THE ORGANISATION AND ACCESSIBILITY OF TECHNICAL INFORMATION IN THE INDUSTRIAL ENVIRONMENT

by Alfred Kenneth Scholes MA (CANTAB)

A Dissertation submitted to the Faculty of Arts
University of the Witwatersrand, Johannesburg,
for the degree of Master of Arts

Johannesburg 1975

PREFACE

The basic reasons for the development of information services are the vast total of man's knowledge and his inability to memorise it all:

> Tout le monde se plaint de sa memoire ....
> (Everyone complains of his lack of memory ....)
>
> La Rochefoucald

> We find a little of everything in our memory; it is a kind of pharmacy or chemical laboratory in which chance guides our hand now to a calming drug and now to a dangerous poison.
>
> Marcel Proust

> Knowledge is of two kinds. We know a subject ourselves, or we know where we can find information upon it.
>
> Dr Samuel Johnson

> A wise man will not go out of his way for information.
>
> Henry David Thoreau

Effective and efficient information services should enable even the not-so-wise individual to find available knowledge without him having to 'go out of his way'.

ACKNOWLEDGEMENTS

The information storage and retrieval systems described in this dissertation were developed by the author for AE&CI Ltd, formerly African Explosives and Chemical Industries, Ltd. The author wishes to thank that company for permission to include information on them and to submit this in the form of a thesis for the degree of Doctor of Philosophy. At the same time he wishes to acknowledge that many of the procedural details have been developed by other workers, supervised by the author and attached to the Information Group, Research Department, of the same company during the period 1963-1975. For this reason the detailed descriptions of the systems have been included in appendixes and not in the main body of the text.

The author also wishes to thank Professor J W Perry, Head of the Department of Bibliography, Librarianship and Typography of the University of the Witwatersrand, and Mr D G Kingwill, Director of Information and Research Services, South African Council for Scientific and Industrial Research, for their guidance and suggestions. Finally, especial thanks are made to Mrs A Klingbiel, Mrs D Speirs and Mrs D Brazier for their very careful typing, and to Mrs P Wood who supervised the reprographic aspects.

DECLARATION

I hereby declare that this thesis has not been submitted to any other university and that it comprises work undertaken by myself alone, except for the assistance acknowledged above.

....

A K SCHOLES

S U M M A R Y

The problem dealt with in this revised dissertation concerns the communication within industry of relevant published, or unpublished and confidential, technical information to those who require it.

In 1962 the author was required by African Explosives and Chemical Industries, Ltd (now AE&CI Ltd) to investigate whether one simple system could be developed which would provide means for the retrieval of technical information irrespective of the type of document containing it. Having determined the objective certain requirements were apparent:

1 The system had to be economically justifiable.

2 It had to be adaptable to changing and to new technical interests.

3 It had to be flexible so as to accommodate information in documents ranging from books and journals to technical reports, items of correspondence and patent specifications.

4 It had to be suitable for the convenient use of manual or mechanized methods.

5 It had to be acceptable to the potential users and permit them to operate it themselves on a 'self-help' basis.

The study set out is not so much a discovery as a pragmatic solution to a problem and it is clear that this solution has potential application in much of South African industry. It is also clear that the systems developed can be used by both large or small industrial concerns, and that they appear to be potentially well suited to the current stage of industrial development in the Republic. This was one of the points stressed when the original proposal for this thesis was submitted to the University. The long period which has elapsed since the first system was developed (1963), as well as the last (1968), and

the fact that all are still in use without major changes, provides evidence that they are both practical and economically viable.

The information presented is set out in dichotomous form. The first part of each Chapter, with the exception of Chapter 6 which deals with the actual systems developed for AE&CI, comprises the author's experience and opinions in conjunction with discussion of relevant published literature. The topics covered start with an introductory consideration of what information science and information services comprise, and this is followed by consideration of the information transfer process and of the needs of information users. The theme is developed to more specific levels and criteria for the design of information retrieval systems, factors which affect the performance of these systems and factors to be considered in the actual selection of a system are discussed in turn. The concluding chapters consider the problems of costing and evaluating and of managing information services; in Chapter 9 what are believed to be novel suggestions for the organisation and management of information services are put forward. These appear to be well suited to activities involving the provision of a service within industry.

In the course of revising this thesis much time was spent in the study of the relevant literature published between 1970 and 1974. It was clear that in the specific field of information retrieval systems the concern has been almost exclusively with highly mechanized systems. In the field of simple manual systems little of value was published so that the majority of the original references were retained. However, in the fields of user studies, costing, management, and probable future developments in Information Science, some very valuable contributions to the literature were published during this period and these have been incorporated.

The AE&CI systems described are all manual card systems and the main finding of the investigation has been that the variation of presentation of information in different types of documents does call for more than one retrieval system, if they are all to be adequately and economically catered for. But it has been demonstrated in practice that these systems can and do have common facets. These are:

1 the use of abstracts, whether informative, indicative or little more than amplified titles,

2 that the same indexing terms can be used for all systems,

3 that they can all be designed so that the costs of using them (ie of finding information quickly) are low, and

4 that they can all be operated so that the users can find information for and by themselves,

5 it is emphasised that it is the use of a common thesaurus of indexing terms which, in current information science jargon, 'integrates' the various systems - ie ties them together to form a comprehensive and more economic whole.

Finally, the author speculates on the future of information systems and sets out a technique which should be of use to those required to set up an information retrieval system.

CONTENTS

Chapter 1 INTRODUCTION

1.1 Scope and arrangement

This dissertation is concerned with the quantitative problems which have arisen from the so-called 'Information Explosion', as well as the qualitative ones. It is concerned primarily with those aspects which affect the sciences and technologies, as opposed to the humanities. It is limited to aspects of science and technology of concern to industrial activity.

As far as possible the arrangement throughout comprises two main sections to each chapter. The first sections are discussions of various subjects in relation to the published literature and to South African industry. The second sections concern the practical application of the same subjects within one industrial concern, namely AE&CI Limited (AE&CI).

1.2 Information services

Use of the phrase 'Information Explosion' is now so common that not only is an explanation unnecessary, it has become a cliché. Nevertheless the existence of this explosion is a truth that is not open to question and the problems which it has engendered are real(1). These problems are both qualitative and quantitative and they concern all the innumerable activities of mankind that are regularly committed to paper; it is difficult today to find any that are not so committed. The exponential growth in the quantity of technical information was to be expected as a direct result of increased population and standards of education, of increased numbers of scientists and technologists, of increased opportunities and pressure for research, and of increased industrial activity everywhere. This increase in industrial activity is itself complex. It has not only led to the appearance of many new sciences and technologies, it has resulted in the rapid development of traditional crafts towards scientific or technological respectability. Examples include farming, tanning, dyeing, pottery, and glass-making. Whatever the source, the total of scientific and technological information (comprising knowledge, experience, hypotheses, opinions, etc) is formidable. It is not practical to consider this information in vacuo since it is valueless to mankind when isolated, and the problems which arise from its existence are, at least initially, concerned with its communication or transfer from mind-to-mind.

At every stage in the historical development of science, information transfer has been an essential requirement. Informal discussions can be traced back to the scientists of ancient Greece, long before the invention of printing, and these are still practised and welcomed since few people are so conceited as to believe their own thinking infallible. A chemist, for example, can scarcely forget Lavoisier's historic error in concluding that oxygen is essential to the formation of an acid - an error that has been perpetuated in German by the name of this element, viz: sauerstoff. The French proverb 'Du choc des opinions jaillait la verité' (Truth springs from the clash of opinions) is as true today as when coined, but the point need not be laboured - without information transfer neither modern science nor technology could have evolved.

In the past it was often noted that the more successful scientists and technologists were usually skilled in either oral or written communication. Until relatively recently, say until the early years of the twentieth century, such personal capabilities sufficed to meet the requirement of scientific and technological communication. Those days are past and they are very unlikely to return. Communication has been called the life-blood of industry by numerous people; far too often, however, this life-blood has been - and still is - in short supply, so that only a few of the people involved in an industry really know why certain steps are taken, or are allowed to suggest what steps might be taken. Without doubt communication down the line has often been neglected so that subordinates feel that they know too little about what is going on to contribute most effectively towards ensuring that their business, research, or other activity is efficient and profitable. It is only since World War II that industry in general has begun to realise the mistakes which have been made, the opportunities and money lost and the time wasted through insufficient in-house communication, where essential ingredients of such communication involve information retrieval or dissemination.

The concept of communication is much wider than that of information transfer per se so that the former term is generic to the latter. However, it is the problems associated with information transfer that have been directly responsible for the activity now commonly called 'information science'. The term 'information science' may well be a misnomer (and indeed this has been the subject of much debate) since it was born out of the technology that produced data and information processing equipment, and as an activity it is perhaps more akin to a technology than to a science. In due course of time it may well develop into a true science, perhaps as a synthesis of concepts and methods originating in the theory of cognitive processes, mathematical linguistics, list processing techniques, logic, and so on. Longmans English Larousse defines science as 'knowledge acquired by careful observation, by deduction of the laws which govern changes and conditions, and by testing these deductions by experiment'. Webster's Third New International Dictionary defines technology as 'the application of scientific knowledge to practical purposes in a particular field'. From this definition of technology it follows that there must be theory underlying the scientific knowledge applied in a particular instance. This is certainly true for the field of communication science. A great deal of effort is being put, and has been put in recent years, into discussing the theory underlying the activity of information science[2-4], in the practical situations encountered in industry. Venn diagrams have certainly proved to be of considerable value. (Note: these have been used in several sections of Chapter 4).

Whether or not information science is a true science or not is debatable, and arguments similar to those often used in this instance apply equally well to numerous other activities; however, consideration of its origin and of attempts to define it are set out below since it is a well known fact in Library and Information circles in South Africa that the term is not understood in much of local industry.

In the late nineteenth century a trend developed towards referring to almost any human activity which could be studied, and around which a body of knowledge had developed, as a 'science'. The field of Librarianship was one such and gradually its accumulated knowledge became known as 'library science'. Many of those who were trained in the traditional natural sciences rather resented this broadening in the usage of the word 'science'. They found it difficult to appreciate that their endeavours to discover the laws underlying naturally occurring phenomena (by performing experiments, formulating hypotheses and theories, and employing the tools of mathematics and formal logic) stimulated others, working outside the traditional natural sciences, to seek comparable discoveries and understanding in their own fields. Certainly there is no prima facie case for not attempting to do so.

Soon after the turn of the century the name 'library science' fell into disuse, being replaced by the term 'documentation'. This latter term was coined by Otlet and La Fonteine [the founders of the 'Institut internationale de bibliographie in 1895, which was the forerunner of today's Fédération internationale de documentation' (FID)] and their definition comprised:

'documentation is a process by which are brought together, classified and distributed all the documents of all kinds of all the areas of human activity'.

At that time, when many librarians were exclusively concerned with the importance of the physical arrangement of documents on shelves, this definition, in conjunction with the writings of Otlet and La Fonteine, must have appeared quite revolutionary. Probably for the first time the subject content of both documents of all kinds and of largely non-verbal records, such as maps, diagrams and pictures, was stressed and emphasised, together with the independence of recorded human knowledge from its form of presentation. The study of this stress on the conceptual as well as the physical aspects of recorded knowledge, particularly in the fields of science and technology, was increasingly referred to as 'documentation' until a new definition by Bradford[5] in the early 1950s appeared to clinch the matter. He defined it by stating that: 'documentation is the art of collecting, classifying and making readily accessible the records of all kinds of intellectual activity'.

While Bradford's definition is brief and to the point, the word 'documentation' unfortunately developed overtones of ambiguity and vagueness, so that when Mooers coined the phrase 'information retrieval' in 1950[6] this was quickly adopted. However it was not long before the limitations of this new term became apparent and a need was felt for something wider, so that the term 'information science' appeared around 1960 and was defined in its scope and aims at the conference on the training of science information specialists, held at Atlanta, Georgia Institute of Technology in 1961. In the literature of the 1950s it was noticeable that the phrase 'scientific information' was frequently used in the sense of referring to the retrieval of information. It is not clear whether the name 'information science' arose spontaneously as a simple inversion of 'scientific information' but use of the former spread very rapidly in North America and subsequently elsewhere. Clearly the last word has not been said on this subject and the proposal by Mikhailov et al[3] in 1966 for adoption of the name 'informatics' appears to be a further step forward. Whether or not this term will come into widespread use still remains to be seen.

Study of the literature of information science shows that numerous efforts have been made to construct mathematical models as an aid to understanding the activity[7-15], but in general, these have not proved to be very helpful in understanding the practical problems of communicating technical information in industry. There may of course be special cases where this is not true but each industrial concern is unique in its range of technical interests, its organisation, its raw materials and products, its markets, and its geographical situation; no two are ever identical.

The result is that each concern has a unique and subjective approach to these problems which is dictated by its technical interests, organisation, etc. That this is factual can be recognised by visiting and studying a relatively small number of companies, and it is in marked contrast to the objective approach to these same problems to be found, and rightly so, in universities, public libraries and other public bodies. Mathematical models have found wide applicability in operations research work and no doubt those that have been set up within industrial companies are valid for each such unique environment. The applicability of each is limited to a particular set of conditions and while there is no a priori reason why this should not also be true for information retrieval purposes, it has proved to be particularly difficult to produce models of wide applicability in this field.

It is of interest to point out that virtually none of the writers who are concerned with a theoretical approach to information science are employed in industry, or claim administrative experience in operating information services or appear to be particularly interested in information problems in industry. One wonders therefore whether this is not an instance where 'the worst use of theory is to make man insensible to fact' as Lord Acton had it.

If it is accepted that recorded attempts to identify and apply the underlying theory of information science have had little <u>practical</u> success so far, then it may not be unreasonable to assume that it is, <u>per se</u>, primarily a pragmatic activity. It is doubtful whether the parameters of this activity can be clearly defined in the industrial environment. There appear to be three basic facets common, to a greater or lesser extent, to all information services. These are:

1) Information storage and retrieval
2) Information dissemination
3) Literature searching and surveying

Over and above these there are numerous ancillary facets and no two services in industry appear to include an identical list, as pointed out by Hoskovsky and Massey[16]. Examples of these are:

1) Industrial Property Services (ie Patents, Trade Marks, Registered Designs and Copyright)
2) Technical writing - eg Product brochures, manuals, etc
3) Editing and/or report writing advisory services
4) Thesaurus construction and control services
5) Translation services
6) Organisation of conferences/meetings/symposia/etc and secretarial services for these
7) Reprographic services.

There are many other ancillary facets of information services in addition to those listed above, but this dissertation is mainly limited to the basic facets. It is concerned particularly with information storage and retrieval systems (IR systems) but includes work done on thesaurus control in so far as indexing for future retrieval is concerned.

It should be noted here that conventional library services are an essential part of any technical information service. However the information scientist tends to cast his net rather wider than the librarian as regards useful sources of information. Not only does the information scientist attach importance to books (which he tends to regard as reference works in entirety, in that they represent knowledge in particular fields up to the time of publication) but he is equally concerned with documents of all kinds. Journals and pamphlets are normally to be found in any conventional library as well as books, but the information scientist is also concerned with such items as preprints and reprints, trade catalogues, standards specifications and patents as well as internally produced documents such as reports, memoranda and technical correspondence. Because of the subjective approach to information which is encountered in industry and which has already been referred to, these latter are very often of extreme importance.

Although industry is developing at a rapid pace in the Republic of South Africa it has a relatively short history. The existence, often on a substantial scale, of various service activities which industry in more advanced countries provides for itself, is not always fully appreciated locally. Certainly until very recently information science has fallen into this category, and the scale and viability of many overseas technical information services is not well known within South Africa, even though many local subsidiaries or associates of major overseas concerns benefit directly or indirectly from the existence of such in-house services. Today in North America, Europe and Japan there is hardly a chemical, pharmaceutical or engineering company of any standing that does not have its own well-developed technical information services. In addition, there are many thousands of small companies throughout the world that have also found it worthwhile to establish their own similar services. These latter are comparable in size with the biggest two to three hundred industrial concerns in South Africa.

The total number of technical graduates, and of experienced people without university technical degrees, who are employed as information scientists is unknown. The author estimates from several discussions with eminent men in the field during 1972 in the United Kingdom that the current total there is about 3 500 and in the United States 12 to 15 000. In West Germany the total is believed to be over 5 000.

In the Republic the total number of information scientists, with or without technical degrees, is estimated to be about 50 so that South Africa still has much development in this field ahead.

1.2.1 Some requirements for future research

The IR systems developed within South African industry so far are all manual systems, or mechanised in a relatively simple manner - eg using optical co-incidence cards. With the sole exceptions of a small system in the field of chemical engineering at Natal University and government systems handling classified information there are no computer-based systems in the Republic. This is not as surprising as it may at first appear to be, partly because of the fact that many of the larger concerns are subsidiaries or associates of major overseas companies, as pointed out in the foregoing section, and partly because much of the rest of South African industry is relatively unsophisticated and comprises small companies. Naturally this is not to say that computerised IR systems will not appear in the Republic within the foreseeable future. Little research appears to have been done within the country on cost/benefit threshold levels for such systems and on whether there is a need for systems to serve specific sectors of industry, made up of numerous small companies, at the national level - eg by the South African Council for Scientific and Industrial Research, or by trade Research Associations. That research is called for is substantiated by the fact that apart from requests for assistance which fall under the heading of Technical Service (ie in connection with raw materials, processing equipment or products) AE&CI alone receives an average of eight requests a month from customers enquiring whether retrospective searching can be done for them in the field of chemistry. This matter has also been discussed with the Information staff of the South African Iron and Steel Industrial Corporation; it is understood that in the field of ferrous metallurgy similar requests from customers are equally prevalent.

A second area in which research is required is related to the special and peculiar conditions of the South African economy and labour force. As a developing country South Africa lags behind the major industrialised countries in a number of ways. With a large uneducated and unskilled labour force numerous labour intensive activities have been retained in the face of demands for the use of more highly mechanised and capital intensive procedures and equipment. There are numerous socio-political reasons for this, and one such has been, and remains, the need to ensure maximum levels of employment. With rapidly rising levels of education and employment opportunities, coupled with a rapidly rising population a situation has developed which is likely to persist for many years; namely in a

very wide range of jobs, including library and information work, fragmentation has developed with a resultant narrowing of the scope of each such job. As an example, an experienced library assistant engaged on inter-library loan work is concerned with all aspects of this job. Currently one may well find this assistant has been replaced by two inexperienced but matriculated Africans; one concerned with out-going requests for inter-library loans and one concerned with in-coming requests. Obviously the whole question of skills, capabilities, capacity for development, opportunities for development, etc, is a very fertile field for research in such conditions.
The results of such research could well be of value in other developing countries, particularly in Africa. At the present accelerating rate of development this pressure for controlled fragmentation will undoubtedly move steadily up the scale and involve work at a higher intellectual level. As an example, there are possibilities for fragmenting the work of a graduate abstracter/indexer for IR systems - _ie_ of separating the work into that of abstracting and that of indexing. This, of course, is not unknown in the developed countries and the writer has seen such separation at, for example, the Battelle Memorial Institute, Columbus, Ohio, and Farbenfabriken Bayer, Leverkeusen, West Germany. However industrial conditions in the Republic will undoubtedly call for local research into both technical and sociological aspects of a more deliberate and formal fragmentation than is commonly encountered in industrially advanced countries.

1.3 AE&CI Limited

A basic purpose of this dissertation is to describe IR systems developed for one industrial concern (AE&CI) and to point out that these can, in part or in toto, be used advantageously by other companies, particularly smaller companies. This being so it is desirable to present some background information on AE&CI, and this is set out below:

very wide range of jobs, including library and information work, fragmentation has developed with a resultant narrowing of the scope of each such job. As an example, an experienced library assistant engaged on inter-library loan work is concerned with all aspects of this job. Currently one may well find this assistant has been replaced by two inexperienced but matriculated Africans; one concerned with out-going requests for inter-library loans and one concerned with in-coming requests. Obviously the whole question of skills, capabilities, capacity for development, opportunities for development, etc, is a very fertile field for research in such conditions.
The results of such research could well be of value in other developing countries, particularly in Africa. At the present accelerating rate of development this pressure for controlled fragmentation will undoubtedly move steadily up the scale and involve work at a higher intellectual level. As an example, there are possibilities for fragmenting the work of a graduate abstracter/indexer for IR systems - ie of separating the work into that of abstracting and that of indexing. This, of course, is not unknown in the developed countries and the writer has seen such separation at, for example, the Battelle Memorial Institute, Columbus, Ohio, and Farbenfabriken Bayer, Leverkeusen, West Germany. However industrial conditions in the Republic will undoubtedly call for local research into both technical and sociological aspects of a more deliberate and formal fragmentation than is commonly encountered in industrially advanced countries.

1.3 AE&CI Limited

A basic purpose of this dissertation is to describe IR systems developed for one industrial concern (AE&CI) and to point out that these can, in part or in toto, be used advantageously by other companies, particularly smaller companies. This being so it is desirable to present some background information on AE&CI, and this is set out below:

1.3.1 A brief history

AE&CI had its origin in the formation in 1895 of the Zuid-Afrikaansche Fabrieken voor Ontplofbare Stoffen with a factory at Modderfontein, Transvaal; after the South African war of 1899 to 1902 this company was taken over by the British South African Explosives Company, in which Nobel's Explosive Company, Ltd, of Glasgow, had a considerable interest[17].

In 1898 De Beers Consolidated Mines, Ltd (De Beers) decided to erect their own explosives factory at Somerset West, Cape Province. This was done and in 1906 the factory hived-off as a wholly owned subsidiary company named 'The Cape Explosives Works, Ltd'. This was followed in 1907 by the establishment of yet another explosives company, this time by Kynoch, Ltd of Birmingham, at Umbogintwini, Natal. These three companies competed until 1918 when Nobel's Explosive Co and Kynoch Ltd formed an alliance in Britain which resulted in the cessation of competition between Modderfontein and Umbogintwini. In 1924 the three South African companies amalgamated to form African Explosives and Industries, Ltd (AE&I), which changed its name in 1943 to African Explosives and Chemical Industries, Ltd - now AE&CI Ltd.

In 1926 several major chemical companies in Britain amalgamated to form Imperial Chemical Industries, Ltd (ICI) and the outcome as far as AE&I was concerned was that this latter company was then jointly owned by ICI and De Beers; a situation which continued unchanged until 1966 when 15% of the total equity capital was offered to the South African public.

1.3.2 Technical information services up to 1963

One result of ICI's participation was that AE&I began to receive formal written but unpublished reports on numerous technical subjects, each containing information confidential to the originators. After 1947, when AE&CI (as it was by then) established its own research department, this traffic in technical reports developed into a two-way stream which increased rapidly. By 1962 the total number of documents involved totalled about 20 000 (by 1973 this had increased to 40 000) and it had become increasingly difficult to find reports by subject content and even by author or serial number at times. The stage had been reached when it was virtually impossible to ensure that <u>all</u> the available information on any given subject could be found. The only available index was a modified Kaiser library catalogue[18] (see Footnote) which had been built up by a succession of staff, most of whom were neither qualified librarians nor technically trained. This was particularly unfortunate since two of the three original companies which had combined to form AE&CI had long enjoyed the benefits of well-stocked libraries as far as published information was concerned. The contents of these two libraries were combined when the AE&CI Research Department was established, and moved to that department. Both the new library and the newly established technical report files relied on the modified Kaiser catalogue referred to on the previous page, until the library changed over to the use of UDC in 1955; after this latter change only the reports - <u>ie</u> the internally produced unpublished documents - were included in the modified Kaiser catalogue.

As the technical report files grew, the percentage of information included which had originated from within AE&CI also grew significantly. Most of this new AE&CI-produced information represented the end product of considerable sums of money invested in research, and as the quantity grew so also did pressure from the staff of the research department increase for its efficient retrieval as and whenever required. It was clear that the modified Kaiser system in use was quite incapable of meeting requirements

---

[Footnote. The Kaiser system was one of the earliest subject indexing systems and was applied mainly within industry[19,20]. Abstracts were prepared but there was no formal control over the indexing vocabulary. The modification used within AE&CI comprised the use of amplified report titles only in place of what today would be called 'indicative abstracts'] for the retrieval of information from the reports in question because of inadequate indexing and abstracting, and a complete lack of indexing vocabulary control. As a direct result of the succession of staff employed on its compilation, and of their lack of technical and librarianship knowledge referred to above, the system was also inadequate for document retrieval purposes.

It must be amphasised that in industry there are stringent commercial security requirements for the contents of internally produced technical report files - ie free access to this information, essential in any conventional library, cannot be permitted. Thus access is restricted to those who 'need-to-know'.

1.3.3 Statement of the problem

In 1962 the author was instructed to investigate at first hand the development of IR systems in chemical and engineering concerns and other organisations overseas, and to put forward proposals which would meet AE&CI's particular problem. At that time the emergence overseas of this new discipline of information science was not understood or appreciated within AE&CI so that the terms information retrieval, information science and information transfer were regarded by management as being virtually synonymous (however extraordinary this may now seem to be). Unfortunately in many of the small companies in South Africa today the same lack of understanding still prevails. [Note: It must be remembered that more than half of the industrial effort in the Republic arises from very small companies when judged by overseas standards].

Essentially the problem was to provide a system which would meet the following requirements:

1) It had to be simple for the research staff to use on a self-help basis - ie research workers had to have direct access, on a 'need-to-know' basis, with a minimum of assistance from the information workers. There was a long tradition of self-help in the library services, and, on the grounds that the research worker should be able to browse through available information

at his leisure and convenience, the same requirement was laid down for the retrieval of information in technical reports.

2) The system adopted had to be comprehensive as regards the Company's current and projected technical interests - ie 'subjective' cover of a scientific or technological field was both preferred and required, in contrast to the 'objective' coverage of each discipline which is normal in university, public and other libraries.

3) High standards of consistency, and also of technical knowledge, on the part of the staff compiling and maintaining any system adopted were considered to be essential for full acceptance of the system by the potential users. This was in keeping with the general high levels of technical competence within the Company.

Chapter 2 INFORMATION TRANSFER AND INFORMATION USERS

In view of society's new-found concern over the uses of science and technology the study of 'information transfer' has become a study in its own right as a branch of sociology. The channels for information transfer are extremely varied, and amongst the more important may be listed (in random order):

1) Governmental economic and military foreign aid programmes.

2) Immigration/Emmigration and foreign student exchange programmes.

3) Staff transfers and secondments within multi-national corporations.

4) Technical meetings, symposia, conferences, etc.

5) Industrial exhibitions, trade fairs, etc.

6) Patents, product and process licensing, sale of 'know-how', etc.

7) Technical literature.

There is nothing mysterious about the factors which are conducive to good information transfer. They include the free flow of information, the encouragement of individual mobility and interpersonal contacts, rewards for seeking, sharing and utilizing new information (which includes the acceptance of and adaptation to change), rewards for risk taking, but above all, easy access to information by individuals.

As pointed out in Chapter 1 the major concern of this study is with the retrieval of information from documents and few would disagree that there is a real need for improved ease of access to knowledge in them. Although he did not in fact do so, T S Eliot could well have had the whole information transfer problem in mind when he wrote the couplet:

> 'Where is the wisdom we have lost in knowledge?
> Where is the knowledge we have lost in information?'

But in industry the need appears in short blunt questions - 'where is it? who has it? what is it? how can I get my hands on it?'

There have been numerous attempts overseas to assess the channels used for information transfer within the field of documentation, but there have been very few within South Africa itself - that by van Houten[21] in the field of electronics being a notable exception. Slater and Fisher, in a very useful report[22], recorded 6 300 acts of use in 104 libraries in the United Kingdon; in addition they list over 40 surveys carried out in that country during the 1950s and 1960s. [Note: Consideration was given to discussion of some of these surveys, particularly those of possible interest to small industrial concerns. However it is believed that this would not be relevant to the main theme of this document - ie information retrieval. The existence of these surveys should however be noted].

2.1 The information transfer process and information seeking behaviour of scientists and technologists in industry

Not only has the quantity of information available to users increased enormously but the spectrum of information relevant to any specific problem has also changed. A generation ago, say around 1930, most subject specialists were able to keep adequately abreast of knowledge in their chosen fields, even if they themselves saw only 70-80% of the published literature in that field. At the same time such specialists usually had the time to gain some knowledge at least in fields totally unrelated to their speciality. Today the same specialists can only keep abreast of knowledge in substantially narrower fields than those of a generation ago; in general many do not expect to today, and are usually quite unable to maintain and improve their knowledge in unrelated fields. They appreciate however that the potential relevance of information in other fields (ie information which could make an impact on their own work) is far greater than it used to be. This arises as a direct result of the well-known breakdown in the barriers between what used to be referred to as the exact scientific disciplines; the borders between disciplines overlap to such an extent that it is frequently impossible now to say where one ends and the next begins - eg chemistry and physics.

However, the matter is not as simple as it may first appear to be. There is now the paradoxical situation that while the various disciplines are themselves no longer as 'exact' as they were previously believed to be - ie the barriers between them have tended to become blurred - communications between many practitioners of these disciplines has become more difficult. This has been well summed-up by D J Foskett[23] as follows:

"...... the vast accumulation of knowledge over the last few centuries has meant that to make new discoveries scientists have to concentrate their interests and devote more intensive study to smaller and smaller fields. This ever-increasing specialization brings with it a growing body of new terms equally specialized, and the circle of colleagues with whom each scientist can communicate freely tends to become narrower. Anyone wishing to understand fully the nature and implications of the scientists' discoveries must first learn these new terms and their definitions, and owing to the vastness of the sum total of knowledge, the number who have the desire and the opportunity to do so for a new field of work constantly diminishes. The process was aptly illustrated by the Conference Lecture by Professor Andrade to the Royal Society Conference on Scientific Information in 1948:

> 'We no longer have men of learning writing for other men of learning, or men of science writing for other men of science, or disciples of exact science writing for other disciples of exact science, or physicists writing for other physicists, or students of optics writing for other students of optics, or spectroscopists writing for other spectroscopists, but rather infra-red spectroscopists writing for other infra-red spectroscopists.'

In the first place, therefore, scientists find difficulty in communicating with each other because the increasing specialization of their work brings linguistic barriers for those not working in precisely the same restricted field."

From this it is concluded that users of information require far greater assistance today from such people as indexers and information systems designers than a generation ago. And it follows that indexers and systems designers require far wider scientific and technological knowledge and, perhaps above all, the imagination to see the major implications, both actual and potential, of the knowledge which is to be processed. Unfortunately the implications are of necessity subjective in industry in that they are limited to those of value or interest to the particular concern. Thus it is clear that the valuable indexer in industry must also have a really good knowledge of the concern employing him, if he is to provide service of a high standard to users. Another way of making the same point is a statement accredited to Raymond Smith by D J Foskett[23] in which the former developed the thesis that all libraries exist to serve organizations, and that their function is simply to 'implement the policy of the organization by means of books and other recorded data'. For industry it is more appropriate to speak of 'information services' than simply of 'libraries', but in either case the statement is correct.

That the majority of information users do not welcome active assistance by information scientists or by technically orientated librarians has also been well established overseas by numerous studies, questionnaires, etc; the author's experience in South Africa confirms that this country differs not at all on this score. Coover[24] has pointed out during discussion of numerous surveys that most scientists require some training at least in literature searching, that they rarely receive it and yet they frequently express lack of confidence in the ability of information scientists or librarians to assist them effectively. Similar views were expressed in 1965 by Professor B H Flowers in a survey for the Great Britain Advisory Council on Scientific Policy[25], by Saul and Mary Herner[26,27], and by Rosenberg[28]. It is also evident that this lack of confidence extends beyond the fields of science and technology, even though the literature outside these fields is far less - eg [illegible]tberschaut[29] has reported a similar situation in social science research.

