observes results given in Table 14 are shown graphically in figures 6 and 7 below. It must be remembered that what these figures show is the effect of using a maximum-recall profile composed of truncated terms linked by logical OR, on the BIO file itself. All documents retrieved are relevant by definition, and precision is therefore a meaningless measure in this context, being always 100%.

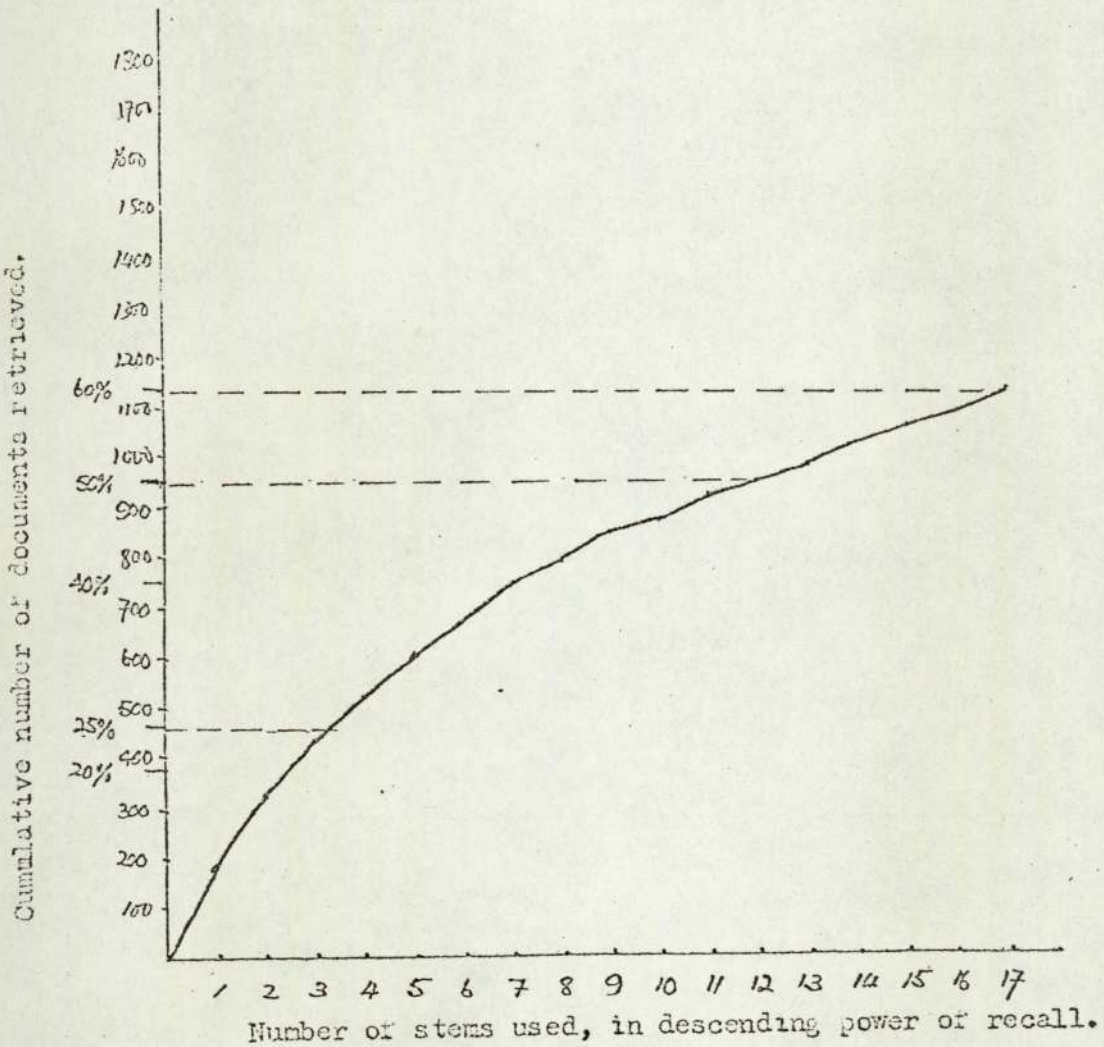


Figure 6. Cumulative number of documents retrieved using stems in descending order of recall potency.

Figure 6 shows that 25% of the file is retrieved by using less than four stems, and 50% is retrieved by less than 13. Beyond about the 50% level, retrieval is becoming increasingly costly in terms of number of stems used, but it is not possible on the

data presented to make an accurate guess at the total number of stems that would be required to retrieve the whole. Examination of Figure 7 illustrates a curious upturn in the productivity of subsequent terms, and it may be that what is happening in this subset of the vocabulary is a specific example of the case in which selection of terms in the manner suggested does not in fact produce the greatest economy in term use. Consider a set of documents each containing some of the terms A,B,C,D,E,F,. If the document set is composed as shown in Table 15 below, then proceeding by selection of the most frequent term, eliminating set members containing it, then

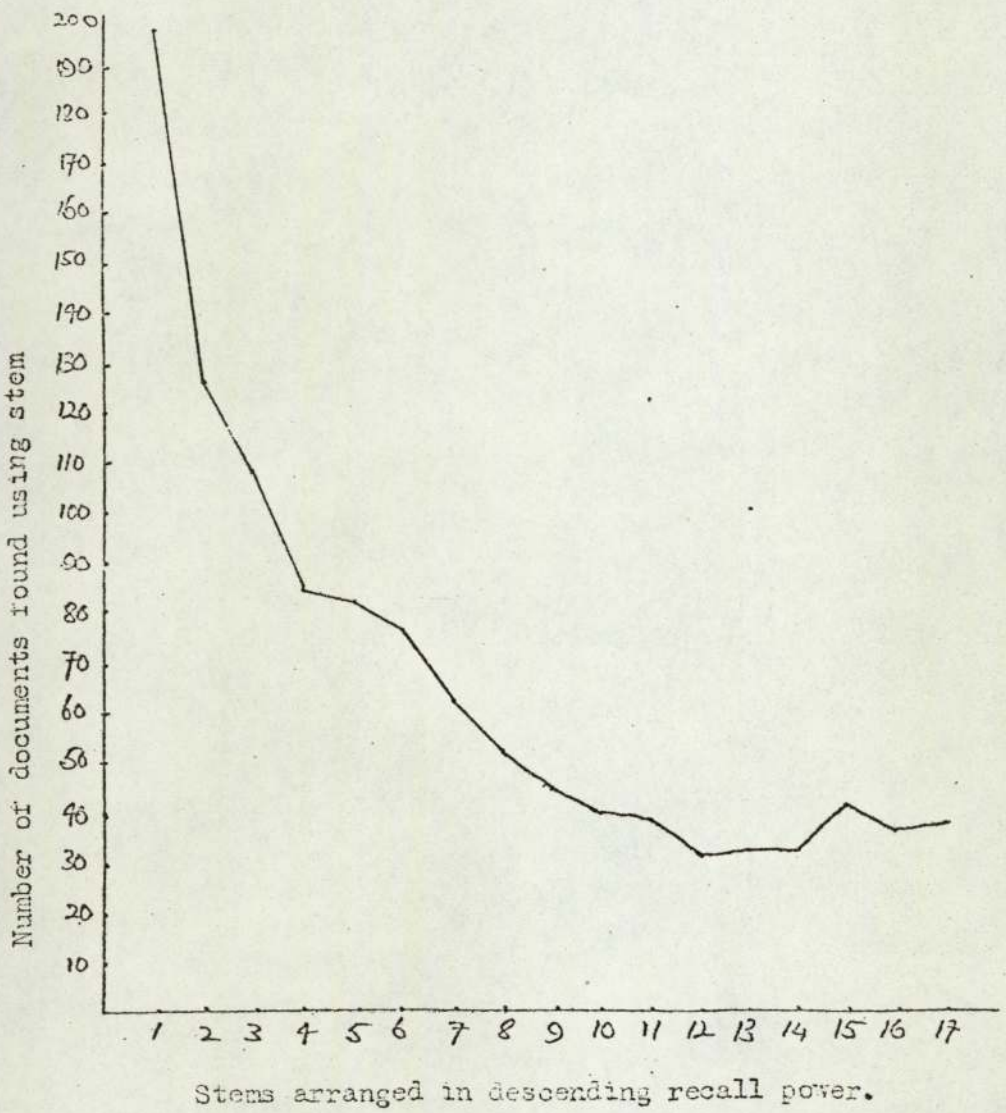


Figure 7. Documents retrieved by stems arranged in descending power of recall.

TABLE 15.

A notional document set						
1.	A	B		D		
2.	A		C	D	E	
3.	A	B			E	
4.	A		C		E	
5.	A	B		D		F
6.	A				E	F
7.	A	B			E	
8.	A		C		E	
9.		B	C			F
10.				D	E	

the set is completely determined by

all A + (B or C or F) + (D or E).

However, this produces a set description containing three terms. The most economical statement, by inspection is

all B + all E.

It can be seen that the problem of specification of a set in this way, with the least possible number of set sub-elements is not easy, particularly in cases like the present, when the number of sub-elements is several thousands and the number of elements in 1874. In terms of the present study, the problem is that of retrieving the whole of the BIO file with the least number of stems.

However, before this problem is explored, it is desirable to decide whether this sort of profile, based on consideration of the frequency of occurrence of words or stems in the literature of biodeterioration, without consideration of the vocabulary of the bases to be interrogated, looks as though it might be optimal

in performance. The profile shown in Table 14, designed to retrieve the largest part of the BIO file with the smallest number of stems (although, as we have seen, not necessarily the smallest number of stems possible) was used to search the 1969 Source Index of the Science Citation Index, a tape base which was already known to carry 1077 items included in the BIO file. The search was simulated by using the Permuterm index to the base.

The results of this simulated machine search are shown in Table 15 below. The costs shown in the second column are the costs for using the stems shown in an ASCA IV Profile, because it is extremely unlikely that the Biodeterioration Information Centre would be able to acquire and search tapes for itself, and the use made of secondary tapes is assumed to be by the purchase of searches or profiles from brokers.

TABLE 15.

			Documents containing stem in both BIO file and ISI cover						
	1	2	3	4	5	6	7	2÷5 8	9
Stem	$ ASCA IV cost	Cumulated cost in dollars	No. of docs	No. of unique decs	Culumated total no of docs	Cumul.% of BIO file	Cumul.% of ISI cover.	Cumul.$ cost per rel.ref.	No of authors recalled from ISI
FUNG-	24	24	111	111	111	5.92	9.92	.22	568
BACTERI-	85	109	77	77	188	10.03	16.80	.58	1721
MICROB-	20	129	60	58	246	13.13	21.98	.52	448
AFLATOXI-	7	136	76	73	319	17.02	28.51	.43	131
STOR-	47	183	47	38	357	19.05	31.90	.51	1046
WOOD-	19	202	42	32	389	20.76	34.76	.52	484
FOOD-	46	248	42	32	421	22.47	37.62	.59	885
GROW-	216	464	63	41	462	24.65	41.29	1.00	3688
ENZYM-	138	602	51	36	498	26.57	44.50	1.21	2494
MICRO ORGANISM-	14	616	38	27	525	28.01	46.92	1.17	201
INSECT-	43	659	37	27	552	29.46	49.33	1.19	708
ASPERGILL-	13	672	44	22	574	30.63	51.30	1.17	197
CONTROL-	253	925	33	13	587	31.32	52.46	1.58	4430*
TOXI-	48	973	42	27	614	32.76	54.87	1.58	912

(Table 15 continued).

SOIL-	93	1066	61	25	639	34.10	57.10	1.67	1514
PRESERV-	12	1078	36	19	658	35.11	58.80	1.64	278
ACID-	628	1706	54	24	682	36.39	60.95	2.50	14000*

These stems are on the Permuterm Index Stop List, and the values shown for number of authors' names retrieved are therefore approximate.

TABLE 15. Retrieval effectiveness and cost of a 'frequent-stem' profile.

Table 15 shows the recall as a percentage of the BIO file and as a percentage of ISI cover of that file; that is, it shows how much can be got of the BIO file references by using the search terms in a profile in the order given. Precision, that is how much of what is retrieved is relevant, cannot clearly be estimated, because as can be seen from Figure 8, against each term-pair shown, only the name of the author appears, and as it is not uncommon for authors to write more than one paper a year in which a given pair of terms appears, so the number of authors shown against a term does not equate to the number of papers found by using that term. Experience suggests that a reasonable approximation is obtained by multiplying the number of authors by 1.25. However, no attempt has been made to correct the figures shown in column 9, to indicate unique authors found by each term; this means that the total of column 9 gives a number greater than the real number of authors found by using the terms shown. As, therefore, the true number of authors found is not known, it has not been corrected to equate a number of documents found. Despite this, it is clear that the 'noise', the number of irrelevant references retrieved, is very large.

It can be seen that over 36% of the BIO file, or nearly 61% of the ISI coverage of the 1969 literature of biodeterioration, is retrieved by the use of seventeen stems only. A study of the cost of using these terms, or the rapid increase in the cost per unique pertinent citation retrieved, emphasises the point

that what has been done is the maximisation of recall by the use of a minimum number of terms; that is to say, the question being answered here is 'How can one obtain the greatest number of relevant references for the least terms?'; this is not necessarily the same as 'How can one obtain the greatest number of relevant references for the lowest cost?'. It would be the same, if the cost of interrogating a service were based on a flat rate for each term in the profile, but this costing procedure is seldom adopted by commercial or cost-recovery services.

