THE EVOLUTION OF AN ON-LINE CHEMICAL SEARCH SYSTEM FOR AN INDUSTRIAL RESEARCH UNIT

A thesis presented for the degree of PhD

at

The University of Sheffield

by

Diane Rosemary Eakin

April 1977

Postgraduate School of Librarianship
and Information Science

BRITISH LIBRARY

IMAGING SERVICES NORTH

Boston Spa, Wetherby
West Yorkshire, LS23 7BQ
www.bl.uk

PAGE NUMBERS CLOSE TO THE EDGE OF THE PAGE.

ACKNOWLEDGEMENTS

The author wishes to express her gratitude to everyone who contributed in whatever way to this study. She wishes to express her deep gratitude to the following individuals without whose willing assistance this study would not have been possible:

Imperial Chemical Industries, Pharmaceuticals Division, who financed this study and provided research facilities.

The Research Department of the above-mentioned Company for their encouragement and keen interest in the study, especially Dr D G Davey and Dr J M Pryce.

The author's work group, Data Services Section, who gave help and encouragement throughout the study, particularly Mr E Hyde.

Professor M F Lynch, Professor of Information Science at the University of Sheffield who supervised this work and whose guidance and enthusiasm throughout this study have been a lasting encouragement.

Dr G W Adamson, Mrs J Radcliffe and Dr T D Wilson who provided assistance and valued advice.

STATEMENT OF AUTHOR'S CONTRIBUTION

Working within ICI, it has been possible to apply the ideas generated whilst investigating this problem, in a practical environment. The author was part of a group and was given programming support where appropriate.

The chemical data bases evolved from the author's own designs, but the programs necessary to implement and use the data bases are largely the products of colleagues. The author's role was to write the specifications or to program pilot applications.

The main part of the study, dealing with the analysis and monitoring of user responses to these data bases, was solely the work of the author. She was responsible for defining the required measurements, collecting the observations, recording the data and performing the final statistical analyses.

TABLE OF CONTENTS PAGE

TABLE OF CONTENTS (Cont.) PAGE

TABLE OF CONTENTS (Cont.) PAGE

TABLE OF CONTENTS (Cont.)

LIST OF FIGURES

LIST OF FIGURES (Cont.) PAGE

LIST OF FIGURES (Cont.) PAGE

THE EVOLUTION OF AN ON-LINE CHEMICAL SEARCH SYSTEM FOR AN INDUSTRIAL RESEARCH UNIT

BY DIANE R EAKIN

SUMMARY

The objectives of this study were to design an information system, using modern computer technology, to meet a research chemist's need for chemical structural information, to quantify the effects of increasing degrees of computer technology on the use made of the facilities, and to relate the use of the service back to the individual chemist, his performance and background.

A computer system was developed based on Wiswesser Line Notation and molecular formula as the chemical structure descriptors. Systems design and analysis were performed so that access to the information could be obtained directly for individual compounds and more generally for classes of compounds.

As the system was being developed, its use by information staff was monitored by constant interaction with the people concerned. Where appropriate, the system was modified to meet information staff's requirements, but a number of precautions had to be introduced to prevent mis-use.

The research chemists' use of the information services was studied retrospectively over a two-year period. In addition to the use made, several other factors were observed for each chemist. These included performance measures and background information on the chemists' research role.

The data showed a steady increase in the demand for the services by the research chemist as the degree of computerisation increased. The use made of the services related closely to the number of compounds prepared by each chemist, but there was no significant correlation between a chemist's success in preparing biologically active compounds and his information use.

The very individual way in which chemists conduct their research was highlighted by the wide range of use of the information facilities and the low correlation with background factors. This makes the design of on-line systems for use by chemists themselves complex and justifies the existence of the information scientist as an interface.

Chapter 1 - Statement of the Problem: Introduction

1.1 Communication in Scientific Research

The problems of effective communication in scientific research are much in evidence today. As inflation takes its toll, so it becomes more important that scientists be aware of the work of other scientists. There is little hope of making significant scientific progress if this does not happen. But how do scientists communicate?