An important point of local interest has been high-lighted by van Houten[21] in that a high percentage of South African industrialists - and it must be remembered that these are small concerns by international standards - welcome any assistance which information scientists or librarians can give their staffs as a matter of principle. Clearly there is a local situation akin to a paradox in that management welcomes what their scientific and technological staffs are, in practice, unwilling to accept. Much has also been published on the question of information user needs and on the complementary aspect of information system objectives; two of the clearest are those by Carleton C Conrad of E I du Pont de Nemours & Co[30] and by David Liston, Jr. of the Battelle Memorial Institute[31]. [Note: A basic objective of all information services is to ensure the utilisation of all literature relevant to a given problem. Unfortunately this ideal is seldom achieved and comments by P J Judge appear to be apposite[32]: 'Current ...... emphasis tends to be on the management of documents; this is the more imperative as it appears that only a part of the mass of literature is frequently cited, although relevant material may be scattered far outside the expected journals. A great number of secondary publications and services has appeared, and to make the most effective use of these the user needs special training - linked also with salesmanship - which is seldom given].

Further discussion on the question of objectives is set out in Chapter 8, but attention is drawn at this stage to A N Nesmayanov's analogy that the work of a technical information service is akin to the first two stages in metal extraction - the first stage is the search for the appropriate ore and the second is its mining and concentration - the third stage of extraction being reserved for the scientist himself[33]. However at this point a clearer idea is required as to what is meant by user needs. For a proper appreciation it is proposed to consider user needs in the wider context of comprehensive information services initially, and only afterwards to look at them from the narrower aspect of information retrieval.

User needs are probably at their most varied and complex in an industrial research department where the staff is discipline orientated. In such environments it is apparent that actual needs are influenced by the individual researcher and his personal information seeking habits as determined, probably by his qualifications and experience. These needs are also influenced by the group of which the individual forms a part, by his level or status within the group (since the higher the level the less is the need, usually, to know the maximum of detail). A further factor, and one which appears to require research, is not a need but a practical problem which seems to be encountered widely, and on which no statistical evidence has been found - namely that the greater the distance of the individual researcher from his sources of information, then the greater the degree of undercommunication with these sources appears tobe. If their nature can be established for such a situation then adequate understanding of them should be easier in other fields of activity, such as production, where staff is mission (target) orientated. It is believed that for an effective information flow in a research environment there are five major and basic user needs, as follows (Note: it is virtually impossible to rank these needs in any meaningful order of increasing or decreasing importance and no attempt has been made to do so):

2.1.1 Users require specific information on a wide variety of topics.

This need includes the mass of technical data and information required by users from day to day. It can often be specified accurately by the users themselves for ready retrieval from IR systems, libraries, report files, patent files, etc.

2.1.2 Users require access to information on subjects which are new to them.

When an individual is required to extend his knowledge he requires appropriate textbooks, review articles, etc. In an ideal situation the information service should have the skills to be able to select those items best suited to the individual's educational standard, experience and intelligence and which are most relevant to the technical problem itself.

2.1.3 Users require current awareness and/or SDI services.

The individual research worker can no longer keep himself up-to-date by his own reading effort in all the fields of interest and importance to him. How extensive or limited these services are depends on a number of factors, not many of which can be measured quantitatively - they include:

1) the size of the research department
2) the spread or homogeneity of its research projects with regard to subjects and to scientific disciplines
3) the average size of project teams
4) the rate of technical advance in the subject fields involved
5) the commercial importance of the subject fields involved to each particular concern
6) the level of technical talent required or available for the research projects and also for the current awareness and/or SDI services themselves.

2.1.4 Users require information which will enable them to avoid the excessive duplication of research.

Presumably this problem arises when a specific research project is being considered and thus it is a conventional information retrieval problem. Experience has shown that close integration between research teams and information services tends to mitigate the problem.

2.1.5 Users require information which enhances the possibility of generating creative ideas.

The generation of creative ideas is the basic reason for the existence of research departments. There is an obvious need to provide an information flow which will maximise the probability of the occurrence of creative ideas that are potentially valuable to the particular industrial concern. To state that this is an essential requirement is a simple matter - to implement it is extremely difficult, and perhaps impossible.

After pointing out the major user needs of information services above it is appropriate to return to the narrower aspect of information retrieval following a request for assistance from a user. Quite apart from whether or not the actual information supplied in answer to a user's request is relevant, and also clearly pertinent to his particular technical problem, the fact remains that many users, probably the majority, have a deep-seated desire to interact with information retrieval systems directly. This means that they prefer to obtain information on a self-help basis, to be able to browse through the available information, and, in many cases, to contribute actively to the improvement of systems. This facility of being able to browse is extremely important and is dealt with at greater length in Chapter 3, Section 3.1.4, pages 41-44.

However, it must be stressed that active voluntary contribution to systems improvement is an entirely different matter to formal studies of user needs. Attention is drawn to the review article by Saul and Mary Herner[27] in which 38 such studies carried out between 1956 and 1966 are considered. The authors point out that when formal studies are carried out or planned it is important to realise that users are unlikely to cooperate willingly for long. This confirms the author's experience that in industry at least, many users show an actual or potential interest in the systems provided for them, and only need a little encouragement in order to express suggestions and criticism. On the other hand, experience has also shown that when a formal study is initiated the longer an interview with a user is, or the longer a questionnaire addressed to users may be, the less factual and worthwhile are many of the answers given to questions asked. It must be admitted of course that this experience may be due to lack of skill in handling such interviews or in drawing up questionnaires; at best, however, valid conclusions are not easily drawn. Certainly though, a few studies based on questionnaires have brought some valuable impossible.

After pointing out the major user needs of information services above it is appropriate to return to the narrower aspect of information retrieval following a request for assistance from a user. Quite apart from whether or not the actual information supplied in answer to a user's request is relevant, and also clearly pertinent to his particular technical problem, the fact remains that many users, probably the majority, have a deep-seated desire to interact with information retrieval systems directly. This means that they prefer to obtain information on a self-help basis, to be able to browse through the available information, and, in many cases, to contribute actively to the improvement of systems. This facility of being able to browse is extremely important and is dealt with at greater length in Chapter 3, Section 3.1.4, pages 41-44.

However, it must be stressed that active voluntary contribution to systems improvement is an entirely different matter to formal studies of user needs. Attention is drawn to the review article by Saul and Mary Herner[27] in which 38 such studies carried out between 1956 and 1966 are considered. The authors point out that when formal studies are carried out or planned it is important to realise that users are unlikely to cooperate willingly for long. This confirms the author's experience that in industry at least, many users show an actual or potential interest in the systems provided for them, and only need a little encouragement in order to express suggestions and criticism. On the other hand, experience has also shown that when a formal study is initiated the longer an interview with a user is, or the longer a questionnaire addressed to users may be, the less factual and worthwhile are many of the answers given to questions asked. It must be admitted of course that this experience may be due to lack of skill in handling such interviews or in drawing up questionnaires; at best, however, valid conclusions are not easily drawn. Certainly though, a few studies based on questionnaires have brought some valuable information on user needs to the fore, such as that of Rosenberg[28]. Rosenberg was able to conclude legitimately that "the ease of use of an information-gathering method is more important than the amount of information expected......"

In studying the literature and in visiting information systems in various overseas countries the author has been struck frequently by the complexity for the user of the systems provided in many industrial and other organisations. In many situations the user is denied the direct access to information that he clearly wants, so that he is forced to pose his questions to others who are usually not scientifically or technologically trained individuals. Only a relatively few writers have stressed adequately the extreme importance of providing easy access to information if system customers - ie the users - are to be satisfied. Rosenberg has done so[28] and Barnes[34] drew attention to the fact that surveys have shown clearly that while users want easy access they very frequently tend to seek references closely related to their basic training - which is understandable even though often illogical. It must also be remembered that to many technologists - more so than for scientists it would appear - ease of access to relevant information means, de facto, ease of access to colleagues, and to informal discussion with them, rather than to the literature. This point has been well made by both McLaughlin et al[35] and Allen[36]. The latter concludes that for engineers:

1) Accessibility is the single most important determinant of the overall extent to which an information channel is used.

2) Both accessibility and perceived technical quality influence the choice of first source.

3) Perception of accessibility is influenced by experience. The more experience an engineer has with a channel, the more accessible he perceives it to be.

4) The rate at which ideas are accepted or rejected is related to the perceived quality of the information provided by a channel. Engineers thus use technical quality as the criterion in a filtering process which compensates, in part, for the neglect of technical quality considerations when selecting an information channel.

Similarly, the vast literature on automated information storage and retrieval systems rarely stresses that users wanting information seldom get easy access to it since the end products of searches are usually document numbers - ie the user still has to consult original documents. This is a point which was brought out in a talk by Miss Maysie Webb on the National Library of Science and Invention. Miss Webb was reported as having stated[37]: ...... 'while computers can pour out lists of titles of articles, authors names and subject indexes, the reader wanting information will ultimately have to see the actual document. This will normally have to be obtained from a library of some kind.' (Note: Much of the literature of information science uses many words far too loosely: eg - in the United States the words 'automated' and 'computerised' appear to be synonymous as far as information systems are concerned).

At this point it is important to distinguish between 'documents' and 'information'. The information scientist, research worker, etc, is generally not interested in original documents at early stages of an investigation. Their interest lies in rapid access to information in documents and, in view of pressures on their time in the competitive environment of industry, any competent pre-digestion of available information, as abstracts of one kind or another, for example, is particularly welcome. This screening stage is invariably followed by study and digestion of those complete documents which are thought to be pertinent, and this in turn is followed by their evaluation. Evaluation may or may not be succeeded by experimental work, by market research, by design studies or numerous other possible activities. Whatever the specific activity, the end product is usually some kind of a report - oral or written - which itself comprises a condensation of the findings of the activity involved. It is on this condensed pre-digested information, whether technical, economic, commercial, etc, or perhaps most commonly, mixtures thereof, that senior management base many important decisions. Clearly the quality of the guidance provided will largely determine how close management will come to the optimum decisions.

Thus information users comprise persons at varying levels in industry right up to that of top management. It is important to understand this process of information transfer in order that it may be carried out as economically and as efficiently as possible. The process itself can be likened to a spiral staircase comprising:

1 - work done
2 - report on the work done (raw information)
3 - condensation of the information in 2 (abstract, summary, etc)
4 - evaluation of the foregoing together with other information
5 - further work done
6 - report on further work done (fresh raw information)
7 - condensation of the information in 6 to guide management in their decision making
8 - decisions by management
9 - plans for development based on decisions in 8
10 - new work done

Since no business can survive on inactivity this chain spirals onwards repeating itself ad infinitum from project to project, interlocking with other similar chains originating in the open literature, in competitive concerns, etc, to form a completely random pattern. If it were possible to observe the whole from a distance it would probably appear very similar to a modern random fibre textile fabric.

As pointed out in the Weinberg report of 1964 to the President of the United States[38] the transfer of information is an inseparable part of many activities. In the case of research and development the report states:

' - individual scientists and engineers, industrial and academic research establishments, technical societies, government agencies - must accept responsibility for the transfer of information in the same degree and spirit that they accept responsibility for research and development itself'.

Thus the question of the need for information for decision making, for evaluation prior to other work, etc, is not at issue. Unfortunately however, all information is not in itself equally valuable, relevant or useable so that it is virtually impossible to predict the actual or potential effect on users of any given piece of information, even when the environment is well known to an observer.

A major source of misunderstanding arises when information scientists maintain that the transfer of information is primarily concerned with written records, or alternatively that this transfer takes place before formal written records are compiled, or instead of such records. As examples of these points of view attention is drawn to the following contradictory statements by two well known American authors, namely J C Costello[39] and D J de Solla Price[40]. Both statements refer to the research environment:

1 Costello:

'In today's complex civilization with its rapidly expanding technologies, there are many (people) working in highly specialized fields. Generally they are individual workers, for all practical purposes completely isolated from others working in the same or related fields with whom they might wish to communicate directly, as in conversation. These individual workers are generating knowledge. <u>The only practical means of transmitting this knowledge from their minds to the minds of others is by static information, or written records</u>. Essentially all technical communications are created with no knowledge of the ultimate prospective consumer. The future recipients of such information may be widely separated from the authors by geography, time, language, and culture. Thus, communication in intellectual pursuits is usually not only static, but also indirect and delayed.

Such communications are generally written in books, journals, periodicals, or internal reports. Other media for information communication include design drawings, maps, drafts, charts, and tape or film records.'

2 de Solla Price:

'What is this special structure of science, and how has it moulded a social structure and communication system to serve its purposes? What is a scientific paper for? Why is it written? Why is it read? A few years ago we would have naively identified the scientific paper with scientific communication. Now we know, thanks to Menzel and to Garvey and Griffiths, that if we are speaking of the communication that is used by each research worker as an inspiration and as a data flow that makes his own work possible, then some 80 per cent of that input comes to him from other research workers at a stage before formal communication and through the informal channels of the grapevine, the conference, the seminar, and the other tentacles of the invisible college. Just as 80 per cent of communication lies outside the scientific paper, I would now like to suggest that 80 per cent of the value and function of the scientific paper lies outside the realm of communication. The paper and communication overlap each other by only 20 per cent. The scientific paper may not be written because somebody wants to read it, any more than a letter to the editor is written in the style of a letter of Uncle George.'

To be fair to Costello the work done by Menzel[41] and by Garvey and Griffiths[42] had not been reported at the time of Costello's publication. Nevertheless it has been recognised in the chemical industry for at least the last 30 years, and as far back as the 19th century in the engineering industry, that informal oral communication is extremely important in information transfer in research, manufacturing and other activities. Long before de Solla Price coined the phrase 'the invisible college' (see Footnote) the same hypothesis was being referred to in one very large chemical concern in the United Kingdom as their 'old-boy-network'. This colloquial description arose because of the Englishman's habit of visiting or telephoning a colleague who might be attached to his own concern, a competitive concern or a University and who might be able to help him with a problem. Invariably, he would open the conversation with 'I say old boy, do you know....' or some such.

Informal information transfer of this 'old-boy-network' kind is more 'open-ended' than its formal counterpart. Scientists and engineers interacting informally are willing to speculate about their work, to discuss their mistakes as well as their successes, and to range over a broad area of interests; many show a bias towards personal confrontation whether this is oral or in writing. In a more rigorous framework, such as in a technical report for example, speculations or discussions may not appear at all or may appear only as minor, unemphasized addenda to specific findings. Additionally, the flexibility of information channels allows the scientist to direct the communication process and to select for himself any specific information which he requires.

[Footnote. The 'invisible college' hypothesis has proved to be of interest to many workers in the information field if the number of times that is cited is any criterion. Although it is a rather vague and elusive concept it is one which many individuals appear to accept with what might be called an intuitive belief in its truth - or at least they accept it without the exercise of their critical and analytical faculties. It is encouraging therefore to note that several studies have made worthwhile contribution to the proper understanding of the concept(43-47). Classic examples of the 'invisible colleges' at work where important new knowledge was known to members well in advance of any formal written communication are:

1) there was fairly widespread awareness in scientific circles in the USA and UK through the international grapevine of the feasibility of uranium fission before World War II - this was a discovery made by Hahn and Strassmann in Hitler's Germany in 1938

2) there is an excellent account of the invisible college at work in the field of molecular biology in J D Watson's book 'The double helix; a personal account of the discovery of the structure of DNA' (48)].

Every researcher has some specific information needs which he may not always be able to express to, say, an information service, but he can usually recognize the information that satisfies those needs. Such needs, which change from time to time, are determined by the subject matter of the scientist's research, his mode of working, his attitudes toward communicating his own work, and the stage of his research. In the case of most formal channels it is not possible to shape information transfer (retrieval) systems to fit all the specific and immediate needs of each user.

It is all the more remarkable therefore that even today many information scientists and librarians appear to believe that the written record far outweighs the importance of oral communication. The truth of the matter is not straightforward since there are numerous variables, the importance of which are weighted to differing extents in differing circumstances.

The major factors appear to be as follows:

1) The individual's level within the organisation concerned (information system user, top management, etc).

2) The type of activity with which an individual is concerned (production, sales research, maintenance, etc).

3) The particular scientific disciplines or technologies involved.

4) The structure, size and complexity of the particular organisations involved.

5) The size and complexity of the particular industry of which 4 above forms a part.

6) The level of development of the national economy in which 5 above is situated.

These factors are set out in an ascending order of generalisation. As an illustration of the weighting of the factors listed, consider the following case histories as applied to industry in the Republic of South Africa - factors 1-6 in reverse order:

6) The national economy of the Republic is at a stage intermediate between full industrialisation and of the so-called 'developing' countries. Thus in the case of plastics materials it is currently able to maintain one manufacturer for each polymer only. These manufacturers are protected against dumping of excess foreign production by tariffs, but local production is supplemented by imported polymers.

5) The plastics industry comprises 25% of the total chemical industry in terms of the value of its products - in 1969 this amounted to R175 million, so that the total value of the products of the chemical industry in the Republic amounted to R700 million that year.

As pointed out by Grant[49] and van Houten[50] of the South African Council for Scientific and Industrial Research, the smaller industrial concerns are very dependent on their raw material and equipment suppliers both for technical service and technical information. For example, AE&CI Limited distribute a monthly bulletin to customers and other interested organisations comprising about 30 abstracts of new information mainly on the processing and applications of the plastic materials polyvinyl chloride (PVC) and polyethylene.

Between 100 and 200 journals are scanned regularly to find this information and it is then carefully selected so as to eliminate items on products unlikely to be made currently in the Republic. Over 100 copies are sent out each month and during 1969 approximately 360 abstracted items were distributed - 1 397 requests were received subsequently for additional information.

4) Requests for additional information from Company A and from Company B were traced back.

Company A is of medium size by South African standards and is concerned with the use of plastic materials for packaging.

Company B is of small size and extrudes plastic materials for the production of consumer goods.

3) The technical staffs of Companies A and B are primarily mechanical engineers who have specialised in plastics technology.

2) The specific activities of the individuals receiving the information distributed were:

Company A : Production and product development
Company B : Production

1) The positions held by the above individuals were:

Company A : Production Manager
Company B : Production/maintenance engineer.

In these case histories there is no major difference between factors 1-3 as far as Companies A and B are concerned. Nevertheless the differences in handling the information communicated to them are of interest:

Company A. The Production Manager receives the bulletin referred to and circulates it to about six persons. All requests for additional information (complete journal articles, etc) are collected and when, in due course, these are received, the Manager's Secretary distributes them as appropriate. The manager relies on his staff to draw his attention to ideas and useful information obtained in this way. On a few occasions meetings of two or more staff have been held to discuss aspects of the information, and action has been taken on it once or twice; in all cases the information is indexed and filed in the Company's small library.

Company B. The bulletin is read only by the production/maintenance engineer. He occasionally requests additional information and when he has found this to be of use he has passed the documents concerned on to his Manager together with oral or written notes. No formal and consistent attempt is made to retain these documents - a few have found their way into correspondence files and the rest are in desk drawers or have been mislaid.

The basic point to note is that in the production environment in smaller companies in South Africa de Solla Price's contentions concerning the transfer of information would appear to apply and to be nearer the truth than Costello's. Attempts to trace back case histories in the sales environment have produced little information of value, probably since technical information is incidental to the main aspect of the activity - ie selling. However, the findings of Grant and van Houten[49,50] do not differ significantly from findings overseas and it appears to be reasonably safe to conclude that local industry has few, if any, unique features in this respect.

In conclusion it must be emphasised that there are no major basic differences between the IR operations carried out by small, medium or large enterprises. This point has been emphasised by B E Holm[51]; any differences are a matter of scale and not of principle.

2.2 Company and user needs within AE&CI Limited

The position as regards the retrieval of technical information up to 1963 has been set out in sub-section 1.3.2. The situation as far as the published literature in the library was concerned appeared to be adequate but by 1961/62 the management of the Research Department had accepted that the modified Kaiser index for internal technical reports must be replaced.

The question was 'by what sort of system should it be replaced?'. Little guidance was available from AE&CI's experience with the Kaiser system since:

1) it had not been thoroughly and consistently applied to the reports available to it,

2) many reports which should have been routed through the existing system never reached it, and

3) for practical purposes the system had broken down - new reports were being filed under accession numbers allocated by the reports file staff, even though they were being received with the originators' identification, which was usually numerical. Once filed many of these reports could not be found.

It must be emphasised that at one time or another virtually all chemical and engineering companies of any size have outgrown their long standing systems for information and/or document retrieval. This happens usually as a result of rapid expansion in size and diversification of interests, thus AE&CI, which experienced considerable growth after World War II, was no exception. And as with many such concerns, when considering what IR system to adopt there was at that time little formal investigation of user needs as discussed in section 2.1. Nevertheless the user needs which were identified by management conformed exactly with the points set out in sub-sections 2.1.1 to 2.1.5, and there was particular concern to ensure that research already carried out within AE&CI and its overseas associates would not be duplicated. At the same time it was required that the system adopted would permit the user to browse and would also meet the Company's long-standing preference for 'self-help' systems.

It was clear at that time, 1962, that the main users of the new system would be (in descending order of user frequency):

1) Research workers,
2) Technically qualified staff in other technically orientated departments (*ie* the Chief Engineer's and Development Departments,
3) Technically qualified staff in sales and technical service-to-customers activities,
4) Production staff concerned with trouble-shooting and factory laboratory staff.

The total of potential system users was approximately 500 technical graduates and subsequent experience has shown that the expected order of user frequency was correct. During the five years 1965 to 1969 the percentages for user frequency were:

1)	Research staff	50%
2)	Technical staff from other technical departments	20%
3)	Technical sales and service staff	15%
4)	Production staff	10%
5)	Others (service departments, managers, associates)	5%
		100%

Some effects of the development of an integrated retrieval system on the information seeking behaviour of different categories of individuals are set out and discussed in Chapter 6.

Chapter 3 CRITERIA FOR THE DESIGN OF INFORMATION STORAGE AND RETRIEVAL SYSTEMS

3.1 Discussion on criteria for IR systems design

Considerable interest was shown during the mid-1960s in the testing and evaluation of IR systems, since it was believed that this would improve knowledge on criteria for the design of systems. More recently this interest appears to have slackened and the total number of papers published has fallen steadily year by year since 1966. The reasons for this are probably that while the best have materially assisted in reducing areas of ignorance, useful findings which can be applied in practical situations have not been numerous. This probably stems from a combination of factors:

1) Oversimplification of the problems involved.
2) Deficiencies in experimental design (eg vide the Aslib-Cranfield Research Project referred to on page 37)
3) Difficulties in undertaking further work under more stringent experimental design requirements.
4) Investigators appear to have exhausted their repertoire of ideas at least temporarily.

If anything is clear from the literature, it is that further insight is required into what to test and how to test it. Whether in fact suitable and satisfactory procedures will be developed in this very difficult matter remains to be seen. As far as this author has been able to determine, nowhere in the literature is there a report verifying anyone else's test results, or developing them further, such as making them more precise.

An obvious possible reason for this is that no two systems are ever identical, and yet the possibilities for major basic differences in IR systems appear to be rather limited. This fact is made clear, for example, by J R Sharp in his survey of IR techniques for Aslib[52]. Thus there are good reasons to expect that suitable and valid tests will gradually come to yield more useful results. [Note: The results of tests on the AE&CI systems are set out in Chapter 6].

However, in this author's opinion far too little has been done and written on the question of the criteria of value in the design of IR systems and adequate emphasis on the fact that systems exist primarily to assist would-be users is lacking. This is the reason for inclusion of this Section in this dissertation.

From a practical point of view within industry, the initial criterion is invariably one laid down by top management after they have been convinced that a system is needed. This is, that an IR system must meet its objective of providing information for decision making purposes at minimum cost compatible with effectiveness. In fact this is usually the overriding criterion, but further discussion is deferred to Chapter 7, in which the economics of information services are considered. However, apart from management's concern with costs, the other essential viewpoints from which an IR system can be evaluated are those of the system designer and of the users. The criteria that the system designer uses are dependent on the one hand on top management's policy (which ultimately determines the intellectual techniques adopted to achieve the objectives, the hardware selected, etc, and so the actual costs). On the other hand, the designer is constrained partly by the size of the collection of documents from which information is to be retrieved, partly by the degree of centralization and geographical scatter of the company concerned and partly by user requirements.

The size of the collection of documents and its expected rate of growth is closely related to the type of IR system eventually adopted in practice and to what is frequently referred to as 'its level of sophistication'. This aspect is discussed in Chapter 4, but it is also tied up with the degree of centralization and geographical scatter of a company. A high degree of centralization and no geographical scatter means that one large IR system will probably be provided, whereas a high degree of decentralization and wide geographical scatter will probably result in the setting up of a number of small independent IR systems.

However, it is in the field of user requirements that, within the limitations imposed by top management policy, the most important criteria are probably to be found. If user requirements are not met, an IR system will not be used, and hence its objectives will not be met. The return on the capital invested in it (and sanctioned by top management) will be inadequate, thus making the project uneconomic and, in short, a failure.

What then are these user requirements? Attention is drawn to some of the numerous lists of requirements which have been published(31,32,39,53-66), but none is as potentially succinct as that of Cleverdon, Mills and Keen(67). They state:

> '.... user criteria .... [are] those factors which are of concern to the users of a system. Such criteria are related to the operational performance of the system and can be listed as follows:
>
> 1) The ability of the system to present all relevant documents (ie recall).
> 2) The ability of the system to withhold non-relevant documents (ie precision).
>
> See Footnote
>
> 3) The interval between the demand being made and the answer being given (ie time).
> 4) The physical form of the output (ie presentation).
> 5) The effort, intellectual or physical, demanded of the user (ie effort).'

[Footnote. The concepts of 'recall' and 'relevance', or 'precision' as the latter is now more commonly known - ie criteria 1 and 2 - were first expressed by Cleverdon and his colleagues in the initial stages of the well known Aslib-Cranfield Research Project (68-70). The initial stage of this project comprised the testing and analysis of an investigation into the comparative efficiency of four indexing systems, namely UDC, Alphabetical, Facet, and Uniterm. The work was carried out at the College of Aeronautics, Cranfield, England and 18 000 documents in the field of aeronautics were indexed. The findings were certainly controversial in some respects and aspects of the design of the experiment were criticised at length by numerous writers (71-81). Cleverdon, on behalf of his colleagues as well as himself, has succeeded in answering reasonably some of the criticism directed against them (67, 82). But whatever the pros and cons of these arguments may be, the twin criteria of 'recall' and 'precision' are now firmly established in the literature of Information Science. Unfortunately the jargon of the Science is still somewhat in a state of confusion - so much so that Charles P Bourne (83) was able to compile a list of 31 terms used in the literature between 1960 and 1966, all of which relate to these two concepts. It is with a sense of relief that many information workers have noted the decrease in use of the term 'relevance' and the even worse term 'relative relevance'. When preparing a list of search terms for a specific search of an IR system such workers know only too well the dangers of including terms 'relevant' to the subject as opposed to being 'precisely' concerned with it].

For purposes of emphasis, clarity and brevity the explanatory word at the end of each of these items has been taken as the main heading, and the headings in the above list have been used for explanatory purposes.

It has been emphasised already in this dissertation (see page 24) that it is extremely important to distinguish between documents and the information in them. It may of course be argued that this is implicit in criteria 1 and 2 above, and that criterion 4 would in any event cover the presentation of information as, say, abstracts. Nevertheless the importance of user preference for information, at least initially in the course of a search, as opposed to the actual documents, is re-iterated. This preference is undoubtedly of direct relevance to criterion 5 above, since browsing through abstracts frequently reduces the effort demanded of the user. Thus to meet the point raised here, Cleverdon and his colleagues would have been nearer to the truth if they had used the phrases 'all relevant information' and 'non-relevant information' with which to conclude the wording of criteria 1 and 2 respectively.

Analysis of the manner in which information retrieved is to be used has been held by Perry[84] to be of considerable importance when designing an IR system. Theoretically any information retrieved is, per se, independent of any uses to which it may be put, but in practice the connecting link between this information and possible uses for it is human thought, conjecture, inspiration, and so on. It is not surprising, therefore, as Perry points out[84] that:

> 'Such analysis (of the manner in which information retrieved is used) may turn out to be surprisingly complex. For example, it is shortsighted to regard the only, or even the most important, use of the literature in scientific research and technical development to be the avoidance of useless repetition of previous experiments and observations. Much time and effort can be wasted in repeating previous research or in reinventing previously patented products and processes. But even more important benefits may be achieved by using information in the literature during

the various stages of planning, conducting and evaluating research, in particular'.

With the foregoing provisos Cleverdon's et al. list of user requirements is considered in detail below; the conclusions drawn are set out in sub-section 3.1.6.

3.1.1 Recall (ie the ability of a system to present all relevant information)

The user of an IR system is primarily concerned with obtaining items of information that contribute to answering a specific subject request. Such subject requests between them can obviously range over the full spectrum of information covered by a given IR system, and the user is fully ··titled to expect that the retrieval for each of his requests ` in fact uncover all the relevant items of information in each instance. This ability of an IR system is known as the 'recall' power of the system and this can be expressed quantitatively by means of a simple ratio, known as the 'recall ratio'. This is usually defined in the literature as $\frac{R \times 100}{C}$, where C is the total number of items of information in a system that are established as being relevant to a specific subject request, and R is the number of these relevant items of information that are retrieved during a search of the system. For example, if there are 20 items of information relevant to a search and abstracted and indexed into a system and if, as a result of a search, 18 of these are uncovered, then the recall ratio for that particular search is said to be $\frac{18 \times 100}{20} = 90\%$. Probably the most important single attribute of an IR system as far as the user is concerned is this recall power, but in fact any figures of this nature are meaningless when considered alone. It is always possible to obtain 100% recall in any search by by-passing the system and scanning all the items of information in it. Further, unless the manager of a system can guarantee that there has been 100% consistency in the indexing of every i·· of information in that system, then comparison between searches is of doubtful value. And, of course, as long as the input to a system requires human intellectual effort, 100% consistency is never obtained. Nevertheless in combination with criterion 2, recall ratio has its uses, as is indicated in sub-section 3.1.6.

3.1.2 Precision (ie the ability of the system to withhold non-relevant information)

An index is basically a filtering device. Its purpose is to reduce the number of items of information that need to be studied as complete documents when a subject request arises and a search is carried out. Since the recall ratio is a measure of the ability of the filter to let through wanted documents, it can be argued that a complementary ratio can be defined as a measure of the ability of the filter to hold back unwanted documents. In practice however it has been found preferable to define precision as the percentage of relevant information in the total information recalled; this 'precision' ratio is defined in the literature as $\frac{R \times 100}{L}$. As in the recall ratio R is the number of relevant items of information retrieved in a search of a system, while L is the total number of items of information retrieved in that search. Reverting to the example given under recall ratio, if in retrieving 18 out of 20 relevant items of information (90% recall) a total of 50 items were retrieved (looked at), then the precision ratio is said to be $\frac{18 \times 100}{50} = 36\%$.

Again, as in the case of recall ratio, figures for precision ratio mean little by themselves. It is claimed that the joint use of recall and precision ratios does appear to afford a meaningful way of gauging the operating efficiency of an IR system and this may well be so. That they may be of value in determining indexing efficacy and hence operating efficiency is apparent from the analysis of tests described in Chapter 6. The actual numerical values obtained for a particular system apply only to that system of course and of this fact there is no doubt. However the conditions and environment of different experiments themselves differ considerably so that it is extremely difficult to determine how worthwhile and meaningful direct comparisons are, as has been clearly shown by both Cuadra[85] and Rees[86-87].

3.1.3 Time (ie the interval between the demand being made and the answer being given)

Generally speaking user response to the time factor is dictated primarily by the urgency, or lack of urgency connected with requests for information. As examples of the two extremes consider a) the doctor who requires information in order to prescribe the correct antidote in a poisoning case where the poison is known, and b) the chemical engineer who requires background information on existing viable processes for the production of a particular chemical. The former needs immediate information with 100% precision while the latter may well be satisfied with 100% recall within a period of, say, three months, or even with a number of key documents such as review articles, within one month and representing perhaps 80% recall. Thus the system designer is of necessity compelled to know and understand the type of request likely to be addressed to an IR system, as pointed out on page 38, together with the reasons for which the information retrieved is likely to be required. If he does not, his system will probably fail, as many have done. For example, in the case of a system concerned with poisons and antidotes it would be foolish to set up a mechanised system that was operable only during local normal office hours; this may appear to be a statement of the obvious yet such foolishness is not unknown.