The usual practice is to relate the charge for the use of a search term to the number of references expected to be retrieved by that term. Variants of this practice arise when logical precision devices are used, and frequently charges are made up of a fixed base charge plus additional charges related to the size of the expected number of references retrieved, but the basic principle of associating cost with number of references delivered holds good. This is clearly true of the ASCA IV system, where the correlation between cost per term and number of authors' names retrieved is 0.99.

In services charging in this manner, one may therefore think in terms of a notional unit cost of retrieving a single reference, and the cost of the use of a term as being that unit cost multiplied by the number of references retrieved. Therefore, the higher the cost per relevant reference retrieved in comparison to the unit cost of retrieving a single reference, the greater is the proportion of irrelevant references retrieved. Consequently, by aiming for the least cost per relevant reference retrieved, one is also reducing the overall cost of using the service or system, by reducing the number of irrelevant references to be inspected and discarded.

The most desirable method of profile compilation, then, is to use stems or terms in order of their forecast cheapness, starting with those which produce the least cost per relevant reference retrieved, rather than using them in order of their recall potency. It might be possible to do this exactly in

the present case, by operating on a stem co-occurrence matrix based on the references which are common to the BIO file and the ISI file, multiple occurrences of a term or stem within the title being taken into account; this would require solution of a similar problem to that posed earlier, of set specification. The situation postulated, however, is that of an information officer compiling a profile without knowledge of the specific items composing the conjunction of the two sets; the emphasis is on developing a forecast best profile, rather than a post hoc device, and therefore an approximate method of forecasting must be used. What is known are the frequencies of occurrence of stems in the BIO file, and the costs of stems in ASCA IV. It is also known that ISI covers approximately 60% of the BIO file (and in this connection it is of interest to compare the last column of Table 14 with Column 7 of Table 15). Co-occurrences of stems within the same title must necessarily be neglected. It the cost of a stem in ASCA IV is noted, and divided by the frequency of occurrence of the stem in the BIO file, and the result multiplied by 100/60 an estimate of the cost per relevant reference retrieved by the use of the stem is obtained. It is not practicable to refine this to an estimate of the cost per relevant unique reference. One then selects the cheapest stem in terms of forecast cost per relevant retrieved citation, uses it, takes the next cheapest, and so on until the expected cost per relevant citation retrieved exceeds a pre-selected cutoff. Table 16 shows the results of using this method of profile compilation on ISI tapes, simulated in the same way as for Table 15.

Table 15 showed that seventeen terms chosen for their high frequency of occurrence in the BIO file produced 682 relevant references at a cost of two dollars and fifty cents for each relevant reference found. This is a somewhat biased result, because 'acid-' is an extremely frequently occurring term throughout the literature of science, as can be seen from its ASCA price of $628, and it is the use of this term which accounts for the very high cost per relevant reference retrieved. If we had stopped the search at the preceeding term, we would have retrieved 658 relevant references for $1.64 a reference. Table

16 shows that forty terms, selected on the basis of estimated cost per relevant reference retrieved, can produce 655 references at a cost of 60.9 cents each, thirty-seven per cent of the cost using the method of Table 15. Further, Table 15 indicates that a total of approximately 19705 authors' names were retrieved by the first profile, which indicates a precision of about 658/(19705 x 1.25)= 2.7%; that is, of every thousand references retrieved, 27 were relevant. The second profile has a precision of about 655/(x 1.25) = . The precision estimates are necessarily crude, but it can be safely assumed that the second profile operates at better than double the precision of the first.

The crude method of forecasting used in Table 16 has obvious defects, which may be exemplified by the results using the stem 'Board-',which occurred 22 times in the BIO file, but only 4 in that part of it covered by ISI, a circumstance perhaps attributable to 'board' being a technological rather than a scientific term, and therefore of comparatively low occurrence in a largely scientific base such as ISI. 'Termite-' is another stem performing well below expectation, for several reasons. 'Wood' performs less well than expected, because many of the documents in which it occurs are already retrieved by preceding terms. Little can be done about this. It is indeed possible that if 'wood' were dropped, many of the documents containing the term would be retrieved by subsequent terms, but to produce a more accurate forecast, taking into account these possibilities, is not possible except in an experimental situation, and on a post hoc basis, which is of no value in the practical situation. As it stands, the method is clear and easily understood, and simple to apply. In practice, the best method of using this approach to profile compilation would be to run the profile composed in this way for a reasonable period of time, monitoring the results, and dropping those terms which appear over time to contribute little to the profile performance.

The terms in Table 16 have been ordered by forecast cost per relevant retrieved reference, starting with the forecast cheapest. The index by which the terms have been ordered is

TABLE 16.

Stem	Cost of stem $	Cumulated cost of stem$	Frequency of stem	Forecast cost per relevant reference.	No. of refs containing stem in ISO BIO cover	No. of unique refs containing stem in ISI/BIO.	Cost per relevant unique reference (2/7) in cents	Cumulated number of relevant refs found.	Cumulated % of BIO file found.	Cumulated % of ISI cover of BIO file found.	Cost per unique relevant ref retrieved overall (3/9) in cents.	No. of authors names listed under stem.
1	2	3	4	5	6	7	8	9	10	11	12	13
AFLATOXI-	7	7	92	12.6	76	76	9.2	76	4.05	6.79	9.2	131
FUNG-	24	31	197	20.3	111	109	22.0	185	9.87	16.53	16.8	568
PRESERV-	12	43	75	26.7	36	35	34.3	220	11.74	19.66	19.6	278
WOOD-	19	62	116	27.3	42	27	70.4	247	13.18	22.07	25.1	484
MICROB-	20	82	112	29.7	60	58	34.5	305	16.28	27.26	26.9	448
PAINT-	7	89	36	34.2	14	9	77.8	314	16.76	28.06	28.3	147
MOLD-	7	96	35	33.3	11	9	77.8	323	17.24	28.87	29.7	265
MICROORGANISM-	14	110	69	33.8	38	37	37.8	360	19.21	32.17	30.6	201
ASPERGILL-	13	123	62	34.9	44	27	48.1	387	20.65	34.58	31.8	197
SPOR-	10	133	47	35.4	38	29	34.4	416	22.20	37.18	32.0	495
BOARD-	7	140	22	53.0	4	3	233.3	419	22.36	37.44	33.4	144
ANTIMICROBIAL-	7	147	22	53.0	14	13	53.8	432	23.05	38.61	34.0	85
PEST-	9	156	28	53.5	11	11	81.8	443	23.64	39.59	35.2	348
TERMITE-	7	163	21	53.5	6	2	350.0	445	23.75	39.77	36.6	40
BEETLE-	8	171	24	55.5	16	15	53.3	460	24.55	41.11	37.2	149
MOIST-	7	178	20	58.3	14	7	100.0	467	24.92	41.73	38.1	290
PENICILLIUM-	7	185	20	58.3	12	11	63.6	478	25.51	42.72	38.7	60
CELLULASE-	7	192	19	61.4	16	8	87.5	486	25.93	43.43	39.5	38
MYCOTOXI-	7	199	19	61.4	14	14	50.0	500	26.68	44.68	39.8	31
SPOIL-	7	206	19	61.4	13	12	58.3	512	27.32	45.76	40.2	24
LIGN-	7	213	18	64.8	15	9	77.8	521	27.80	46.56	40.9	183
ATTACK-	7	220	18	64.8	5	3	233.3	524	27.96	46.82	42.0	175
MICROFLORA-	7	227	18	64.8	11	7	100.0	531	28.34	47.45	42.8	39
FERMENT-	11	238	27	67.9	20	19	57.9	550	29.35	49.15	43.3	199
HERBICID-	7	245	17	68.6	13	3	233.3	553	29.51	49.42	44.3	199
DETERIORAT-	7	252	16	72.9	2	0		553	29.51	49.42	45.6	38
COLEOPTER-	11	263	25	73.3	22	17	64.7	570	30.42	50.94	46.1	149
CLOSTRIDIUM-	9	272	20	75.0	17	8	112.5	578	30.84	51.65	47.1	129
PECT-	7	279	15	77.7	5	4	175.0	582	31.06	52.01	47.9	152
STERIL-	7	286	15	77.7	8	7	100.0	589	31.43	52.64	48.6	246
STREPTOMYCE-	7	293	15	77.7	12	10	70.0	599	31.96	53.53	48.9	100
IDENTI-	7	300	15	77.7	10	3	233.3	602	32.12	53.80	49.8	1347
STOR-	47	347	100	78.3	47	31	151.6	633	33.78	56.57	54.8	1044
MEAT-	9	356	19	78.9	14	5	180.0	638	34.04	57.01	55.8	171
TAXONO-	8	364	16	83.3	12	9	88.9	647	34.53	57.82	56.3	211
CEREAL-	7	371	14	83.3	10	3	233.3	650	34.69	58.09	57.1	101
GROUNDNUT-	7	378	14	83.3	11	0		650	34.69	58.09	58.2	37
TIMBER-	7	385	14	83.3	4	1	700.0	651	34.74	58.18	59.1	59
CITR-	7	392	14	83.3	6	3	233.3	654	34.90	58.45	59.9	310
FUSARIUM-	7	399	14	83.3	9	1	700.0	655	34.95	58.53	60.9	90
FOOD-	46	445	87	88.1	42	22	209.1	677	36.13	60.50	65.7	885
BIODEGRADA-	7	452	13	89.7	8	8	87.5	685	36.55	61.22	66.0	12
COCKROACH-	7	459	13	89.7	8	8	87.5	693	36.98	61.93	66.2	88
TENEBRIO-	7	466	13	89.7	12	6	116.7	699	37.30	62.47	66.7	40
TEXTILE-	7	473	13	89.7	2	1	700.0	700	37.35	62.50	67.6	110
FUMIGA-	7	480	12	97.2	9	3	233.3	703	37.51	62.82	68.3	52
PEANUT-	7	487	12	97.2	10	0		703	37.51	62.82	69.3	83
CONTAMINAT-	9	496	15	100.0	7	4	225.0	707	37.73	63.18	70.2	210

Performance of terms selected on a basis of forecast least cost per relevant citation retrieved.

calculated as follows:

$$\frac{\text{cost of the term in the service being searched}}{\text{frequency of occurrence of term in BIO file} \times \text{\% coverage of BIO by service searched}}$$

and the terms are used in ascending value. In the ISI file, cost of term used is directly related to frequency of occurrence of the term in the file, so that frequency of occurrence could be substituted for cost without altering the ranking of the terms. Similarly, the process of multiplying the frequency of a term in BIO by the percentage cover of BIO by ISI is an attempt to estimate the number of relevant documents containing the term in the ISI file. If the number of relevant documents containing the term in the ISI file is substituted for the denominator, then the expression becomes $\frac{\text{total number of documents in which the term occurs}}{\text{number of relevant documents in which the term occurs.}}$

This, arrived at by a different route, is the inverse of the 'term specificity' measure suggested by Barker, Veal and Wyatt of the UK Chemical Information Service Research Unit (12). They suggested using 'all terms above a chosen specificity value.... as search terms in a simple one-parameter search of the appropriate fixed file.' In the UKCIS study 'the first set of relevant items was obtained by carrying out a search with an intellectually constructed profile'. The authors continued, 'However, we could equally well have started off with a set of relevant items provided by the user'. The difference between the approach adopted here and the UKCIS approach is that in the latter case, the assumption was implicitly made that all the references in the 'set of relevant items' were contained in the system being searched, which is not so in the BIO case. However, the use of inverted term specificity, and the ranking of search terms by taking the lowest-rated first is clearly the same as using term specificity and taking the highest rated terms first, so that the method developed in the present study is not different from the UKCIS method. That essentially the same method should be arrived at by different routes, and show roughly the same measure of success when applied to two different bases, one

subject-oriented and one multidisciplinary, suggests that the method is of general validity and practicality.

Searching other bases

In an earlier part of this study, it was shown that substantial portions of the BIO file were contained in Chemical Abstracts - Condensates (Chemical Abstracts), Chemical Titles, BA-Previews (Biological Abstracts), MEDLARS (Index Medicus) and Food Science and Technology Abstracts. It was hoped that the next step would be to design profiles which would retrieve those portions from the files containing them, in order to discover the numbers of irrelevant items retrieved at the same time and thence to estimate the precision available; and in addition, by running the profiles designed, to estimate the cost per relevant reference retrieved, at various levels of recall. One methodological difficulty which would have arisen in applying this approach would have been the problem of defining the file to be searched. If, for example, part of the target file (the BIO portion contained by a service) is covered in one year's issue of a service, and another part in the next, then it is clear that among the references retrieved by a successful profile would be a number of references from years before and after the BIO file base date (1969), some of which would be relevant. This would entail either having the retrieved material assessed for relevance if it came from outside the base year, or, better, eliminating all retrieved references except those from the base year. This would have been feasible, although it might have complicated the costing, but as it happened, no back files of the services concerned were available, and therefore, this part of the study had to be abandoned.

Instead, profiles were calculated, and searches run on current files, the product of the searches being assessed for relevance by the Biodeterioration Information Centre. In this way it is possible to estimate the precision for a specificity-based profile, but not the recall, although very crude guesses may be made. The cost per pertinent citation retrieved by the profiles

tested can be estimated, but again, this cannot be done over a range of recall estimates. The full methodology as originally proposed requires firstly that the files for the appropriate period, that containing the 1969 material in this case, be available, or alternatively, that sufficient funds be available for a relatively large number of searches to be run on each service, with profiles of varying length, in order to collect estimates at various recall levels. In the present case, although a number of searches were run, funds did not permit running of a series as suggested.