In the days of the monastic scholar, the handwritten manuscript served as the medium for communication. This proved satisfactory until the fifteenth century when the large increases in population and in literacy required an improved means of communication. The problems were eased by the introduction of the printed book - it was estimated that over two million books were issued between 1460 and 1500 (1). Scholars began to use the book as the chief means of exchanging facts and ideas, and the availability of the book provided the background for the rapid growth of science.

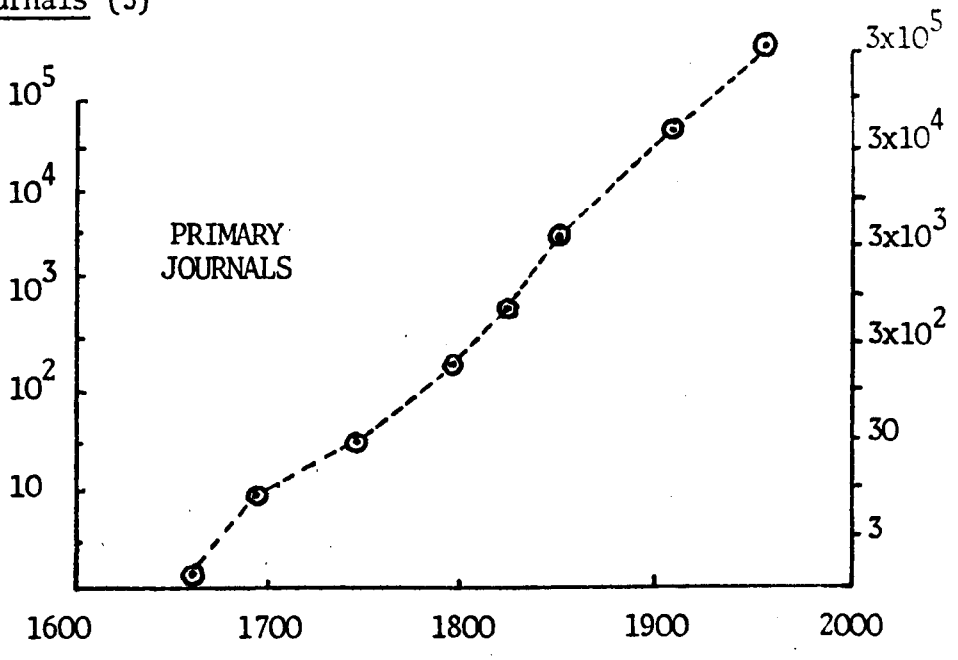
By the seventeenth century new information was being obtained so quickly that even books were becoming too slow a medium for effective communication. As a result, societies began to be formed so that groups of scientists could directly exchange facts and ideas. Gradually the proceedings of these meetings began to be printed so that records of the discussions were available. The first recognised scientific journal, the Journal des Scavants, appeared in January 1665 (2). During the eighteenth century many more scientific societies published transactions and proceedings, and this facilitated an interchange of information amongst the various disciplines. Figure 1 plots the growth in the world's scientific journals since their inception in 1665 and illustrates the startling growth rate (3,4). The importance of the journal in scientific communication is undisputed (as illustrated by the fact that journals have continued in approximately the same format for over 300 years).

However, the journal is more than a medium for the transfer of information, it also conveys prestige and recognition (5). There is intense competition to publish new ideas in science, and a primary reward for a scientist is to have his work respected by competent fellow workers (6). Scientists in industry are limited as to the amount and type of scientific publication and presentation they can make. Publication in the external literature can, therefore, not be the only reward. One might, however, expect to see some parallels in their productivity within internal company information systems, and the way in which other scientists use the main body of literature.