3.1.4 Presentation (ie the physical form of the output)

Logic is frequently not applied as it should be in selecting the preferred form of presentation of the output from an IR system. This is particularly true in the industrial environment. Too frequently management has somewhat reluctantly agreed that an IR system must be set up for their staff, and have insisted on the constraint that it must be as inexpensive as possible to establish and maintain. It is not unreasonable to expect some qualification such as 'as inexpensive as possible to establish and maintain as may be compatible with user requirements'; frequently such qualification is not forthcoming. Alternatively having committed their concern to investment in a computer many managements overseas have succumbed to arguments such as: 'now that we have it and have spare time on it, why not let us have a computerised IR system?' Again, user requirements are frequently not considered adequately, perhaps because of a need to justify the investment in a computer or because of enthusiasm for its potential application.

In either of the two extremes described above, both of which are far too common, the output from IR systems is nearly always in the form of document numbers only. This arises in the former case because this is the obvious way to keep the system input costs to the minimum - ie the intellectual effort necessary for indexing, and hence time and the cost of salaries, is as low as it can be. In the latter case, not only does the computer lend itself to the matching of numbers corresponding to documents and index terms, but the total costs of operating a computer, many of which are hidden costs, are so high that it becomes very difficult to justify any other presentation of output.

Thus numerical output may comprise sequential document numbers or the 'code' numbers associated with conventional library classification systems (either numeric or alpha-numeric). (Note: In some circumstances and in order to improve the security of confidential information, the output may be in the form of 'idiot' numbers, which require a further matching in order to obtain the actual document numbers). The main factor overlooked in such systems is that while recall is probably high, precision is not, even granted consistent indexing. There is no 'filter' between the system and the user to eliminate information irrelevant to a specific subject request, and the user has to be his own filter. The intellectual effort required for the perusal of the documents is time consuming and as users are nearly always the relatively highly paid staff, the output costs for such systems are invariably high.

The extraordinarily rapid advances made in computer technology make it very difficult to predict for how much longer this state of affairs is likely to continue. The recent development of large-scale integrated circuits coupled with miniaturisation trends holds out the prospect of very low cost computers having substantial 'memory' capacity - a basic requirement for computerised IR systems - becoming available well before 1980. Facilities for low cost experimentation will undoubtedly result in a tremendous upsurge in the use of computerised IR systems; this subject is dealt with further in Chapter 9.

Paradoxically it is usually the very large industrial concerns that practice the extremes referred to above. It is they who are usually faced with quantitatively big information retrieval problems, the cost of solving which may frighten management, so that the problems are not always examined as thoroughly as they ought to be. It is also only large concerns that have computers with sufficiently large memory units to enable this type of IR system to be set up.

But what of the smaller concerns such as make up the bulk of South African industry? Firstly, the very small concerns are unlikely to have any formal IR system and to depend on equipment and raw material suppliers for much of their technical information needs. Secondly, experience shows that in medium sized concerns when a system is established the foregoing arguments are more likely to be appreciated, for reasons that are not evident. The result is that a number of systems are currently boing developed in the Republic where the output is in the form of expanded document titles, fuller indicative abstracts or informative abstracts. The implications are that although the input costs to such systems are higher, since greater initial intellectual effort is needed, the output costs are lower and user satisfaction will be greater. The oldest such system in the Republic appears to be that of AE&CI Limited which is described at length in Chapter 5.

The importance of presentation which permits browsing is considerable, and it is to be expected that it is a factor that all users of IR systems will rate highly. Browsing is a natural activity for man that begins in early childhood and continues throughout life. People are accustomed to browsing in books, magazines, newspapers, etc, and while such activity may often be random, when specific information is being sought it is necessary for individuals to filter out what is required (or what can be eliminated) and what is most probably relevant (or irrelevant); this has been well discussed by Morse[88]. The reason is that no matter how carefully an IR system is designed, how consistent the indexing of information fed into it may be, or how carefully the user phrases his questions, there is always an element of subjective judgement. To try and avoid completely such subjective judgement and the stimulation of ideas by the use of imagination is to emasculate, or at least devitalize an IR system for many users. Unfortunately many modern IR systems are designed on the false hypothesis that there is no difference between 'data' and 'information'. The factual material that comprises data can be numbered, stored and retrieved whenever it is required without any great difficulty. Subject information is far more difficult to identify positively so that there is no overlap between concepts. This becomes clear when structuring a thesaurus and deciding where the borderline lies between synonyms and near-synonyms, and this of course is subjective judgement again. Thus it is highly desirable to design IR systems more imaginatively than data retrieval systems.

3.1.5 Effort (*ie* the effort, intellectual or physical, demanded of the user)

In 1949, G K Zipf proposed a hypothesis, which in the USA at least, has become known as Zipf's Law[89]. He stated that individuals, when choosing between several possible paths all leading to the same goal, base their decision upon the single criterion of 'least average rate of probable work'. This, of course, is simply a formalised way of emphasizing that people are naturally lazy. When faced with a problem, each individual will instinctively seek the most accessible path for solving it - *ie* the path that requires the minimum immediate effort. That this is equally true in the field of information science has been shown by Mentzel *et al*[90] and also by Allen[91], and the implications are extremely important. Improving the subject coverage of an IR system, or of a library for instance, will not of itself lead to increased usage of these services. Unless the information in report files, library holdings, etc is easily accessible the money invested in the selection, purchase and/or storage of these documents will probably be largely wasted. Scientists and engineers will not be attracted to systems, libraries, etc simply by improvements in the quantity and quality of material contained there.

In a sense, these same systems and libraries, must go to the users by ensuring that information is as accessible as possible to those who need it, when they need it.

The perception of accessibility by the user - *ie* that one path will lead to the goal more easily than another, appears to be entirely a matter of experience. Excessive zeal on the part of the system designer to 'sell' his ideas to users is frequently met by indifference. But let the user try out a system for himself and realise that a) untrained in its intricacies as he is, even he can use it successfully and that b) little intellectual or physical effort is required, then he will subconsciously weigh this against other less satisfactory experiences. The moral is clear; in designing an IR system, aim to present inf rmation to the user in such a way that with minimum effort he *can* understand any information retrieved - don't present it so that if he will only make an effort, then he *ought* to be able to understand it.

3.1.6 Conclusions

The conclusions to be drawn from the foregoing are as follows:

3.1.6.1 It has been pointed out in Chapter 2 (page 22) that users have a deep-seated desire to interact with IR systems directly, meaning that they prefer to obtain information on a self-help basis. This implies the need by users to realise that they cannot both interact with a system and yet make no intellectual effort whatsoever to obtain the answer to a problem. Once this is understood then the student of information science can appreciate that the value of recall and precision lies more in the testing of existing systems, with a view to improving them, than as criteria in the design of new systems. Lancaster[92] has pointed out that systems can be tested (each within its own particular environment and not for purposes of comparison) to determine those techniques which produce a high precision ratio with minimum sacrifice of other user requirements, such as recall, time, presentation, and effort. In this instance Lancaster's findings and opinions as above are fully acceptable, but unfortunately elsewhere[93] he has stated:

> '......... we must obviously base our recall and precision figures on the raw results delivered by the index and not upon these results after some human intervention, such as a "weeding" or "screening" operation (ie a second-level filtering) by a human operator.'

As yet no one has successfully avoided intervention by human operators.

Cleverdon has correctly pointed out[68] that assuming consistency in indexing, as recall improves quantitatively in a given search of a system, so does precision tend to decrease. Surely, however, this is simply emphasizing that the user cannot have things both ways? If he phrases a wide question when he should be more specific, then he will find that the IR system he is using has a high recall ratio but a low precision ratio. If he phrases a narrow highly specific question to the same system when he ought to be more generic he will find that the system now has a high precision ratio and a low recall ratio. In other words his risk of missing useful information is much greater.

In any event too much depends on each individual's subjective judgement of what is relevant to a particular information problem, and what is not. It would be unreasonable to believe that the information in documents has a fixed relevance to an expressed problem for all users at all times. Records 'inform' different people in different ways and varying judgements are applied to the same information, depending on the different purposes for which it is required.

3.1.6.2 Such factors as time, presentation and effort all interlock. The way in which they do so is entirely dependent on the requirements for any given IR system in any given industrial concern, which, as has been pointed out already, means in any one of numerous possible unique environments.

The common thread is the matter of the accessibility of information in records, and in system design this appears to be one of the most important criteria. The importance of the time within which information becomes accessible after a subject request has been formulated is dependent on the value attached to the information sought by the particular concern involved, as well as on the actual mechanics of a system (hardware, etc). In general the value attached to the information sought is far more important than the mechanics of a system in industrial situations, although this is not at all obvious from studies of the literature. The form of presentation of the output of an IR system largely determines whether it is immediately accessible to the user or not - eg for purposes of filtering, a document number does not give a user immediate access to information whereas an informative abstract does do so. The psychological importance of this on users can hardly be over-stressed and is clearly directly concerned with user effort also.

With the proviso that users must make an initial effort to formulate the right question for any particular subject request addressed to any IR system, it is axiomatic (as pointed out in sub-section 3.1.5) that users will always make the minimum effort necessary to obtain access to the desired information. For any IR system to achieve optimum effectiveness - ie to justify the money invested in it by providing information when needed, and thus to be used - it should preferably be usable on a self-help basis. It must provide direct access to information (as opposed to document numbers or references only) and sufficient information for the user to undertake with confidence filtering of the information presented - ie it must provide browsability and adequate partial substitution for original documents.

To summarise then, there are several simple basic criteria to be taken into account when designing an IR system. In essence they are little more than formalised common sense; but so are all the present day industrial management systems. The effort to formalise these criteria, however simple they may be, appears to be the most effective way of emphasizing their existence. This is an exact parallel with management systems.

3.2 Criteria for an AE&CI system

At the time when a new IR system was being considered for AE&CI - 1962 - such criteria as 'recall' and 'precision' had just been reported in the literature for the first time, and were still hardly known in the Republic. Nevertheless the objectives of and criteria for the proposed system were reasonably clear[94]; generally speaking the criteria were the same as would be considered important today and which have been discussed in the foregoing sections of this Chapter.

The objectives were, firstly, that technical information already within the Company and still of value - that is primarily but not exclusively unpublished information - should be retrievable at an economic cost. Secondly, all useful technical information generated within the Company or received from its business associates after the system became operative should automatically find its way into the system - again, at an economic cost.

The specified criteria were that the retrieval process should be simple and quick and that the system should be convenient to use on a self-help basis. This matter of 'self-help' was in fact traditional within the Company and probably arose because of the relative scarcity of sources of technical information within South Africa during the first half of this century. Certainly it was accepted, almost without question, that users are generally much more interested in getting the information which they are seeking than they are in learning about extended use of a system.

Certainly it has to be admitted that there was no systematic approach to the determination of user needs and, with hindsight, it was unfortunate that these were not surveyed formally. While the system has been shown to be meeting current user needs (see Chapter 6) the lack of statistical data which pre-dated the system ad[illegible] has made it very difficult to draw valid conclusions on its true effects on Company staff in their information seeking habits.

One criterion taken into account and which is one that all industrial concerns must consider, concerned the sources of information to be used for the system. (Note: This is a subject on which the literature of information science offers little guidance, although that of librarianship most certainly does). This criterion actually comprised a constraint on the proposed Company system in that it was decided not to include, and thus duplicate the availability of, published information picked up by the commercial abstracting and indexing services. Examples of such services, of which there are about 1 800 covering all the sciences and technologies, are Chemical Abstracts, Biological Abstracts and Engineering Index.

On the one hand no industrial concern can cover the published literature as completely as a commercial service can do so at an acceptable cost level - eg Chemical Abstracts service collects its material from over 12 000 periodicals plus many tens of thousands of patent specifications. On the other hand, many industrial concerns were very critical in the early 1960s of the tardiness with which many abstracts were published by these commercial services. For instance, in the case of Chemical Abstracts many items were included from three to twelve months after publication of the original article, etc. For some of the smaller services this delay was anything up to two years, and as a result, some of the larger chemical and engineering companies overseas were carrying out much abstracting and indexing of this published material themselves, on the grounds that such delays were not acceptable in highly competitive environments.

As the level of development of industry in the Republic a decade ago had not reached the levels of complexity of much of that in the bigger and more industrialised countries, it was submitted at that time that adequate coverage of the published literature could be obtained by AE&CI by subscribing to many of the commercial abstracting services covering subjects relevant to its technical interests. At the same time, it was concluded that in its capacity as a customer of such services, the Company should put whatever pressure it could on these services to improve (which is a proper activity for the purchaser of any kind of service). In fact, similar pressure from industry in the USA in particular has resulted in substantial improvements in many of them. For example, the majority of abstracts published in Chemical Abstracts in 1970 have appeared five to ten weeks after publication of the original article, etc.

Thus the above constraint has been fully justified as far as published information in technical journals, etc is concerned. However it was appreciated that certain information is not picked up by the commercial services wherever it is published or is only published locally and yet is of importance to South African industry. Such sources of information include manufacturers pamphlets from all countries, many of which include useful technical data, journals published in Southern Africa, South African Government Gazettes and Regulations etc. Thus it was proposed that useful technical information, statutory requirements, etc published by such sources would be included in the Company system - in addition to the main sources of information, which comprised the raison d'être for the system - ie unpublished technical reports produced by the Company and its business associates.

Chapter 4 FACTORS AFFECTING THE PERFORMANCE OF IR SYSTEMS

4.1 Discussion on factors

Before technical information can be retrieved, a complex series of operations must take place:

1 Information must be recorded in documents.
2 Documents must be stored in some known and accessible place.
3 Sufficient bibliographic information must be recorded so that this serves to identify each document.
4 Characteristics of the information in each document must be classified or indexed in some preferred way, and may also be abstracted.
5 For practical purposes, limitations and control of classification or indexing must be imposed on the system.
6 The potential user must formulate a question expressed in terms of either bibliographic information relevant to the documents, or of characteristics of the information in the documents.
7 If the question posed in step 6 has been correctly phrased, then a search of the system should yield all appropriate documents or information (or both) that are in the system.

The factors which affect the performance of an IR system are contained in steps 3 to 5 inclusive and in step 7. However, step 3 is a simple matter and in the case of published documents will probably comprise conventional bibliographic information as used in libraries. In the case of unpublished documents, such as internally produced technical reports (frequently referred to today as 'in-house' publications) identification is usually limited to a sequential number, the author's name, issuing authority or department, and date of issue. This step is, therefore, not considered further.

It is apposite to refer again to the difference between document and information retrieval systems. Any system based entirely on bibliographic information as in step 3 above is a document retrieval system. Closely akin to this are systems designed for the retrieval of data only. A system based primarily on step 4 above, which of necessity also includes bibliographic information for the purposes of document identification, is an information retrieval system.

There are five major factors that affect IR system performance namely:

1	Abstracting	
2	Indexing	Step 4
3	Sub-indexing devices	
4	Indexing vocabulary (Thesaurus) control	Step 5
5	Hardware used	Step 7

Factor 5 is complicated by the fact that the hardware used is particularly involved with user requirements. Nevertheless it appears to be more appropriate to deal with this matter in this Chapter rather than in Chapter 3.

Before discussing the above factors in detail, it is desirable to consider why it is that information science has virtually rejected the traditional classification approach for IR systems, and has adopted alphabetical subject indexing almost exclusively.

4.1.1 Alphabetical subject indexing versus classification

Few subjects are more controversial between information scientists and librarians than that of classification.

The history of classification is long. It appears to have its roots in Aristoteiian logic and can be traced back to an initial application by the Greek philosopher Porphyry. The value of classification for the natural sciences was first appreciated in the eighteenth century. Linnaeus in botany and Haller in physiology realised that the grouping of objects into classes enabled the characteristic features of each class to be ascertained. Such systems are still being developed, for chemical notation for example, and they are extremely valuable. However, their value lies in their application to the organisation of 'systematized facts', or data, rather than to the less precise concepts of 'technical information'.

The requirements of IR systems in industry are, as has already been stressed, that they must operate successfully in unique environments where information is rarely indexed or requested objectively - it is the subjective viewpoint of the particular company concerned that is important. As far as the information scientist is concerned, classification is a device which enables documents (books, etc) on like subjects to be grouped together on library shelves. For this purpose it is very useful in facilitating browsing; no doubt many librarians will regard such a viewpoint as being extremely superficial, although they themselves are not unanimous in this as has been well aired by Luther H Evans [95]. Barbara Kyle, for many years librarian of Aslib, pointed out [96] that the 'catalogues of special libraries are primarily finding tools, and not repositories of exemplary cataloguing'. (Note: The literature of classification is vast and its history is an honourable one. For those who wish to study it D J Foskett has set out a most concise yet comprehensive summary in Aslib's Handbook of Special Librarianship and Information Work [52] while a book by R F Kennedy[97], late City Librarian of Johannesburg, and W C Berwick Sayer's manual[98] both cover the subject extremely well.

The information scientists' objections to classification systems are fundamental. They can be summarized as follows:

1) All classification systems involve the use of a coding system, which may be numerical, alpha-numeric or alphabetical. In all instances these codes do not provide the direct access to information which is a user requirement and which has already been stressed at length in this dissertation.

2) Classification systems provide logical means for the hierarchical organisation of knowledge, subject by subject. Such organisation automatically involves the use of a framework, which in turn imposes a rigidity and lack of flexibility. It can be argued, and correctly, that the process of structuring a thesaurus, list of key-words (call it by whichever term is preferred) is also a form of classification. However, the presentation remains completely alphabetical and a structured thesaurus is as open-ended as an unstructured one. This means that new terms can be added and unwanted terms be deleted as required. The disadvantages and delays in the preparation of revised internationally available tables for such classification systems as UDC are too well known to need stress here (see example at end of this sub-section on page 55).

Similarly the rigidity that the use of such tables implies is of necessity based on the assumption that there are barriers between scientific and technological disciplines which make each a self-contained entity. Again, it has been stressed already that such barriers have in fact been broken down to a substantial extent, and as scientific research continues, more and more interrelationships are being discovered. Thus in attempting to use a conventional classification system one item of information can logically go into one of several classes depending on the judgement of the classifier concerned. If, in order to ensure adequate coverage, the information _is_ put into several classes (the so-called analytical entries of library systems), then the total length of the code (_ie_ of the class and sub-class numbers, etc) is both confusing and totally unacceptable to the scientist or technologist user. The reason for this is that the minds of people naturally work 'alphabetically' (_ie_ by subject name) rather than 'systematically' (_ie_ logical class numbers); this is a scientific fact the importance of which, in the self-help environment of industry, cannot be gainsaid.

Dewey and UDC are nineteenth century classification systems drawn up at a time when the search for new knowledge was far less concentrated and much more a matter for individuals than it is today. The delay between the discovery of new knowledge and its application was usually lengthy as a matter of course. Thus it is an extraordinary situation to find that work has been going on all over the world to adapt these older systems for use in IR systems, by means of repairs, amendments, new editions, etc[68, 99-105]. In addition, new systems have been developed since World War II which, admittedly, do overcome several disadvantages of the older systems. Of these, Vickery's Faceted Classification is the best known[106-108]. This system does not set out to provide lists of subjects contained by a process of logical division from _summum genus_ (the generic level of the information scientist) to _infima species_ (the specific level). Instead it attempts to analyse the phenomena that occur in nature into sets of homogeneous categories or 'facets', and to list in each category the names of the things, processes and so on, that are appropriate to that category. There is no central or international body attempting to regulate or up-date categories since each user decides his own. After the categories have been chosen and their contents listed, they and their contents are arranged in some preferred sequence and fitted with a code or notation (usually alphanumeric) that allows items from various different facets to be combined as may be appropriate to the information in a particular document. Such systems certainly overcome the rigidity of classes, and also there need be no delays in expansion which may be called for in up-dating. Nevertheless they still fail to meet the basic user requirement of plain language alphabetical presentation, since a code is essential to them. No classification system is as simple as alphabetical subject indexing from the all-important point of view of IR system users.

As one example of many possible difficulties in applying UDC because of slowness in up-dating classification tables, the subject of Aerosols can be cited. A suitable number is available for Aerosols <u>per se</u> as the physical chemistry of the suspension of particles in media, namely 541.182. However, after nearly a decade of marketing aerosol products there is currently still no suitable number available for the process and technique of packaging these products in cans. Thus the librarian is forced to devise his own numbers which, most likely, will not agree with the official numbers when such eventually become available. To the information scientist, who, because of his technical background, is primarily a scientist, it is small wonder that the French documentalist G Cordonnier should have referred to the UDC system as 'un monstre préhistorique'.

4.1.2 <u>Abstracting</u>

The purpose of abstracts is to serve as an intermediate screening device between the IR system user and the file of original documents. To the extent that abstracts meet this purpose they are substitutes for the original documents. Well written and technically accurate abstracts meet user requirements for direct access to information, for browsing and so on, and these are the basic reasons for the use of abstracts in IR systems. No IR system can operate without some form of indexing but plenty of systems - document retrieval systems - operate without abstracts. On the face of the matter the necessity for abstracts is not so apparent, particularly to the layman. But any system that excludes all abstracts of some kind or other is, by definition, not an information retrieval system.

Industrial management, when considering the setting up of a retrieval system, is inclined to query the necessity for abstracts on the ground of the additional costs involved - ie input costs to the system. There are two factors to be considered.

Firstly, while the input costs may be lower for a document retrieval system, output costs are higher. With no screening device available the user is forced to at least original documents; these higher costs are cumulative since they are applicable each time a particular document is looked at. Frequently the total number of documents to be examined in this way is large and the task is time consuming. The number of documents rejected as not relevant to a particular problem, after time has been spent on obtaining and looking at them, is often considerable. Clearly much depends on both the competency of indexers in their choice of indexing terms and the competency of users in phrasing their questions, if the percentage of 'false drops' is to be kept within reasonable limits. In systems where abstracts are provided it is usually possible for the user to find available relevant information without achieving quite the same level of precision in his question, and also to modify his question during the browsing process in the light of the information found. Usually the screening process is fast and completely subconscious.

Secondly, the effort required of an indexer in order to index a document accurately and in depth means that he has to study it thoroughly. Assuming that a well prepared author summary or abstract is not already available, which it frequently is in industrial conditions, the additional effort required for the preparation of an abstract is relatively small. It is estimated to add not more than 25% to the cost of indexing and that this expenditure is fully recovered in time saved when abstracts are studied during a search. These savings are cumulative in that they are achieved each time an IR system is used, as far as each abstract looked at is concerned.

The point made above that in industrial conditions a good author summary is often already available requires some amplification. In the preparation of, say, a research report, it is essential to provide a comprehensive, clear and concise summary, and this should preferably be done by report authors. A good summary is required since it is often the only part of the document that a manager will look at carefully. This may be because of the time factor, but it is true that decisions by management for further action are frequently taken largely on the contents of a summary when there is reasonable familiarity with the project reported. An abstractor can very often use a good summary verbatim for inclusion in an IR system; whenever this happens the words 'summary' and 'abstract' are synonymous.

There are two main types of abstracts, the indicative and informative, although some writers claim that there are more [39,109], and these have been described widely in the literature. Probably as simple and clear an explanation as any of the two types is that by David Liston, Jr, of Battelle Memorial Institute who writes as follows [110]:

> <u>The Indicative Abstract</u>
>
> An indicative abstract describes what a document "talks about" without recording "what it says". The indicative abstract does not include data, facts, observations, or conclusions that are discussed in the document. Instead, it records that such data, facts, etc, are subjects discussed in the document. This type of abstract is intended as a tool to help decide whether the original document is actually of interest to the searcher, in which case the original document must be consulted for the actual information it contains.

The Informative Abstract

It is the intent of a good informative abstract to include all pertinent data, facts, observations, or conclusions presented in the original source document. If this is accomplished adequately, it is theoretically unnecessary to refer to the original document from that point on.

Examples:

The following two sentences show how the same information would be covered in indicative and informative abstracts respectively:

1 INDICATIVE PRESENTATION:

The steady pressure required for extrusion of 99,9% pure lead, over a range of extrusion ratios, and at two ram speeds, was determined and expressed as a formula.

2 INFORMATIVE PRESENTATION:

At extrusion ratios (E) from 4 to 175 ram speeds of 1 and 15 cm/minute, the steady pressure (P) required for extrusion of 99,9% pure lead was found to be given by $P = \log E + K$, where K is a constant depending on ram speed.

For purposes of information retrieval there is no doubt that the informative abstract is always preferable. However, it is not always practical to prepare them. In the case of typical research reports and of many papers and articles presented respectively to meetings or published in the journals of learned societies the major subject is relatively 'narrow' and is dealt with in detail. It is reasonably easy to prepare a good informative abstract for such documents.

However, when the subject matter is 'broad', as one expects to find in such documents as annual reports, progress reports for all the work going on in, say, a research and development department, review articles, symposia, manuals, or books, etc, the situation is entirely different. In these instances an informative abstract would of necessity be pages long, which is not acceptable. Thus there is no alternative to the preparation of indicative abstracts unless these documents are to be fragmented and the parts treated as being separate documents. This is done where a particular document is of especial value to a company, but the sheer quantity of items to be processed into most IR systems is sufficient to prevent such treatment as a matter of routine. Other reasons for the preparation of indicative abstracts are, firstly, when time is short and the abstracting must be completed as quickly as possible and, secondly, when it is desirable to process a document into an IR system even though the contents are somewhat trivial or are of marginal interest only.

4.1.3 Constructing abstracts

Probably the most thorough investigation into criteria for good abstracts was carried out by Harold Borko and Seymour Chatman in the early 1960s. They reported as follows [111]:

> 'Distilling the essence from the 130 manuals of instructions, we list as our criteria for an adequate abstract the following.
>
> Content
>
> Purpose: A statement of the goals, objectives, and aims of the research or reasons why the article was written. This statement should be included in both the informative and indicative abstracts.
>
> Method: A statement about the experimental techniques used or the means by which the previously stated purpose was to be achieved. If the techniques are original or unusual, or if the abstract is informative, more detail should be included.
>
> Results: A statement of the findings. The informative abstract tends to be more quantitative than the descriptive abstract.
>
> Conclusions: A statement dealing with the interpretations or significance of the results.
>
> Specialized content: Certain subject-matter fields require that the abstract contain specialized

information. Medical journals, for example, require that the abstract contain details of diagnosis and treatment, drug dosages, etc, where applicable. In writing or evaluating abstracts in these fields, the specialized requirements must be considered.

The abstract is to be brief (100-500 words) and non-repetitive.

Form

More variation is possible in the form of the abstract. Therefore, form criteria must be applied judiciously and cautiously. The major criteria are clarity of content and conciseness of expression. Having recognized the existence of great variation in form, we can nevertheless specify general differences between the informative and indicative abstract.

	Informative	Indicative
1	Active voice	Passive
2	Past tense	Present
3	Discusses the research	Discusses the article which describes the research.'

The author's personal experience has shown that the process of constructing abstracts can be broken down into a number of elements, each of which comprises several steps that can be enumerated. These apply to any document, as follows:

1) Find what the document is about. The necessity to examine a document closely does not mean that it has to be read from cover to cover; this is usually a waste of time. The objective is to pick out the important aspects and/or information as quickly as possible. There are suitable techniques for studying each type of document:

Example 1 - a technical report:

Note the title, read the contents list and summary, if these are present, read the preface or introduction, skim the body of the report noting tables, graphs, illustrations, etc. Finally read the conclusions reached and recommendations made; if these are not clearly set out, read the final paragraphs where they will probably be found.

Example 2 - a patent specification:

Note the title and patentee, read the introductory paragraph and the consistory clause, skim the body of the specification noting examples and any drawings. Finally read the claims as set out in numbered sequence at the end of the specification.

2) Decide whether an indicative or an informative abstract is more appropriate and then prepare a draft of it. The following points should be borne in mind and ought to be apparent from the study of each document:

(a) Objective and scope of the investigation, work done, etc.

(b) Methods and materials used.

(c) Results obtained, conclusions reached, recommendations made, etc.

(d) Leave out background and elementary basic information. Also omit well-known and very simple or trivial techniques, methods, procedures, facts, results, etc.

3) Check the draft to ensure that its three main elements of Content, Arrangement and Language are right:

(a) <u>Content</u> The facts given and the information set out must be accurate.

(b) <u>Arrangement</u> The information should appear in logical order and the train of thought should be easy to follow. There should be no jumping back and forth between topics, and no statements which cannot be understood until a later sentence has been read.

(c) <u>Language</u> The abstract should be easy to read. It should be free from long and involved sentences, grammatical mistakes, and unnecessary words. The words selected should always be the simplest that will convey the intended meaning, so that the contents of each sentence are obvious at first sight. Once the draft has been checked as above it then becomes the actual abstract.

The consistent application of the above procedures will ensure the preparation of good abstracts. However, it must be admitted that of all scientific writing, abstracts are probably the dullest and most impersonal both to write and to read. They epitomize the fact-grinding side of science and technology and

Example 2 - a patent specification:

Note the title and patentee, read the introductory paragraph and the consistory clause, skim the body of the specification noting examples and any drawings. Finally read the claims as set out in numbered sequence at the end of the specification.

2) Decide whether an indicative or an informative abstract is more appropriate and then prepare a draft of it. The following points should be borne in mind and ought to be apparent from the study of each document:

(a) Objective and scope of the investigation, work done, etc.

(b) Methods and materials used.

(c) Results obtained, conclusions reached, recommendations made, etc.

(d) Leave out background and elementary basic information. Also omit well-known and very simple or trivial techniques, methods, procedures, facts, results, etc.

3) Check the draft to ensure that its three main elements of Content, Arrangement and Language are right:

(a) <u>Content</u> The facts given and the information set out must be accurate.

(b) <u>Arrangement</u> The information should appear in logical order and the train of thought should be easy to follow. There should be no jumping back and forth between topics, and no statements which cannot be understood until a later sentence has been read.

(c) <u>Language</u> The abstract should be easy to read. It should be free from long and involved sentences, grammatical mistakes, and unnecessary words. The words selected should always be the simplest that will convey the intended meaning, so that the contents of each sentence are obvious at first sight. Once the draft has been checked as above it then becomes the actual abstract.

The consistent application of the above procedures will ensure the preparation of good abstracts. However, it must be admitted that of all scientific writing, abstracts are probably the dullest and most impersonal both to write and to read. They epitomize the fact-grinding side of science and technology and their flow is never ending. But they are essential, since they provide a much needed conspectus of each science and technology. And it should be noted that while many humanists have been shy to accept the necessity for abstracts for many years in their fields of interest they are now being increasingly used in all the humanities. The humanists also are all too familiar with their own information explosion.

In conclusion, one aspect of abstracts of published material should be noted. It is sometimes argued that informative abstracts should frequently be longer than they usually are, so that more details can be included and so make it even less necessary to refer back to the original document. Even if time and the additional costs were to permit this, it is not a legally acceptable procedure since such abstracts would be more akin to 'digests'. The copyright acts of many countries make it clear that digests are prohibited without the consent of the copyright owner, on the grounds that they tend to decrease the sales of the published material in its original form. The probability is that this is true for the Republic of South Africa also; the Act itself does not specifically cover the point and in the absence of any relevant court decisions within the Republic it is not a clear cut issue [112].

4.1.4 Subject indexing

Indexing can be regarded as a two-step process:

1) It is necessary to decide what are the essential technical ideas or concepts in a document that are of interest to the particular industrial concern involved. If these items of information are to be retrievable in the future, then the concepts must be recorded.

2) The selected concepts have to be recorded in a standard and consistent manner. These standards are laid down when each IR system is designed, and they include such aspects as the degree of exhaustiveness of indexing to be aimed at as well as the level of specificity of indexing. These particular aspects are clearly important in governing the recall

and precision capabilities of an index. The actual standards adopted are closely concerned with the economics of operating IR systems, the time available or worth spending on each document, and so on.

The general principles of indexing are well covered in the publication 'USA Standard Basic Criteria for Indexes' [113]. There should, however, be full appreciation of the fact that good indexing is not a simple task that can be given to just anyone to do. It is an intellectual process, which in the industrial field requires scientific or technological knowledge, as well as reflection, judgement and discrimination.

It has been emphasized already that subject indexing is preferred for IR systems in industry. Study of the literature of librarianship shows that there are two main types of subject indexing, each of which is known by several different names. Of the various possibilities the names 'alphabetico-classed' and 'alphabetico-specific' appear to be adequately descriptive for the purpose of this dissertation.