Three searches were run on three consecutive issues of each of Chemical Titles, CA Condensates (even numbers), CA Condensates (odd numbers) and BA Previews (strictly, in this case, two searches on BA Previews and one on BioResearch Reports). In each of these cases, frequency counts were available of the occurrence of terms. For MEDLARS, which is a controlled-language base, it would have been necessary first to collect the indexing terms assigned to each reference of the BIO file known to have been covered by MEDLARS, but this information was not available, and therefore, in order not to omit this service entirely, two searches were run, one using the single heading BIODEGRADATION, and the other locating all references which included the headings (BACTERIA or FUNGI) and MATERIALS.

Chemical Titles.

One of the search aids available for Chemical Titles is a KLIC (Key Letter in Context) Index, which lists all combinations of three characters occurring more than times in a number of issues of Chemical Titles, together with their frequencies of occurrence. A section of this is shown in Figure 9.

	FUN	DF
	FUNGAL	20
ANTI -	FUNGAL	17
	FUNGI	30
	FUNGICIDES	20
	FUNGUS	12

SUL	FUR	493
SUL	FUR - 32	8
SUL	FUR - 35	12
SUL	FUR - NITROGEN	6
FUR	FURAL	19
	FURAN	121
	FURANOSE	15
	FURANOSIDE	6
	F RANOSIDES	6
	FURANOSYL	18
	FURANS	17
SUL	FURATION	7
	FURFURAL	19
	FURFURYL	13
SUL	FURIC	136
SUL	FURIZATION	8
DE - SUL	FURIZATION	21
	FURNACE	68
	FURNACES	24
	FURO	7
	FUROSEMIDE	8
SUL	FUROUS	6
	FURTHER	91

Figure 9. Part of the UKCIS KLIC Index,

Using this listing, each term occurring more than four times in the BIO file was taken and its frequency of occurrence in the BIO file divided by its frequency in the KLIC. The terms were then ranked in descending value of the ratios obtained, and the first sixty were used to comprise a profile for search of Chemical Titles. This profile is shown in Table 17. It was used to search the current Chemical Titles file on the 30th July, 13th August 1973, and 29th August. The resulting references were then evaluated by the Biodeterioration Information Centre, where the staff classified each reference as being 'relevant', 'irrelevant' or 'possibly relevant'.

Table 18 summarises the results of the searches. The profile has been divided after every tenth term, so that Section A represents the first decade, B the second, and so on.

TABLE 17

Profile term.	Rel.	Poss Rel.	Not rel	Profile term.	Rel.	Poss Rel.	Not rel
A				D			
Preserv	4	2	4	Food	4	2	17
Fung	12	6	12	Fresh		1	14
Taxono		6		Year	1		11
Wood	4	5	2	Entero	1		10
Microb	8	4	19	Mite			3
Deteriorat	1			Sewage			3
Steril	2	1	2	Wash			6
Cockroach			1	Pathogen	1		4
Microflora			1	Aspergill		3	16
Cellulase		1	1	Moist	1	1	10
B				E			
Peanut		1	7	Conserv		1	14
Tropical	1		5	Pect			7
Malathion				Penicillium	1		3
Meat	2	1	4	Contaminat		2	12
Microorganism	10	4	14	Fish		2	12
Attack	1		5	Poles			2
Paint	5	1	7	Cure	1		3
Marine		6	12	Cut			18
Xylan	2	1	3	History			7
Incidence			5	Mold			26
C				F			
Spor			9	Extracellular			11
Pine			15	Prevent			4
Stor	3	1	41	Streptomyce			7
Anti-microbial				Clostridium			9
Fusarium			3	Old			3
Textile	1		4	Herbicid	4		6
Humidity			1	Flour		1	1
Raw			8	Weevil			
Future			4	DDT	1	2	3
Refrigerat			2	Myceli			2

Chemical Titles profile and item results

TABLE 18.

Summary of results of Chemical Titles searches.							
Totals	Relevant No.	%	Possibly Relevant No.	%	Not Relevant No.	%	Total
Section A	31	31.6	25	25.2	42	42.8	98
Section B	21	21.6	14	14.4	62	63.9	97
Section C	4	4.3	1	1.1	87	94.6	92
Section D	8	7.3	7	6.4	94	86.2	109
Section E	2	1.8	5	4.5	104	93.7	111
Section F	5	9.3	3	5.6	46	85.2	54
Grand Total	71	12.7	55	9.8	435	77.5	561

Some terms were not considered for inclusion in the profile used on a priori grounds; for example, 'studies', 'Report', 'method', 'relationship' and similar terms, which may from time to time appear to be of high specificity, are effectively meaningless as content indicators, rather like prepositions or articles. It could be argued that a term such as 'studies' appears more frequently in the Biodeterioration literature than in, say, the literature of chemistry as a whole (which tends to be confirmed by the term's high specificity as calculated from the frequency of occurrence of terms in CA Condensates). However, use of the term in a Biodeterioration search is extremely unproductive (in the CA-Condensates (odd series) searches, the term found 1 relevant document, 4 possibles and 327 irrelevant), and B.C. Vickery suggests that although the term may be of relatively low occurrence in the chemical literature as a whole, there may be subfields of chemistry in which the term is frequent. In practice, this amounts to saying that such terms occur in all literatures, and occasionally appear to be of high specificity by chance. There seems to be no way of identifying and eliminating such terms by algorithmic means, but they can be largely ruled out by inspection. If the intervention of human judgement is ruled out by the necessity for a purely algorithmic

approach (as, strictly, it should be in the present study), then the terms can more positively be identified by a few trial searches, although this could be relatively costly. None of the profiles derived by specificity calculations should be assumed to be fixed and immutable, because the method is not sufficiently precise. The correct way to use such profiles would be to try them for a number of searches, and then to withdraw the non-productive terms. All that is claimed for this method is that it is a guide to the production of profiles which will yield the greatest number of relevant items for the lowest cost per relevant item retrieved; but it is no more than a guide.

Tables 17 and 18 show the effects of such a profile as applied to Chemical Titles. The first conclusion that one may reach is that three searches, on six weeks' stock of Chemical Titles is not really enough to provide very firm conclusions. It would be unwise, for instance, to suppose that the average number of relevant documents found per search would be 71/3, and that therefore the annual number to be found by the profile as suggested would be 24 x 26, that is 624; the three searches yielded respectively 37, 20 and 14. Chemical Titles covered 435 items of the 1969 literature, and there is no strong evidence that the Biodeterioration literature has increased by about 50% over the period 1969-1973. It may have so increased, and it is also possible that the BIO file represents substantially less than the whole of the relevant literature of 1969, but it must be appreciated that we have, unfortunately, no guide to the recall level of the profile as used. All that can really be said is that the percentages shown of relevant references in Table 18 are equivalent to the precision levels of the profile at particular points, and that the precision seems to decline (although, again, three searches only do not provide good enough evidence) as specificity declines, as was predicted. The distinctions between sections A and B, and B and C, are clear enough, but it seems likely that a six-month trial would be needed before any positive conclusions could be drawn.

The costs of searching Chemical Titles were calculated by adding a Basic Subscription charge of £35, a charge of 2p for every Search Unit, and a charge of £5 for every 150 units output on card or for every 300 output on paper or every 500 output on Abstract Numbers only. Output units are items retrieved by search, so that this element of the cost is directly related to the size of the drop. Search units are a function of processing time, and varies with the logic used. The actual searches carried out were paid for as trial searches, not as part of a normal running programme, and so the cost per relevant reference retrieved is not directly calculable. Table 19 shows the search and translate times incurred in the three searches; the total time incurred by running the profile as a full sixty-term search is shown, followed by the timings for the first twenty terms (Sections A plus B), Section C, Section D, Section E and Section F. Totals of sub-searches will not bear an exact relationship to the totals of the full searches, because in performing the full profile search, references are found by any of the terms in the profile, whereas in the sub-searches, the timings have been adjusted to allow for the fact that by using a later section, some documents would have been retrieved by earlier profile sections.

TABLE 19.

Search times for CT profile, and sub-profiles.	
Profile	Timing for all 3 searches
Full	68 seconds
Section A + Section B	25
Section C	10
Section D	12
Section E	11
Section F	11

This allows us to say that the full profile cost, very crudely (and the figures are worth no more than crude estimates), one second per relevant reference retrieved, A + B cost about ½

second per relevant reference, C cost 2½ seconds, D cost 1½ seconds, E cost 5½ seconds and F cost about 2 seconds.

The cost estimates of a year's run of the full profile provided by UKCIS were £212 for card output, £132 for paper output and £97 for abstract number only output. In the present context, paper output would probably be the most suitable, because the increased cost of paper as against abstract number can be offset against the time required to check each number against the appropriate abstract in Chemical Abstracts printed version. Calculated in the same way, assuming that one can take the totals for the 3 searches performed, divide by 3 (to get an average for one search) and multiply by 26 (to get an estimate for a year's run), costs for A + B would be
£35 +((25 seconds/3)x 26 x2)pence + ((98/3)x26)/300 x£5 =£53.48.

These estimates imply that one is therefore comparing the retrieval of 615 relevant references for a cost of £132 with the retrieval of 450 for £54, approximately. These gives estimates of roughly 21½ pence per relevant reference retrieved, using the full profile, with a precision of 12.7%, and using sections A and B only, 12 pence per relevant reference retrieved with a precision of 26.7%. The 'possibly relevant' items have been neglected in these estimates, because the opinion of the Biodeterioration Information Centre is that they are more likely to be irrelevant than relevant.

Unfortunately, the relative recalls of these estimates is unknown, and given the disparity between the coverage figure for CT for 1969 and the estimate of items retrievable by the full profile for 1973, the recall cannot be estimated. One can only say that there are indications that as terms are used in descending order of specificity, the overall cost per relevant citation retrieved tends to increase, as does the proportion of irrelevant items retrieved.

Chemical Abstracts Condensates - even issues.

The even issues of CA Condensates cover macromolecular, applied and physical chemistry (see page 22) and 127 Bio file items were found in the even sections, as against 693 in the odd. A listing of term frequencies similar to the UKCIS KLIC, although based on words rather than word-fragments, has been produced in Germany and a copy is held at Nottingham. These frequencies were used to produce profiles for the even and odd series, in the same way as described above. The resulting even profile is shown, with its performance, in Table 20 below, and Table 21 summarises the results of the CA Condensates even searches. A longer list of profile terms was used in the three searches carried out on this base, but because the actual profiles run were of even greater length, and included several very low-productivity terms (relationship, studies, property, British, contribution, sole, method, compare) which are removed from the profile as shown, no timings or costs can be given. Crude estimates can be made, but the reliance to be placed on such estimates is not great.

TABLE 20.

Profile term	Rel.	Poss Rel.	Not Rel.	Profile term	Rel.	Poss Rel.	Not Rel
A				B			
Microbiolog	4		4	Fusarium	1		
Aflatoxin				Larva			
Aspergillus				Cereal			
Genus			1	Cockroach			
Clostridium			1	Groundnut			
Preserve	2		1	Pseudomonas	1	1	
Salmonella	1			Cellulase			
Taxonomy			1	Pathogen			2
Stores	1			Poultry			2
Fresh	1		6	Microflora	1		

Chemical Abstracts - Condensates (even issues) profile and item results.

TABLE 20. (continued)

Profile term	Rel.	Poss Rel.	Not Rel.	Profile term	Rel.	Poss Rel.	Not Rel
C				D			
Washed				Filtrate			1
Fung	8			Infect			5
Antigungal		1	1	Diet			
Catalase				Fumigation			
Soy		1		Stimul			24
Sterility			1	Bacillus			
Thiobacillus	3			Yeast	2		3
Nutrition			1	Bacteriological	1		1
Beef				Egg			1
Culture				Escherichia			
E				F			
Mildew				Insecticide		1	1
Staphylococc				Biochemistry			
Tomato				Herbicide			
Isolated			1	Wheat			1
Biosynthesis				Microorganism	1	1	3
Meat			1	Relationship			1
Peanut			1	Metabolism			
Insect			1	Dehydrogenase			
Biology				Inactivation			
DDT	1		1	Wine			
G				H			
Biodegrad	12	1		Streptomyces			
Ferment			1	Enzyme	1		13
Adult				Food			26
Incubation				Ecology			1
Market			2	Fish	1		14
Metabolic			1	Antimicrobial			1
Sweet				Protease			
Milk			4	Variet			2
Respiration	1		1	Citrus			
Flour				Fruit			3

TABLE 20. (continued)

Profile term	Rel.	Poss Rel.	Not Rel.	Profile term	Rel.	Poss Rel.	Not Rel.
J				K			
Creosote				Species			17
Maintain			1	Attack			4
Bacteria	4		4	Proteolytic			
Purifying	1		19	n-alkanes			
Moth				Rice			3
Potato				Pesticide		1	6
Agar			1	Deterioration		1	3
Artificial			20	Disinfectants			
Kernel				Cottonseed			
Tolerance			1	Incidence			2
L				M			
Antifouling			1	Destruction			5
Year			12	Anaerobic	2		3
Disease			6	Tropical			1
Old			5	Alga	3		9
Pest				Hygiene			5
Conservation			14	Toxic	1		50
Inoculation			7	Wood	1	1	63
Organism			2	Apply			4
Produce			42	Marine	1		14
Chip			1	Tissue			8

Chemical Abstracts - Condensates (even issues) profile and item results.

TABLE 21.