1) Alphabetico-classed indexing

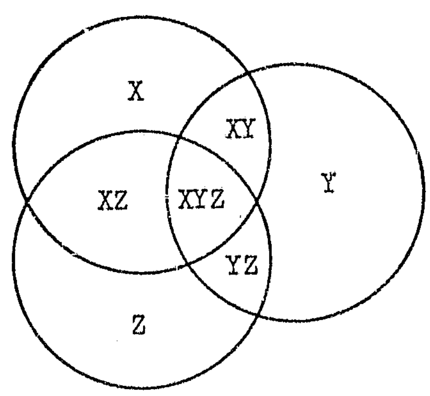
Alphabetico-classed indexing has been applied to library catalogues for many years, particularly in the United Kingdom where it is the basis of the British Technology Index developed by Coates [114-116]. There are many of these indexing systems, some of the well-known ones being those of Cutter [117], Bliss [118] and Kaiser [18] - these and others have been well described in the literature by many writers, a good example being Berwick Sayers book 'Manual of classification' [98]. The development of some systems led to the publication of lists of subject headings, particularly in the United States; probably the best known list is that of the Library of Congress [119]. It also led to the appearance of publications providing guidance on the preparation and development of such lists, such as those by D J Haykin [120,121].

Generally speaking, alphabetico-classed indexing has not found favour for IR system use, on the grounds discussed in sub-section 4.1.1, on classification versus alphabetical subject indexing. These systems are concerned primarily with the arrangement of books on shelves for the convenience of library users, and it is very unlikely that they could oust 'alphabetico-specific' indexing in IR systems.

2) Alphabetico-specific indexing

As pointed out above, the publication of lists of subject headings for alphabetico-classed systems has long been popular in the United States. Thus it is hardly surprising that the advantages to IR system users of alphabetico-specific indexing should have been appreciated from the early days of information science in the 1940s in that country. The American approach to IR system design is generally practical and user orientated and the simpler name 'subject indexing' has become the standard term of information science in place of the librarian's more definitive term of alphabetico-specific indexing. Since this dissertation is concerned primarily with information science as distinct from librarianship, the term 'subject indexing' will be adhered to.

There are essentially two types of subject indexing in use for IR systems, known as post-coordinate and pre-coordinate. Before briefly discussing them however, it is desirable to consider the logical principle underlying the whole matter of subject indexing in coordinate systems for information retrieval. This is best done by illustration using a Venn diagram (after the economist John Venn, 1834-1923) thus:



For example, if the circular area X represents the concept 'diesel engines', Y the concept 'valves' and Z the concept 'fuel', then the intersections of these circular areas represent the sub-sets of items of information on two or more of these concepts. Thus, in this example these sub-sets comprise:

XY	=	diesel engine valves
YZ	=	fuel valves
XZ	=	diesel engine fuel
XYZ	=	diesel engine fuel valves

The individual circular areas X, Y and Z may, of course, represent generic or specific concepts but the areas of intersection will always represent less generic (or more specific) concepts. This can be clearly seen in the above example where XY, YZ and XZ are all more specific than X or Y, Y or Z and X or Z respectively. Similarly, XYZ is more specific than XY, YZ or XZ.

The above example also serves to illustrate the basic difference between post-coordinate and pre-coordinate systems. In post-coordinate systems the indexer will record only the three concepts X, Y and Z, leaving the system user to coordinate them in any way that he may wish at the time of carrying out a search. In pre-coordinate systems the indexer will carry out the coordination and record the concepts XY, YZ, XZ, and XYZ, or only those relevant to the information in a particular document - eg, he might record XY, YZ and XYZ only for an item of information on diesel engine fuel valves, where the diesel fuel itself was not discussed at all. Both systems have their advantages and disadvantages.

4.1.4.1 Post-coordinate indexing

Early workers in the field of information science, shortly after World War II, saw the latent possibilities of a conceptual approach to the problems of storing and retrieving technical information [6,122-125]. This approach was first recognised by librarians as being a means for dealing with the problems of bibliographic analysis, but the pressures generated by the post-war information explosion were needed to force its development and practical application. The first such successful commercial application appears to have been that by the US Armed Services Technical Information Agency who installed Mortimer Taube's so-called 'Uniterm' system on an experimental basis in 1952 [126]. This was followed very quickly by other inverted indexing systems and all were known at that time as 'concept coordinate indexing systems', since no distinction was being made between them and pre-coordinate systems as such.

The Uniterm system has been adequately described in the literature [127,128] and the only development that has taken place in it has been that the original conception of the 'unit-term' was sometimes applied too strictly so that artificial situations were created. For an example refer again to the concept of 'diesel engines'. This can be broken down to the concepts of 'engines' and 'oil fuel' so as to distinguish them from engines fueled by petrol. However, the terms 'diesel engine' and 'petrol engine' are widely understood and each comprises one concept.

The Uniterm system of coordinate indexing differs in no way in principle from other more 'sophisticated' versions. This sophistication is primarily concerned with the type of hardware used, together with sub-indexing devices; both of these aspects are discussed in later sections.

The advantages of post-coordinate systems, together with discussion of refinements, etc have been stressed at length in the literature [128-132]. Nevertheless, these include the fact that the searcher can phrase his question as any preferred combination of the indexed concepts. This is particularly valuable where IR systems cover limited, and probably highly specialised subject fields in considerable depth. The users of such systems are invariably familiar with the said subject field, which means that they usually have little trouble in phrasing questions so that the precision of searches is high (see Chapter 3, Section 3.1.2). However, when post-coordinate indexing is applied over wider fields by indexers who cannot possibly have expert knowledge in all the fields covered, the level of indexing tends to be more generic (less specific). This tendency is relative to the degree of interest in the subject field in each particular industrial concern. But where the IR systems are large, reduction in precision becomes noticeable for the following reasons:

1) The subjective opinions of the indexer as to the intent and content of the indexed material are dependent both on his subject knowledge and his knowledge of the particular company's interests.

2) The reason for the search as conceived and phrased by the system user - <u>ie</u> his interpretation of the search language of the system.

3) The understanding of 2) by the go-between in the case of highly mechanised systems - which many large systems are - where the user is not permitted to search on a self-help basis.

4) The user's acceptance of the output of the system in terms of his original conception. Where time delays occur between phrasing questions and receiving answers, considerable changes in the user's thinking and his conception of his requirements may have occurred.

Efforts to overcome the problems which these aspects engender led to the development of the sub-indexing devices discussed in sub-section 4.1.5.

4.1.4.2 <u>Pre-coordinate indexing</u>

It has been noticed that some writers refer to conventional library alphabetico-classed indexing systems as being the same type of system as the pre-coordinate systems of information science [67,133]. Strictly speaking this may not be incorrect, nevertheless it is a tendency which, if encouraged, can only add confusion to the already very confused jargon of both information science and librarianship. The term 'pre-coordinate indexing' should be reserved for that type of alphabetico-specific indexing where the indexer of information in a document carries out the coordination of the concepts to be recorded. As pointed out already this enables a clear and concise distinction to be drawn between post- and pre-coordinate indexing.

The major advantages of the pre-coordinate index have been made clear already - they are user convenience in that he can browse in these indexes and operate the systems himself, together with the fact that system output costs are low, since he can quickly decide whether he needs to look at any original documents. It cannot be denied that the input costs for pre-coordinate systems are higher than for post-coordinate, but the overall costs of input plus output are no higher. In addition it must be remembered that in a post-coordinate system coordination must be carried out every time the same information is required. In a pre-coordinate system this is done once only at the time of indexing. Thus, however frequently the same information is sought in this type of index minimum time and costs are involved in finding it.

In industry the adoption of a pre-coordinate system offers an advantage which is often overlooked, but which is perhaps of particular concern to small and medium sized companies. If it is accepted that thesauri of indexing terms are necessary for the successful operation of an IR system (see sub-section 4.1.6), then the pre-coordinate system offers the advantage that the thesaurus of this type can be built up as the system develops. Should it become very large, it can be structured in exactly the same way as a post-coordinate thesaurus. In a post-coordinate system it is essential to develop a fully structured thesaurus before operation commences.

In using the pre-coordinate type of system the user is constrained by those terms which the indexer has selected. In practice in systems where the standard of indexing is good, where abstracts are included and where there is simultaneous display of all the indexing terms used for each document on every card, then this constraint is of no real importance. This is illustrated in the systems outlined in Chapter 6.

4.1.5 Sub-indexing devices

All three sub-indexing devices discussed in this

Section were developed primarily to overcome deficiencies and difficulties of post-coordinate systems.

The continuing increase in subject specialization in the sciences led directly to the need for more detailed, or 'deeper' indexing in order to simplify searching. But it has long been known that when such deeper indexing is applied in post-coordinate systems, then the possibility of retrieving irrelevant information or documents is also increased (frequently referred to as 'false drops'). To overcome this the two sub-indexing devices known as 'links' and 'roles' were developed.

Briefly, a link is a symbol - usually alphabetical - used to prevent the unwanted coordination of terms. For example, a report concerned with the efficiencies of catalysts in nitric acid production could be indexed under the terms:

efficiency	nitric acid
catalyst	production

In a search for information on the production of catalysts the above report would also be retrieved and would be a 'false drop'. To prevent this sort of unwanted retrieval the terms can be 'linked' as follows at the time of indexing:

efficiency - A	nitric acid - B
catalyst - A	production - B

Roles indicate the context in which each term is used, and for this reason are often known as 'role indicators'. In the English language a combination of two concepts may have a certain meaning, whereas inversion of the same two concepts provides an entirely different meaning. For example consider the terms 'Venetian blind' and 'blind Venetian' - normally a post-coordinate system cannot distinguish problems of syntax of this type. Thus roles reduce unwanted retrievals by enabling the searcher to coordinate the terms searched in their right context. Numerical symbols are invariably used for roles:

Packaging of polyethylene	-	packaging	-8
		polyethylene	-9
Polyethylene for packaging	-	polyethylene	-8
		packaging	-4

(Note: the precise meanings of the roles used in this example are given in Appendix I, which lists the roles developed by the Engineers Joint Council (EJC)).

The pros and cons of using links and roles or either links or roles, together with information on the resulting improvements in efficiency, or lack of it, and the increase in indexing costs, has been discussed exhaustively in the literature (134-146). The comments which follow in sub-sections 4.1.5.1 and 4.1.5.2 respectively, are limited therefore to aspects which appear to have received scant attention to date.

More recently, a third sub-indexing device, known as 'weighted term searching' has appeared. This has developed for use in computerised post-coordinate IR systems and is a means for indicating the importance of different concepts in the context of the information in each document. Thus it is used to ensure the increased relevance of information retrieved and is in fact a sub-indexing device although this is not obvious from its name.

4.1.5.1 Links

Generally speaking, the use of links has not found favour, since the cost of rejecting false drops arising from their non-use is usually less than the additional indexing costs involved. In fact, links came into being largely to deal with a problem which earlier experience of librarians in subject indexing should have ensured was widely understood. When subjects are fragmented into segments, such as is done in post-coordinate indexing, then the possibility of unwanted coordination is always present. This was known to many librarians and the introduction of links appears to have been an example of reinvention of the wheel, this time by information scientists (see Footnote).

It is obvious that the use of links has no applications in pre-coordinate indexing since the tail of a pre-coordinated term serves the same purpose (eg, Nitric acid - production).

It would appear that the major worthwhile applications for links are in a few types of collections of very specialized documents. One such is in the field of chemical patent specifications where the structure of each document is extremely complex. Barbara A Montague of E I du Pont de Nemours & Co has done much work in this field (140). She points out that links can be very useful to differentiate between chemical compounds discussed only in the body of a patent specification, and those which are specifically covered in the claims at the end of specifications.

Another potentially valuable application is in the field of industrial progress reports. These documents are usually issued regularly at monthly or quarterly intervals and are consequently known as 'period' reports. They may, for instance, cover all the numerous research topics currently in hand in a research department. In addition, these reports may be the only source of reported information on all or many of these topics, thus they will cover the total history of each in a fragmentary manner and over periods of time which may be considerable. Since numerous aspects of many subjects are likely to be worth indexing in these reports, the possibility of unwanted coordination taking place is far greater than in those reports which deal with one subject in detail. The use of links ought to substantially facilitate searching in a collection of period reports of this type; it is an aspect that should repay closer investigation.

[Footnote. At this point it is perhaps appropriate to quote a statement by Jesse H Shera, Dean of the School of Library Science at Western Reserve University, Cleveland, Ohio (147): 'Information science, then, is not antithetical to librarianship; on the contrary, the two are natural allies, and librarians should not reject this new intellectual relative, nor should the information scientist discredit the librarian. Both have made and will continue to make mistakes, and if the librarian is the more guilty it is only because he has had more time in which to err. The laws of chance prohibit innovation without error. For the moment at least the librarian and the information scientist may speak with different tongues - new concepts require a new terminology - but eventually a consensus and a common understanding will be achieved'].

4.1.5.2 Roles

Roles have been used on a much wider scale than links, but here again there is suspicion at least that information science may have reinvented the wheel; Artandi and Hines [134] contend that roles are indistinguishable from the sub-headings found in many conventional subject indexes. Probably the n difference is that a limited number of roles are used with defined meanings - see Appendix I - as opposed to a complete lack of formalisation for sub-headings.

However, there is one extremely important aspect of this device which has not, as far as can be ascertained, been reported in the literature at all. The literature has been concerned with syntactical problems so as to reduce the number of unwanted coordinations in post-coordinate systems. In doing so, it has been overlooked that roles can be applied to pre-coordinate systems equally as well as to post-coordinate systems, and that in either case the indexer is presented with an extremely valuable tool to help him to maintain indexing consistency.

Order and consistency in indexing systems give rise to predictability, which in turn gives rise to dependability and user confidence. Consider for instance the subject index to Chemical Abstracts (CA), the biggest of all the commercial abstracting services. A user may not necessarily agree with the way in which a given chemical compound or phenomenon is indexed in CA, but he often knows how CA does it and where to find it in the CA indexes, because generally speaking this service has adhered to the same formalised procedures for many years [148-152].

P Zunde and Margaret E Dexter of the School of Information Science, Georgia Institute of Technology, have listed the factors which influence human indexing [153]. They divided them into three categories, namely, pragmatic, environmental and semantic factors. Amongst the semantic factors they list various indexing aids, such as thesauri and dictionaries. They could well have included role indicators since adherence to a list, such as the EJC list, compels the indexer to analyse the major concepts of each document within their context. A good indexer may do this subconsciously, but this formalised approach ensures that important aspects of the information in a document are not overlooked. At the same time a list of roles like the EJC list is so broad that it does not provide any constraints when indexers are faced with describing new concepts; it can be used in any industry.

4.1.5.3 Weighted term searching

It is typical of both post-and pre-coordinate indexing that all indexing terms maintain an equal level of importance. In pre-coordinate systems where browsing by the user through the index is possible this fact is of little consequence. The user's mind automatically relates (or weights) items of information to the question asked and sub-consciously ranks them in order of importance. In a post-coordinate system, particularly in large ones and ones where neither roles nor links are used, weighted term searching is of especial value in that when a search is carried out, it can guide the searcher as to 'how little to read'. In conventional post-coordinate indexing a concept is either considered to be worth indexing or it is not - this process can be represented as a 'yes' or 'no' assessment. However, with computerised IR systems it becomes possible to use the binary algebra of George Boole, known as Boolean algebra, to weight the value (as information) of concepts in a document and to rank them in order of importance as far as the particular company whose system it is is concerned. The idea is a good one but it is not proposed to deal with it in detail in this dissertation, on the grounds that this searching technique can be used only in post-coordinate IR systems that have been computerised. There are few such systems in the Republic of South Africa today, and little prospect of any big ones before the end of the 1970s. There are various refinements being reported in the literature for this technique already, but no one appears to know the real costs involved in applying the technique itself or the refinements. Theoretically, the principle of weighted term searching is sound provided that a company does not change its major interests over a relatively short period. Quite clearly, the indexer applies his subjective judgment, based on his knowledge of his employer's interests, when weighting terms, and if these interests change rapidly these judgments will be outdated equally rapidly. The technique has been fully described in the literature [154-158] and the articles by Brandhorst [154] and by Iker [155] in American Documentation are particularly clear.

4.1.6 Thesaurus control

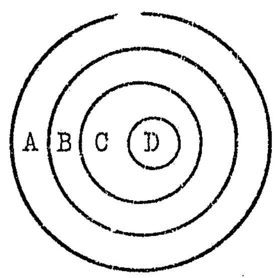
Much has been published on the subject of thesauri, the need for them, the different types and how to set about constructing them [149,159-165]; there appears to be little justification for reviewing these publications here and the interested reader is simply referred to them. Again there has been a proliferation of terms in the jargon of information science which have added little or no clarity to the whole subject. For practical purposes terms such as 'authority lists' 'descriptor lists', etc tend to confuse the non-specialist and to differentiate at too fine a level for the needs of industry. For the purposes of this dissertation only the following terms are used:

1) Lists of subject headings as developed for library catalogues. The purpose of these lists is to indicate what whole documents in a collection deal with a particular subject - 'subject' being used in a broad sense (generic level). They are intended for widespread application.

2) Lists of words developed for the retrieval of information and specifying the subject contents of whole documents. These lists are referred to as thesauri, whether they are structured or not; and whether they are drawn up for post- or pre-coordinate IR systems.
   Apart from the relatively few published thesauri (references [166-180] list a variety of examples) those developed in industry are generally intended for application to one IR system only - ie to each concern's own system.

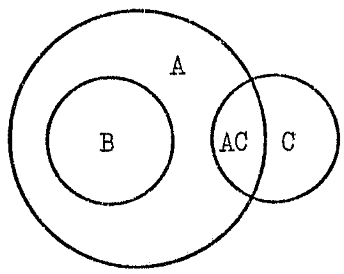
The first use of the word 'thesaurus' in connection with information retrieval appears to have been made by H P Luhn of IBM [181] in 1957. Luhn, however, was thinking in terms of 'idea groupings' such as is found in Roget's Thesaurus, but use of the term for the alphabetical listing of words so as to show hierarchical relationships for post-coordinate indexes was adopted very soon thereafter. The hierarchical relationship of words means simply the relating of words according to their generality or specificity of meaning. For example, the word 'chemicals' is very generic, the words 'inorganic chemicals' are less generic, the words 'sodium salts' are even less generic and the chemical substance 'sodium chloride' is specific.

These relationships are perhaps more easily appreciated by referring again to mathematical sets illustrated by Venn diagrams, thus:



A = Chemicals
B = Inorganic chemicals
C = Sodium salts
D = Sodium chloride

Consider also:



A = Explosives
B = Commercial explosives
C = Pyrotechnic substances
AC = Explosive pyrotechnic material

It is clear therefore that all hierarchical relationships can be illustrated in this way.

Today in industry the term 'thesaurus' has become widely accepted and has itself become less specific. It now includes lists of both post- and pre-coordinate terms, whether hierarchically arranged (ie structured) or not - examples of each type are set out in Appendix III. However, it may well be asked why a thesaurus of any kind is necessary in the first place, and above all, if thesauri have been published in a field of interest to a particular company, why is it apparently still necessary for that company to develop its own.

The necessity for a thesaurus to control the indexing terms used in any IR system arises from the nature of language itself. Categories, classes, hierarchical arrangements, etc do not exist in nature. In his efforts to understand his environment man arbitrarily selects certain characteristics and features of the universe and groups them together. However, it has been realised for some time that in selecting features for grouping there is no single clear-cut pigeon-hole for each one; this break-down between scientific disciplines has been referred to already in Chapter 1. This state of affairs is reflected in man's language, since it is both imprecise and ambiguous. A word or expression can have a number of meanings (homographs) and connotations, while several words can have the same meaning (synonyms) or nearly the same meanings (near-synonyms). Meanings change over a period of time and often vary from one individual to another; language grows and adapts to new discoveries, and changes in response to forces that are not clearly understood. While all this prevents monotony in the daily use of natural language, it also makes more difficult the process of precise communication. Thus in the case of IR systems, the purpose of thesaurus control is to relate words and concepts precisely and thus to maximise the likelihood that the indexer will be successful in leading the searcher to the information he is seeking.

There are, however, a few large IR systems in existence where the indexing vocabulary is still not controlled by means of a thesaurus. Probably the largest of these are to be found at the Battelle Memorial Institute in Columbus, Ohio. Battelle's earlier 'philosophy' of information science has a great deal to commend it and has been well and succinctly described in the literature [182-186]. Ralph A Darby, currently Head of Battelle's Information Systems Division, admitted in 1969 in personal discussion with the author that it had become necessary to introduce control of the indexing vocabulary for their newer systems. Thus the one major advocate of IR systems without thesaurus control in the sphere of technical information has apparently been compelled to change his ideas.

The extent of the prevalence of synonyms and near synonyms in all the sciences and technologies is not generally appreciated by the layman, but it can be clearly seen if one or two of the published thesauri cited as references are examined (eg, 166 and 169). In the field of organic chemistry alone there appears to be an average of five possible names for each substance [187,188] and the situation is far worse in such disciplines as pharmaceuticals and pesticides where large numbers of trade names are also involved [189,190], while the various branches of engineering are no better [191].

The reason that each company finds it highly desirable to build up its own thesaurus is directly related to the statement on page 5, Chapter 1, that: 'Each industrial concern is unique in its range of interests .....'. Published thesauri always include numerous terms of no interest to a specific company since the compilers of such works must endeavour to cover a discipline or technology objectively and comprehensively. Similarly published thesauri do not, and cannot include terms of local interest only, nor trade names, especially those peculiar to one country, and often they do not include coverage that is sufficiently detailed at the specific level to meet the requirements of one company. As an example consider the case of a South African manufacturer of commercial explosives: the EJC Thesaurus [169], which is the one most relevant to this subject field, covers the whole technology of explosives, including military explosives, with 49 terms. The actual requirements of this particular manufacturer are for very specific indexing terms which in fact total over 500 for commercial explosives alone.

Considerable argument has developed as to whether a thesaurus should include the maximum number of apparently useful terms or the minimum number, together with many cross-references [192-196]; the views of R Moss [195] on the importance of using the minimum number of apparently useful terms are clearly set out and worth studying. Probably the truth lies in between the two extremes but, again, much depends on the peculiar interest of each industrial concern. Loukopoulos [197] has pointed out that the widespread use of computers in the United States for post-coordinate IR systems has led to indexers freely assigning numerous terms to each document. In many instances this amounts to 'redundant' indexing such as happened in the early days of Uniterm indexing. In other words, excessive indexing may minimize the necessity for the user to phrase his search question explicitly, but it also leads to re-assessment of the search question after retrieval because of the large number of unwanted coordinations (false drops) obtained. Thus it is obvious that 'lazy' indexing by using too many words in order to be sure of covering all aspects leads to increases in system output costs and to the waste of users' time. Effort by indexers to analyse each document and to reduce the total number of terms required to cover the major concepts of interest are well worthwhile in ensuring user satisfaction. This confirms the hidden truth of Pascal's statement in the letter to a friend, namely that: 'I have made this letter longer than usual because I lack the time to make it shorter'. It requires more thought and is always more difficult to be concise and explicit than to ramble on, and this applies equally well to building up a thesaurus.

4.1.6.1 Constructing a thesaurus

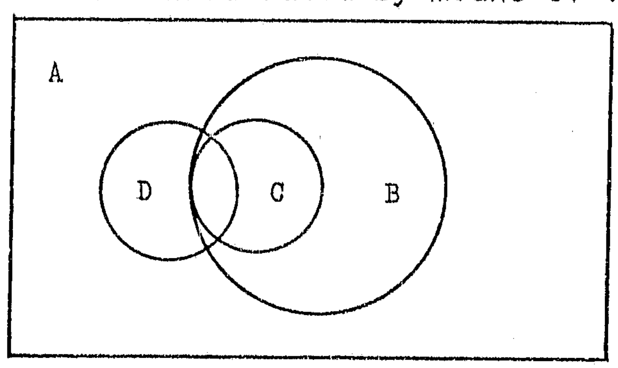
Reference has been made already to the prevalence of synonyms in virtually all fields of science and technology. It is essential to index all the information on one subject under one term, and to provide adequate cross-references from other terms to the one preferred, otherwise much information can be lost. To select a so-called 'preferred' term it is best if possible to consult an expert in one's own company, since he will know what term is in use within the organisation. If there is no expert at hand it may be possible to obtain the desired information from suppliers of raw materials or equipment, or from customers, as may be appropriate, before consulting dictionaries, encyclopaedias, published Thesauri, etc. As an example, in plastics technology four expressions are commonly encountered in the literature to express one concept, viz:

Foamed plastics
Blown plastics
Expanded plastics
Cellular plastics

Enquiries in the South African plastics processing field quickly established that the locally preferred term is 'expanded plastics'. In building up a particular thesaurus this expression was adopted, but 'see' references (*ie* cross-references) were included under each of the other three terms.

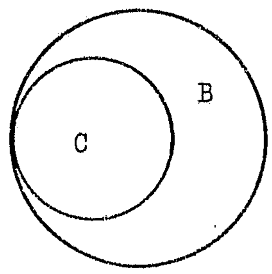
Attempts have been made in some scientific fields to standardise terminology, particularly in chemistry. However, these standards often go against strong conventional preferences or entrenched local but well understood terms. Since the IR system designer, or thesaurus compiler, is faced with the practical difficulty of providing a service to his colleagues within a company or other industrial organisation, it is essential that he should accept most of their preferences in order to develop their confidence.

In addition to entrenched preferences are the limitations of interest apparent in each industrial concern. For example consider a company that manufactures chemicals for the control of insects, *ie* insecticides. There are other means, such as non-chemical means, for controlling insects, but these would not be of direct interest to a chemical company. Suppose that one of the insecticides manufactured by this same company is specific to the control of aphids, which are a family of insects - *ie* it is an aphicide. Again there are known means for controlling aphids other than by the use of chemicals; these would also not be of direct interest to the same company. Again this whole matter is best illustrated by means of Venn diagrams, thus:



A = Insect control
B = Insecticide
C = Aphicide
D = Aphid control

For the practical requirements of this particular company 'insect control' would be treated as being a synonym of 'insecticide', and 'aphid control' as being a synonym of 'aphicide'. The terms 'insect control' and 'aphid control' would be retained in their thesaurus as cross-references (known as 'see' entries), eg 'insect control, see insecticides'. The entries in the thesaurus for indexing purposes would simply be 'insecticides' and 'aphicides', illustrated thus:



B = Insecticide
C = Aphicide

It is obvious in this example that insecticide is the more generic term.

Another difficulty is that the names of technical processes often take a long time to settle finally, but in the meantime words for indexing these concepts are often required for inclusion in a thesaurus. An example is the 'bulking' process applied to synthetic fibres to give the yarn softness to the touch, apparent elasticity and varied thermal insulation properties. This process is variously called 'bulking', 'crimping', 'false twisting', 'texturing', and, most recently, 'texturising'. In cases like this the thesaurus compiler should select one term, again preferably with guidance, and provide cross-references for all the others.

The question as to whether or not a thesaurus should be structured is entirely a matter of the scope of the subject field(s) to be dealt with. If it is a relatively small field that is unlikely to develop rapidly there is no case for going to the trouble of providing a structured thesaurus - a straight-forward alphabetical list will suffice. However, when an IR system has to cover many fields, some at generic level and some at very specific levels, sooner or later it becomes essential to provide a structured thesaurus. The reason is that so many terms are required for large systems that no one can remember and appreciate the relationships between all the words involved, so that synonyms, and particularly near-synonyms creep in unnoticed.

In the case of post-coordinate systems all but Uniterm systems (ie, small systems) are of necessity mechanised in some way. As soon as mechanisation is required it is essential to draw up a thesaurus before the system can operate, although terms can still be added or deleted as necessary. In the case of pre-coordinate systems the indexing vocabulary can be developed as required. In either case structuring of a thesaurus can only be carried out when many of the terms required are known. Examples of unstructured and structured post- and pre-coordinate thesauri are set out in Appendix III, together with some 'trees' of relationships between terms. In the process of structuring, various levels of generality and specificity become obvious and their diagramatic presentation as 'trees' is extremely helpful in showing relationships and in ensuring that nothing important is overlooked.

A table of recommended procedures for developing a structured thesaurus is set out below in Table 1, while examples of structured terms together with 'scope notes' are illustrated in Appendix III.

[Note: The examples set out in Appendix III are intended only to illustrate various possibilities. Some of these were procedures tried out and used within AE&CI and are referred to in Section 4.2]

TABLE 1

PROCEDURES FOR DEVELOPING A STRUCTURED THESAURUS

Step	Description	Notes
1	Select subjects to be dealt with	Establish priorities
2	Collect terms for each subject	Sources: Existing thesauri Dictionaries Encyclopaedias Discussion with experts
3	Clarify meanings of all terms collected	Only necessary where doubt exists
4	Identify all synonyms	Decide on preferred terms
5	Identify all near-synonyms	Decide which are to be treated as a) actual synonyms b) being other terms
6	Arrange in hierarchical order with the most generic at the top and the most specific at the bottom	To determine the order for any two related words, A and B: if "all B is A, but not all A is B" then B is more specific than A, and vice versa. All such relationships can be illustrated by Venn diagrams.
7	Draw up a tree of relationships	One level at a time from the most generic downwards - see Examples in Appendix III
8	Put terms that do not fit the tree to one side	Include in thesaurus as 'related terms'
9	Decide which 'tree' suits the particular company's interests best	It is often possible to draw up 'trees' of the same terms in different ways
10	Write out the relationships of each word in the order in which printing is to be done	Use the following symbols: UF = Use for. BT = Broader term NT = Narrower term U = Use RT = Related term
11	Decide whether scope notes are necessary - if they are then add them in parenthesis	Scope notes may be required to indicate homographs eg, Crane (bird) Crane (mechanical device)

4.1.7 Hardware and formats

For any IR system to meet the requirement of maximum acceptability to users it is essential for system designers to consider what equipment (hardware) is to be used, as well as the format of presentation.

Equipment tends to be outdated rapidly today and it is not considered that a dissertation of this type should go further than to indicate broad guidelines. When the literature of information science is studied it is clear that many workers in the field have become overly concerned with theory at the expense of practice. In some instances the use of mechanical or electronic equipment appears to have become a goal to be achieved rather than a tool to be used to ensure more effective and less expensive service. It would appear that it has been forgotton sometimes that equipment such as a computer is basically a high-speed sorting aid (and an expensive one) when employed for information retrieval. Under the conditions of South African industry it is quite clear that no company will establish a highly sophisticated and computerized system initially. It is axiomatic that mechanical equipment should be the simplest that will meet current and foreseeable requirements and the various possibilities have been well described in the literature [198-201].

Generally speaking mechanical equipment is required for post- rather than for pre-coordinate systems, whereas the reverse is true as far as the format of presentation is concerned. This arises since most post-coordinate systems provide a reference number as their output - ie a document or abstract number for example. In the case of pre-coordinate systems where abstracts are included, which is the case for most such systems, presentation is virtually limited to the use of cards of varying sizes or of one or other of the microforms (microfiche, aperture card, microfilm, etc). Thus the major requirement is to decide on the form of reproduction (ie typing, duplicating, etc), and on the physical arrangement of the following (198,199):

1 Index terms:	Should all the terms relevant to each item of information be reproduced on each card, or only one term per card? Where on the cards should index terms be printed?
2 Abstracts:	Are informative or indicative abstracts to be provided? Are 'extracts' from any particular documents to be used?
3 Bibliographic information:	What is the minimum information that will serve to identify each document? Is it necessary to indicate authors? What information must be included to pinpoint the actual whereabouts of original documents?

4.1.8 Conclusions

The main conclusions to be drawn from the foregoing on the factors affecting IR system performance are as follows:

1 When an industrial concern recognises that it has accumulated, and is continuing to accumulate valuable technical information which cannot be located easily and consistently, serious consideration should be given by management to the establishment of an IR system.

2 Whatever system is established it should be user orientated and should therefore be an information retrieval system as opposed to a document retrieval system.

3 Where a general IR system is required to permit the retrieval of a wide variety of scientific and technological information from a wide variety of documents stored in different places, a pre-coordinate indexing system is most desirable. Such systems have low information output costs and permit browsing.

4 It is preferable to provide informative type abstracts for unitised types of documents that cover one subject in considerable detail (_eg_ technical reports) and indicative type abstracts for documents that cover a wider range of subjects in less detail (_eg_ books).

5 Where it is desired to retrieve information in a relatively narrow subject field which should be indexed in considerable detail, then a post-coordinate indexing system is preferable. This should be operated with the simplest equipment that will meet current and foreseeable requirements - _eg_ a Uniterm system will handle up to about 500 references comfortably, an optical coincidence system up to 20 000 references and only if the collection exceeds this figure should more elaborate (and more expensive) equipment be considered.

6 For both pre- and post-coordinate indexing systems strict control of the indexing vocabulary by means of a thesaurus is essential. As a system grows and by the time it includes upwards of 20 000 items of information, the thesaurus should preferably be structured.

4.2 Factors considered by AE&CI

While it can be stated quite categorically that all the factors discussed in section 4.1 above were not identified by AE&CI in 1962, nevertheless enough of them were carefully considered for - in retrospect - an adequate appraisal to have been made in the selection of an IR system which, it was believed, would meet the Company's requirements. These are discussed below and included the question whether abstracts were required or not, and if they were then which type, the type of indexing to be adopted, whether or not a thesaurus of acceptable indexing terms was necessary, and whether or not either or both of the sub-indexing devices of 'links' and 'roles' were to be used. (Note: Weighted-term searching was unknown at that time).

In 1962 the author was sent on an overseas study tour to investigate the current state-of-the-art as regards IR systems, and reported his findings to AE&CI's management in January 1963 [202]. During this tour some thirty visits were made to industrial concerns, learned societies and government organisations in the United Kingdom, continental Europe, Canada, and the United States and even at that time it was found that most of the concerns visited had established, or were in the process of establishing, mechanised IR systems. Some of the older ones were using tabulating cards sorted by electro-mechanical sorters but most were venturing into the use of computerised systems at considerable cost.

In 1962 there were very few electronic computers operating in the Republic of South Africa, and certainly there were none in the chemical industry. For this reason alone the possibility of establishing a computerised IR system in AE&CI within the then foreseeable future was dismissed, but there were other reasons also, including cost and the total absence in the Republic of anyone experienced in establishing and operating any mechanised IR systems for technical or other information. The possibility of developing a system based on tabulating cards and electro-mechanical sorters was considered and dismissed for the following reasons:

1 These systems were primarily document retrieval systems.
2 Since the cards were sorted by electro-mechanical means they were subject to damage and therefore had to be renewed after about fifty to one hundred runs through the sorters. This increased costs and involved the repetition of clerical chores. It also made the cost of reproducing informative or even indicative abstracts on the cards far too high, and since the cards were punched according to a code, the possibility of obliterating figures, decimal points or symbols in an abstract was an important disadvantage.

3 Users did not have direct access to information.

4 There was a definite practical limit of around 90 000 documents which could be handled reasonably by these systems. However any system which included even 50 000 documents tended to be cumbersome, since the fastest sorters could only handle 30 000 cards per hour, and these required considerable maintenance.

The alternative to a mechanised system was, obviously, a manual one and well established manually operated systems were seen in operation at the Battelle Memorial Institute at Columbus, Ohio, which appeared to come close to meeting AE&CI's needs. One of these, namely the Defence Metals Information Center (DMIC) system, had at that time been in use for twelve years and contained several million items of information, any one of which could be located within minutes. These Battelle systems have been well described in the literature [182-186] and it was evident to the author that they operated at cost levels acceptable to AE&CI (see Chapter 7) and enabled their staff to browse and to find information themselves rapidly. Thus it was decided to follow Battelle-type procedures as far as possible, but to try to make provision for future mechanisation with minimum ensuing disruption. The following conclusions were reached and procedures based on these conclusions were adopted (see Chapter 6):

1 Abstracts The system would be concerned primarily with technical reports generated by the Company and its overseas associates. Most of these reports included summaries, many of which were, for practical purposes, identical to informative abstracts. Thus, since the use of informative abstracts fitted in well with the Company traditions of self-help for information retrieval and of browsing by staff, it was decided that the system would include them in preference to indicative abstracts. It was also believed that to keep searching costs (time)

down, the system must provide sufficient information for the abstracts to act as acceptable substitutes for the original documents in the majority of instances. Again, this appeared to confirm that informative type abstracts were preferable.

The whole matter of costs is difficult since the provision of internal services, within industrial concerns at least, accountants tend to make their assessments in terms of actual direct costs, rather than in terms of direct costs plus savings which cannot be quantified accurately from their very nature. For any information storage and retrieval system the total cost of operating the system is the cost of getting information into it (input) plus the cost of searches to retrieve the same information (output). As pointed out in Chapter 3 AE&CI had operated a rather primitive retrieval system for its technical reports prior to 1962 which was based on the Kaiser classification system [18]. In this instance the input costs were very low as no abstracting was done and the indexing was inadequate - averaging between one and two concepts per report. Unfortunately no records were kept of the actual costs but at that time the time wasted by highly paid research staff in attempting to retrieve information from the system was substantial. The complaints concerning the ineffectiveness of the system and the time wasted in trying to use it had convinced management that the output costs were extremely high. The only record available is that during 1961 one research chemist reported that he carried out six searches during the month of July and that he spent six and a half hours of his time on these searches. Of these six searches two yielded positive results (ie some information was found) and four negative results. Subsequently a more experienced colleague was able to show by means of memory (and not by using the system) that one of the two searches with positive results was most incomplete and that two of the four searches with negative results should have yielded positive results.

This type of situation was sufficiently common for management to accept the case for a retrieval system with higher input costs, in order to reduce staff frustration and misleading output results even in the absence of accurate data. [Note: actual cost figures for the system adopted are set out in Chapter 7].

2 Type of indexing Virtually all the mechanised systems referred to above were using post-coordinate indexing. It was considered that this would prove to be cumbersome for a manual system because of the expected frequency of use of such terms as 'analysis', 'production', etc, and because any need for system users to personally 'coordinate' terms was unlikely to be acceptable to them - a situation which management was anxious to avoid if possible.

In addition, the use of post-coordinate indexing makes it essential to draw up a thesaurus before an IR system is established. Since this would have involved an unacceptable delay of two to three years before an AE&CI system could be established, in view of the very wide technical interests of the Company and the large number of scientific and technological disciplines which would have to be covered, the use of post-coordinate indexing was rejected. Finally, there appeared to be no obvious or major objections to the use of pre-coordinated indexing, and consequently it was decided to adopt it, if only because a thesaurus could be built up as indexing terms were required. This was in fact done as explained below.

3 Thesaurus In the Battelle DMIC system no indexing vocabulary control was exercised. But for an AE&CI system it was concluded that such control was essential in view of the prevalence of synonyms in the field of chemistry, and also because of the possibility that the system might be mechanised at some future time. Experimental runs showed that the construction of a pre-coordinate based thesaurus would be no more

difficult than a post-coordinate one would be.

As stated in Section 5.2 it was decided to process the existing collection of reports (ie the backlog) into the system in addition to current accessions. It was agreed that this would be done as quickly as possible and a total of R20 000 was earmarked for expenditure on salaries over a five year period for this specific purpose. The system started up with two full-time abstractors/indexers in March 1963 and this number was built up to a staff of five by July 1964 and this complement was maintained until 1967. By this time it was clear that the backlog would have been worked through completely within an acceptable period of a further two years if the complement was run down by wastage (ie resignations, transfers, etc). By mid-1969 the whole backlog of approximately 30 000 reports had been examined and, of these some 5 000 had been discarded as not being of sufficient value to warrant processing.

By the middle of 1964 the input rate to the system exceeded 700 reports per month and this was coupled with a demand for new thesaurus terms and cross-references of, at one stage, almost 1 000/month. From the beginning a control system was introduced whereby:

1. Each abstractor/indexer was encouraged to consult all available published thesauri and technical dictionaries, and, at the same time to consult the numerous subject specialists on site.
2. Regular meetings of all abstractor/indexers were held with their supervisor as Chairman. Each term was considered for its appropriateness and potential degree of interest to the Company as a technical subject. This helped to avoid the adoption of near-synonyms and also to high-light gaps in specific subject areas. These decisions led to the development of an indexer's guide, much of which is set out in Appendix II.
3. It was to be expected that arguments would develop

and the Chairman made the final decision when necessary, always consulting the potential users of the system (ie the subject specialists) whenever necessary.

This system worked satisfactorily with a strong Chairman, but could well have resulted in difficulties with a weaker personality.

The decision to use the same thesaurus for indexing technical correspondence (CORIS) resulted in the addition of remarkably few new terms - 48 in the first three years - while the extension to the Library catalogue (LIBRIS) resulted in the addition of 530 new terms between mid-1967 and the end of 1968. This was a surprisingly low figure at first glance but in fact it was not so, as the thesaurus had developed to approximately 29 000 entries - including 'see' and 'see also' entries - by the end of 1968. Currently the total is approximately 30 000 entries while the rate of addition fell to 155 for 1970 and has fallen to four/month since the early part of 1971.

It should be clear from the foregoing that the process of developing the AE&CI Thesaurus was relatively expensive although these costs were spread over six years. Subsequent maintenance of the thesaurus has become a very low cost operation as a part-time job for one nominated science-graduate Information Officer; in 1973 the cost was approximately R50 per month.

4. 'Links' and 'Roles' The use of links was rejected from the beginning on the grounds that the use of pre-coordinated indexing terms made them entirely unnecessary. There has been no reason to change this opinion

On the other hand it was decided to use role indicators for two reasons. Firstly it was believed that the additional divisions of subjects which they introduced could prove useful in saving the system user's time when the total number of cards on a specific subject exceeded about twenty. Also, if means for mechanising the system were to be developed then it was expected that roles would assist in reducing 'false drops'. Secondly it was believed that the additional thought required by the indexer in applying roles would be a discipline which would assist him to be consistent.

In fact experience has shown that the system users seldom use the roles as a searching aid, but that their presence does not appear to confuse users, nor to obstruct them in any way. On the other hand the majority of indexers, when questioned, gave it as their considered opinion that their indexing was more consistent as a result of having to apply roles, since this requirement forced them to consider the concepts to be indexed more carefully and more thoroughly. As the use of roles slowed the indexing process, it added to indexing costs and so an attempt was made, therefore, in 1968 to test this belief in their value.

Unfortunately while the details of this test were worked out with the aid of a mathematical statistician they were not recorded properly and the author is thus dependent on memory. The test comprised using two indexers to index the same 25 reports independently, applying role indicators. Two other indexers repeated this process without using roles, but indexing the same 25 reports. The results were analysed and, taking into account and discounting the selection of slightly varying ranges of terms between indexers - ie by concentrating on terms common to each pair of indexers - it was found that the consistency when roles were used was significantly better than when they were not. [Note 1: 'Significantly' is used in the statistical sense. Note 2: A more recent attempt to repeat this experiment so as to provide proper and complete details for this thesis have been foiled by the fact that only one experienced part-time indexer is now employed for TIRSS].

Whether or not the test should have been conducted on a larger sample is probably arguable, and all that can be said now is that the statistician concerned appeared to be satisfied as to the validity of the findings. On the basis of this test, and on the continued belief by successive indexers in their value as an 'indexing discipline' the use of roles has been retained. This belief by indexers which has been consistent since 1963, and continued by indexers who were not in the employ of the Company at the time of the test, is perhaps circumstantial evidence in favour of the retention of roles.

Chapter 5 ON SELECTING A SYSTEM

5.1 General considerations

The criteria to be considered for the design of IR systems and the factors which affect their performance have been considered in some detail. It is hoped that this detail will assist those persons in industry who may be faced with the task of designing practical IR systems for their own particular companies, each with its own unique problems. However it is also desirable to provide guidance from a broader and more general point of view for the benefit of management.

Generally speaking, industrial management within the Republic has become aware during recent years of growing information transfer problems within its own organisations. Some managements have realised that the provision of a technical library alone is insufficient to solve these problems. These same managements have usually heard of information science but are often and understandably a little wary of it, since it has become surrounded by a mystique which even extends to the concept of IR systems. This mystique is completely unjustified but it has probably arisen because of the fairly frequent articles in technical journals which describe or refer to the so-called 'information explosion'. These articles frequently suggest that the only way to handle this explosion is to set up a large, and expensive, computerised IR system based on post-coordinate indexing [(203-207)]. It is not suggested that there is anything wrong with this approach in the right context nor with the fact that much of the literature of information science itself is taken up with various aspects of similar systems or by highly academic discussions.

This has been particularly unfortunate in its effects in the smaller industrial countries, such as South Africa, where the process of industrialisation has only proceeded rapidly since World War II. Nevertheless there is a real need by the many relatively small concerns in this country for simple practical systems as has been established, for example by van Houten (21). There has been to date an almost total lack of local training facilities in the simple practical techniques of information science which can be applied at relatively low cost and on a small scale. Admittedly a certain amount of information has been published in the South African technical press, none is recent and it appears that very little of it has been of value as a guide to management. The reason for this is that this published information has mainly been addressed to a very limited industrial readership and not to management itself. Examples include articles which have appeared in 'South African Chemical Processing'[94,208], in 'South African Libraries', (the official journal of the South African Library Association)[209-213], in Proceedings of the Southern Africa Regional Symposium on Scientific and Technical Information, 1968,[214-224] and in the Proceedings of the 1969 Conference of the Library Association [225,226].

The only article published locally on information problems which comes close to being addressed specifically to management appears to be that by F G Hill, then Technical Director of Rand Mines Ltd [227]. Unfortunately, it must be recorded that so far the South African mining industry has been extremely conservative in its approach to technical information services, and Hill's article apparently fell on stony ground.

As a guidance to management in the selection of an IR system, attention is particularly drawn to three publications (references 228-230), by R G Breadmore, J C Costello and Eugene Wall respectively. In addition, attention is drawn to the subject of the 'systems analysis' approach to the selection of an IR system; this technique has been expounded in recent years by a number of writers and attention is drawn to articles by R E Nance [231] and R R Segel [232]. All five of these papers are concerned with the problems of the (by world standards) small organization or company. They illustrate the importance and value of a systematic approach to assessing and analyzing these problems of information storage and retrieval. Finally, and for purposes of emphasis, the following points are set out in some detail as general guidance to management.

5.1.1 Small scale operation

It is appreciated that there are distinct differences in the internal organization of large and small companies. Large companies have found it necessary to equip themselves with numerous specialized services, and, in general it is they who have developed the techniques of present day IR systems.

It is risky to assume that all of those services which work well on a large scale can be made to work equally well on a small-scale. Nevertheless, it must be emphasized that in the case of IR systems, the principles outlined in Chapters 3 and 4 are equally applicable on small or large scales. The scale is more likely to determine the extent to which mechanical or electronic aids are used than anything else.

The primary task when investigating information transfer problems is to determine the types and quantities of documents to be processed. From this it can be decided whether more than one system is necessary; the important point is to select the simplest system which will meet user requirements. Invariably, the simpler the system the less costly it is to establish and maintain.

5.1.2 The absence of any one ideal system

It must be understood fully that there is no one ideal system that will meet all the requirements of all companies, or even all the requirements of one company. Nor is there any fixed performance value for any type of IR system. Rigby, a consulting information scientist who has regular direct contact with many smaller companies in the United States [233], has pointed out that:

'........ It would be useful to dispel for ever the fallacy so often promulgated that any one system .... is the only system, or a useless system ....'

It has been stressed several times in this dissertation that no two concerns have identical interests, etc, and that their requirements of an IR system are not identical either. Each situation has to be considered on its own merits.

5.1.3 Building user confidence

The following is an extract from the report written by the author in January 1963 and referred to in Chapter 4, section 4.2 (202): this extract appears to be relevant to this Chapter:

'Overseas practice draws a distinction between information and document retrieval and one of the disadvantages of mechanised systems as commonly operated is that they lead the searcher to documents only. A considerable intellectual effort is required to phrase questions with the precision necessary to avoid subjecting the searcher to a flood of documents, while at times it is quite impossible to phrase questions precisely. It is well known also that when searching for information, the technical man is anxious to be led to relevant information which may help him to crystallise his thinking as well as suggest new lines of thought, and it is concluded that this is an extremely important aspect of information retrieval. User requirements must be taken into account and many overseas companies have set up internal Technical Information Committees to ensure that their systems are developing in accordance with the requirements of their staff as well as in efficiency. These Committees serve another purpose in that they assist in developing confidence in the system provided. Experience in the USA and in Europe is that staff will not abandon expensive and inefficient private filing systems until their confidence in a company system has been won. This confidence cannot be achieved by managerial edict and it is necessary to 'sell' the system adopted to the users.'

Probably the most significant statements in the above extract are in the last two sentences, which have been underlined. Subsequently experience showed that the most satisfactory way to 'sell' systems to potential users is to let selected key personnel (see Footnote) find information for themselves which they already know to be in some specific document on the same site. Every time such personnel carry out a search successfully they not only accept use of a system for themselves but they tend to convince others of its value. The implications of this observation are twofold:

(a) Firstly, while not essential, it is usually highly desirable to process an existing collection of documents into a new IR system, working backwards in time so that the most recent additions - ie those most likely to be remembered and to be referred to frequently - are processed first. From the start-up date new accessions must obviously be processed as received.

(Note: For the sake of brevity, a collection of documents already in existence before the start-up date of an IR system will be referred to from this point onwards as 'a backlog').

The advantages of working through a backlog are:

(i) An opportunity is provided to cull the collection and thus get rid of any items that are no longer of value.

(ii) User confidence that the system can provide useful and valuable information is established considerably sooner than if it is decided only to process documents received after the start-up date.

[Footnote: Wherever a group or team of people work together on research and other projects, it is often found that one of them is consulted regularly by his colleagues as being likely to know, or to know where to look to find the answer to a problem. Such potential users should be selected as 'key personnel' in aiding the 'sale' of IR systems to others, whenever they can be identified. These people are regularly referred to in the literature of information science as 'the gatekeepers' (91,234-238) and it is of interest to note that this concept was first conceived by Francis Bacon in 'New Atlantis' (239), published in 1627. Although Bacon calls them 'depredators', 'pioneers' and 'compilers' the basic idea is his].

(iii) A worthwhile return on the investment made in setting up an IR system is achieved very much sooner. In the case of a big system, this return may be achieved right from the time it is made available to staff for their use - see (b) below - instead of after a period which may extend to two or three years.

(b) Secondly no system should be made available for general use until some agreed percentage of recent backlog (as well as new accessions received during what is known as the 'start-up period') have been processed into a system. The start-up period may comprise any time interval, but is likely to be between three and eighteen months depending on:

The size of the document collection,

the extent to which access to backlog is required,

the rate at which new accessions are received,

the number of abstracters/indexers available,

the skill and experience of abstracters/indexers.

5.1.4 Conclusions

The conclusions set out at the end of Chapter 3 and particularly at the end of Chapter 4 provide considerable guidance in determining the most suitable IR system for the special circumstances in a given industrial concern. However no guidance can cover every unusual exceptional circumstance. In such instances a company should consult other industrial concerns or organisations in the same scientific or technological field, or, if available, a consultant in information and library science.

5.2 Factors considered by AE&CI

AE&CI paid particular attention in designing an IR system to the problems of building user confidence as rapidly as possible.

The question of whether or not to process the then existing collection of reports (*ie* the backlog) in addition to the current accessions was considered carefully along the lines set out in sub-section 5.1.3. After considerable discussion with senior Research and Development staff it was decided that it was essential to process the backlog but that it would be tackled as follows:

1 The subjects of greatest technical importance to the Company would be dealt with first.

2 Each subject would be worked through in reverse time order - _ie_ the most recent reports would be processed first.

3 The rate of technical advance in each subject would be taken into account in determining how much of the backlog would be processed, _eg_ explosives technology is, understandably, rather conservative, so that with the exception of trivia virtually all the backlog would be processed. In the case of pesticide technology it was extremely unlikely that the arsenical formulations of the 1930s would ever be used again - it was decided, therefore, not to process any of these.

In 1963, when the present AE&CI system was started up, it was impossible to find trained technical abstractors/indexers in the Republic. Thus it was necessary to recruit and train staff for this work and it was decided that since high standards of technical accuracy were required only chemists or chemical engineers would be employed for this purpose. Further, it was decided that the system would not be made available for general use until all the current receipts of new reports could be handled and, in addition, about 10% of the backlog had been processed. Two abstractors/indexers were taken on in March 1963 as stated in Section 4.2 to start the system selected, and this staff was increased gradually to five. The system was made available for general use when the above target was reached after 15 months - _ie_ in July 1964; the progress made was considered by management to have been satisfactory.

All the abstractors/indexers were given intensive and personal training by the supervisor of the system. This training consisted of a thorough understanding of the objectives and of the system itself, as well as practice in the analysis of the contents of documents for the production of technically accurate and clear and concise informative abstracts.

1 Company experience indicated that the system to be adopted must be an information retrieval system as opposed to a document retrieval system. Further it was required that users should have direct access to the information in it on a self-help basis and should be able to browse at leisure through it.

2 Since it was expected that many of the system users would be unfamiliar with, or inexperienced, in the subjects on which information would be sought, it was concluded that document titles alone would be inadequate and that abstracts would be required. If abstracts were to be provided it was concluded that these should be informative rather than indicative, so as to reduce the 'look-up time' by providing sufficient information for the user to decide whether or not he really needed to refer to the original documents.

3 It was also clear that some strict form of indexing control would be most desirable in view of the extreme prevalence of synonyms and near-synonyms in all sciences and technologies. It was agreed that the need for a system was so urgent that any initial requirement for a fairly complete thesaurus before start-up could not be met. This meant that post-coordinate indexing had to be avoided since not only would an initial thesaurus be required, but in its only known manual application - ie in the Uniterm system - the quantitative limitations for convenient operation were too severe. For the above reasons pre-coordinate indexing was selected as being the more suitable. A thesaurus could be built-up as the system developed and grew, and by the time it reached a substantial size there did not appear to be any major obstacle to conversion to some form of mechanised procedures.

4 As indicated already in 1962/63 there were no computers

available in the South African chemical industry. As an IR system was required immediately this automatically ruled out a computerised system, quite apart from any question of economics or efficiency.

5 A mechanised IR system based on the electro-mechanical sorting of tabulating cards was considered. This was rejected in view of overseas experience, particularly in Union Carbide Plastics Co of Elizabeth, New Jersey. That company had had 14 years experience of a system of this type by 1962 and was able to point out that the tabulating cards used had to be replaced after 50-100 sortings as a result of physical damage. In addition the speed of the fastest sorters (which were the most expensive) was of the order of 30 000 cards/hour - in practice this meant that these systems were limited to a total size of 50 000 - 90 000 cards but were found to be cumbersome even at the lower figure. Finally, such systems require the initial preparation of a thesaurus, just as computerised systems do, and since this would have involved a delay of about two years it was not acceptable.

6 A mechanised system based on the use of optical coincidence cards was also considered. These systems were just being developed at that time and it appeared that while a thesaurus was necessary before they could operate, the practical quantitative limits were even more severe than for systems using electro-mechanical sorters; the upper limits were about 30 000 - 40 000 documents although a total exceeding 20 000 made the system rather tedious to operate.

7 None of the systems referred to above in 4 to 6 appeared to be intended for general operation on a self-help basis, with the possible exception of one based on optical coincidence cards. It was also evident that if none of these was to be adopted, then the Company system would have to be a manual one which could, perhaps, be automated in some way in the future should this ever become desirable and economic.

8 In view of the urgency of the need for AE&CI to have its own IR system, and to have it operating as soon as possible, and owing to total inexperience within the Company of the economics of such systems it was too difficult to attempt to estimate cost-effectiveness before the selected system was in operation. The system selected is described in Chapter 6 and in Appendix V, together with the actual starting-up and system maintenance costs. This selection was based on the author's overseas visits in 1962, which have been referred to already [202], coupled with a conviction, along with T J Allen [91], that the human being was still the most effective source of information. Thus it appeared that a manual card system which permitted browsing and self-help by people would best meet AE&CI's requirements.

Chapter 6

THE DEVELOPMENT OF AN INTEGRATED IR SYSTEM: A CASE HISTORY

The following is an account of the IR systems developed by AE&CI Limited and of their integration. The whole is, to the best of the author's knowledge, unique at least in so far as manual systems are concerned, in the way in which information in internally produced technical reports, technical correspondence and the contents of substantial technical libraries have all been brought within its scope. However the system in use by the Defence Metals Information Center (DMIC) at the Battelle Memorial Institute, Columbus, Ohio (see section 4.2) provided the basis from which AE&CI's first system was developed. This system was given the acronym - technical information retrieval and storage system.

6.1 TIRSS

In mid-1964, by which time about 6 000 documents had been processed, TIRSS began to come into formal use and rapidly received widespread acceptance.

It was, and still is, an entirely manual IR system based on the use of informative abstracts and pre-coordinated indexing terms. It is of interest to note that the introduction of metrication to the Republic of South Africa was anticipated and from its inception ISO A5 cards, 148 x 210 mm, were used. The system set out to include various types of documents and also standardised bibliographic information on each card to ensure uniformity of presentation and adequate document identification and location. The bibliographic information is followed by an informative abstract, or in certain instances verbatim extracts from the original document. Finally, all the indexing terms selected are included on each card, together with the appropriate role indicators. The system is described in detail in Appendix V and information is included on the preparation and filing of the cards, as well as on guidance provided for the abstractors/indexers.

By mid-1974, ie after approximately ten and a half years of operation, about 75 000 documents had been examined for possible inclusion in TIRSS. Of these 31 000 had actually been found to be worth including and the average number of cards had remained constant from the time of start-up at twelve per document - ie the system includes 370 000 cards.

Study of the details of the system shows that it offers considerable advantages to the user along the lines stated as being desirable in earlier chapters of this dissertation. No evidence has been found that there are any practical quantitative limits or other limitations to the total size of the system. At the current rate of input it is expected that TIRSS will contain about 480 000 cards by 1980 and it is believed that it will be just as simple and quick for the users to locate information then as it has been to date. However it is unlikely that the floor space required for the storage of this number of cards in conventional cabinets will be acceptable by that date and this aspect is discussed in sub-section 6.1.3.

From the inception of the system to mid-1974 approximately 10 500 searches had been made in it. The time taken to carry out searches varies considerably and is dependent on numerous factors, such as the complexity of the technical information sought, the degree of exhaustiveness of a search, the technical competence and mental outlook of the searcher, and so on. A record of 100 consecutive searches during February and March 1970 showed that the shortest search took two and a half minutes and the longest three hours nine minutes. Identical tests were carried out in 1973 and 1974 when the average search times were 19 and 16 minutes respectively. Thus it appears that the steady increase in the number of cards has had little or no effect on search time.

While the factual details of TIRSS are set out in Appendix V as already stated, the following additional comments are apposite:

6.1.1 Complex questions

A claim frequently made by the designers of mechanised

By mid-1974, ie after approximately ten and a half years of operation, about 75 000 documents had been examined for possible inclusion in TIRSS. Of these 31 000 had actually been found to be worth including and the average number of cards had remained constant from the time of start-up at twelve per document - ie the system includes 370 000 cards.

Study of the details of the system shows that it offers considerable advantages to the user along the lines stated as being desirable in earlier chapters of this dissertation. No evidence has been found that there are any practical quantitative limits or other limitations to the total size of the system. At the current rate of input it is expected that TIRSS will contain about 480 000 cards by 1980 and it is believed that it will be just as simple and quick for the users to locate information then as it has been to date. However it is unlikely that the floor space required for the storage of this number of cards in conventional cabinets will be acceptable by that date and this aspect is discussed in sub-section 6.1.3.

From the inception of the system to mid-1974 approximately 10 500 searches had been made in it. The time taken to carry out searches varies considerably and is dependent on numerous factors, such as the complexity of the technical information sought, the degree of exhaustiveness of a search, the technical competence and mental outlook of the searcher, and so on. A record of 100 consecutive searches during February and March 1970 showed that the shortest search took two and a half minutes and the longest three hours nine minutes. Identical tests were carried out in 1973 and 1974 when the average search times were 19 and 16 minutes respectively. Thus it appears that the steady increase in the number of cards has had little or no effect on search time.

While the factual details of TIRSS are set out in Appendix V as already stated, the following additional comments are apposite:

6.1.1 Complex questions

A claim frequently made by the designers of mechanised post-coordinate systems is that they permit complex questions phrased at a very specific level to be answered. As an example consider the following question, which contains eight concepts for coordination:

'What information is available on fast kinetic analysis methods, such as stroboscopic pulse radiolysis systems, for detecting transient light-absorbing species formed in chemical solutions by individual fine structure pulses from a 40 MeV (or bigger) linear accelerator?'

If a post-coordinate IR system has been well designed then the above claim is well-founded, provided that the indexers did in fact index all the concepts at a very specific level. However, the ability of a system to answer a question such as the above is not the only criterion, or even the criterion of greatest importance in system design.

Firstly, the very complex and very specific question comprises only a small percentage of the total questions asked even in the largest industrial concerns in the United States. In 1969 at E I du Pont de Nemours & Co the author was advised that such questions represent not more than five percent of the total in their central research department, which is the department where most such questions originate. The author was also given the same figure during a visit to the same Company in 1962, so that there is no evidence of any tendency for such questions to grow as a percentage of the total volume. In any event, it is often overlooked that multi-concept questions can readily be broken down to a number of questions, each based on fewer concepts, especially by experts in the technical field concerned.

Secondly, in the average South African company of today, and even in the larger ones, technical staff rarely ask questions of the type under discussion. Thus it would be quite uneconomic to develop a large-scale IR system within South African industry as it is today on the basis that it should be able to answer questions of this nature. Whether or not there is a need for one such system as a service to the nation is an entirely different matter and is not considered in this dissertation.

The author's experience to date has been that a pre-coordinate system, such as TIRSS, is capable of answering the vast majority of questions addressed to it. Where a need exists for detailed indexing at a very specific level in a narrow specialised technical field, then a post-coordinate optical coincidence IR system will meet these requirements satisfactorily at low cost. In either case, as pointed out at the end of Chapter 5 each user of an IR system has very personal needs, but perhaps instinctively, most users tend to break-down complex questions themselves, so that either pre- or post-coordinate systems should equally well be able to meet their requirements.

6.1.2 Thesaurus control

It was pointed out in Chapter 4, sub-section 4.3.2, that the thesaurus for use in a pre-coordinate indexing system can be built up as the system grows and develops. However, the bigger it gets the more frequent indexing staff changes are, then the more essential does strict control of additions to the thesaurus become. The reasons for this are that the bigger the thesaurus the more difficult does it become for an indexer to remember all the terms relevant to one subject, and the greater the staff turnover is, the more difficult does it become to maintain reasonable indexing consistency. More and more near-synonyms tend to creep in until drastic pruning action becomes necessary. It was believed by 1969 that the most satisfactory solution to this problem would be to 'structure' the thesaurus - cf the examples of unstructured and structured pre-coordinate thesauri set out in Appendix III.

A pilot scheme was initiated in 1969 and the example of a structured pre-coordinate thesaurus in Appendix IV is taken from it. This work was done in order to assess the effectiveness of carrying out structuring and to assess the magnitude of the task and its particular problems. After drawing-up 25 structured trees at generic levels and 50 at specific levels it was realised that the task was extremely time consuming and expensive; it was estimated that to structure the whole thesaurus would require some two man-years work at science graduate level, and at a cost of R20 000 between mid-1970 and mid-1972. It was concluded that this expenditure was not justifiable in terms of cost-benefit particularly in view of an urgent requirement to reduce any expenditure which added to the Company's overhead costs. However the technique demonstrated that in areas of subject uncertainty and confusion it provided a valuable means for the clarification of real indexing needs. As such, the technique has been applied regularly as needed and has proved its value by clearing up many indexing fields which had become untidy. During the four years of its use it has resulted in the moving of approximately 1 800 indexing terms to 'see' and 'see also' entries, rather than in the complete elimination of these terms. The effect has been that the thesaurus has become easier to use and the number of lightly posted terms has been reduced significantly.