Totals	Relevant No.	%	Possibly Relevant No.	%	Not Relevant No.	%	Total
Section A	9				14		23
Section B	3		1		4		8
Section C	11		2		3		16
Subtotal A-C	23	40.9	3	6.4	21	44.7	47
Section D	3				35		38
Section E	1				5		6
Section F	1		2		6		9
Subtotal D-F	5	9.4	2	3.8	46	86.8	53
Section G	13		1		9		23
Section H	2				60		62
Section J	5				46		51
Subtotal G-J	20	14.7	1	0.7	115	84.6	136
Section K			2		35		37
Section L					90		90
Section M	8		1		162		171
Subtotal K-M	8	2.7	3	1.0	287	96.3	298
Grand Total	56	10.5	9	1.7	469	87.8	534

Summary of results of CA-Condensates (even issues) searches.

Again, it seems that as terms are used in descending order of specificity, the 'noise' increases, and the return in terms of relevant citations retrieved, tends to fall off, so that cost per relevant reference retrieved will rise. Table 5 suggests that it would be worth restricting search to the few sections contributing more than, say, ten items a year in 1969; this

effectively means searching only those sections dealing with cellulose and other wood products, coatings and inks, and sewage and wastes. These three sections accounted for slightly more than 58% of the CA-C even/BIO file coverage. However, this only amounted to 74 BIO file items, and, as the costing algorithm is the same as for CT, would cost at least 50 pence, and probably more, for each relevant item found, the basic cost of subscription to the service being still £35.

TABLE 22.

Profile term	Rel.	Poss Rel.	Not Rel.	Profile term	Rel.	Poss Rel.	Not Rel.
A				B			
Biodegrad	1		1	Microbiolog	7	2	16
Isolated	2		120	Celeoptera			2
Preserve			6	Wood	1		11
Board			1	Stores			2
Paint	3		6	Equipment			3
Creosote				Fabric			
Carbon	2	6	383	Spoilage	5		6
Poles				Cellulolytic		1	3
Purify			5	Microbe	3		10
Ship			1	Dimensions			1
C				D			
Deterioration	2	1	2	Sewage			1
Valuable			1	Cellulase	2	2	6
Aflatoxin	5	3	10	Attack			3
Basidiomycetes				Corrosion			5
Groundnut			3	Enterotoxin	1		3
Influence	3	2	169	Inoculation			3
Great			2	Market			2
Weather			1	Nut			3
Rot	1		8	n-alkanes			
Sole				Microflora			

Chemical abstracts - Condensates (odd issues) profile and item results.

TABLE 22. (continued)

Profile term	Rel.	Poss Rel.	Not Rel.	Profile term	Rel.	Poss Rel.	Not Rel.
E				F			
Finish				Estimate			4
Clean			1	Thiobacillus		1	1
Hazard			6	Future			4
Keeping				Tenebrio		3	
Washed				Decay	1		8
Tribolium	3		1	Cutting			3
Genus			15	Economic			7
Construction			5	Cockroach	6	2	5
Term			28	Apply		1	1
Mycelium			4	Wool	1		2

Chemical Abstracts - Condensates (odd issues) profile item results.

TABLE 23.

Summary of results of CA-Condensates (odd issues) searches.

Totals	Relevant No.	%	Possibly Relevant No.	%	Not Relevant No.	%	Total
Section A	8	1.5	6	1.1	523	97.4	537
Section B	16	21.9	3	4.1	54	74.0	73
Section C	11	5.2	6	2.8	196	92.0	213
Section D	3	9.4	2	6.8	27	84.4	32
Section E	3	4.8			60	95.2	63
Section F	8	16.0	7	14.0	35	70.0	50
Grand Total	49	5.1	24	2.5	895	92.5	968

Tables 22 and 23 above present similar results for the searches on odd issues of CA-Condensates, which cover biochemistry and organic chemistry. Table 5 shows that the bulk of the material relevant to Biodeterioration in this file is found in the first twenty sections, those covering Biochemistry. Microbial

biochemistry is the richest section, but relevant material is scattered throughout. Table 22 shows that the poor precision of Section A, which should be the highest performer, is largely attributable to the use of 'carbon-' as search term, although, overall, no term is notably useful on its own. The sample searches do not, as has been noted above, provide sufficient evidence for firm generalisation, but it seems not unlikely that simple specificity alone as applied to this file is not very useful as a guide to profile construction. On present showing, it appears that a good proportion of the literature of Biodeterioration might be found in CA-Condensates odd issues, but retrieving it is likely to be both difficult and costly.

BA-Previews.

Table 24 and 25 present the profile used to search BA-Previews and its performance. It was derived by use of 'A guide to the vocabulary of biological literature', which is effectively 'a compilation of the words used most frequently in the citations of biomedical research literature indexed by BIOSIS since late 1959'. (13) A sample of this publication is given in Figure 10.

TABLE 24.

Profile term	Rel.	Poss Rel.	Not Rel.	Profile term	Rel.	Poss Rel.	Not. Rel.
A				B			
Antimicrobial	2		13	Spoilage	4		
Paint		1	5	Cellulase			7
Afla	8	1	17	Build			12
Corrosion				Termite	2	1	
Dermestidae				Psychrophilic			
Cellulolytic	1	1		Organoleptic			
Textile			2	Pectic			2
Fillet				Sitophilus	7		1
Cosmetic			3	Valuable			1
Board		1	2	Deterioration	1		5

BA - Previews profile and item results.

TABLE 24. (continued)

Profile term	Rel.	Poss Rel.	Not Rel.	Profile term	Rel.	Poss Rel.	Not Rel.
C				D			
Wood	10	1	50	Aspergillus	1	7	39
Thermophilic			3	Ferro			8
Sole			2	Decay	1	1	8
Pole			2	Tenebrionidae			4
Log			5	Microbiolog	2	5	16
Lignin			7	Detergent			13
Fungistatic		1		Ascospore			
Sclerotium			2	Coat			11
Mold		1	8	Biodegradation			
Refrigeration			1	Degradation	2	1	43
E				F			
Keeping			4	Sporulation		1	17
Biosynthesis	1		94	Finish			5
Cockroach	3		8	Bacteri	6	8	361
Nut	1	3	207	Malathion	1		3
Stor	6	1	73	Additive	1	3	17
Filtrate			1	Attack	1		10
Actinomyces	1		13	Tribolium	1		1
Fungicid	2		18	Pheromone	2		9
Rot	1		66	Wash		1	25
Surfactant				Penicillium	2	3	14

BA-Previews profile and item results.

TABLE 25.

Totals	Relevant No.	%	Possibly Relevant No.	%	Not Relevant No.	%	Total
Section A	11	19.3	4	7.0	42	73.7	57
Section B	14	32.6	1	2.3	28	65.1	43
Section C	10	10.8	3	3.2	80	86.0	93
Section D	6	3.7	14	8.6	142	87.7	162
Section E	15	3.0	4	0.8	484	96.2	503
Section F	14	2.8	16	3.3	462	93.9	492
Grand Total	70	5.2	42	3.1	1238	9 .7	1350

Summary of results of BA-Previews searches.

†FUNGI 33,562
see also ASCOMYCETES; BASIDIOMYCETES; DEUTEROMYCETES; LICHEN; MICROORGANISM; MILDEW; MYCOLOGY; MYCORRHIZA; PHYCOMYCETES; YEAST; FUNGI SYSTEMATICS
Fungi Imperfecti —
see DEUTEROMYCETES
†FUNGI SYSTEMATICS (BT6) 25,805
/includes taxonomy, classification, and distribution/
†FUNGICIDES 2,609
see also PESTICIDES
†FUNGISTATIC 220
†FUNGOIDES 185
†FUNNEL 122
†FUR 424
†FURAN 240
†FURAZOLIDONE 199
†FUROSEMIDE 534
see also LASIX
†FUSARIUM 2,493
see also DEUTEROMYCETES
†FUSION 729
see also MELTING
†FUTURE 1,451
†G 12,473
†GADUS 239
see also OSTEICHTHYES; TELEOST
†G 1,191
also ACQ ; ADD NCRE/

Figure 10. Section of listing of frequency of BIOSIS terms.

BA-Previews searches (including Bioresearch Index searches) are offered experimentally at £25 per search, and therefore no serious costing estimates for this profile can be made. The same general comments apply to these searches as apply to those done on CA-Condensates.

However, one important point must be made, since it affects the performance of specificity-based profiling. Search of Chemical Titles is performed using the titles of the documents held on the file only. However, both CA-Condensates entries and BA-Previews entries include added keywords, which are searched at the same time and in the same search as the terms present in the titles of the documents. Many of these assigned keywords will relate to relatively minor parts of the documents to which they are assigned, and might reasonably be assumed to add to recall, but to erode precision, to an unknown extent. Like must always be compared to like, and profile terms derived by specificity measures must always be derived by relating frequency of terms in the 'known relevant' collection to frequency of the same type of terms in the file to be searched. That is to say, in the present context, words-in-titles should be used. When approaching a file organised by a controlled language, the terms applied in that language to the 'known relevant' file should be compared with the total assignation of terms in the searched file. Otherwise, the results, as can be seen, are bad.

MEDLARS

MEDLARS is a service which is organised on the basis of a controlled language, and ideally this service should have been approached with a knowledge of the full indexing which had been applied to those BIO file items the service had covered. The frequency of occurrence of these terms in the BIO subset should then have been related to the frequencies of assignation in the full MEDLARS file. Strictly speaking, the whole BIO file should have been indexed using the Medical Subject Headings, in order to be strictly comparable with the methods used for other services, but it is conceivable that MeSH headings do not exist for some of the concepts included in the full BIO file. However, it was not possible to discover the full list of headings assigned to BIO file documents covered by MEDLARS, and the whole methodology had to be abandoned. In order not to give up the service entirely, two searches were run, both with output restricted to

a stated number of items. In one search, limited to an output displayed of 450, (BACTERIA or FUNGI) and MATERIALS was used as search statement. This resulted in 67 relevant, 13 possibly relevant and 370 not relevant, a precision (taking 'relevant' only) of 14.9%. In the second search, only one term was used, 'Biodegradation'. This is not the same as biodeterioration, but fifty items containing this term were displayed, of which 25 were relevant, 6 possibly relevant, and 19 not relevant, a precision of 50%. Once again, of course, no figure for relative recall can be deduced. Nor can any estimate of the cost per relevant citation retrieved, because of the flat-rate charge for MEDLARS in the United Kingdom.

Conclusions.

Before discussing the results of this investigation, it must be pointed out that the findings and conclusions relate to a specific situation, and that this situation is not necessarily a very common one. This study has been discussing the ways in which a Specialised Information Centre can acquire its reference base, and the needs of such a centre differ from those of an individual user of information, and possibly from those of a normal industrial or governmental information centre. The Biodeterioration Information Centre attempts (as do most SICs) to collect comprehensively within its field, which is broad, not very easy to define, and whose literature is scattered thinly through a large number of publications. Its requirement is for 'all but not only', in Fairthorne's words, as opposed to the requirement of an individual with a request or current need for information on a narrower, more specific topic, who more generally wants 'only but not all'. The choice between these alternatives may not be very palatable, but it is inescapable. The Centre, that is, gives greater weight to recall than to precision. However, it must give some weight to precision, in searching for references because it is absolutely constrained by a finite budget, and must keep the proportion of irrelevant references collected as low as possible so that it does not dissipate its funds buying unwanted material, and paying for the intellectual labour required to reject it. It must, in simple

terms, get as much as it can of what it wants, and as little as possible of what it does not want, for as low a price as possible, and if the cost of comprehensive collection is greater than its resources, it must necessarily be satisfied with less than comprehensiveness.

The nature of its reference requirement is very complex, more so than that of the individual information user. As an illustration of this point, the MEDLARS search team at the British Lending Library were presented with a list of the 463 BIO file items covered by Index Medicus, and were asked for advice on the profile which would retrieve as many of these items as possible. The response was that an appropriate search statement would be extremely complex, and would retrieve many thousands of unwanted items. Another illustration was provided by a search of a sample of ISI tape, using a fifty-term profile linked with logical ORs; two references were found to contain three of the profile terms, but neither of these references was relevant.

The fundamental problem when searching the secondary services is that they are not indexed from the viewpoint of biodeterioration; this statement includes the statement that the natural language indexing provided by titles is not sufficiently specific. No concept can be identified within a retrieval system unless the concept is explicitly or implicitly present in the indexing language, and the appropriate term or terms from the language have been applied to documents containing the concept.

Machine search of reference files is only appropriate when a very large file has to be searched in a complex way, or in a way for which no printed-index facilities are available. The first case does not seem to apply to Biodeterioration because, although the large files exist, the 'complex way' of searching them is ill-defined and apparently likely to produce much unwanted material. The second case does not apply because printed versions of the appropriate services do exist, in usable form. The secondary service which carried the greatest number of references relevant to Biodeterioration in the present test was found to be the ISI

file. This exists in a printed form as the Source Index, which could be scanned manually, but at enormous cost in time, and with the probability of great inaccuracy induced by human fatigue and fallibility. It also exists in another form, which is more easily handled; this form is Current Contents. It has been shown that a search statement which is potentially capable of capturing roughly half of the relevant material covered by ISI can be constructed simply on specificity ratios, but it is evident that capture of the other half of the coverage would only be possible at very greatly increased cost, which would be unacceptable to the Centre. Again, the quantity of irrelevant material retrieved would be unacceptable. Sophistication of the profile to improve precision would degrade recall, and probably increase cost still more. A human scanning Current Contents undoubtedly has a search statement in mind when scanning, but the statement is only partly explicit, and the effect is that each item scanned is classed as relevant, possibly relevant or irrelevant on apparently subjective grounds. The process by which the mental 'profile' is composed, and the manner in which it is manipulated, are not well understood at present, and existing profiling techniques are still relatively crude attempts to produce models of this mental process. It would be instructive (but is beyond the scope of this study) to take a large number of documents of stated relevance, and explore in each case the reasons why they are selected as relevant. For present purposes, it is sufficient to state that scanning Current Contents is the 'best' way of collecting relevant references from ISI services, for the Biodeterioration Information Centre.