6.1.3 Storage space for TIRSS

The 370 000 A5 cards are stored in 22 filing cabinets and the total space currently required (including storage and room for searchers to work, browse, etc) is 30 square metres. It may well be asked whether the additional space which will be required in the future can be provided economically. The answer is that at some date, which has still to be decided, it will become more economic to transfer the total system to microforms. A procedure for this has already been worked out and can be implemented at short notice. It is intended to microfilm all the existing cards onto 16 mm film and to insert this into A6 sized film-jackets by subject and by role indicator. The film-jackets will comprise the master index, and microfiche copies will be provided as and when necessary. Examples of both a film-jacket and a microfiche copy are set out in Appendix VI.

The advantages of the proposed procedure are expected to be:

(1) Savings in storage space of 95%

(2) Savings in storage equipment

(3) Easier control over the total systems from a security aspect

(4) Improved file integrity

(5) Improved user convenience as far as ensuring that all available information on a given subject is assembled on one or two sheets of film

The potential disadvantage is that the user will require access to a microfiche viewer and/or a reader/printer. But in fact TIRSS users are already very familiar with the use of microforms in general and any disadvantage is no more than potential.

As far as the costs of maintaining the system on microforms as opposed to the present card methods are concerned, there should be a reduction in the total of about 10%. The cost of filming, processing the film and inserting it into film-jackets should be more than balanced by the avoidance of any need to type onto offset-litho masters, to use the expensive once-through typewriter ribbons necessary for best results from this process and to duplicate the requisite number of cards in each instance. As at mid-1974 the current cost of maintaining TIRSS was approximately R9 500 pa but owing to the current high rate of inflation little actual change is expected in 1975; the details as to how the total is made up is set out in Chapter 7.

6.1.4 Investigations into the evaluation of TIRSS

The whole question of evaluating IR systems is a difficult one and after much discussion with users comprising informal sessions with 21 individuals in groups of two and three, and all in the Research Department, it was appreciated that nearly all of them were anxious to please. It was realised that it is not difficult to steer people to express the kind of opinions one would like to hear, which, of course, are merely a reflection of ones own opinions and not in the least objective.

An earlier questionnaire which was completed by 266 users over a two month period in 1967 had indicated that the vast majority found the system easy and quick to use and believed that well over 90% of the information in the system was being retrieved when searches were carried out. The bias of this approach was not appreciated for some time, but when this was pointed out a more objective test was devised in 1973 with the assistance of a mathematical statistician.

The basic requirement for the test was to have a method which would determine the efficiency of TIRSS in retrieving information that had been processed into the system, ie its recall efficiency, and its efficiency to sift out the relevant information from the non-relevant, ie its precision or relevancy.

A set of test questions were prepared as follows:-

a) 105 reports known to be in the TIRSS system were picked at random. Of these, 34 concerned Explosives, 27 Plastics, 30 Chemicals and Chemical Engineering, and 14 Analytical Chemistry.

b) An abstract of each, with the list of index terms omitted, was handed to nine senior members of the technical staff in the Experimental Groups of Research Department, all of whom were knowledgeable in the subject fields covered.

c) These individuals were asked to prepare questions relevant to the most important information in the documents, with the instruction that the questions should, as far as possible, not be a re-phrasing of the titles.

d) It was suggested to them to phrase the questions from the point of view of someone who was going to do further research on the subject matter of these reports. (Note: None of the questions, which are all listed in Appendix VII, was edited afterwards).

6.1.4.1 Search procedure

Three groups of volunteer searchers with different technical backgrounds and subject specialization agreed to take part in the test:

- two graduate Information Officers with considerable experience of indexing
- three Senior Information Assistants with considerable experience of literature searching but with no indexing experience
- seven graduate Research Officers from different experimental groups, with Company service ranging from about 3 months to 3 years.

The Information Officers were each given 52 test questions and each set of questions covered all the different subject fields dealt with; the Senior Information Assistants were given

35 questions each and the Research Officers were given 15 questions each; in these latter cases the questions were limited to their respective fields of subject specialization.

Each searcher was asked to follow the instructions set out below. In order to be quite sure that each understood what was required of them a 'worked' example of question and its analysis was provided:

'Instructions

a) Select relevant index terms from the Company thesaurus to cover all the concepts in a question.

b) Consult the TIRSS cards and modify the search programme if necessary.

c) Enter the selected thesaurus terms opposite each question.

Notes:

1 All terms selected should be related to the questions.

2 Total time allocated: a maximum of a half hour per search programme.

3 There are documents in the system for each question.

4 Two factors are being measured; recall and relevancy. A very broad selection of terms usually improves percentage recall but lowers percentage relevancy. The two together determine the efficiency of the system and/or the search programme.

The worked example provided was based on the following TIRSS card

PROGRAMME NO.: AER 1/CP32

REPORT NO	: AERD 0301/A	DATE: 28.10.1970 027901
AUTHOR	: C.M. STANDER, Research Department, AE&CI	
TITLE	: THE ELECTROLYTIC REMOVAL OF IRON FROM ALUMINIUM SULPHATE SOLUTIONS (5pp.)	
AVAILABLE	: Central Report files, Research Department, AE&CI	

ABSTRACT

A mercury cathode was used to electrolytically remove iron from aluminium sulphate solutions down to a level of 25 ppm of Fe in a stagnant cell. Current efficiencies of 20% were obtained at current densities of 0.5 kA/m^2, using a potential of 7 V. It was found that for each cell there is an optimum anode to cathode gap, which is not necessarily the smallest gap. Various aspects of cell design are discussed.
Recommendations. A model cell should be built in which materials of construction can be tested. The current/temperature relationship for a technical cell should be calculated. The effect of cooling upon the current density should be evaluated. The specification for iron in aluminium sulphate solution should be established. The cost of incorporating the proposed process in the aluminium sulphate plant at Modderfontein should be evaluated.

INDEX TERMS

ALUMINIUM SULPHATE SOLUTION -PURIFICATION -4
ANODE/CATHODE GAP -6
ANODES (LEAD) -4
CATHODES (MOLTEN METAL) -4
CURRENT DENSITY -9
CURRENT EFFICIENCY -7
ELECTRODES - MATERIALS (CONSTRUCTION) -9
ELECTROLYSIS -4
IRON REMOVAL -4

Question: What materials of construction can be used for electrolytic cells to remove iron from alum solutions?

CONCEPTS	INDEX TERMS
1 Materials for constructing electrolytic cells.	ELECTRODES - MATERIALS (CONSTRUCTION)
2 Removal of iron from alum solutions	ALUMINIUM SULPHATE SOLUTION PURIFICATION
	*ELECTROLYSIS
	IRON REMOVAL

* Note: The term 'Electrolysis' is an example of a very broad or generic term, which, in this case will retrieve the required information but will at the same time retrieve much non-relevant information. By looking at the TIRSS cards filed under this term the searcher should be able to decide for himself whether the term is worth including in the search programme.'

Physical retrieval of the information was a clerical operation and was done by simply pulling out all the cards filed alphabetically under the selected index terms for each question. Copies were made of all the cards withdrawn and arranged in batches under each question. From these the efficiency of the system in terms of recall and relevance was determined.

6.1.4.2 Determination of efficiency

As pointed out in Cha er 3 the operating efficiency of an IR system depends upon its performance as regards recall and precision. Recall ratio equals $\frac{100\ R}{C}$, where C equals the total number of documents in the collection which are relevant to the questions put to the system, while R equals the number of those relevant documents retrieved in a single search; precision ratio equals $\frac{100\ R}{L}$, where L equals the total number of documents retrieved in a single search.

Question: What materials of construction can be used for electrolytic cells to remove iron from alum solutions?

CONCEPTS	INDEX TERMS
1 Materials for constructing electrolytic cells.	ELECTRODES - MATERIALS (CONSTRUCTION)
2 Removal of iron from alum solutions	ALUMINIUM SULPHATE SOLUTION PURIFICATION
	*ELECTROLYSIS
	IRON REMOVAL

* Note: The term 'Electrolysis' is an example of a very broad or generic term, which, in this case will retrieve the required information but will at the same time retrieve much non-relevant information. By looking at the TIRSS cards filed under this term the searcher should be able to decide for himself whether the term is worth including in the search programme.'

Physical retrieval of the information was a clerical operation and was done by simply pulling out all the cards filed alphabetically under the selected index terms for each question. Copies were made of all the cards withdrawn and arranged in batches under each question. From these the efficiency of the system in terms of recall and relevance was determined.

6.1.4.2 Determination of efficiency

As pointed out in Chapter 3 the operating efficiency of an IR system depends upon its performance as regards recall and precision. Recall ratio equals $\frac{100\ R}{C}$, where C equals the total number of documents in the collection which are relevant to the questions put to the system, while R equals the number of those relevant documents retrieved in a single search; precision ratio equals $\frac{100\ R}{L}$, where L equals the total number of documents retrieved in a single search.

a) Recall ratio

The determination of recall efficiency was based on the success or failure in retrieving the source documents on which the questions were based. If one or more terms chosen in the search programme matched any of the terms selected in the indexing of the source document, the search was rated as successful. The recall ratio can thus be calculated as

$$\frac{\text{Number of source documents retrieved} \times 100}{\text{Number of source documents in system}}$$

= % Recall

b) Precision ratio

In this test, two types of precision ratios were determined: document relevance and indexing relevance; the one measuring the number of relevant documents retrieved and the other the number of relevant index terms selected in the search programme.

b-i) Document relevance

Copies of all the TIRSS abstracts which were retrieved under the test questions were handed to the nine senior staff who were responsible for preparing the original search questions. They were given the following guidance on how to judge relevance:

'Please indicate the order of relevance of the attached TIRSS cards in relation to the "Source documents" (ie the TIRSS cards on which the test questions were based).

1 First order relevance - as useful as the source document from the point of view of someone wanting to do further research on the subject of the question.
2 Second order relevance - of some use for doing further research on the subject of the question.
3 Third order relevance - of no use for doing further research on the subject of the question.'

Document relevance was thus calculated as

$$\frac{\text{Number of relevant documents retrieved} \times 100}{\text{Total number of documents retrieved}}$$

= % Document relevance

b-ii) Indexing relevance or term relevance

Indexing relevance is the ratio between the number of relevant terms, ie those that were successful in retrieving the source documents, and the total number of terms selected. Thus,

$$\frac{\text{Number of relevant terms selected} \times 100}{\text{Total number of terms selected}}$$

= % Indexing relevance

6.1.4.3 Test results

In the analysis of the test results a number of comparisons were made with the object of determining whether there were any significant differences which would indicate weaknesses in the operation and effectiveness of TIRSS.

Using the variables recall and precision the following comparisons were made:

1 Overall comparison between the subject fields of Plastics, Chemistry and Chemical Engineering and Explosives.
2 Comparison between Information Officers, Senior Information Assistants, and Research Officers in the different subject fields.
3 Comparison between the different types of searchers ie totalling all the results obtained respectively by the Information Officers, Senior Information Assistants, and Research Officers.
4 Comparison between the different subjects, ie totalling all the results obtained respectively in the fields of Plastics, Chemistry and Chemical Engineering, and Explosives.
5 Comparison between Indexing relevance and Document relevance, ie totalling all the results obtained respectively in the fields of Plastics, Chemistry and Chemical Engineering, and Explosives.

Table 1

Summary table of the main results

Searchers	IO's	SIA	RO's	IO's	SIA	RO's	IO's	SIA	RO's
Recall	72,0	72,0	70,4	82,5	74,4	66,7	79,4	76,5	88,2
Indexing Relevance	51,5	30,6	48,1	52,7	45,2	44,2	47,8	47,4	52,6
Document Relevance 1st + 2nd order	35,5	31,3	37,1	48,8	50,7	50,7	52,1	50,4	53,5
Document Relevance 1st order	22,7	19,2	28,1	33,9	35,5	35,7	27,5	27,7	26,9
	PLASTICS			CHEMISTRY AND CHEM ENG			EXPLOSIVES		

IO's = Information Officers

SIA's = Senior Information Assistants

RO's = Research Officers (3 experimental groups)

Table 2

Summary table for recall

Subject field	% Success
Plastics	71,4
Chemistry and Chemical Engineering	74,6
Explosives	81,6
Total for all subjects	76,2

Table 3

Summary table for indexing relevance

Subject field	% Success
Plastics	41,1
Chemistry and Chemical Engineering	47,5
Explosives	49,6
Total for all subjects	46,3

6.1.4.4 Statistical analysis

The mathematical statistician suggested the use of the Chi squared test in order to test for independence between the various attributes of the following tables - thus the significance or otherwise of the differences in percentage success was established:

a) Comparison between Information Officers (IO), Senior Information Assistants (SIA), and Research Officers (RO), in the different subject fields

i) Recall

Subject field	% Success			Significance
	IO's	SIA's	RO's	
Plastics	72,0	72,0	70,4	Significant at the 2% level
Chemistry and Chemical Engineering	82,5	74,4	66,7	No significant difference
Explosives	79,4	76,5	88,2	No significant difference

ii) Indexing Relevance

Subject field	% Success			Significance
	IO's	SIA's	RO's	
Plastics	51,7	30,6	48,1	Significant at the 2% level
Chemistry and Chemical Engineering	52,7	45,2	44,2	No significant difference
Explosives	47,8	47,4	52,6	No significant difference

iii) Document Relevance
Total relevance (1st order and 2nd order)

Subject field	% Success			Significance
	IO's	SIA's	RO's	
Plastics	35,5	31,1	37,1	No significant difference
Chemistry and Chemical Engineering	48,8	50,7	50,7	No significant difference
Explosives	52,1	50,4	53,5	No significant difference

iv) First order relevance

Subject field	% Success			Significance
	IO's	SIA's	RO's	
Plastics	22,7	19,2	28,1	Significant at the 5% level
Chemistry and Chemical Engineering	33,9	35,5	35,7	No significant difference
Explosives	27,5	27,7	26,9	No significant difference

b) Comparison between different types of searchers Recall, Indexing Relevance, and Document Relevance

Totals under all subjects	% Success			Significance
	IO's	SIA's	RO's	
Recall	78,8	74,5	75,2	No significant difference
Indexing Relevance	50,9	40,4	48,5	No significant difference
Document Relevance - Total (1st and 2nd order)	45,3	47,0	48,3	No significant difference
Document Relevance - 1st Order	28,4	26,1	29,8	No significant difference

c) Comparison between different subjects : Recall, Indexing Relevance, and Document Relevance

Totals under all searchers	% Success			Significance
	Plastics	Chemistry & Chem Eng	Explosives	
Recall	71,4	74,6	81,6	No significant difference
Indexing Relevance	41,1	47,5	49,6	No significant difference
Document Relevance : Total	34,0	49,9	52,0	Significant at the 0,1% level
Document Relevance : 1st order	22,3	34,7	27,4	Significant at the 0,1% level

d) Comparison between Indexing Relevance and Document Relevance (1st and 2nd order relevance)

Totals under all searchers	% Success		Significance
	Indexing Relevance	Document Relevance	
Chemistry and Chemical Engineering	47,5	49,8	No significant difference
Explosives	49,6	52,0	No significant difference
Plastics	40,9	34,0	Significant at the 10% level
Plastics (omitting the SIA-Plastics)	50,0	36,1	Significant at the 1% level

6.1.4.5 User evaluation of TIRSS

Finally: all the searchers were asked to evaluate TIRSS on the basis of their experience during the test by completing the questionnaire set out below. The objective was to identify specific weaknesses in the system:

Questionnaire

The following headings were used in the questionnaire and the document included a request for the use of examples to support statements in the spaces provided between the headings

1. The Thesaurus
    1.1 Exhaustiveness (ie availability of index terms to cover all the required information).
    1.2 Precision and accuracy (ie availability of index terms to describe the required information precisely and correctly)
    1.3 Cross-references (ie adequacy of lead-in terms, as "see" and "see-also" references, to guide the searcher to the preferred thesaurus terms.
    1.4 Further comments and/or suggestions for improvements
2. Indexing
    2.1 Exhaustiveness (ie insufficient, or overdetailed and redundant terms)
    2.2 Precision and accuracy (ie choice of terms to describe the information precisely and correctly)
    2.3 Further comments and/or suggestions for improvements
3. The test questions
    3.1 Too general
    3.2 Too detailed
    3.3 Not easily understood
    3.4 Misleading
    3.5 Incorrect
4. Physical aspects of the system
    4.1 TIRSS cards (ie usefulness/adequacy of bibliographic details, abstract, index terms, lay-out, etc)
    4.2 Filing (ie ease of finding the required cards, mis-filing, adequacy of guide cards, etc)
    4.2 Thesaurus lay-out and format
    4.4 Further comments and/or suggestions for improvements
5. Overall comments on the system

From the questionnaire and a study of the items of information retrieved during the searches, the reasons for search failures or non-recall of source documents can be summed up as follows:

1. Inadequacies in the indexing - 50% of the failures
2. Inadequacies in the questions - 20% of the failures
3. Inadequacies in the thesaurus - 20% of the failures
4. Other reasons - 10%

Two principal factors affect the operational performance, measured by recall and relevance ratios, of any retrieval system. These are the degrees of specificity and exhaustivity of indexing.

Specificity, at the input stage, is a property of the index vocabulary or of the thesaurus. The more specifically a concept can be described, (ie the degree to which the exact species rather than a containing genus is recognised) the higher the relevance ratio of searches involving this concept. The term polyethylene for example is a more specific term than polyolefines or polymers.

Exhaustiveness, the second factor, is dependent not on system vocabulary but on a decision by the indexer. Exhaustive indexing means the recognition and description of each item of information in a document considered to be worth indexing and it follows that a high level of exhaustiveness will result in a high recall ratio and vice versa.

During the Aslib Cranfield investigation (67-70) it was found that an inverse relationship existed between recall and precision and it can be said that high exhaustiveness in indexing results in high recall and precision, while high specificity in the index language results in high precision and low recall. This implies that once an optimum level of performance has been achieved for a system, any method used to improve recall must inevitably reduce precision; conversely any method to improve precision must reduce recall.

a) Recall ratio

For this test the recall figure was calculated as the percentage of success in retrieving the source documents on which the test questions were based. It was assumed that the same figure should hold for all relevant documents in the system (Note: this hypothesis was shown to be true during the Aslib Cranfield Project [68].

The total recall ratio for all subjects was 76% which compares favourably with the published results of tests conducted elsewhere. TIRSS was designed and developed as a self-help information retrieval system for use by all technical staff; this test confirms that this belief is justified since it performed equally well in different subject fields and when searchers with different levels of qualifications and experience use it.

b) Precision ratio

i) Document relevance

To determine whether a document is relevant or not depends to a large extent on the needs of the enquirer. These may vary, depending on the knowledge he already has of the subject, his depth of interest in the problem, etc. Any judgement on the relevance or otherwise of a document is therefore highly subjective. However, the operating efficiency of a system depends on how well it satisfies the requirements of its users, and the retrieval of "relevant" documents is meaningful only in the context of being useful to the particular users of the system.

In this test the nine senior members of the technical staff who were responsible for preparing the search questions were also asked to determine the relevance of the retrieved information, the assumption being that their requirements would be sufficiently representative of the requirements of all technical staff.

ii) Indexing relevance or term relevance

In preparing a search programme the ideal is to select index terms which will match the terms selected during the original indexing of the documents containing the required information. This means that the searcher "indexes" his search question in the same way as the original documents were indexed for input into the system. In practice, however, inappropriate terms and redundant terms may be selected, or appropriate terms may be overlooked - either at the input stage, or at the output or search stage, or both. The purpose of determining indexing relevance was, therefore to identify possible weaknesses in the original indexing, or in the adequacy of the thesaurus to lead the indexer and the searcher to the appropriate index terms.

iii) Comparison betwee document relevance and indexing relevance

The determination of indexing relevance introduced in this test is believed to be a novel concept in the testing of information retrieval systems and the possibility had to be considered whether there was a correlation between indexing relevance and document relevance, the latter being the accepted technique of measuring the precision ratio of a system.

In the test there was no significant difference between total document relevance and indexing relevance (averaging 51% and 48% respectively) except in the field of Plastics where document relevance was 35%. The difference between the two types of relevances for Plastics is significant at the 1% level.

If indexing was done at the same level of specificity as the concepts dealt with in the documents then it can be expected that indexing relevance will be the same as document relevance. The reason for this is that when a document is indexed the indexer first decides which concepts in it should be indexed, and these are then translated into appropriate index terms. The terms selected should match the degree of specificity of the concepts. If the concepts are more specific than the index terms describing them, document relevance will be lower than indexing relevance. To give an example: if a document dealing with the molecular structure of plasticers is indexed under the term "Molecular structure", instead of under "Plasticers - Molecular structure", a high proportion of non-relevant documents will be retrieved.

The investigation showed that the specificity of the indexing terms used matched the specificity of questions addressed to the system for both explosives and chemistry and chemical be overlooked - either at the input stage, or at the output or search stage, or both. The purpose of determining indexing relevance was, therefore to identify possible weaknesses in the original indexing, or in the adequacy of the thesaurus to lead the indexer and the searcher to the appropriate index terms.

iii) Comparison betwee document relevance and indexing relevance

The determination of indexing relevance introduced in this test is believed to be a novel concept in the testing of information retrieval systems ahd the possibility had to be considered whether there was a correlation between indexing relevance and document relevance, the latter being the accepted technique of measuring the precision ratio of a system.

In the test there was no significant difference between total document relevance and indexing relevance (averaging 51% and 48% respectively) except in the field of Plastics where document relevance was 35%. The difference between the two types of relevances for Plastics is significant at the 1% level.

If indexing was done at the same level of specificity as the concepts dealt with in the documents then it can be expected that indexing relevance will be the same as document relevance. The reason for this is that when a document is indexed the indexer first decides which concepts in it should be indexed, and these are then translated into appropriate index terms. The terms selected should match the degree of specificity of the concepts. If the concepts are more specific than the index terms describing them, document relevance will be lower than indexing relevance. To give an example: if a document dealing with the molecular structure of plasticers is indexed under the term "Molecular structure", instead of under "Plasticers - Molecular structure", a high proportion of non-relevant documents will be retrieved.

The investigation showed that the specificity of the indexing terms used matched the specificity of questions addressed to the system for both explosives and chemistry and chemical

engineering. In the field of plastics, however, the terms tend to be too generic, and the index terms in the thesaurus are, therefore, broader than the concepts dealt with in the documents. In practice this means that proportionately a higher percentage of irrelevant items of information will be retrieved when a question is asked in this field.
The reasons for this discrepancy could be:

i) Lack of adequate technical knowledge of the indexers to select appropriate index terms to match the specific concepts in the documents

and/or

ii) Because reports on plastics deal with so many different and very specific aspects of information, the indexers are reluctant to create very many new terms.

One obvious way in which to overcome this problem would be to index more specifically in the field of plastics and it was decided as a result of the test to review the existing thesaurus terms and to consult more closely with plastics technical staff in future when documents are to be indexed. In this manner any changes to the thesaurus can be made gradually and as and when the need arises.

c) <u>User evaluation of TIRSS</u>

The use of role numbers with index terms was found to be of little value to the searchers and was also unreliable for retrieval purposes. The reason for this is the difficulty in explaining their unambiguous and proper use. The only real value that role numbers have is that their use introduces a systematic method of indexing so that they will for this reason be retained in the TIRSS system.
User reaction during the test was generally favourable. Certain aspects of the system, however, need improvements and these were found to be:

i) More specific indexing in the field of plastics as discussed above

ii) More thesaurus aids should be provided in the form of lead-in terms or cross references to guide the searcher from the many synonyms and related terms used in technical language to the preferred index terms.

d) <u>Conclusions</u>

1) There is no significant difference in the recall efficiency of TIRSS between the different types of searchers viz: Information Officers, Senior Information Assistants, and Research Officers in the experimental Groups.

2) The overall operational efficiency of TIRSS compares favourably with published results of tests conducted elsewhere.

3) The proportion of relevant documents retrieved by all searchers in the plastics field is significantly lower than in the other subject fields.

4) There was no significant difference in the proportion of relevant documents retrieved by Information Group staff as compared to Research Officers in the Experimental Groups <u>except</u> in the field of plastics where both graduate and non-graduate Information Group staff had a significantly lower performance.

5) Approximately 50% of the search failures were because of inadequacies of indexing; 20% because of poor questions; 20% because of inadequacies in the thesaurus, and 10% because of a combination of other factors.

6) The use of role numbers with index terms was found to be of little value to the searchers and was also unreliable for retrieval purposes, but worth retaining for the systematic methodology which their use introduces to the process of indexing.

The above answers would appear to be satisfactory and they were accepted as such. However, it was realised that the overall picture could be misleading to some extent, since few of the users who completed the questionnaire had had any previous experience of using any IR system - <u>ie</u> their standards for the comparison of TIRSS with other systems were virtually non-existent. Nevertheless it was agreed as far as the Company was concerned, since the test results were apparently as good as for any published results of other systems, that it was meeting requirements, and that this was the most important factor. (Note: The subject of costs is dealt with in Chapter 7).

6.2 CORIS

By the beginning of 1966, TIRSS and its thesaurus had both grown substantially, and at this time it was decided to investigate the Research Department's procedures for indexing and filing correspondence. In view of the fact that generation of the TIRSS thesaurus had cost a considerable amount of money it was decided to try and use it for the indexing of technical correspondence. The investigation showed that there were good reasons to change the then existing procedures and that it should be a practical proposition to adopt similar pre-coordinate indexing, using the existing thesaurus. A system known as CORIS - correspondence information system - was then set up, and this is outlined in Appendix VIII.

An important aspect of the whole system is that both incoming and outgoing correspondence are culled or 'weeded' vigorously when new so that not more than 15% of the total is included in the system. The other 85% of all the incoming and outgoing correspondence is destroyed after three months; the overall embarrassment effect of this drastic action on the department has been negligible. (Note: In view of the high cost of weeding old correspondence files it was decided not to include any correspondence that pre-dated mid-1966 in the new system).

In effect it is only that small percentage of correspondence which contains technical information of potential long term value, together with items containing legal information, which are retained. (It should be noted that all correspondence dealing with company policy and personnel matters is totally excluded from CORIS and is dealt with in other ways).

The problem of storing large quantities of old correspondence has been successfully eliminated with this system and further relief has been obtained by microfilming all retained items over two years old. The filming is done once a year - eg all the items for the year 1970 were filmed at the beginning of 1973. The microfilm is loaded into film-jackets each of which corresponds to an original file number.

It was expected that the inclusion of CORIS would result in the addition of a substantial number of new terms to the thesaurus, many of which might well be commercial rather than technical. However it soon became clear that the non-technical terms tend to occur more frequently in that large percentage of correspondence items which are of ephemeral interest and which are never included in CORIS. After the first four years the total number of new terms in the thesaurus which had been added to meet the requirements of CORIS was only 160, and of these only 28 are not strictly technical; the current increase is at the average rate of one per month.

Finally, of the 40 000 items currently indexed in CORIS, about 600 contain technical information of sufficient value to warrant their inclusion in TIRSS as well.

6.2.1 Comment on experience with CORIS

The system currently averages eight searches per working day, a figure which has remained constant during the last four years. After eight complete years of operation it has been found that the time required to locate a document is very dependent on the clarity and specificity of each request received.

The system has not been tested for the effectiveness of name searches since these are, by their very nature, precise and unambiguous. However it has been tested by subject searches. 100 successive requests were recorded and it was found that 43 of these were clear and precise and that the retrieval time for the relevant documents varied between one and a half and three minutes, with an average of two minutes. In 57 cases the request was either vague or inaccurate and the retrieval times varied between five and forty-five minutes, with an average of eight minutes. In six of the 100 searches it was found that the document requested had not been retained and included in CORIS but in all the other instances the items required were located successfully; this matter of permitting subjective judgment by the originators and recipients of correspondence for its retention or destruction must almost inevitably lead to some mistakes when a firm policy of 'if-in-doubt - then destroy' is insisted upon. In this test only one of the six items destroyed was of any real importance - this situation has arisen so seldom, that no changes appear to be needed as yet. Thus the system meets the objective of ensuring the speedy retrieval of correspondence satisfactorily. (Note: The subject of costs is dealt with in Chapter 7).

6.3 LIBRIS

The success of the use of the TIRSS thesaurus for CORIS resulted in efforts to find still further uses for it. As already mentioned, AE&CI's central library in the Research Department had adopted UDC for its catalogue in 1955. Generally speaking this change had not been really satisfactory for a number of reasons:

1 the previous experience of the library staff involved in the application of UDC had been limited,

2 the technical knowledge of the library staff had consistently been almost nil, and in all fairness to them it must be emphasised that in a country like South African it is virtually impossible to find individuals who are technically qualified or experienced and <u>also</u> qualified or experienced in librarianship,

3 the technically qualified library users were, in general, reluctant to ask the library staff for assistance when seeking information, at least partly because of their realisation that the latter had no technical knowledge,

4 library users were mainly unwilling to try and understand the catalogue, and those few who did do so were highly critical of the standard of classification,

5 as a result of the foregoing the majority of users tended to browse along shelves, in what looked like the right section, and to hope for the best. They probably missed 60% of the relevant information in the available books and 100% of that in the miscellaneous publications, such as pamphlets, translations, etc.

TIRSS itself has proved to be a success from the point of view that users were prepared to search it on a self-help basis, and it was believed that a similar system could be provided for the library itself. As a result, LIBRIS - library information system - was set up in the middle of 1967.

As pointed out at the beginning of this chapter AE&CI had enjoyed a long history of well stocked libraries so that management was well aware of the value of efficient library services. Jain, of Purdue University has correctly pointed out [240]:

> 'An adequate library is not only a basis of all teaching and study, it is an essential condition of research, without which additions cannot be made to the sum of human knowledge. A research library plays a very important role in the efficient running and the advancement of the institution of which it is part and parcel. The function of a research library is to enable inquirers to identify (information in) library material relevant to their inquiries .....'

At that time the contents of the library amounted to some 10 000 books and 10 000 pamphlets and the existing UDC catalogue comprised the usual Subject, Author/Title and Classified sections. As in the case of TIRSS, and for the same reasons, it was decided to reprocess all the existing book stock into the new system. The opportunity was taken to weed-out some very out-dated books and many pamphlets which had proven to be of temporary value, so that by mid-1972 the backlog of items to be reprocessed had been eliminated. In addition, it had proved possible to include the holdings of the various factory libraries so that for the first time a central catalogue has been built up on a company-wide basis.

LIBRIS is a simpler system than TIRSS. It was appreciated that it is virtually impossible to prepare informative-type abstracts for books, and the 'easier-to-prepare' indicative abstracts are used. Similarly no role indicators are used with the index terms since the contents of most books are too diffuse and at too generic a level. Nevertheless it was found that a considerable amount of information can be set out on an A6 size index card (105 x 148 mm) - certainly sufficient to provide a clear indication of the subject scope of a book. The details of the system are set out in Appendix IX, but the following additional comments are considered to be apposite:

6.3.1 Shelving

It was clear that LIBRIS itself could make no provision for an acceptable shelving procedure, since virtually all IR systems depend on some form of sequential numbering for location of the documents concerned. However much potentially useful information may be missed as a result of browsing along the shelves of any library, many users like to do this. In addition, while it is true that one copy of a book which is concerned with, say, five to ten technical subjects, can be shelved in one place only, in general the grouping of books on like subjects is still possible, is useful and sometimes valuable. Thus there appeared to be good reason to retain the UDC system, at least in a simplified form, for shelving purposes. This was done, and all classifications are limited to a maximum of six digits; this has proved to be satisfactory for these said shelving purposes.

6.3.2 Abstracting

Most of the abstracts have been written by staff with technical knowledge but limited technical qualifications, and their work has proved to be satisfactory. It is clear that the preparation of indicative abstracts is not as time consuming nor as intellectually exacting as for informative abstracts. The material for an adequate abstract can often be found in a review which led to purchase or in the introduction or preface or, occasionally, in the publisher's comments on a book jacket.

6.3.3 Indexing

Indexing has been done by qualified librarians and also by the same staff as have prepared abstracts. It is frequently possible to select terms from chapter headings alone and, on the whole, the level of indexing is far more generic than in TIRSS, and this is to be expected as has already been pointed out. At the same time, in addition to generic indexing being appropriate for books because of their nature, a lower level of technical knowledge suffices, which in turn affects the selection of staff required for this work.

The following comments by an experienced librarian with little technical knowledge who was intimately concerned with the start-up of LIBRIS are reported verbatim [241]. They were written after six months experience of the system, by which time she had become a competent indexer:

> 'It is more difficult to index books than it was to classify according to a set scheme. I could understand enough of a book on a very technical subject to know where it fitted in a classification scheme, and to make added entries for any subjects dealt with which were not covered by that classification number. Now I have to understand what the book is about before I can make index terms for it - especially in the field of organic chemistry. Nevertheless, for the unscientific indexer, it is easier to decide on terms for a book once a few have been indexed on that subject; discussion with technically minded colleagues is a great help, as were the chemistry lectures provided'.

Here is clear acknowledgement that technical knowledge is of considerable advantage to the indexer, but that with some training in a technical subject (in this instance ten lectures in elementary chemistry were provided), with opportunity for discussions with technically qualified individuals and with experience, the 'unscientific indexer' can learn the job within a reasonable time.

6.3.4 Comment on experience with LIBRIS

While the primary objective of setting up LIBRIS was to improve the library service to company staff, it also offered the important advantage of ensuring further additional use of the thesaurus which had already been developed. It is of interest to note that the effect on the thesaurus was the addition of 530 terms between mid-1967 and the end of 1968, as stated in Chapter 5 and of a further 320 new terms since then, mainly generic and mainly technical. No especial difficulties have been encountered.

The effect on library users of the new catalogue appears to have been considerable; the system is well-liked and is used extensively. Users have found that they can quickly obtain a comprehensive idea of the resources of the library on a given subject as well as on the information in individual books, etc. They can also judge whether they need to ask to see a book currently out on loan. While no accurate figures can be provided for the increase in effective use of the library, the actual total usage increased by 25% during the three years mid-1967 to mid-1970, ie from approximately 4 800 items issued on loan to 6 600, and to 8 600 in 1973. There has been no corresponding increase in the numbers of potential users employed - in fact there has been a decrease of 12% - so that it is obvious that it is being used more by the same people. This, it can reasonably be inferred, has meant an increase in the return on the investment made by the Company in publications. While no cash value can be attached to this increase it is believed to represent a substantial sum.

Efforts were made to determine more realistically the effects of introducing LIBRIS, in addition to the purely quantitative increase in loans referred to above. Only for one subject - the management/personnel field (UDC 65.01 to 651.08) could pre-LIBRIS records be found concerning usage. These records showed that in 1967 there were 73 books listed in the catalogue and that of these 29 - ie 39.7% - were out on 'permanent loan' in various offices. An accurate picture of the usage of those items on permanent loan could not be obtained but enquiries suggested that it was very low. Of the balance of 44 items only nine - 20.4% - were borrowed over the six month period from January to June 1967; one of the nine items was read by four individuals and two by two the important advantage of ensuring further additional use of the thesaurus which had already been developed. It is of interest to note that the effect on the thesaurus was the addition of 530 terms between mid-1967 and the end of 1968, as stated in Chapter 5 and of a further 320 new terms since then, mainly generic and mainly technical. No especial difficulties have been encountered.

The effect on library users of the new catalogue appears to have been considerable; the system is well-liked and is used extensively. Users have found that they can quickly obtain a comprehensive idea of the resources of the library on a given subject as well as on the information in individual books, etc. They can also judge whether they need to ask to see a book currently out on loan. While no accurate figures can be provided for the increase in effective use of the library, the actual total usage increased by 25% during the three years mid-1967 to mid-1970, ie from approximately 4 800 items issued on loan to 6 600, and to 8 600 in 1973. There has been no corresponding increase in the numbers of potential users employed - in fact there has been a decrease of 12% - so that it is obvious that it is being used more by the same people. This, it can reasonably be inferred, has meant an increase in the return on the investment made by the Company in publications. While no cash value can be attached to this increase it is believed to represent a substantial sum.

Efforts were made to determine more realistically the effects of introducing LIBRIS, in addition to the purely quantitative increase in loans referred to above. Only for one subject - the management/personnel field (UDC 65.01 to 651.08) could pre-LIBRIS records be found concerning usage. These records showed that in 1967 there were 73 books listed in the catalogue and that of these 29 - ie 39.7% - were out on 'permanent loan' in various offices. An accurate picture of the usage of those items on permanent loan could not be obtained but enquiries suggested that it was very low. Of the balance of 44 items only nine - 20.4% - were borrowed over the six month period from January to June 1967; one of the nine items was read by four individuals and two by two individuals - the rest were read once only and of the nine items none was more than four years old. By 1972 the number of items in the LIBRIS catalogue falling within the scope of the same UDC numbers was 320. Of these 147 were books and 49 - ie 30.0% - were out on permanent loan. During the six months February to July 1972 some 24 - 25% - of the remaining 98 books were borrowed; one item was read by nine individuals, two by seven, six by six, five by four, and five others by two individuals. Of course the samples are small and in a field such as 'management/personnel', which is subject to succeeding fashionable techniques, it is difficult to draw firm conclusions. The more detailed indexing which is an essential part of the LIBRIS system (an average of 7.4 terms/book in this particular instance) should in theory make access to material of interest significantly easier, but on the basis of these samples it cannot be claimed that this did in fact happen. On the one hand the total number of staff working on the library site had not increased between 1967 and 1972 which suggests a greater interest in the management/personnel field. On the other hand increasing competitiveness in South African industry and the appearance of various schools of business administration attached to local universities had resulted in an increased awareness of the need for greater knowledge in these fields. One interesting fact came to light which was that while in the first sample the oldest book read was four years old, in the second sample the oldest was five years old. Thus none of the items loaned out in the first sample were borrowed during the time of the second sample five years later, and obviously this raises the question as to how long it is worthwhile to retain books and other material in an industrial library on non-technical subjects.

A detailed examination of the indexing in a technical field was carried out to see whether any useful conclusion could be drawn [Note: As pointed out above there were no detailed records of usage in any technical field which pre-date LIBRIS, so that it has not been possible to make meaningful comparisons of a 'before-and-after' nature]. In view of the findings concerning the quality of indexing in the field of plastics for TIRSS, as set out in sub-section 6.1.4, it was decided to examine an aspect of this field. The subject of polyvinyl chloride (PVC) was chosen because this was the first plastic material of interest to AE&CI and the library collection was bigger than for any of the other plastics of interest to the Company.

The results of the examination were as follows:

1. LIBRIS cards:

1.1 Total documents indexed under PVC = 176, comprising
68 books,
99 pamphlets, and
9 translations into English
of foreign language journal articles.

The relevant cards are set out in Appendix X. Reference was made above to the use of simplified UDC numbers for the shelving of books and these numbers appear in the top right hand corner of the cards. In 58 of the 68 books (ie 85%) the primary topic was classed under UDC 678 and sub-divisions - ie macromolecular materials, rubber and plastics. Thus indexing of PVC where it was a minor topic, but which is important to AE&CI, appeared in 15% of the total books. It is certainly not claimed that conventional analytical entries would not have taken care of this need, but it is believed that the LIBRIS-type subject entry system permits high specific indexing to be done relatively more easily.

1.2 The total number of indexing entries for the 176 documents on the 409 cards in the PVC section was 5066. These entries are set out in Appendix XI - the terms used multiplied by the number of times each was used = 5 066.
The maximum number of indexing terms for one document was 23 for a book. On the other hand only one term was used for one document and two terms for each of five documents; all six of these documents were pamphlets.

2.. THESAURUS

The total number of index terms in the thesaurus concerned with PVC was 138. In addition there were 24 'see' references - eg 'POLYVINYL CHLORIDE AUTOCLAVES - SEE AUTOCLAVES' and 9 'see also' references - eg POLYVINYL CHLORIDE LAMINATES - SEE ALSO LAMINATES'. As shown in Appendix V the thesaurus includes a column of symbols to indicate whether information on a given subject can be found in TIRSS(T), LIBRIS (L) and/or CORIS (C). In no instance did a term have 'L' as the only entry, which would indicate that the particular term had been added to the thesaurus to meet the specific needs of the Library. In fact only 64 of the 138 terms were posted with 'L' entries - ie 46.4% and this is not really surprising. The thesaurus was designed originally for indexing very specific research reports in depth - in general books cover very much wider fields and their indexing needs can be met adequately with less specific terms. Examination of the cards illustrated in Appendix X tends to confirm this for books. In the case of pamphlets some have very general subject coverage and are indexed accordingly while others contain highly specific and detailed information and are indexed almost as though they were research reports.

6.3.5. Test of the retrieval effectiveness of LIBRIS

In testing the retrieval effectiveness of LIBRIS it was decided to use some of the same questions used in the testing of TIRSS - see sub-section 6.1.4 - and which are listed in Appendix VII. In view of the less favourable results obtained in the TIRSS test for the indexing of reports concerned with plastics it was decided to examine this field. Not all of the questions listed were suitable for a search of published information since many of them were too specific and concerned with AE&CI products. However of nine questions concerned with PVC, eight were considered suitable (Nos 1, 6, 70, 71, 73, 91, 92 and 93 - see below), and of eight questions concerned with polyethylene, two were considered suitable (Nos 27 and 88). The question as to who should be involved in the actual searching was discussed with the same statistician as had been involved in the TIRSS test. Consideration was given to using technical graduates with plastics technology experience, to using technicians with similar experience and to using information scientists or library staff. However, after examining the questions and the broad scope of the documents in the library which appeared to be relevant to the questions, the statistician concluded that the outcome would most probably be very similar whoever did the searching. The reasoning was that the books and other material were so very much broader in scope when compared to research reports and so broadly indexed that anyone comparing the terms used in the questions with the terms available in the thesaurus would have very little scope for displaying considerable knowledge of the subject field or a basic lack of knowledge of it. In an attempt to overcome probable objections to this reasoning, and pragmatic approach, three individuals were asked to prepare lists of search terms for the same questions. The individuals and their background were as follows:

Individual A - a research chemist with five years' experience of plastics,

Individual B - a technician working in a plastics technical service laboratory with two years' experience,

Individual C - an information officer with no technical qualifications but six years' experience with the literature of plastics.

After compilation of the lists of indexing terms by the individuals A, B and C the author carried out the search with the results set out below. The relevance to each question of the documents traced was finally confirmed by examining them :-

Question 1: How does plasticizer absorption behaviour of suspension PVC depend on the nature of the plasticizer?

Thesaurus terms selected by:			LIBRIS cards found which appeared to be relevant	Check on the actual relevance of documents
A	B	C		
1. Plasticizers - absorption	Plasticizers - absorption	Plasticizers - absorption	Nil	-
2. Plasticizer/PVC compatibility	-	-	One pamphlet	Not relevant - concerned with a proprietary brand of plasticizers
3. Polyvinyl chloride - plasticization	-	Polyvinyl chloride - plasticization	One book	Relevant
4. Plasticization	Plasticization	Plasticization	Four books	All relevant
5. -	Absorption	-	Nil	-

Question 6: What polymer properties control the rate of absorption of hot plasticizer by granular PVC?

Thesaurus terms selected by:			LIBRIS cards found which appeared to be relevant	Check on the actual relevance of documents
A	B	C		
1. Plasticization	Plasticization	Plasticization	Four books	Two relevant
2. -	Absorption	-	Nil	-
3. Plasticizers - absorption	Plasticizers - absorption	Plasticizers - absorption	Nil	-
4. Polyvinyl chloride - plasticization	-	Polyvinyl chloride - plasticization	One book	Relevant
5. Polyvinyl chloride - granular	Polyvinyl chloride - granular	-	Nil	-

Question 27: What materials have been tried as clarity modifiers for manufacture of polyethylene film grades?

Thesaurus terms selected by:			LIBRIS cards found which appeared to be relevant	Check on the actual relevance of documents
A	B	C		
1. Polyethylene film	Polyethylene film	Polyethylene film	Two books and three pamphlets	One book and two pamphlets relevant
2. Plastic films	Plastic films	Plastic films	One book	Not relevant
3. Polyethylene - properties (optical)	-	-	One pamphlet	Not relevant

Question 70: What PVC copolymer properties influence their processing characteristics in plasticized applications?

Thesaurus terms selected by:			LIBRIS cards found which appeared to be relevant	Check on the actual relevance of documents
A	B	C		
1. Polyvinyl chloride (plasticized)	Polyvinyl chloride (plasticized)	Polyvinyl chloride (plasticized)	One book	Relevant
2. Polyvinyl chloride (properties)	Polyvinyl chloride (properties)	Polyvinyl chloride (properties)	Two pamphlets	Both relevant (published by an associated company)
3. Copolymers (properties)	Copolymers (properties)	Copolymers (properties)	One book (same book as in 1 above)	Relevant

Question 71: How can the kinetics of the interaction of plasticizer with granular PVC be determined?

Thesaurus terms selected by:			LIBRIS cards found which appeared to be relevant	Check on the actual relevance of documents
A	B	C		
1. PVC (granular)	PVC (granular)	PVC (granular)	Nil	-
2. Kinetic calculations	-	-	Nil	-
3. Kinetics	Kinetics	Kinetics	Nil	-

Question 73: How does rate of absorption of plasticizer by granular PVC vary with temperature?

Thesaurus terms selected by:			LIBRIS cards found which appeared to be relevant	Check on the actual relevance of documents
A	B	C		
1. -	Absorption	Absorption	Nil	-
2. Plasticizers - absorption	Plasticizers - absorption	Plasticizers - absorption	Nil	-
3. Plasticizer/PVC compatibility	-	-	Nil	-
4. Polyvinyl chloride - plasticization	-	Polyvinyl chloride - plasticization	Nil	-
5. Plasticization	Plasticization	Plasticization	Five books	One relevant
6. Polyvinyl chloride (granular) -	-	-	Nil	-
7. Plasticizers (volatility)	-	-	Nil	-

Question 88: What techniques are available to improve the dispersion of pigment particles when compounding polyethylene?

Thesaurus terms selected by:			LIBRIS cards found which appeared to be relevant	Check on the actual relevance of documents
A	B	C		
1. Polyethylene - compounding	Polyethylene - compounding	Polyethylene - compounding	Nil	-
2. Polyethylene - pigments	Polyethylene - pigments	Polyethylene pigments	One pamphlet	Not relevant
3. Pigments dispersion	-	-	Three books	None relevant (all concerned with paints)

Question 91: How may the chain branching of PVC be determined?

Thesaurus terms selected by:			LIBRIS cards found which appeared to be relevant	Check on the actual relevance of documents
A	B	C		
1. Polyvinyl chloride (analysis)	Polyvinyl chloride (analysis)	Polyvinyl chloride (analysis)	Two books	One relevant
2. Chain branching	Chain branching	Chain branching	One book	Not relevant
3. Chain branching (measurement)	Chain branching (measurement)	-	Nil	-
4. -	Chain branching (reactions)	Chain branching (reactions)	Nil	-

Question 92: How does heat stability of suspension PVC depend on granulating agent used?

Thesaurus terms selected by:			LIBRIS cards found which appeared to be relevant	Check on the actual relevance of documents
A	B	C		
1. Polyvinyl chloride (stabilization)	Polyvinyl chloride (stabilization)	Polyvinyl chloride (stabilization)	Six books and one monograph	Two books and the monograph relevant
2. Polyvinyl chloride (granulation)	Polyvinyl chloride (granulation)	Polyvinyl chloride (granulation)	Nil	-
3. -	Polyvinyl chloride (granular)	-	Nil	-
4. Stabilizers (heat)	-	-	One book	Relevant

Question 93: Is there a precise method for determining particle size distribution of PVC latices?

Thesaurus terms selected by:			LIBRIS cards found which appeared to be relevant	Check on the actual relevance of documents
A	B	C		
1. Polyvinyl chloride particle size	-	Polyvinyl chloride - particle size	Nil	-
2. Polyvinyl chloride - particle size - analysis	Polyvinyl chloride - particle size - analysis	Polyvinyl chloride - particle size - analysis	Nil	-
3. Particle size analysis	Particle size analysis	Particle size analysis	One book	Relevant
4. Particle size determination	Particle size determination	Particle size determination	One book (same book as in 3 above)	Relevant

To draw very firm and clear conclusions from this test would be unjustified. Nevertheless there appear to be several tendencies which are worthy of comment:

1. The Library staff who carried out the LIBRIS indexing tend to use more generic terms as pointed out earlier, and this appears to be adequate for the retrieval of documents such as books; in view of the comparatively wide scope of most technical books (as opposed to 'subject' reports) this is not surprising and is sensible.

2. In the test the graduate tended to use similar generic terms to the non-graduates but, in addition, he also used some highly specific terms. Again in view of his experience in the field of plastics this is not surprising.

3. Of the questions used in the test it appears that No 71 was probably too highly specific. The graduate used in the test - individual A - not only agreed with this opinion but expressed the opinion that relevant information was in fact unlikely to have been published, and so would not be found in an 'open' library.

To-date there has been little reason to doubt the efficiency of the LIBRIS system and heavy usage by staff tends to confirm this. In these days of major developments and changes there is no room for complacency however, and, coupled with continuing diversification of AE&CI's technical interest, and hence continually increasing purchases of publications the whole system will require regular testing in the future. No doubt the scale of testing, and the thoroughness, will need to be stepped up; it is proposed to undertake this testing at three to four yearly intervals.

Certain practical and minor disadvantages in the presentation currently in use are known. These concern the filing equipment more than anything else and are expected to disappear when the whole catalogue is transferred to microforms - probably to film-jackets as in the case of TIRSS.

6.4 Integration of the systems

From the foregoing sections of this Chapter it is clear that the common factor linking TIRSS, CORIS and LIBRIS is the thesaurus. About the time that LIBRIS was developed in mid-1967 the up-dating of the thesaurus had been shifted from an expensive typing and re-typing operation to the use of tabulating cards, which were run through an IBM 360 system printer at three monthly intervals. The opportunity which this technique presented to add additional information at low cost was appreciated. Thereafter the symbols 'T', for TIRSS, 'C', for CORIS, and 'L', for LIBRIS were used to indicate the whereabouts of information on any given subject.

An extract from the thesaurus is shown below in order to illustrate this point:

ORGANIC COMPOUNDS - NOMENCLATURE	L
ORGANIC COMPOUNDS - NUCLEAR MAGNETIC RESONANCE	L
ORGANIC COMPOUNDS - OXIDATION	LT
ORGANIC COMPOUNDS - PATENTS RETRIEVAL	T
ORGANIC COMPOUNDS (PEROXIDE)	CLT
ORGANIC COMPOUNDS (PEROXIDE) SEE ALSO PEROXYDICARBONIC ACID, ESTERS, ETC	
ORGANIC COMPOUNDS (PHOSPHORUS)	CLT
ORGANIC COMPOUNDS (PHOSPHORUS) - INSECTICIDES SEE INSECTICIDES ORGANIC COMPOUNDS (PHOSPHORUS)	
ORGANIC COMPOUNDS - POLAROGRAPHIC DETERMINATION	T
ORGANIC COMPOUNDS - POLYMERIZATION	L

Provision has been made for the inclusion of a maximum of twelve such symbols. The incorporation of others, such as 'E' for information in the technical files held in the Chief Engineer's department, is expected shortly after numerous delays arising from staff shortage.

In sub-section 4.1.8 the point was made that where it is required to retrieve very detailed information, or data, in a narrow subject field, a post-coordinate indexing system is preferable to a pre-coordinate system.

One such post-coordinate system in the field of gas chromatography is in use and has been incorporated into the thesaurus as shown in the following extract.

CHROMATOGRAPHY	CLT
CHROMATOGRAPHY - BIBLIOGRAPHY	LT
CHROMATOGRAPHY (COLUMN)	LT
CHROMATOGRAPHY (GAS)	CLT
SEE ALSO TERMATREX SYSTEM : FOR ACCESS CONTACT GROUP LEADER - ANALYTICAL GROUP - RESEARCH DEPARTMENT	
CHROMATOGRAPHY (GAS/SOLID) SEE CHROMATOGRAPHY (GAS)	
CHROMATOGRAPHY (GEL PERMEATION)	CLT
CHROMATOGRAPHY (ION EXCHANGE)	L

As more systems are developed, as more sections of AE&CI are brought actively within the scope of the overall IR effort and as more records are converted to microform for storage, so it will become worthwhile to distribute copies of the thesaurus more widely to technical staff. The distribution of duplicate sets of microforms of TIRSS, for example, to the major centres of AE&CI may also become economic and worthwhile as the company grows. Integration between AE&CI systems and those of overseas associated companies is practicable and is gradually being implemented.

At present the thesaurus is the key to perhaps 60% of the useful written technical information available within AE&CI; as integration proceeds it is confidently expected that this percentage will increase considerably.

One such post-coordinate system in the field of gas chromatography is in use and has been incorporated into the thesaurus as shown in the following extract.

CHROMATOGRAPHY	CLT
CHROMATOGRAPHY - BIBLIOGRAPHY	LT
CHROMATOGRAPHY (COLUMN)	LT
CHROMATOGRAPHY (GAS)	CLT
SEE ALSO TERMATREX SYSTEM : FOR ACCESS CONTACT GROUP LEADER - ANALYTICAL GROUP - RESEARCH DEPARTMENT	
CHROMATOGRAPHY (GAS/SOLID) SEE CHROMATOGRAPHY (GAS)	
CHROMATOGRAPHY (GEL PERMEATION)	CLT
CHROMATOGRAPHY (ION EXCHANGE)	L

As more systems are developed, as more sections of AE&CI are brought actively within the scope of the overall IR effort and as more records are converted to microform for storage, so it will become worthwhile to distribute copies of the thesaurus more widely to technical staff. The distribution of duplicate sets of microforms of TIRSS, for example, to the major centres of AE&CI may also become economic and worthwhile as the company grows. Integration between AE&CI systems and those of overseas associated companies is practicable and is gradually being implemented.

At present the thesaurus is the key to perhaps 60% of the useful written technical information available within AE&CI; as integration proceeds it is confidently expected that this percentage will increase considerably.

Chapter 7 COSTS AND THE EVALUATION OF INFORMATION SERVICES

7.1 Costing information services

It has been stressed in this dissertation that the technical information service requirements of no two industrial concerns are ever identical, and this divergence of requirements extends further to the nature and complexity of technical information itself and to the uncertainty and very personal nature of each user's needs. Thus it is hardly surprising that when the question of the economics of providing such services is considered, clear cut guidance on costing is not readily available in the literature.

In an effort to clarify the problem, consider an information service as though it was a manufacturing concern in which the product, 'information', is produced in a range of forms. Raw material input is obtained from various information sources and is studied, screened, reviewed, evaluated, and disseminated, or stored in a way that it can be retrieved for later dissemination. The information disseminated is the product, which is distributed to the customer - *ie* to the originator of a request. At the same time the marketing function is necessary in order to obtain new customers, and the after-sales service function to ensure that the product meets requirements. Thus management of this operation appears to be analogous to that of a manufacturing concern. However, in an actual manufacturing concern one of the corner stones of the enterprise is the cost-accounting system which supplies management with reliable and accurate information. Much of the control and decision-making exercised by management is dependent on such information - and it is just at this point that the analogy appears to break down, since until recently there has been an almost total absence of ideas and means for measuring the cost effectiveness of information services.

One may well ask why this has been so? It appears to have resulted from the common procedure of financing library operations as a part of overheads - *ie* to indirect expenditure that cannot be traced to specific services, processes and items. Long before industry became concerned to provide its own library services, municipalities, governments and universities had run them, and these were financed traditionally by indirect sources rather than through charges to the users of their services. Thus it is not surprising that when industry started to set up its own libraries it adopted a similar method. It is hardly less surprising that when information services began to develop they also followed suit, since many of them developed around libraries.

However, pressures from accounting departments, etc coupled with rising costs led, by the 1960s, to a better appreciation of the reasons why information services need to be costed. The reasons include the following:

7.1.1 Budgeting: Normal cost-accounting principles apply to budgeting for an information service and, in general the three major aspects of 1) salaries, 2) materials - <u>ie</u> books, journals, microforms, paper, etc and 3) overheads present no major problems. The common method of arriving at a probable figure for materials is to extrapolate from previous years, allowing for changes in the range of technical interests, publishers prices, etc. [Note: cost analyses for libraries are well covered in the literature and need no comment here]. The total expenditure permitted undoubtedly varies very widely between different concerns and is frequently based on subjective factors - <u>eg</u> the value which top management put on information services. The specific environment in which the service is provided is also an important factor - <u>eg</u> the need for an information service is invariably more apparent in a scientific research environment than it is in a technology based company producing a range of consumer products. In the case of large concerns the total cost can be substantial - <u>eg</u> during the period 1964 - 1973 the figure in AE&CI varied between 9 and 12.5% of the annual research budgets of between R1m and R1.5m [see Section 7.3 for further explanation of these figures].

The question of overheads is relatively simple as far as the costing of information services is concerned since the total charge is invariably dictated by the actual company costing system in use. In fact the charge is calculated by the accountants and imposed on each and every company activity impartially; it comprises such costs as a share of building depreciation and services such as building maintenance, lights, water, heating, sewerage, telephones, subsidised catering, and so on. Thus the conscientious manager of an information service has little or no say in, or control over overhead charges levied against him, but it is obviously important for him to know of what they consist since they can add from 25 - 50% to his direct costs.

Once a budgeted figure has been spent - ie after the end of the particular financial year - the actual costs can be related to each specific activity, such as the actual buying of publications, inter-library loans, indexing, abstracting, etc in order to arrive at unit costs for each. Once such cost analyses have been done it becomes possible to evaluate each activity, to seek ways of increasing efficiency, reducing costs, etc, and to prepare more realistic budgets for future years.

7.1.2 Allocation of resources. Managers of any activity need to plan ahead and know only too well that projects planned several years in advance seldom work out exactly as planned. Conditions, circumstances and pressures change and so do the demands for and of services such as information services. Thus there is a frequent need to trade off between projects and activities, while staying within the overall current budget, and the normal basis of such trade is rands and cents. Experience soon shows that to do this successfully - ie to use money as a common denominator - it is necessary to know the unit costs for the activities involved.

7.1.3 Long range planning: From one aspect long range planning is an extension of Allocation of Resources, but it can go further. The job of planning ahead infers advance knowledge of company development plans, such as for new plants, products or processes, the withdrawal from certain product fields and markets, and so on. This knowledge is clearly necessary so that new or modified services, materials and staff may be available at the right time and so that the budget itself may be realistic.

7.1.4 Intellectual curiosity: As a reason for costing information services intellectual curiosity is a sound one. Curiosity of this type is a real reflection of interest in the whole job and of a determination to manage it well and efficiently. It is only by undertaking cost analyses that the true cost of a service becomes known so that it becomes possible to make comparative value judgments.

To return to the analogy with manufacturing concerns, companies and information services have products, but only in the former case have these usually been sold on the open market. This, it is believed, is a major reason why many information services are still not as cost conscious as they might otherwise be and, in fact, will have to become. However, with the development of commercially operated information services, where the product is sold, much work has been done on developing accurate cost-accounting systems for them. In these systems each activity is itemized as a direct cost and, as an example, one can cite the Aerospace Research Applications Centre (ARAC) operated by Indiana University, Bloomington, Indiana. When ARAC was stated in the early 1960s it was financed entirely by US Government Federal Funds, primarily by NASA since its objective was to promote interest in, and use of declassified information (the so-called 'spin-off') arising from the enormous aerospace research effort. A cost-accounting system gradually became more and more necessary as it was realised that Federal Funds would be withdrawn over a period of time and that it could only expect to continue to exist as a self-supporting entity. Since ARAC's services are sold to anyone who cares to buy them it was appreciated that the existing and perhaps superficial cost treatment was inadequate. Detailed studies followed and accurate cost information obtained, which is clearly essential for the preparation of any worthwhile balance sheet where the proper matching of expenses and revenue is required. By mid-1967 a system had been worked out and was tested over a period of six months at that time before its adoption[242]. By mid-1969 ARAC was almost completely self-supporting and was in a position to say exactly how much a given search of their IR system had cost them, with the customer being invoiced accordingly[243]. Numerous other studies for similar specialized technical information services and for commercial abstracting services have also been reported in the literature[244-257]. Similarly a few industrial concerns in the USA published figures on their costs in the late 1960s and early 1970s for the preparation of abstracts, for searching their IR systems and for disseminating information[257-260], while a few studies have compared the costs of carrying out the same tasks by different techniques (eg informative and indicative abstracts) or using significantly different equipment (eg optical coincidence and a computer)[261-267].

Study of the information published in the above references justifies comment on two points :

7.1.5 With the exception of commercial information and abstracting services where the end product is offered for sale (eg ARAC, Chemical Abstracts Service, etc) the figures reported are of little significance in themselves outside the organisation reporting them. This is because the variables between any two such organisations are always considerable. Nevertheless these figures do show that it is possible to cost both the input and output of an IR system with reasonable accuracy; they also offer guidance to information workers in other organisations who may wish to carry out similar costing and help to ensure that significant factors are not overlooked. With the steadily increasing quantitative size and the complexity of the equipment used in many IR systems, the necessity for accurate cost information for each unit process of input and output is also increasing steadily. This information is essential for the efficient management and economic operation of an IR system; without it management is little more than guesswork - uninformed and therefore perhaps somewhat irresponsible. Unfortunately it is all too clear in industry that some concerns which operate highly mechanised IR systems do not know the true costs of their systems.

7.1.6 There is little that is inherently unusual or difficult about cost-accounting both the 'input' and 'output' of IR systems. Freeman, Director of Science Information Exchange of the Smithsonian Institution, appears to be one of very few authorities who disagrees with the above statement. He states[252] that 'standard business accounting practices are not applicable to documentation and scientific information centers and services without considerable modification'. It is difficult to see the basis for Freeman's assertion, particularly as he does nothing to substantiate it. However he does make it clear that he is not confusing the difference between the cost of getting information into or out of an IR system with attempts to evaluate it. This is an entirely different and much more complex matter - and one that is considerably more difficult to come to grips with. The difference has been pointed out well by C R Blunt et al of

H R B-Singer, Inc (a subsidiary of the Singer Company) in a reported attempt to produce a general model for simulating IR systems[268]:

'The operating cost of an information system is the sum of the operating costs of each function (eg, data collection, input preparation, storage, retrieval and presentation) plus the maintenance and support costs incurred to maintain the operations. Initial costs may include expenditures for research, development, equipment purchases and personnel training. Although cost determinations involve a reasonable direct accounting of expenditures, value determinations are a more complex problem. The value of an information system and its costs are not necessarily in proportion nor are they measured in the same manner. Costs can be quantitatively denoted at every stage of processing from collection to output; however, the value of an information system can only be assessed in a qualitative manner.'

It follows that if Freeman[252] was in fact concerned with evaluation rather than cost-accounting, then his comments are acceptable. For example, it is extremely difficult to set a capital value on a collection of abstract cards, especially as this value may appreciate or depreciate in part or in whole with time.

It has been made clear in earlier chapters of this dissertation that the objectives of and methods used by each IR system determine the actual input and output costs, and the relation between them, for each such system.

7.2 Evaluating information services

So far this chapter has been concerned with the relatively straightforward matter of costing information services; the question of evaluating them is far more difficult. It is understandable that top management in industry should want to know what benefits are arising from substantial investment in such matters as the purchase of publications, searching for information in them or disseminating new information. However like all sociological activities, services of this kind do not lend themsleves readily to evaluation in terms of cash and in fact their evaluation tends to be subjective. In industry the selection of information for inclusion in an IR system is itself subjective, as has been emphasised, while the intellectual job of matching the information stored in a retrieval system with the information that is wanted at a particular moment is also partly a subjective one. Perhaps most variable of all is the purpose for which information is required at all and its importance to the concern wanting it.

Nevertheless intellectual curiosity, as well as managerial pressures, have resulted in attempts to provide tools for assisting in evaluating aspects of these services, such as Westat Research Inc's Guide to the evaluation of document retrieval systems(269). F W Lancaster has clarified the fact that there is a clear difference between 'cost-effectiveness' and 'cost-benefit' for an information system(270):

'We can evaluate an information retrieval or dissemination system from any of the following viewpoints.

1. How well the system is satisfying its objectives, which will usually mean how well it is satisfying the demands placed upon it. Here we are evaluating the _effectiveness_ of the system.

2. How efficiently (in terms of costs) it is satisfying its objectives. This is _cost-effectiveness_ evaluation.

3. Whether the system justifies its existence (_ie_ the system worth). In evaluating system worth, we are concerned with _cost-benefit_ relationships.