Scan of Current Contents could have provided (and probably did) about 60% of the BIO file. Use of all other services examined could not have added more than a further 17%. About 23% of the BIO file was not covered by any service. This means that whatever means were used to collect the uncovered 23%, plus Current Contents scan, would have been covered at least 83% of the BIO file, and very probably more, because the process of finding the uncovered portion would almost certainly have picked up some of the references outside ISI cover but cover by other secondary services. The 'other methods' included scan of a number of minor

primary journals, and, much more important, scan of some secondary services produced by the Commonwealth Agricultural Bureaux. It is unlikely that the results obtained by these methods can be improved on at present.

In view of the results obtained by searching Chemical Titles, it might have been advisable to include search of this service, but unfortunately it is no longer available. Similarly, it would be worthwhile to search MEDLARS using the term Biodegradation; however, there is no particular pressure on the Biodeterioration Information Centre to be highly current, and it might therefore be equally useful to use this term in searching Index Medicus manually.

In the present study, only simple truncated terms have been used in searching secondary services. A number of other forms of entry exist, such as citation linkages, CROSS codes, systematic indexing and so on, but the relatively limited coverage of the services examined does not give any grounds for expecting the use of these services to perform better than the reference collection method recommended above.

During the course of the study, it began to seem likely that a curve could be drawn relating recall to cost per unique relevant reference retrieved, for search of a computerised base. This curve is illustrated in Figure 11, Curve A.

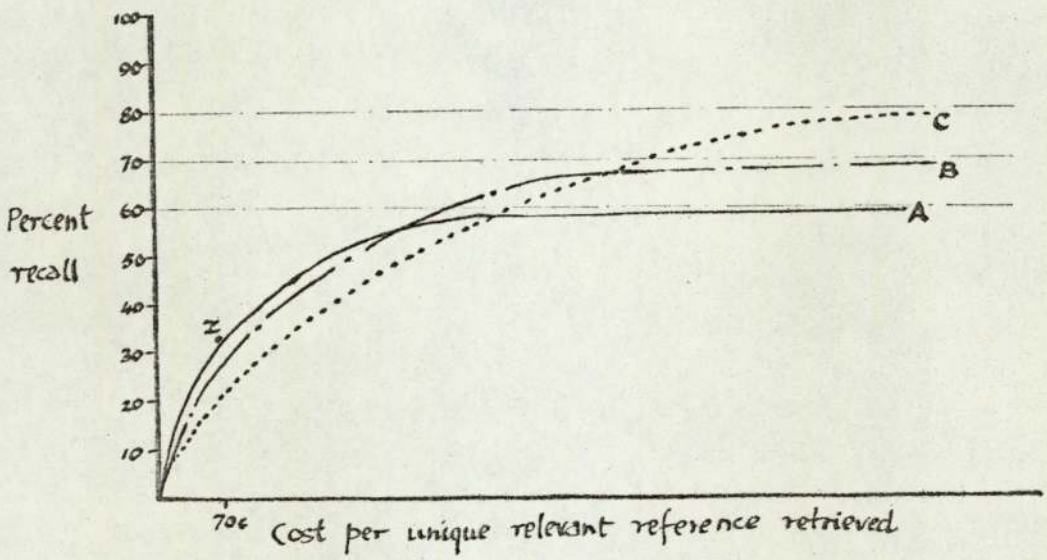


Figure 10. Relation between cost and recall.

'Recall' in this context means recall expressed as a percentage of the total number of references published in a year or other fixed time period; here it can be taken as percentage of the BIO file. Curve A can be taken as representing what happens when the best base, ISI, is searched. The curve is asymptotic at 60% recall, because the ISI file only covered 60% of the BIO file references. The section of the curve between the origin and point Z approximates to the use of the profile shown in Table 16, and the curve thereafter is based on educated guesswork. Curve B shows what might be expected to happen if two services are searched; there is a period at the bottom of the curves where cost per unique reference is higher because of the incidence of duplication. The asymptote is higher, at 70%, because 70% is the BIO file coverage of ISI and CA combined, and somewhere there is inevitably a cross over point where use of two services is cheaper than the use of one. Curve C represents the situation where three services are searched. In the present case, no matter how many services are searched, the asymptote is never higher than 77% recall. Very little quantification has been possible, because of difficulties and cost considerations mentioned in earlier sections, but such evidence as has been found tends to support this hypothesis. It would be interesting to carry out sufficient tests to prove or disprove this hypothesis, and to put real quantities on the axes of the graph, but empirical evidence suggests that until the majority of bases can be searched much more cheaply than at present, or are structured in such a way that relevant references cost less, and until coverage of all subject is higher, the research would be interesting rather than helpful.

Lastly, most of the work in this study has been conditioned by the fact that 23% of the literature of Biodeterioration is not covered by existing computerised secondary services. Other earlier work (5,6) suggests that secondary services generally miss about this amount of material. The present study suggests that what is not picked up cannot be written off as trivial or of little value. Either, therefore, it should be admitted than existing services tend to be oriented towards the

needs of scientists only, and steps be taken to introduce adequate services for the academically humbler but economically more productive technologist, or coverage of the world's literature by the existing services should be improved. One can only get out of a system what is put into the system, and generally speaking one can economically only get out a part of it. No amount of system refinement, and no mode of access, on or off line, can get out what is not there. For the individual user, who often does not require completeness, this may not matter, but for the Specialised Information Centre, it does.

One obvious conclusion to be drawn from the study reported here is that the establishment of the Biodeterioration Information Centre was completely justified, in that it provides a service which is not provided by any other single body and is imperfectly provided by a combination of other bodies. The direct costs of Centre operation are of the order of £9840 for the period 1st August 1973 - 31st July 1974, excluding printing costs of a new bulletin dealing with Waste Materials, and including costs of the enquiry service; the population served (i.e. the number of copies of Biodeterioration Research Titles, the successor to IBBRIS, distributed) numbers approximately 800. Where need groups of similar size can be found to exist, and where their pattern of reference need is comparably interdisciplinary, the provision of Specialised Information Centres, operating in much the same way as the Biodeterioration Information Centre, appears at present to be the best way of providing information support.

Given the existence of a network of secondary services organised along the existing disciplinary lines, it seems likely that a number of areas of need which do not fit easily into the disciplinary classification may find themselves effectively underprivileged as regards information support. In view of the difficulties of providing support from the existing services which have been illustrated by the present study, it is suggested that in planning future information service networks, provision be made for the establishment and support of a network or array of small specialised information centres, in order to bridge gaps in existing provision, and to mediate between

special interest groups and the major disciplinary services.

References.

1. President's Science Advisory Council. Science, government and information ("The Weinberg Report"). U.S. Government Printing Office, Washington. 1963.

2. ROBERTSON, S. and REYNOLDS, R. Five Specialised Information Centres OSTI Report No. 5050, OSTI, London 1969.

3. MARTYN, John. Notes on the operation of Specialised Information Centres. Aslib, London 1970.

4. MARTYN, John. Evaluation of specialised information centres. Information Scientist 4 (1970) p.123-34.

5. MARTYN, J. and SLATER, M. Tests on abstracts journals. Journal of Documentation 20 (1964) p212-35.

6. MARTYN, John. Tests on abstracts journal: coverage, overlap, indexing. Journal of Documentation 23 (1967) p.45-70.

7. LYNCH JT and SMITH GDW. Scientific information by computer Nature 230 (March 1971) p.153-156.

8. Aslib Research Department. Information on tape. Aslib, London 1971.

9. SCHIPMA, P.B., WILLIAMS M.E. and SHAFTON, A.M. Comparison of Document Data Bases. Journal of the American Society for Information Science. 22 (1971) p.326-332.

10. VICKERY, Brian C. Techniques of Information Retrieval Butterworth, London.

11. Institute for Scientific Information. ASCA (L) IV. Philadelphia 1969.

References (contd)

12. BARKER, F.H., VEAL, D.C., and WYATT, B.K. Towards automatic profile construction. Journal of Documentation. 28 (1972) p.44-55.

13. Biosciences Information Service of Biological Abstracts. A Guide to the Vocabulary of Biological Literature. Biological Abstracts, Philadelphia.

APPENDIX A.

43 JOURNALS PRODUCING 50.4% OF BIO FILE JOURNAL REFERENCES.

JOURNAL TITLE	NO.OF REFS.
Applied Microbiology	83
Agricultural and Biological Chemistry	48
Journal of Economic Entomology	40
Canadian Journal of Microbiology	40
Material und Organismen	38
Mikrobiologiya	37
Archiv für Mikrobiologie	35
Journal of Stored Products Research	33
Journal of the Association of Official Analytical Chemists	30
Journal of Fermentation Technology	38
Process Biochemistry	27
Phytopathology	26
Food Technology	22
Labdev Journal of Science and Technology	22
Journal of Agricultural and Food Chemistry	22
Fette, Seifen, Anstrichmittel	21
Annales de l'Institut Pasteur	20
Zeitschrift für Lebensmittel	20
Biblioteka Musealnictwa I Ochrony Zabytkow Ser.B	19
Nature	18
Mycopathologia et Mycologia Applicata	17
Journal of the American Oil Chemists Society	16
International Biodeterioration Bulletin	16
Journal of Food Science	16
Pest Articles and News Summaries (PANS)	15
Chemistry and Industry	14
Transactions of the British Mycological Society	14
Folia Microbiologica, Praha	14
Forest Products Journal	14
Holz als Roh - und Werkstoff	14
Journal of the Science of Food and Agriculture	14
Mycologia	14

(Appendix A continued)

Journal of Applied Bacteriology	13
Journal of Bacteriology	13
Journal of the Institute of Wood Science	13
Biotechnology and Bioengineering	12
Bulletin of the Japanese Society of Scientific	12
Canadian Journal of Botany	12
Journal of General Microbiology	12
Postepy Mikrobiologii	12
Soil Biology and Biochemistry	12
Plant Disease Reporter	11

Appendix B

ALPHABETICAL LIST OF JOURNALS SUPPLYING REFERENCES TO THE BIO FILE.

JOURNAL TITLE	NO.OF REFS.
Acarologia	1
Acta Agricultura Scandinavica	1
Acta Botanica Neerlandica	1
Acta Entomologica Bohemoslovaca	3
Acta Microbiologica Polonica	10
Acta Mycologica	1
Acta Pathologica et Microbiologica Scandinavica	1
Acta Physiologica Scandinavica	1
Acta Societatis Botanicorum Poloniae	4
Acta Veterinaria Academiae Scientiarum Hungaricae	1
Advances in Chemistry series	2
Advances in Microbial Physiology	2
Agricultura	1
Agricultural and Biological Chemistry	48
Agricultural Food Chemistry	3
Agriculture	1
Agronomy Journal	1
Alimenta	5
Amateur Gardening	1
American Dyestuffs Reporter	2
American J. Botany	3
American J. Hospital Pharmacy	2
American J. Veterinary Research	1
American Paint J.	3
American Zoologist	2
Analytical Biochemistry	1
Analytical Chemistry	1
Analyst	4
Angewande Botanik	1
Aminal Behaviour	1
Annales de Biologie Animale, Biochimie, Biophysique	1
Annales de l'Institut Pasteur	20
Annales de la Nutrition et de l'Alimentation	1

(Appendix B continued)

Annales de Phytopathologie	1
Annals of Applied Biology	4
Annals of the Entomological Society of America	2
Annals of Phyopathological Society of Japan	3
Annual Review of Entomology	1
Annual Review of Microbiology	1
Anti-corrosion	2
Antonie van Leeuwenhoek J of Microbiology and Serology	2
Anzeiger Schädlingskunde Pflanzenschutz	2
Applied Microbiology	83
Archiv für Hygiene ū Bakteriologie	5
Archiv für Lebensmittelhygiene	2
Archiv für Mikrobiologie	35
Archiv für Pflanzenschutz	1
Archiva Veterinaria	1
Archives Internationales Physiologie et Biochimie	4
Archives of Biochemistry and Biophysics	5
Archives of Environmental Health	1
Archives Roumaines de Pathologie Experimentale et de Microbiologie	2
ASB Bulletin	1
Australasian Engineer	1
Australian Food Manufacture	1
Australian Forestry	1
Australian Journal of Botany	1
Australian Journal of Experimental Agriculture and Animal Husbandry	1
Australian Power Engineering	1
Australian Timber Journal	1

B

Bacteriological Reviews	1
Baker's Digest	2
Baywood Courier	2
Bericht über die Tätigkeit der St. Gallischen Naturwissenschaftlichen Gesellschaft.	1

(Appendix B continued)

Bericht der Deutschen Botanischen Gesellschaft	7
Berichte Ohara Instituts Landwirtschaftliche Biologie	4
Biblioteka Musealnictwa I Ochrony Zabytkow,	19
BIBRA Bulletin	2
Bi-Monthly Research Notes	7
Biochemical and Biophysical Research Communications	2
Biochemical Journal	3
Biochemical Pharmacology	1
Biochimica et Biophysica Acta	9
Biokhimiya	1
Biologia Plantarium	1
Biological Bulletin	3
Biologie du Sol	1
Biopolymers	1
BioScience	4
Biotechnology and Bioengineering	12
Bitki Koruma Bulteni	1
BLF Forschungsergebnisse	1
Boden Wand Decke	1
Boletim da Sociedade Portuguesa Ciencias Naturais 2a serie	1
Bollettino dell'Instituto Pathologia del Libro	1
Bollettino Museo Civico Venezia	1
Brauwissenschaft	1
British Columbia Lumberman	1
British Corrosion Journal	1
British Journal of Cancer	1
British Journal of Experimental Pathology	1
British Journal of Pharmacology and Chromotherapy	1
British Poultry Science	6
Brot ü Gebäck	2
Bulletin of the Academy of Sciences USSR - Biological Series	1
Bulletin of Environmental Contamination and Toxicology	2
Bulletin of the Faculty of Agriculture, Hirosaki University	1

(Appendix B Continued)

Bulletin of the Government Forest Experiment Station	1
Bulletin of the Japanese Society for Scientific Fisheries	12
Bulletin de la Société Botanique de la France	2
Bulletin de la Société de Chimie Biologique	1
Bulletin Trimestriel de la Société Mycologique de la France	3

C

Canada Courier	2
Canadian Entomologist	3
Canadian Forest Industries	3
Canadian Institute of Food Technology Journal	7
Canadian Journal of Biochemistry	1
Canadian Journal of Botany	12
Canadian Journal of Chemistry	1
Canadian Journal of Microbiology	40
Canadian Journal of Pharmaceutical Sciences	2
Canadian Journal of Plant Science	4
Canadian Journal of Zoology	3
Canadian Medical Association Journal	1
Canadian Mining and Metallurgical Bulletin	1
Canadian Plant Disease Survey	1
Caribbean Farming	1
Cellulosa e Carta	1
Cereal Chemistry	5
Cereal Science Today	1
Ceres/Vicosa, Brazil	1
Ceska Mykologie	1
Chemical and Engineering News	2
Chemische Rundschau	1
Chemistry and Industry	14
Chemistry in Britain	1
Chemotheraphy	1
Chimie - Actualites	3
Chemie Industrie, Genie Chimique	2
Ciencia Cultura	2
Coatings	1

(Appendix B continued)

Color and Paint	1
Commercial Fisheries Review	1
Commonwealth Phytopathological News	1
Comparative Biochemistry and Physiology	1
Comptes Rendus Academie Science Paris, Ser D.	6
Comptes Rendus Séances Société Biologie	2
Confectionery Manufacture and Marketing	1
Confectionery Production	1
Conserva	2
Corrosion Prevention and Control	2
Corrosion Science	1
Corrosion, Traitement, Protection, Finition	1
Cronache Chimica	1
Crop Science	2
Cuoio	1
Current Science	3

D

Dansk Skovforenings Tidskrift	1
Deutsche Lebensmittel - Rundschau	8
Deutsche Textiltechnik	2
Developments in Industrial Microbiology	2
Diaita	1
Drevarsky Vyskum	2
Drug and Cosmetic Industry	2

E

Ecology	3
Engineering News	1
Entomologia Experimentalis et Applicata	1
Entomologicheskoe Obozrenne	1
Entomologist's Monthly Magazine	1
European Chemical News	1
European Journal of Biochemistry	1
Experientia	6

F

FAO Plant Protection Bulletin	1
Faber ü Lack	2
Farming World	1
FDA Papers	1
Feedstuffs	3
Fette, Seifen, Anstrichmittel, Ernährungsindustrie	21
Fleischerei	1
Fleischwirtschaft	7
Florida Entomologist	1
Flussiges Obst	1
Folia Microbiologica/Praha	14

(Appendix B continued)

Folio Forestalia Polonica, Ser.B	1
Food in Canada	1
Food Chemical News	1
Food Cosmetics Toxicology	4
Food Engineering	4
Food Manufacturer	4
Food Preservation Quarterly	2
Food Processing Industry	1
Food Technology	22
Food Technology in Australia	2
Food World	2
Forest Products Journal	14
Forest Science	1
Forestry Chronicle	2
Friesia	1

G

Gas ū Wasserfach	1
Geologische Rundschau	1
Gesundheitstechnik	1
Getreide ū Mehl	1
Ghana J. Science	1
Gidrobiologicheskii Zhurnal	1
Gidroliznaya Lesokhimicheskaya Promyshlennost	2
Giornale di Microbiologia	1
Gordian	3

H

Hemijska Industrija	1
Highlights of Agricultural Research	1
Hindustan Antibiotics Bulletin	2
Holz als Roh-ū-werkstoff	14
Holz Zentralblatt	1
Holzforschung	9
Holzforshung ū Holzverwertung	2
Holzindustrie	6
Holztechnologie	1

I

ILZRO Research Digest	1
Indian Forester	1
Indian Journal of Experimental Biology	1
Indian Phytopathology	1
Indian Rubber Bulletin	2
Industries Alimentaires et Agricoles	1
Information Chimie	1

(Appendix B continued)

International Biodeterioration Bulletin	16
International Shipbuilding Progress	1
Internationale Revue Gesamten Hydrobiologie	1
Investigacion Perquera	1
Investigacion Tecnicos Papeleros	1
Ion (Madrid)	1
Irish Journal of Agricultural Research	3
Israel Journal of Agricultural Research	1
Israel Journal of Botany	3
Israel Journal of Chemistry	1
Izvestiya Akademii Nauk SSSR, Ser.Biologicheskaya	1
Izvestiya Vysshikh Uchebnykh Zavedenii Lesnoi Zhumal	2

J

J. Agricultural Chemical Society Japan	4
J. Agricultural & Food Chemistry	22
J. Agriculture (South Australia)	1
J. Agriculture University of Puerto Rico	1
J. Allergy	1
J. American Dietetic Association	1
J. American Oil Chemist's Society	16
J. American Veterinary Medical Association	1
J. Antibotics	6
J. Applied Bacteriology	13
J. Applied Chemistry	3
J. Applied Ecology	1
J. Applied Radiation and Isotopes	1
J. Association Official Analytical Chemists	30
J. Bacteriology	13
J. Biological Chemistry	1
J. Cell Science	1
J. Chromatography	3
J. du Conseil, Conseil International Permanent pour l' Exploration de la Mer	2
J. Dairy Science	2
J. Economic Entomology	40
J. Elisha Mitchell Scientific Society	2
J. Experimental Botany	1
J. Experimental Zoology	1
J. Faculty Agriculture, Kyushu University	1
J. Faculty Science, University of Tokyo	1
J. Fermentation Technology	28
J. Fisheries Research Board of Canada	3
J. Food Science	16
J. Food Science and Technology	2
J. Food Technology	7
J. General Applied Microbiology, Tokyo	2
J. General Microbiology	12
J. Horticultural Science	1
J. Hygiene	5
J. Institute of Brewing	1

(Appendix B continued)

J. Institute of Wood Science	13
J. Institution of the Rubber Industry	1
J. Investigative Dermatology	1
J. Insect Physiology	5
J. Invertebrate Pathology	4
J. Japan Tappi	2
J. Marine Biological Association	7
J. Matsusaka Women's Junior College	1
J. Medicinal Chemistry	2
J. Milk & Food Technology	8
J. Oil & Colour Chemists' Association	8
J. Organic Chemistry	2
J. Paint Technology	3
J. Pharmaceutical Sciences	4
J. Pharmacy Pharmacology	2
J. Polymer Science	1
J. Royal Society of Arts	1
J. Rubber Research Institute of Malaya	1
J. Science of Food and Agriculture	14
J. Society of Cosmetic Chemists	1
J. Society of Dairy Technology	1
J. Stored Products Research	33
J. Water Pollution Control Federation	2

K

Kakao Zucker	1
Kaltetechnik-Klimatisierung	2
Karstenia	1
Kieler Meeresforschungen	2

L

Labdev J Science Technology	22
Laboratory Animals	1
Laboratory Equipment Digest	1
Laboratory Practice	9
Lantbrukshögskolans Annaler	1
Lebensmittel Wissenschaft Technologie	1
Life Sciences	1
Lloydia	1
Lubrication Engineering	3

M

Magyar Allatorvosok Lapja	1
Makromolekulare Chemie	1
Malayan Forester	1
Manufacturing Chemist Aerosol News	9
Manufacturing Confectioner	1

(Appendix B continued)

Marine Biology	4
Material Organismen Material Organismen, Beihefte	38
Materials Protection	5
Materials Research Standards	1
Meddelelser Fra det Norske Skogforsøksvesen	1
Medycyna Doswiadczalna Mikrobiologia	1
Meilland Textilberichte	1
Memoirs Faculty of Agriculture, Kagoshima University	1
Metallvereinigung Vorbehandlung	1
Metron	1
Metal Finishing	3
MGA Bulletin	1
Microbiologia Española	2
Microbiologia Parazitologia si Epidemiologie	1
Microbios	1
Mikologiya Fitopatologiya	4
Mikrobiologija	2
Mikrobiologiya	37
Milling	2
Mitteilungen Floristisch-Soziologischen Arbeitsgemeinschaft	1
Mitteilungen Gebiete Lebensmitteluntersuchung Hygiene	2
Mitteilungen Obstbau	1
Morfologia Normala Patologica	1
Müanyang Gumi	1
Mukomol'no-Elevatornaya Promyshlennost'	1
Mycologia	14
Mycopatholgia et Mycologia Applicata	17
Mykosen	1

N

Nachrichten Chemie Technik	1
Nachrichtenblatt Deutschen Pflanzenschutzdienst	1
Nahrung	2
Natur Museum	1
Nature	18
Naturwissenschaften	1
Naturwissenschaftliche Rundsehau	1
Netherlands J Plant Pathology	2
Neue Verpackung	1
New Pathologist	4
New Scientist	6
New Technology	2
New Zealand J Agricultural Research	1
Non-ionizing radiation	1
Nordisk Veterinaer Medicin	1
Ngtt Magasin Botanikk	1

(Appendix B continued)

O

Oberflache - Surface	1
Ocean Engineering	2
Ochrona Zabytkow	1
Oecologia	1
Oikos	1
Oil, Paint and Drug Reporter	1
Oleagineux	4
Opakowanie	1
Osterreichische Botanische Zeitschrift	2

P

Pacific Science	1
Paintindia	1
Paint Manufacture	1
Paint, Oil Colour Journal	2
Paint Technology	4
Paint and Varnish Production	1
Pakistan J. Forestry	1
Pakistan J. Forestry	1
Paper Tape Journal	4
Papier	2
Papiermacher	1
Papir Celulozo (Czech)	1
Pathologia et Microbiologia	2
P.A.N.S. Pest Articles and News Summaries	15
Peintures, Pigments, Vernis	7
Pest Control	5
Pharmaceutica Acta Helvetiae	2
Pharmaceutical J	4
Philippine Forests	2
Philosophical Transactions of the Royal Society, Biological Sciences	1
Photochemistry and Photobiology	1
Physiologia Plantarum	3
Phytochemistry	2
Phytopathology	26
Phytopathologische Zeitschrift	1
Phytoprotection	1
PIRA Paper and Board Journal	1
Plant and Soil	7
Plant Disease Reporter	11
Planta	2
Plaste u Kautschuk	1
Plastics Australia	1
Postepy Mikrobiologii	12
Poultry Science	1
Prace Instytutów i Laboratoriów Badawezych Przemslu Spozywczego	6

(Appendix B continued)

Prace Instytutu Przemyslu Organicznego	2
Prace Instytutu Technologii Drewna	5
Praktische Schädlingsbekämpfer	1
Prikladnaya Biokhimiya Mikrobiologija	2
Proceedings American Society of Civil Engineers Journal. Sanitary Engineering Division	1
Proceedings American Wood Preservers Association	5
Proceedings Australian Biochemical Society	1
Proceedings Chemical Specialities Manufacturing Association	1
Proceedings of the Indian Academy Sciences	1
Proceedings of the Institute of Food Science and Technology	2
Proceedings Kansai Plant Protection Society	1
Proceedings National Academy of Sciences	1
Proceedings Royal Irish Academy	1
Proceedings Royal Society	1
Proceedings Society of Analytical Chemistry	1
Proceedings Society of Experimental Biology and Medicine	1
Proceedings Soil Science Society of America	2
Process Biochemistry	27
Przemysl Spozywezy	1
Przemysl Fermentacyjny Rolny	3
Pulp Paper International	1
Pulp Paper Magazine Canada	3
Pyrethnum Post	1

Q

Qualitas Plantarium et Materiae Vegetabiles	1
Quarterly Reviews	1
Quick Frozen Foods	1

R

Radiation Research	1
Recorder CSIR ACCRA	4
Review of Applied Mucology	1
Revista Agricultura Subtropicale Tropicale	1
Revista Agroquimica Technologia Alimentos	1
Revue de l'Association Technique Industrie Papetière	1
Revue de l'Ecologie Biologie Sol	2
Revue Française Corps Gras	1
Rivista Agricoltura Subtropicle Tropicale	2
Roczniki Chemii	1
Roczniki Nauk Rolniczych	1
Roczniki Panstwowego Zakadu Higieny	3
Royal Society Health J.	1
Rubber and Plastics Age	1
Rybnoe Khozyaistvo	1

(Appendix B continued)

Schweizer Archiv für Angewandte Wissenschaft ü Technik	1
Schweizensche Wascherei -Zeitung	1
Science	6
Science & Culture	3
Science Journal	1
Seifen Ole - Fette - Wachse	2
Soap & Chemical Specialities	5
Soap, Perfumery, and Cosmetics	2
Soil, Biology and Biochemistry	12
Soil Science	1
South Africa Citrus J.	1
Southern Pulp and Paper Manufacturer	1
Span	2
Starke	2
Studies in Museology	1
Studies in Speleology	1
Studii si Cercetari Biochimie	1
Successful Farming	5
Sumarski List	1
Sumimoto Electric Technical Review	1
Surface Coatings	2
Susswaren	4
Svensk Papperstidning	4
Sylwan	1