Cost-effectiveness, then, deals with the relationship between level of performance (effectiveness) and the costs involved in achieving this level. There may be several alternative methods that could be used to obtain a particular performance level, and these can be costed. _Cost-benefits_ refers to the relationship between the benefits of a particular product or service and the costs of providing it. Generally speaking, benefits are more difficult to measure than performance (effectiveness) except that, in a commercial sense, benefits equate with return on investment. The expression _cost-performance-benefits_ relates to the entire interrelationship between costs, performance (level of effectiveness), and benefits.'

King & Caldwell(271) have derived a model for use in exploring cost-effectiveness factors. This can be of use when the question of deciding between different potential retrieval or dissemination systems arises, provided it is treated as an indicating guideline and not a tool on which to lean heavily. In this regard the theoretician in the information field would do well to remember the warning by E S Quade(272):

'It is important to remember that all analysis of choice falls short of scientific research. No matter how we strive to maintain standards of scientific inquiry or how closely we attempt to follow scientific methods, we cannot turn cost-effectiveness analysis into science. Its objective, in contrast to that of science, is primarily to recommend - or at least to suggest - policy rather than merely to understand and predict.'

Cost-benefits under South African industrial conditions are invariably evaluated on the basis of managerial experience and judgment rather than by the use of any formal techniques, models, etc. To date it is clear from the very wide contacts which the author has with South African industrial concerns operating technical information systems that there has been little or no pressure for a more formal approach. Neither are there any signs of an impending change, so that it can be assumed that the whole matter is either not seen as of sufficient importance to warrant closer investigation or, in general, managements are reasonably satisfied with their judgments (however smug this may appear to be to full-time workers in the field).

Clearly the problem of cost-benefit evaluation is being viewed from a sociological basis and it is common practice for managers to ask key-personnel questions and to use any concensus of opinion obtained in this way as an indication of probable return on investment.

At this point it is as well to be reminded that many scientists and technologists consistently prefer, and often place greater reliance on personal (person-to-person) communication with colleagues, recognised authorities, etc - the invisible college hypothesis - than on communication through the information in documents(255,273). The Engineers' Joint Council, publishers of the Engineering Index, which abstracts the bulk of new information in the various fields of engineering technology, estimates that their service meets only 10% of the total information needs of engineers(274). If the same estimate holds good for commercial abstracting services for the sciences and also for the

IR systems set up within industrial organisations, then it becomes extremely difficult to justify the existence of such services on straightforward economic grounds. However, in spite of such, perhaps, pessimistic estimates, IR systems continue to be developed and to prove their value in ways which cannot often be calculated in terms of rands and cents. This is not really surprising since services such as IR systems are closely related to the overall welfare economics of a company.

Consider some of the more important aspects of the whole matter:

7.2.1 In scientific circles there is general agreement that adequate awareness of all relevant technical information can be decisive in determining the success of research and development projects. G J Brockis and P F Cole of British Petroleum have pointed out that, generally speaking, there are two aspects to the provision of technical information in the research and development context(275):

'One of these is creative in character and reflects the ability of information to stimulate new lines of investigation and fresh approaches to existing lines of investigation. The second, a defensive aspect, is concerned with guarding against the unintentional duplication of research which has already been performed elsewhere and of avoiding losses resulting from failure to take into account all available relevant information.

The problem of evaluating the creative aspect appears almost insurmountable. Indeed, only in those exceptional circumstances where the discovery of specific information can be identified as the stimulus which originated a line of investigation leading directly to a new product of process, or some identifiable economy, is any evaluation possible.'

(Note: See also Chapter 2, sub-sections 2.1.4 and 2.1.5 on creative ideas and the duplication of research, respectively).

Many of the published attempts to evaluate the defensive aspect as usually reported(275-277) are concerned with unintentional duplication of research. The figures quoted are usually on a national basis, and consequently very large; they are claimed to be based on reasonable assumptions. Figures for particular industries or for specific companies within an industry are just not available, and in a competitive economy this is hardly surprising. However individuals employed in the research and development fields can nearly always recall instances where information retrieved from an IR system had one or more of the following effects:

1 it changed the course of the work
2 it shortened the time required to conclude a job
3 it pointed out difficulties which had not been fully appreciated
4 it led to benefits and successes additional to the original objective
5 it led to a preferred method of tackling a given problem (often by analogy).

All of the foregoing can sometimes be allocated a monetary value, but only after completion of the particular job. Another aspect is that information is sometimes retrieved after the completion of a job, when it is realised that earlier retrieval would have affected the outcome in one way or another as above.

7.2.2 It is not only in the sphere of research and development that the two aspects of creativity and defence are apparent. Requests for information in connection with 'trouble-shooting' on plants, technical service for customers, specialised subject fields within an organisation (eg packaging, materials handling, etc) can all show both aspects. Perhaps the reason that research and development has received more attention is that the value potential is higher there than in the more mundane, and usually much smaller scale of problem in trouble-shooting, technical service, etc. It is certainly no easier to attach a cash value to many of these smaller scale problems.

7.2.3 Some investigators have suggested that the easiest way to evaluate IR systems is to arrange an internal charging system for each and every search[92, 255, 278-282], while Ramey[273] would apparently base his evaluation on the cost per answer obtained from an IR system or the cost per question satisfied as opposed to the cost per search. Such proposals are really for cost-accounting systems which enable operating costs to be determined, but contribute nothing to telling what the organisation concerned is really getting for the money spent. One point worth noting is that Ramey in particular acknowledges that not every search is successful and that the successful ones could reasonably carry the total cost of all searches. In practice this is exactly what does happen. However, as with the other investigators he appears to have missed the point that an item of information, even when of use, is rarely of equal value with any other item of information that is also of use to the organisation retrieving it.

7.2.4 Thus, to sum up the whole subject of costing and the evaluation of information systems, the valid conclusions are few in this difficult subject field and appear to be as follows:

1 An industrial concern establishing an IR system for the first time should ensure that provision is made for cost accounting it. This is highly desirable for purposes of good management and, while it may appear to be a statement of the obvious, it has in fact been done by very few companies in South Africa.

2 Generally speaking, in the practical environment of industry it is impossible to evaluate the 'creative' aspect of retrieving information from documents.

3 It is sometimes possible to evaluate the 'defensive' aspects of information retrieval, but only retrospectively after the completion of a job. The accuracy of such evaluation must of necessity be dependent on what cost records have been kept - whether the evaluation is actually worthwhile will depend on the merits of each case.

7.3 Costing and evaluating AE&CI systems

The figures quoted for technical information service costs within AE&CI in sub-section 7.1.1 require explanation.

1 The total costs of 9 - 12,5% of the research budgets comprise both direct costs (salaries, wages and materials - ie purchases of publications, paper, machines, etc) and indirect costs (overheads).

2 The services themselves are provided to the whole company, to its local and overseas associates and subsidiaries and, to some extent, to its customers. They are housed physically in the Research Department since the staff of this department collectively comprise the biggest single user - ie about 35% of the total. These services include document repositories or storage facilities (ie the central library, report files and correspondence files), information dissemination services (ie current awareness and selective dissemination of information (SDI)), the IR systems described in this dissertation, and various other ancilliary services such as technical writing (preparation of company technical brochures for production, machine operating and maintenance manuals, technical report writing advisory service and training courses), literature searching service, translation service, and repro[illegible]ic services (photocopying, offset-printing and microfilming).

3 Of the total information services costs (salaries, materials and overheads) for 1973 of R212 000 the costs for input to the three IR systems amounted to R22 300, made up as follows:

TIRSS	:	9 300
CORIS	:	4 300
LIBRIS	:	8 700
		R22 300

These figures are discussed in the following sub-sections:

7.3.1 TIRSS

During 1973 a total of 1 192 items were processed into TIRSS - of these 832 were technical reports and 360 were other items (important correspondence, journal review articles, Government Gazettes concerning technical matters, etc). The costs involved in this processing were:

1 the time taken to abstract and index each item

2 the time taken to edit each abstract and check for consistency in indexing

3 a share of the cost of maintaining the thesaurus

4 typing and checking the typescript

5 duplicating the appropriate number of cards

6 filing the cards

7 a share of the cost of a supervisor.

Thus the total cost per item was $\frac{9\ 300}{1\ 192}$ = R7,80, made up as follows:

		R	%
1	Abstracting and indexing	4,20	53,8
2	Editing	0,35	4,5
3	Share of thesaurus	0,20	2,6
4	Typing and checking	1,80	23,2
5	Duplicating	0,80	10,2
6	Filing	0,05	0,6
7	Supervision	0,40	5,1
		R7,80	100,0

This figure comprises for each item a very small fraction of one percent of the average cost of a research programme undertaken by AE&CI. It is considered to be a small premium to have to pay for ensuring the prompt retrieval of technical information arising from the money invested in each research programme.

It is not known how this figure compares with similar totals in other industrial concerns since most of the few figures which have been published refer to US companies. Experience has shown that worthwhile comparisons with costs in North America are extremely difficult to make since salary structures there have for long been completely different to those in all other parts of the world.

7.3.2 CORIS

As pointed out in Chapter 6, Section 6.2 only a small percentage of the total number of items of correspondence received in and originating from AE&CI's Research Department are considered to contain technical information of sufficient long term value to warrant their inclusion in CORIS. In 1973 a total of 8 000 items are estimated to have passed through the hands of the correspondence clerk. Of these 1 208 items - ie 15% - were considered to be of sufficient value to warrant inclusion in the system.

In the handling of correspondence numerous minor steps are involved which do not merit listing and analysis. Thus the overall cost per item handled was $\frac{4\ 300}{8\ 000}$ = 54 cents each. However the correspondence clerk spent approximately 50% of her time processing the 1 208 items into CORIS, 30% on the balance of 6 800 items and 20% on other work (microfilming, searching, etc). Thus a truer reflection of the costs of handling correspondence would be that each item processed into CORIS cost $\frac{4\ 300 \times 50}{1\ 208 \times 100}$ = R1,79 while each item not so processed cost $\frac{4\ 300 \times 30}{6\ 800 \times 100}$ = 39 cents.

No comparable figures have been traced in the literature but the CORIS system works satisfactorily and meets its objective of ensuring that important correspondence can be retrieved when it is needed. Thus the premium of R1,79 per item is considered to be reasonable.

7.3.3 LIBRIS

During 1973 a total of 1 156 non-journal items were added to the library and of these 881 were books and 375 were pamphlets. The costs involved in processing these were:

1 the time taken to prepare an indicative abstract and to index each item

2 the time taken to classify each book using simplified UDC or to number each pamphlet in accordance with the internally developed system in use

3 a small share of the cost of maintaining the thesaurus

4 typing and checking the typescript

5 duplicating the appropriate number of cards

6 display, shelving, etc

7 a share of the cost of a supervisor.

Thus the total cost per item was $\frac{8\ 700}{1\ 156}$ = R7,53, made up as follows:

			R	%
1	Abstracting and indexing		3,81	50,6
2	Classifying		0,62	8,2
3	Share of thesaurus		0,05	0,7
4	Typing and checking		1,80	23,9
5	Duplicating		0,80	10,6
6	Display, shelving, etc		0,05	0,7
7	Supervision		0,40	5,3
		Total	R7,53	100,0

At this point it is of interest to compare the percentage costs for TIRSS and LIBRIS and to note that major items, particularly abstracting and indexing, show no significant difference.

It has proved possible to obtain some cost figures from outside AE&CI for purposes of comparison, although in each instance the library in question requested that its name should not be disclosed. It is clear that the procedures and standards of thoroughness varied between the organisations concerned but in all instances the costs disclosed are those incurred from the time of receipt through to shelving each item.

TABLE 3

COMPARATIVE COSTS OF PROCESSING BOOKS IN LIBRIS AND FOR SOME ORGANISATIONS OUTSIDE AE&CI

Organisation	Comment	Cost per item, 1973
1 AE&CI	Non-conventional LIBRIS system	R7,53
2 A large iron and steel industrial corporation	Classification by simplified UDC - no abstracting	approximately R7
3 A large Government supported research organisation	Very detailed classification by UDC - no abstracting	between R10,50 and R12
4 A large public library	Volumes in the Reference Section only - Dewey. Efficient but rather perfunctory treatment	approximately R4

From these figures it would appear as though the costs of the LIBRIS system are close to the average for larger organisations - ie those which tend to have high overhead costs. It is not unreasonable to assume that in smaller organisations the true costs for a LIBRIS system would be (statistically) significantly lower.

Chapter 8 THE MANAGEMENT OF INFORMATION SERVICES

8.1 Discussion

Although this dissertation is concerned primarily with the storage and retrieval of technical information this is but one , rt of the whcle activity covered by the phrase 'technical information services'. Provided that there is good understanding of a whole activity it is a valid scientific approach t› subject a part of it to intensive study and development. However this type of approach is not ally appropriate to consideration of such matters as the management of information services; a broader approach is desirable and has been adopted for this chapter.

Firstly, and perhaps most important of all, there is no difference between the management of a technical information service and the management of any other activity within industry. The personal qualities and abilities expected of a manager naturally vary somewhat between, say, production, sales and service jobs - the name 'technical information services' makes clear the category into which a manager of this type of activity falls. Thus it is especially important in the management of such services to have clear objectives relevant to the provision of a service, whether this is to a small or large community within an organisation.

The search for greater managerial efficiency in the increasingly complex environment of industry has led to the development of various techniques. Mason(283) has discussed systems analysis, operations research and market research in a clear and simple manner, as well as the general applicability of Organisation and Method (O & M) studies and Management by Objectives (MBO). The literature of industrial and personnel management covers these subjects in extenso and they do not require elaboration here.

Nevertheless every manager should want to improve his performance and basic aspects of how to tackle this are set out concisely in a paper by J W Humble(284). Since this is a subject which does not appear to the author to be of sufficient concern to many managers of information services in South Africa, the following points made by Humble are stressed:

1. In attempting to improve the performance of managers lasting success depends primarily on the interest and attitude of mind of the Chief Executive. It is he who must think through the objectives of the business and provide the practical facilities for them to be achieved. He must develop a genuine sense of participation amongst his managers and be receptive to their constructive criticism. He must concentrate on accomplishment, rather than promotion and be willing to delegate responsibility and encourage risk taking.

2. It follows that every manager has a dual responsibility: - to develop himself, and to create a demanding environment which will stimulate the self development of his subordinates. The span of executive command is the span of direct personal responsibility for subordinate growth.

3. There must be a sustained effort to clarify objectives for the company and to reconcile with these the personal (job and emotional) goals of individual managers. This unified sense of purpose within a business, concentrated on what really matters, can be a formidable tool of profit improvement and cost reduction.

4. The deep involvement of managers in defining key results, securing change and reviewing success is fundamental to the successful management of any activity.

5. The focus is on results. It is concerned with measurable benefits for manager and company, not with vague generalisation and platitude. Thus, it is not enough to agree what a man must do, but rather what he must achieve; not enough to analyse a problem, but rather to produce an action plan to begin to solve it; not enough to issue instructions but to associate with them measures of accountability.

'Improving Management Performance' is dynamic in philosophy, looking at tomorrow's opportunities rather than yesterday's mistakes. It follows that forms and procedures are not sacred. Flexibility in ensuring that the company's human physical and financial resources are concentrated in the areas of greatest opportunity is dominant.

6. This concern with results changes the emphasis from personal style of management and personality factors to the achievements of a manager. It is the middle road between the view of human beings as 'just another machine' and the extreme school of 'Human Relations' which emphasizes social factors to the detriment of technical and economic factors in business.

7. The improvement of performance and people is regarded as a continuing process, for the business and all its managers, not a shot in the arm programme for a chosen few.

All of the points made above are relevant to the management of technical information services, but if any one is especially pertinent it is perhaps that 'forms and procedures are not sacred'. A common fault in many information retrieval and library situations is to regard the system in use as sacrosant so that it becomes an objective in itself to maintain it, and to do so without changes. Nothing could be further from the true objective of providing a service which is both effective and efficient in the continually changing conditions within industrial organisations. Further, L F Urwick[285] has warned that a business cannot solve its problems merely by buying a technique or system:

> 'Every business enterprise is a living organism, with its own traditions, its own climate of opinion, its own special make-up. Every situation is different. And every kind of system has to be custom-built to the individual business'.

The unique character of 'the individual business' has been stressed throughout this dissertation and the point for a manager to remember is that having bought a system, or even having designed one, it is the practical requirements of the company which dictate its possible modification and the way in which it is applied.

In South African industry it is very probable that the management of many information services will be a part-time activity for an executive who already 'performs a multiplicity of functions, and often carries a heavy work load' as pointed out by Myatt[286] in a paper on the special conditions in small companies in North America. Similar conditions prevail in the Republic where, in fact, a very high proportion of industrial concerns are small by world standards. But whatever the scale of operation may be, the rapid pace of scientific and technological progress, and the complexities of effective management at top and middle management level demand information which is current and relevant, comprehensive and factual - and nearly always in a predigested form. This is true of problem solving and also of management's planning and control function.

This poses a challenge, which, if it is to be met, necessitates that all aspects of the information transfer process within an organisation be understood thoroughly, and also that each step in this transfer process be carried out as economically and as efficiently as possible. For technical information services to be really successful it is necessary for the management of the service to be fully committed to this challenge. This commitment is

1 implicit in the tasks of collecting, storing, retrieving, and disseminating, as appropriate, both documents and/or the information in them - certainly in activities as industrial research and development such services are a support operation in the search for, and selection of lines of research and new product opportunities, and

2 necessitates a formal 'information policy'. This has been well summed up by Kjeld Klintøe as follows[287]:

'An enterprise can only afford to make use of a very limited part of the enormous amount of information available. It must therefore state a policy of information containing the following:

1 Definition of the concept and the objectives of the enterprise
   - Competition policy (price or quality?)
   - Manufacturing policy (mass production, serial production, or manufacture of investment goods?)
   - The nature of the market wanted (local, regional or global?)
   - Policy of technological development (to keep pace with development, to keep pace and occasionally lead, or to be in front all the time?)

2 Description of the nature of information which must be procured
   - Experience gathered by staff members
   - Information on customers' members
   - Reports (on science and technology, finance, management, organisation, etc)
   - Literature (journals, primary and secondary)

3 Description of the organisation of the internal information service
   - Centralized (with the use of advanced equipment)
   - Centralized-Decentralized (with participation of the staff)
   - Decentralized (making use of the responsibility of individual staff members).'

A B Westerman of the Battelle Memorial Institute points out that the most critical factor for any technical information service is its objective(288), and one of the most important objectives of any such service is to aid corporate management in optimum decision making. As such it is one of a number of such aids available to industrial management.

H C McDaniel of Westinghouse Electric Corp has pointed out(289):

'It doesn't really matter where a manager stands in the organization structure -- bottom or top -- he has to make decisions. And usually when he makes a decision he commits his company in some degree to a course of action. If his decision is a right one, his company prospers and flourishes. But, if it is a wrong one, his company suffers accordingly. And here's tragic evidence that wrong decisions are made with surprising frequency: Only one-third of the first hundred companies in this nation (USA) in 1909 were on that list fifty year later (1959).

How large a role inadequate management information played in this economic plague will never be known. But the chances are, it was larger than imagined.

Today, many companies are faced with this problem: not enough of the information available is organized for management use in decision-making. And while there are perhaps as many reasons for the predicament as there are companies faced with it, it's generally recognized that the origin of the problem lay in the inability of a static information system to meet the needs of a dynamic organization structure. .... And with these new jobs has come new responsibilities, new decision-making authorities, and reshaped reporting relationships -- all of which have combined to generate new demands for information -- information that is usually missing in existing systems. Hence, the great hue-and-cry about the "information crisis".

Your stake in this crisis is as great as anyone's because you frequently produce the technical information management so sorely needs to round out its data and intelligence. The question is: Do you produce this in the form most useful for management decision-making purposes? ....

In all of the definitions of the word "decision" that I've read -- and this includes the dictionary -- there exists a central idea: finding a possible course of action by selecting one alternative from among two or more alternatives. My dictionary says it's "the act of deciding or settling a dispute or question by giving a judgment".

commits his company in some degree to a course of action. If his decision is a right one, his company prospers and flourishes. But, if it is a wrong one, his company suffers accordingly. And here's tragic evidence that wrong decisions are made with surprising frequency: Only one-third of the first hundred companies in this nation (USA) in 1909 were on that list fifty year later (1959).

How large a role inadequate management information played in this economic plague will never be known. But the chances are, it was larger than imagined.

Today, many companies are faced with this problem: not enough of the information available is organized for management use in decision-making. And while there are perhaps as many reasons for the predicament as there are companies faced with it, it's generally recognized that the origin of the problem lay in the inability of a static information system to meet the needs of a dynamic organization structure. .... And with these new jobs has come new responsibilities, new decision-making authorities, and reshaped reporting relationships -- all of which have combined to generate new demands for information -- information that is usually missing in existing systems. Hence, the great hue-and-cry about the "information crisis".

Your stake in this crisis is as great as anyone's because you frequently produce the technical information management so sorely needs to round out its data and intelligence. The question is: Do you produce this in the form most useful for management decision-making purposes? ....

In all of the definitions of the word "decision" that I've read -- and this includes the dictionary -- there exists a central idea: finding a possible course of action by selecting one alternative from among two or more alternatives. My dictionary says it's "the act of deciding or settling a dispute or question by giving a judgment".

When a manager is faced with choosing one course from among two or more alternative courses, he must have a reason for making a choice. Unless he knows the objective to be achieved -- regardless of what the objective might be -- how can there exist a basis for determining the soundness of choice? From this, I think we've got to conclude that while there can be many motivating reasons for making a decision, the primary motive is the goals and objectives of his company.

From the foregoing, two conclusions can be drawn: First, the decision-maker cannot give a judgment without data and intelligence, facts and competent opinions; and second, since you are responsible for a vehicle of communication used by managers, you must design, organize, and structure these vehicles for the purpose intended: decision-making.'

However, there are other objectives for technical information services that are also of considerable importance. In the environment of research and development Dr H Skolnik of Hercules, Inc has succinctly drawn attention to several[290]:

'Underlying most functions and activities of the Technical Information Division is the philosophy of freeing the Hercules scientists as much as possible for accomplishing their technical objectives. If we increase their productive reading of the technical literature, we increase their research productivity. If we provide them with answers to their questions of the technical literature when and to the degree they want, we increase their research productivity. When we are successful in these objectives, we are managing information operations and services successfully.

Management of an information operation requires both an awareness of the larger goals which are being supported and also a continuing rapport with higher management or the sponsor and with those whom the facility serves. Successful management implies creativity in the design of systems, proper attention to the component parts, control over the contents, efficiency in the use of people and time, and subject area knowledge important to the environment being served. It further implies the recruitment of qualified personnel and the establishment of an environment that encourages growth with suitable rewards and challenges; it implies methods for measuring the contributions of the information operation and of its personnel.

A key management function is communication with those who finance the information facility and with those who are served by it. Good communication is needed for controlling current operations and planning for the future, for determining objectives from a number of alternatives open, for efficient use of the resources that may be available, for reaching rational decisions. <u>Management must be committed to the success of its parent organization or to the goals of its sponsor</u>.'

In Chapter 7 reference was made to the almost impossible task of costing the 'creative' aspect of information services. This does not mean however that stimulation of the creative aspect is not an important objective of the management of information services, for indeed it is. James Hillier of the RCA Laboratories has the following comments on this subject(291):

'In other words, the individual in a creative pursuit will have learned from experience the channels through which he has the highest probability of obtaining a useful piece of information and will tend to select those channels whenever a choice is involved.

To be effective in this framework, an information service will have to be able to demonstrate by actual performance that it provides a more profitable input channel to the creative individual. Moreover, it is the judgment of the individual rather than that of management that will prevail.

These thoughts begin to raise some questions as to the possible effectiveness of an information service.

Let us now turn to consideration of the process of creative action and to its impact on information flow to the creative individual. It is generally agreed that one of the essential conditions for the initiation of a creative action or idea is that the necessary information, or deductions therefrom, come together in time and in a single mind. It is also obvious by definition that these coincidences cannot be programmed - that is, neither the creative idea nor the information necessary to spark it can be specified in advance. This leads to the immediate conclusion that creative action, and therefore research, is probabilistic in nature. Thus, to the limitations and uncertainties imposed by a limited human communications system, we have added the further uncertainties of timing and combination in the creative individual's mind.

Management's responsibility in this situation becomes clear. In order to maintain or increase the productivity of the laboratory, it must find means of increasing the probability of creative action by the technical staff. Management can do this by increasing the individual probabilities involved. There are some techniques for accomplishing this, at least in part. (See 'A theory of communication in a Research Laboratory'[292] - also by James Hillier).

The specific question before us here, however, is that of the value of a scientific information service group placed either in series or in parallel with the scientist-information-retrieval link to preselect or filter the information flowing from the scientific literature. On the one hand, we have the need to find a way to display more of the ever-increasing mass of published scientific information to the creative individual in spite of the fixed limit on his communications channel. On the other hand, such control as there is over the probable occurrence of a creative action is in the hands of the creative individual and depends upon his lifetime of experience and upon his intimate personal goals. His success as a creative individual is a measure of his ability to optimize these probabilities without any a priori knowledge of the specific idea he will encounter or of the specific information needed to arrive at the idea.

Conclusion - It becomes clear from this brief look at the major ingredients of the problem of providing information services to a research laboratory, that management's considerations must go beyond the simple provision of extra brains to perform a function that has grown to exceed the capability of the individual.'

However there is no doubt that many research workers have what is perhaps a natural aptitude for, or inclination towards providing an environment in which their creative abilities have full play. This inclination appears to be more subconscious than deliberate and to introduce an element of what might be called 'dreaminess' before the familiar pattern of logical thought takes over. This condition has been well illustrated by Richard M Willstätter, the German Nobel prize-winner, who elucidated many of the problems associated with chlorophyll, plant pigments, etc. He wrote in his autobiography[293]:

'The use of books and periodicals was a habit I had learned and practiced early. Reading a good book at leisure, a book with some bearing on what I have hoped, felt, worked on, experienced, and lost, puts me in a solemn and excited mood, like looking at the restful sweep of a great, wide landscape. Often I have to put my book aside so that I can seize and hold a thought which suggests itself dimly behind the printed pages and gains clearer content and full shape only on being followed. It is only rarely that such traces of ideas betray a recognizable relationship with the subject of my reading, the pursuit of which they hinder. Various ideas are born of some sort of excitement through reading; often they are the beginnings of scientific experiments, or comments on a manuscript, or they even bear on events long gone and forgotten -- in short something which, remote from the sentences of my book, spins out or even fulfills my plans, impressions, and oft-upset desires.'

It appears reasonable therefore to contend that the creative research worker should acquire the habit of reading widely. No information service can do this for him but what it can do is to provide well prepared abstracts focused on and around his subject field - not as a substitute for reading but, as Hillier[292] has pointed out, as a guide through the 'information explosion' so that he can avoid much time-consuming searching himself.

For at least the first of the above objectives to be achieved - ie the assistance which an information service can provide in optimum decision-making - there are a number of basic factors to be considered. Firstly, and probably the most important one in a country such as South Africa, where industry is still young but expanding rapidly, is the environment in which information operations are conducted. To provide information that is relevant, useful, etc, it is necessary for corporate management to support the activity unreservedly. Certainly in research and development work they should create an environment that seeks the active participation of the information service staff in each phase of the activity itself. In response to this participation in the mainstream of the research, the staff of the information service should commit themselves to providing information that is useful and of high quality, and to ensuring prompt response to changes in information user requirements.

Secondly, as far as may be economically possible, user needs should always override the continued maintenance of any systems and techniques currently in use when such systems no longer meet requirements. The consistency and uniformity of the information available to users from a familiar and long used system has its merited place in any information service. Nevertheless, when the system no longer meets user needs it should be modified or changed, and the organisation of the service should be sufficiently flexible to permit this with minimum disruption. By the same token as long as a system does meet user requirements then there is no justification for changing it. These facts are also clearly implied in Humble's comments on improving the performance of managers referred to earlier in this Chapter[284].

Thirdly, it follows from the foregoing that systems cannot exist for their own sake. They exist for the benefit of the users and these latter should be integrated into the service, not merely as customers but whenever possible as planners and designers. Involvement and participation by users in the design and continuing operation of a system not only help to identify information needs, but also develop a commitment by the user to ensure greater or continued efficiency. (See also reference in Chapter 2 to user involvement). Thus the user works to maximise the return on the investment in the system, however complex or simple it is.

Fourthly, as has been stressed throughout this dissertation, industry rarely has an objective and impartial approach to information on any subject - whether this is scientific, technological, economic, or commercial. The relevance and value of a given piece of information is never the same for any two concerns. Further, all information, unfortunately, is not in itself equally valuable, relevant or usable. This means that it has to be studied, screened, reviewed, evaluated, and so on, all in the context of the interests and requirements of the concern collecting it, before it can be properly used.

In view of the points made above and because of the variable nature of the responsibilities and activities of information services in different concerns, it is not possible to specify management guidelines and principles which are universally applicable and appropriate. However, there are some general observations which can be made:

8.1.1 Without exception, major decisions which affect the direction and progress of a business are based on information of some kind or other - usually on mixtures of kinds, such as technical and economic for example. This information arises from many sources, of which formal information services are but one, and the higher the level at which decisions are made, the more pre-digested and concentrated is the form in which information must be presented. The more important the decision, the less likely it is to depend directly on raw information direct from some system or service - nevertheless the soundness of a decision will frequently be related to the extent to which, and the way in which information is collected, processed and distributed. Thus by definition information systems are integral to the decision-making process; any question as to the need for information is not at issue and the point serves to emphasise the opinions expressed by McDaniel and quoted at length earlier in this Chapter(289).

8.1.2 The extent to which information services are integrated into the total decision-making process is very variable. This variation is directly related to the investment that top management are prepared to make in providing formal information services, or, from a different point of view, to the services rendered to the information users. It is obvious then, but perhaps not mentioned frequently enough that the overall objective of any information service is service; specifically service of high quality and efficiency. Whenever this objective is achieved and maintained, management and user confidence in the corporate role of information services is reinforced and their collective and individual involvement extends well beyond mere organisational recognition and token use. Active participation in the mainstream by the information staff is expected, and it is apparent to the author that it is this participation which is to be found in effective efficient services. In turn such participation tends to stimulate users to seek influence in the design and operation of information systems, and this gradually leads to the establishment of mutual confidence between the information worker and the information user.

Experience has shown that it is only when team spirit is evident between information worker and information user that natural communication barriers are eliminated, or at least greatly reduced. When an information service is mentally fragmented by the users so that the IR system, the main library, etc are regarded as being units on their own, there are always communication barriers. Under these conditions both the service and its users are the losers.

8.1.3 Much has been written about the relationship of adequate, rapid and accurate communication to the achievement of management in general and of information services in particular(4,23,32,40,84,289,294-296). The emphasis on effective communication stems from the recognition that people are, after all, the working parts of any organisation, and that a participating role stimulates contribution. Certainly in industrial research, the successful researcher is one who not only devises well-thought out experiments but is able to communicate the results clearly and concisely both orally and in writing. B C Brookes, who is reader in Information Science at University College, London said in an unpublished paper and quoted in a letter to the author(297):

> 'Workers who cannot talk or write coherently about their scientific work, who are good with their hands, clever at their experimental work, but inarticulate, are not scientists'.

He would call them technicians or mechanics, since research work serves very little purpose unless the results are communicated to others. The important point from the information service aspect is that if the individuals who comprise the staff of such services cannot communicate effectively then they will not be able to assist effectively in the decision-making process (point 8.1.1 above) nor be accepted as an integral part of a team striving towards some clear cut goal (point 8.1.2 above).

8.1.4 In conclusion, the desirability of the accurate cost accounting of information services, as discussed in Chapter 7, is emphasized again. The manager of a technical information service usually has little control over the actual salaries paid by his employer nor over the costs of all the sources of information (purchases of publications, etc) or of internal overheads or of the equipment used. His control is rather over the quantity and quality of the labour used and the materials and equipment purchased. One thing is certain however and that is that all these costs have been rising steadily, and at the same time information services within each industrial concern have also been