T

Tappi	6
Tenside	2
Tethys	1
Tetrahedron	2
Tetrahedron Letters	1
Textile Research Journal	3
Textilveredlung	1
Timber	4
Tin and its uses	1
Toxicon	1
Trans. British Mycological Society	14
Trans. Institute in Chemical Engineers	1
Trans. Institute of Marine Engineers	1
Trans. Mycological Society of Japan	1
Trans. Society of Occupational Medicine	2
Travaux Centre Recherches Etudes Oceanographiques	1
Tribology	1
Tribune du Cebedeau	1
Tropical & Geographical Medicine	1
Tropical Forestry Industries	1
Tropical Science	1
Tropical Stored Products Information	4
Trudy Vsesoyuenogo Nauchno Issledovatel' skogo Instituta Zashchity Rastenii	1
Turrialba	1

(Appendix B continued)

U

Ukrains' Kii Biokhimichnii Zhurnal	2
Umschau Wissenschaft Technik	1
Undersea Technology	1

V

Verfkroniek	1
Veroffentlichungen Instituts Meeresforschung Bremerhaven	1
Verpackungs - Rundschan	4
Veterinariya	1
Veterinary Record	2
Virology	1

W

Wascherei Technik Chemie	1
Water Research	1
Weed Research	1
Weed Science	3
Werkstoff Korrosion	1
Wochenblatt Papierfabrikation	1
Wood / London	7
Wood Fiber	1
Wood Industry Tokyo	1
Wood Preserving	2
Wood Science	3
Wood Science News	1
Wood Science & Technology	1
Woodworking Industry	2
World Crops	1
World Review Pest Control	5

Z

Zaschita	1
Z.Allegmaine Mikrobiolgie	10
Z.Angewandte Entomologie	3
Zeitschrift Chemie	1
Z.Gesamte Hygiene Grenzgebiete	1
Z.Lebensmittel - Untersuchung - Forschung	20
Z.Naturforschung	2
Z.Pflanzenernährung Bodenkunde	2
Z.Pflanzenkrankheit Pflanzenpathologie Pflanzenschutz	1
Z.Vergleichende Physiologie	1
Zentralblatt Bakteriologie, Parasitenkimide, Infektionskrankheiton Hygiene I.orig.	2
Zh. Mikrobiologii Epidemiologii Immumobiologii	1

APPENDIX C.

Frequency of words in BIO file titles.

FUNGI	170	DEGRADATION	32
EFFECT	145	DEVELOP	40
STUDY	122	WATER	40
WOOD	109	MATERIAL	39
SOIL	98	USE	38
STORES	94	PAINT	38
AFLATOXIN	93	SPORE	37
ACID	91	YEAST	36
FOOD	82	TOXIC	34
GROWTH	82	CHEMICAL	33
PRESERVE	78	HYDROCARBON	33
BACTERIA	71	PROPERTY	33
PRODUCE	71	COMPOUND	32
ACTIVITY	70	TEMPERATURE	32
CONTROL	70	ACTION	31
METHOD	68	BIOLOGICAL	31
ASPERGILLUS	63	CHANGE	31
ENZYME	63	DECOMPOSITION	31
MICROBE	61	FUNGICIDES	31
PRODUCTION	60	GRAIN	31
MICROORGANISM	59	INVESTIGATION	30
ISOLATED	56	VARIETY	30
INFLUENCE	54	METABOLISM	29
TEST	53	EVALUATION	28
MICROBIOLOGY	51	COMPARE	27
RESISTANCE	49	FACTOR	27
CULTURE	48	FERMENT	27
PART	48	PAPER	27
SPECIES	48	PROBLEM	27
TREAT	48	PROCESS	27
DETERMINE	46	STRAIN	27
INSECTA	45	SYSTEM	27
RELATIONSHIP	45	UTILIZATION	27

(Appendix C contd.)

FORMATION	26	CLOSTRIDIUM	20
LABORATORY	26	DETECTION	20
MARINE	26	INHIBITION	20
PROTECT	26	MOISTURE	20
RESEARCH	26	PENICILLIUM	20
COLEOPTERA	25	PINE	20
FISH	25	PURIFY	20
NATURAL	25	REDUCTASE	20
QUALITY	25	TECHNIQUE	20
ROT	25	ASSOCIATION	19
SOURCE	25	ATTACK	19
AGENT	24	CONTENT	19
BEETLE	24	GERMINATION	19
GENUS	24	OCCURRENCE	19
BACILLUS	23	SEED	19
BOARD	23	SPOILAGE	19
CORROSION	23	ANTIBIOTIC	18
ENVIRONMENT	23	BIOCHEMISTRY	18
INDUSTRIAL	23	CHARACTERIZATION	18
INSECTICIDE	23	MICROFLORA	18
MEDIUM	23	NITROGEN	18
MOULDY	23	TOXIN	18
PSEUDOMONAS	23	TYPE	18
TERMITES	23	APPLE	17
ANTIMICROBIAL	22	APPLICATION	17
MEAT	22	BEHAVIOUR	17
PROTEIN	22	CARBON	17
HERBICIDE	21	CELLULOLYTIC	17
OIL	21	COAT	17
RADIATION	21	FRUIT	17
ROLE	21	TWO	17
CELL	20	WASTE	17
CELLULOSE	20	CONDITION	16

(Appendix C contd)

DERIVATIVE	16	NUTRITION	14
DETERIORATION	16	ORGANIC	14
DIFFER	16	PARTICLE	14
DISTRIBUTION	16	POSSIBLE	14
PESTICIDE	16	PREPARE	14
SALMONELLA	16	REDUCE	14
SUBSTANCE	16	STERILITY	14
TAXONOMY	16	TIMBER	14
AFFECT	15	WHEAT	14
CONTAMINATION	15	ALGEE	13
CONTRIBUTION	15	AMINO	13
EXTRACT	15	ASPECTS	13
FREST	15	CAUSE	13
FUSARIUM	15	CELLULOSE	13
GAS	15	CONTINUOUS	13
IDENTIFICATION	15	EFFECTIVENESS	13
LARVA	15	HAET	13
LOW	15	HYDROLYSIS	13
RESIDUE	15	MECHANISM	13
RESULT	15	PEANUT	13
SURFACE	15	PEST	13
TRANSFORM	15	PLASTIC	13
AIR	14	STRUCTURE	13
ANALYSIS	14	PREVENT	13
BIODEGRADATION	14	RAPID	13
CEREAL	14	REPORT	13
COCKROACH	14	SUSCEPTIBILITY	13
DECAY	14	BLUE	12
FEED	14	CANDIDA	12
GROUNDNUT	14	CERTAIN	12
LOSS	14	CONSERVATION	12
MICROORGANISM	14	CULTIVATE	12
MYCOTOXIN	14	DDT	12

(Appendix C contd)

DRY	12	ABSORPTION	10
EXAMINATION	12	AGAR	10
EXTRACELLULAR	12	AQUEOUS	10
FIELD	12	ANIMAL	10
GAMMA	12	ANTIFOULING	10
INFECT	12	FACTERIOLOGY	10
IRRADIATION	12	BIODETERIORATION	10
LINGNIN	12	BORER	10
MILK	12	CITRUS	10
MOLD	12	FROZEN	10
NOTE	12	COSMETIC	10
NUTRITION	12	CURE	10
TEXTILE	12	DERMESTID	10
THREE	12	ECOLOGY	10
ACTINOMYCETE	11	EGG	10
BACTERIUM	11	INDUCE	10
BLACK	11	IRON	10
COTTON	11	LAYER	10
DESTRUCTION	11	LIGHT	10
DETERGENT	11	MALATHION	10
EXPERIMENT	11	METHYL	10
FLOUR	11	MILDEW	10
HIGH	11	OBSERVATION	10
MEDIA	11	PATHOGEN	10
MERCURY	11	PECTIC	10
ORGANO	11	POPULATION	10
PIGMENT	11	POTATO	10
SAMPLE	11	POULTRY	10
SIGNIFICANCE	11	PRECEDURE	10
SODIUM	11	RAT	10
SPORULATION	11	REFER	10
STABLE	11	SEA	10
THERMOPHILIC	11	SELECT	10
THIN	11	SEPARATE	10

(Appendix C contd)

SOLUTION	10	ACTIVE	8
SOUTH	10	BIOLOGY	8
STAPHYLOCOCCUS	10	BROWN	8
SURVICE	10	CHIP	8
SYNTHESIS	10	CHROMATOGRAPHY	8
WORK	10	COUNT	8
ADDITIVE	9	FIVE	8
ASSAY	9	FOODSTUFFS	8
ATHMOSPHERE	9	GERMAN	8
BIOSYNTHESIS	9	HUMIDITY	8
CHARACTERISTIC	9	HYGIENE	8
COMPOSITION	9	IMPROVED	8
CONTAIN	9	METABOLITE	8
CORN	9	MINERAL	8
DAMAGE	9	MITE	8
DEHYDROGERASE	9	MODEL	8
FIBER	9	MYCOFLORA	8
GREEN	9	NITRATE	8
LIQUID	9	NORTH	8
MEASURE	9	ORGANISM	8
CXIDE	9	OXIDATION	8
PRESSURE	9	OXYGEN	8
PROTEASE	9	PACKAGING	8
PULP	9	PHYSIOLOGY	8
PYRETHRIN	9	PRESENCE	8
REACTION	9	PROTEOLYTIC	8
SALT	9	RED	8
SCLEROTIUM	9	RESPIRATION	8
SITOPHILUS	9	SIMPLE	8
STANDARD	9	STATE	8
STREPTOMYCES	9	TISSUE	8
TIME	9	TRIBOLIUM	8
TROPICAL	9	ULTRAVIOLET	8
XYLAN	9	VALUABLE	8

(Appendix C contd)

VEGETABLE	8	GLUCOSE	7
WASHED	8	HAZARD	7
WHITE	8	HISTORY	7
YEAR	8	IMPORTANCE	7
ADDITION	7	INCIDENCE	7
AGE	7	INTERACTION	7
ALKYL	7	INVITRO	7
ANAEROBIC	7	LIFE	7
ANTIFUNGAL	7	LIPID	7
APPROACH	7	MODIFICATION	7
BARLEY	7	MOTH	7
BASIDICMYCETES	7	MUTANT	7
BUILD	7	N-ALKANE	7
CATAIASE	7	ORANGE	7
COMBINE	7	ORGANOLEPTIC	7
COMMERCIAL	7	PHENOL	7
COMPLEX	7	PHENYL	7
COMPONENT	7	POTENTIAL	7
CONCENTRATION	7	RATE	7
COTTON SEED	7	RAW	7
COUNTRY	7	REQUIRE	7
DIGEST	7	RESIN	7
DISEASE	7	RESPONSE	7
ELECTRON	7	REVIEW	7
ENTEROTOXIN	7	SENSITIVE	7
ESTER	7	SEVERAL	7
ESTIMATE	7	SEWAGE	7
EXPOSE	7	SMALL	7
FILM	7	SOLE	7
FLUORESCENCE	7	SOY	7
FOREST	7	SPECIAL	7
FUMIGANT	7	STAIN	7
GEL	7	STRENGTH	7
GLASS	7	STREPTOCOCCUS	7

(Appendix C contd)

SUBSTRATE	7	INFESTATION	6
SUGAR	7	INFORMATION	6
SULFATE	7	INOCULATION	6
SYMBIOTIC	7	KEEPING	6
TENEBRIONIDAE	7	KINETICS	6
WALL	7	LAKE	6
WOOL	7	MAINTAIN	6
ZINC	7	MEANS	6
ACETATE	7	MICROSCOPE	6
APPARATUS	6	OBTAIN	6
APPLY	6	OCHRATOXIN	6
BANANA	6	PERSISTENCE	6
BASE	6	PH	6
BEEF	6	POLLUTION	6
BIPHENYL	6	REDUCTASE	6
CARBOHYDRATE	6	REFRIGERATE	6
COMMON	6	RICE	6
CONFUSED	6	RING	6
CONSTITUENT	6	SOLUBLE	6
CREOSOTE	6	SOLVENT	6
CYANIDE	6	SPECIFIC	6
DEPOSITES	6	STIMULUS	6
DIET	6	SUBMERGED	6
DRUG	6	SULFUR	6
EQUIPMENT	6	SURFACTANT	6
ETHYLENE	6	SURVEY	6
FABRIC	6	SYNTHETIC	6
FERROBACILLUS	6	TENEBRIO	6
FILTRATE	6	TOLERANCE	6
FLY	6	VARIATION	6
FORMULATION (formula)	6	WINE	6
FOUR	6	YIELD	6
FUMIGATION	6	ABILITY	5
FUTURE	6	ADULT	5
INACTIVATION	6	AID	5

(Appendix C contd)

ALKALOID	5	HYLOTRUPES	5
ALPHA	5	INCUBATION	5
ANTIBACTERIAL	5	ION	5
APPEAR	5	KERNEL	5
AROMATIC	5	LATEX	5
ARSENIC	5	LOGS	5
ARTIFICIAL	5	LONG	5
ASCOSPORE	5	MALE	5
ASSESS	5	MARKET	5
BAG	5	MATTER	5
BENZENE	5	MEMBRANE	5
ECTRYTIS	5	MATABOLIC	5
BREAKDOWN	5	METAL	5
BRITISH	5	MILL	5
CLEAN	5	MIXED	5
CONSTRUCTION	5	MIXTURES	5
COOLING	5	MYCELIUM	5
COPPER	5	NUCLEASE	5
CRYSTALLINE	5	NUMBER	5
CURCULIONIDAE	5	NUT	5
CUTTING	5	OLD	5
DESCRIPTION	5	PARTICULAR	5
DIMENSIONS	5	PERMEABILITY	5
DISINFACTANT	5	PHEROMONE	5
ECONOMIC	5	PLATE	5
ESCHERICHIA	5	POLES	5
FATTY	5	POST-HARVEST	5
FILLETS	5	RECENT	5
FINISH	5	RELEASE	5
FUNGISTATIC	5	PHIZOSHERE	5
GLYCOL	5	RODENTIA	5
GREAT	5	ROOT	5
HONDAMYCIN	5	SAPWOOD	5

(Appendix C contd)

SEX	5
SHIP	5
SKIN	5
SPRAY	5
STAGE	5
STARCH	5
SUBSTITUTED	5
SWEET	5
TERM	5
THERMAL	5
TOMATO	5
TROGODERMA	5
VERTICILLIUM	5
VITAMIN	5
WEATHER	5

APPENDIX D.

Frequencies of stems in BIO file titles.

BIO full sample word-stem-frequency.

FUNG-	207	DETERMIN-	46
EFFECT-	158	RELAT-	45
PORUDCT-	139	METABOLI-	42
BACTERI-	130	DEVELOP-	41
WOOD-	124	VARI-	41
STUD-	122	MATERIAL-	40
MICROB	112	BIOLOG-	39
STOR-	103	HYDROCARBON-	39
SOIL-	99	USE-	39
ACID-	94	DEGRAD-	38
AFLATOXI-	93	PAINT-	38
FOOD-	92	FORM-	38
GROW-	83	CHEMI-	37
PART-	81	YEAST-	36
ACTIV-	79	TEMPERATURE-	34
PRESERV-	77	PROPERT-	33
CONTROL-	71	COMPOUND-	32
INSECT-	70	GRAIN-	32
METHOD-	67	ACTION-	31
CELL-	66	CELLULO-	31
ENZYM-	66	CHANG-	31
ASPERGILL-	63	GEN-	31
TOXI-	59	DECOMPOS-	30
MICROORGANISM-	57	CARBO-	30
ISOLAT-	55	INVESTIGAT-	30
INFLUENC-	54	FERMENT-	29
TEST-	53	FISH-	29
RESISTAN-	49	PEST-	29
SPECIES-	49	EVALUAT-	28
SPOR-	48	PROTECT-	28
TREAT-	48	SYSTEM-	28
CULTUR-	47	TERM-	28
WATER-	47	UTILI-	28

(Appendix D contd)

CHARACTERI-	27	ANTIMICROBIAL-	22
COMPAR-	27	MEAT-	22
FACTOR-	27	RADIATION-	22
PAPER-	27	RAT-	22
PROBLEM	27	AIR-	21
QUALIT-	27	DETECT-	21
STRAIN-	27	HERBICID-	21
PLANT-	26	LIGN-	21
PROCESS-	26	ROLE-	21
LABORATOR-	26	ASSOCIAT-	20
MATINE-	26	CELLULOSE-	20
NATUR-	26	CLOSTRIDIUM-	20
PROTEIN-	26	INHIBIT-	20
RESEARCH	26	MOIST-	20
ROT-	26	PINE-	20
SURFAC-	25	REDUC-	20
COLEOPTER-	25	TECHNIQUE-	20
SOURCE-	25	ATTACK-	19
APPLI-	24	CONTENT-	19
AGENT-	24	GERMINAT-	19
PUR-	24	NITROGEN-	19
INDUSTR-	23	NUT-	19
BACILL-	23	OCCUR-	19
BOARD-	23	PECT-	19
CORRO-	23	PENICILLIUM-	19
ENVIRONMENT-	23	SPOIL-	19
MOULD-	23	WASTE-	19
OIL-	23	AMIN-	19
ORGAN-	23	ANTIBIOTIC-	18
OXID-	23	BEETLE-	18
PSEUDOMONAS-	23	BIOCHEMI-	18
SEED-	23	COTTON-	18
TERMITE- (see term-)	22	MICROFLORA-	18

(Appendix D contd)

MYCOTOXI-	18	CONTRIBUTI-	15
NUTRI-	18	COUNT-	15
ORGANO-	18	FUSARIUM-	15
TYPE-	18	GAMMA-	15
ACTINOM-	17	GAS-	15
APPLE-	17	HEAT-	15
BEHAVIO-	17	LARVA-	15
CONTAMINAT-	17	PLASTIC-	15
DIFFEREN-	17	RESULT-	15
FRUIT-	17	SEA-	15
LOW-	17	STERIL-	15
PARTICL-	17	STREPTOMYCE-	15
TAXONO-	17	IDENT-	14
TWO-	17	MECHANI-	14
COAT-	16	BIODEGRADA-	14
ADD-	16	CEREAL-	14
CONDITION-	16	CONTINU-	14
DERIV-	16	CULTIVA-	14
DETERIORATI-	16	DDT-	14
DISTRIBUTION-	16	DECAY-	14
FRESH-	16	GROUNDNUT-	14
LIP-	16	LONG-	14
RESID-	16	LOS-	14
SALMONELLA-	16	METHYL-	14
SOLU-	16	POSSIB-	14
SUBSTANCE-	16	STRUCTUR-	14
TRANSFORM-	16	TIMBER-	14
EXTRACT-	15	WHEAT-	14
ACET-	15	COCKROACH-	13
AFFECT-	15	PREPARATI-	13
ALG-	15	ASPECTS-	13
ANALY-	15	CONSERV-	13
CAUS-	15	FUMIGA-	13
CITR-	15	HIGH-	13

Appendix D (contd)

HYDROLYS-	13	MEDIA-	11
PEANUT-	13	PACK-	11
PREVENT-	13	PHENOL-	11
SCIEROTI-	13	PIGMENT-	11
SUSCEPTIB-	13	PRESEN-	11
TENEBRIO-	13	SIGNIFICANCE	11
TEXTILE-	13	SODIUM-	11
THREE-	13	STAB-	11
BLUE-	12	THERMOPHIL-	11
CANDID-	12	THIN-	11
CERTAIN-	12	AGAR-	10
CONTAIN-	12	ANIMAL-	10
ENTERO-	12	ANTIFOULING	10
EXAMINATION-	12	BIODETERIORATI-	10
EXTRACELLULAR-	12	BOR-	10
FAT-	12	COMPOSIT-	10
FEED-	12	COSMETIC-	10
FIELD-	12	DERMESTID-	10
INFECT-	12	DETERGENT-	10
IRRADIAT-	12	ECOLOG-	10
MEDIUM-	12	EGG-	10
MERCUR-	12	FLOUR-	10
MILK-	12	INDUC-	10
NOTE-	12	LAYER-	10
PULP-	12	LIGHT-	10
RAPID-	12	MALATHION-	10
SAMPL;	12	MILDEW-	10
SOFT-	12	MIX-	10
ABSOR-	11	OBSERVATION	10
AQU-	11	PATHOGEN-	10
BLACK-	11	POPULATION-	10
DESTR-	11	POTATO-	10
DRY-	11	POULTRY-	10
EXPERIMENT-	11	PROCEDURE-	10
FLOUR-	11	REACTI-	10
FROZEN-	11	REFERENCE-	10

(Appendix D contd)

SELECTI-	10	FILT-	8
SOUTH-	10	FIVE-	8
SPECIF-	10	GEL-	8
STAPHYLOCOCC-	10	HYGIEN-	8
SURVIVAL-	10	IMPROV-	8
SYNTHESIS	10	MEASURE-	8
WORK-	10	MINERAL	8
ASSAY-	9	MITE-	8
ATHMOSPHER-	9	MODEL-	8
BENZ-	9	MYCOFLORA-	8
BIOSYSTHE-	9	MITRAT-	8
CHIP-	9	NORTH-	8
CORN-	9	OXYGEN-	8
CRYSTAL-	9	PHYSIOLOG-	8
DAMAG-	9	PROTEOLYTIC-	8
DEHYDROGEMA-	9	PSYCHROPHYLIC-	8
FIB-	9	RED-	8
GENE-	9	RESPIRAT-	8
GREEN-	9	SALT-	8
IRON-	9	SENSITIV-	8
LIQUID-	9	SEPARATION-	8
PRESS-	9	SIMPL-	8
PROTEASE-	9	STATE-	8
SITOPHILUS-	9	SUBSTRAT-	8
STANDARD-	9	SULF1	8
TIME-	9	TISSUE-	8
TROPICAL-	9	TRIBOLIUM-	8
ALKAL-	8	ULTRAVIOLET-	8
CURE-	8	VALUE-	8
CUT-	8	VEGETA-	8
CYAN-	8	WASH-	8
DISINFECT-	8	WHITE-	8
ETHYL-	8	YEAR-	8

(Appendix D contd)

AGE	7	MOTH-	7
ALK-	7	MUTANT-	7
ANAEROBI-	7	N-ALKANE-	7
ANTIFUNGAL-	7	ORANGE-	7
APPROACH-	7	PHENYL-	7
BARLEY-	7	POTENTIAL-	7
BASIDIO-	7	RATE-	7
BOTRY-	7	RAW-	7
BROWN-	7	REQUIR-	7
BUILD-	7	RESIN-	7
CATAL-	7	RESPONSE-	7
CHROMATOGRAPH-	7	REVIEW-	7
COMBIN-	7	SEVERAL-	7
COMMERCIAL-	7	SEWAGE-	7
COMPONENT-	7	SMALL-	7
CONCENTRAT-	7	SOLE-	7
COUNTR-	7	SOY-	7
DIGEST-	7	SPECIAL-	7
ELECTRON-	7	SPECTR-	7
ESTERS-	7	STAIN-	7
ESTIMAT-	7	STRENGTH	7
EXPOS-	7	STREPTOCOCC-	7
FILM-	7	SUGAR-	7
FOREST-	7	SYMBIO-	7
GERMAN-	7	THIOBACILL-	7
HAZARD-	7	WALL-	7
HISTORY-	7	WOOL-	7
HUMIDITY-	7	APPARATUS-	6
GLUCOSE-	7	AROMA-	6
IMPORTANCE-	7	BANANA-	6
INCIDENCE-	7	BEEF-	6
INTERACTION-	7	BIPHENYL-	6
IN-VITRO-	7	CARBOHYDRATE-	6
LIFE-	7	COMMON-	6

(Appendix D continued)

COMPLEX-	6	REFRIGERAT-	6
CONFUSED-	6	RICE-	6
CONSTITU-	6	RING-	6
CREOSOTE-	6	RODENT-	6
DEPOSITS-	6	SHELF-	6
DIET-	6	SHIP-	6
DRUG-	6	SOLV-	6
EQUIPMENT	6	STIMUL-	6
FABRIC-	6	SUBMERGED-	6
FERROBACILL-	6	SULPH-	6
FOUR-	6	SURVEY-	6
FUTURE-	6	SWEET-	6
GLASS-	6	SYNTHETIC-	6
INACTIVAT-	6	TOLERANCE-	6
INFEST-	6	WEATHER-	6
INFORMATION-	6	WEEVIL-	6
INOCUL-	6	WINE-	6
KEEP-	6	YIELD-	6
KINETIC-	6	ADULT-	5
LAKE-	6	AID-	5
LATEX-	6	ALPHA-KETOGLUTAR-	5
MEANS-	6	AMYL-	5
METAL-	6	ANTIBACTERIAL-	5
MICROSCOP-	6	APPEAR-	5
MODIF-	6	ARSENIC-	5
OBTAIN-	6	ARTIFICIAL-	5
OCHRATOXIN-	6	ASCOSPORE-	5
PERSISTENC-	6	ASSESS-	5
pH	6	BAG-	5
PHYSIC-	6	BASE-	5
PIP-	6	BREAKDOWN-	5
POLLUTION-	6	BRIT-	5
QUANTIT-	6	CONSTRUCTION-	5
REDUCTASE-	6	COOL-	5

(Appendix D continued)

CURCULONIDAE-	5
DESCRIPTION-	5
DIMENSION-	5
DIMETH-	5
ECONOM-	5
ESCHERICHIA-COLI-	5
FILLETS-	5
FINISH-	5
FLY-	5
GLYCOL-	5
GREAT-	5
HONDAMYCIN-	5
HYLOTRUPES-	5
INCUBAT-	5
ION-	5
KERNELS-	5
LOGS-	5
MAINT-	5
MARKET-	5
MALE-	5
MATTER-	5
MEMBRANE-	5
MILL-	5
MYCELI-	5
NUCLEASE-	5
NUMBER-	5
OLD-	5
PARTICULAR-	5
PERMEA-	5
PLAT-	5
POLES-	5
POLYMER-	5
POST-HARVEST	5
RECENT-	5
RELEASE-	5
RHIZOSPHERE-	5
ROCK-	5
ROOT-	5
SAPWOOD-	5
SEX-	5
SKIN-	5
SOLID-	5
SPRAY-	5
STAGE-	5
STARCH-	5
STEROID-	5
SUBSTIT-	5
THERMAL-	5
TOMATO-	5
TROGODERMA-	5
VERTICILLIUM-	5
VITAMIN-	5

Appendix E

Sample section of the Permuterm Index 1969.

Sample section of the ISI Source Index 1969