Many feel that this intense competition to establish priority in science is the main cause for the exponential increase in research communications, and for the continuing need for today's scientist to keep abreast of the literature (7). Indeed, today's scientist wishing to remain up-to-date in his awareness of latest developments in his particular field and who wishes to make new contributions to science, faces unprecedented problems. He cannot hope to scan all the available literature for relevant communications or for answers to particular queries.

However, far-reaching changes are again beginning to emerge, much the same as the development of the book in the fifteenth century. The application of computer technology to the publication, dissemination and retrieval of scientific and technical information, is changing the patterns of scientific communication.

Figure 1 - Increase in the world's total number of scientific journals (5)



The objective of the study was to design an information system using modern computer technology to meet user's needs for the storage and retrieval of internally-generated chemical structure information. In meeting the user's needs, it tried to establish the effect of increasing degrees of computer technology on the use of the information facility by the user population, and to measure the acceptability of the changing service. The use of the service was related back to the individual, his background and performance in the hope that relationships could be found between the group of users and the level of computerised information facility.

There had been a number of studies into the communication behaviour of scientists all pointing to the immense variation in individual behaviour. To study the effect of a group of scientists on an information service, it was necessary to be aware of the fundamental differences in scientist's behaviour.

Crane showed that this was influenced by the scientist's individual identification with a particular set of research problems (8). Perrucci and Rothman dealt with technologists rather than scientists, and their findings showed there to be considerable differences in the knowledge and awareness of the individual technologists (9). How do these differences arise and what effect do they have on an individual's attitude to the use of information services?

Various studies have been aimed at establishing a detailed understanding of communications in specialised research communities. Allen studied the communication flow in various research and development laboratories, working with small organisations where every member of the system was identifiable (10). He showed that there were key communication people in a given working unit. These individuals tended to use external information sources much more frequently than their colleagues, tended to communicate more freely with others in the same office or department, and tended to form a closer communication network amongst themselves. He named these people "communication stars", and felt that the transfer of much of the technological information in a department depended on these "stars".

In this study an equivalent research group was being studied - the Research Department at ICI Pharmaceuticals Division, a large research team of biologists, chemists and biochemists. The principal user group being considered was the Chemistry Department. The number of research personnel, around 80, was fairly constant, although there was a major expansion over the five years prior to the study. The number of new

people was therefore small. However, additional problems do exist in the integration of new scientists. The impact of new personnel was examined by Gertsgerger(11). He looked at the communication patterns of 44 new engineers in a total working population of 117. They were found to participate less in discussion with other members of the group, but talked amongst themselves and others outside the organisation. Perhaps they would prefer to use an information service than to ask fellow engineers?

The Chemistry Department at ICI Pharmaceuticals Division was divided into nine sections, each with a specific research topic, and a formally-defined section leader. Smith examined the formally-defined team leader and his relationships with other members of the research group (12). He analysed 418 scientists and engineers engaged in petroleum research in a large laboratory, and found the more informal problem-solving approaches there were between leaders and members the better the research teams performance in terms of papers published and patents obtained.

All these studies (8-12) point to the fact that an analysis of the information needs and use of a research team must bear in mind the environment of each individual scientist. Factors such as age, length of time within a team, role within the team, research topic and relationships with team leader can all affect a scientist's performance and perhaps his information gathering habits.

1.2 Information Use by Scientists

Knowledge about a scientist's use of the various information channels and the nature of the information being sought was limited.

A scientist may approach various information channels depending on the type of information he seeks: for example, current, specific or exhaustive sources (13). Menzel examined the behaviour of some 161 polymer chemists working in 36 institutions of various types and showed that scientists put emphasis on different channels for different uses (14), i.e. deliberate searches, current awareness, general information, and accidental acquisition. From his results Menzel postulated that a perfect or closed system may actually reduce the flow of scientific information. Conversely, it could be argued that a scientist would not have to rely so heavily on accidental acquisition if his main channels were more productive. When designing a system, one should be aware of the different uses to be made of the system and make some allowances for these. A further problem is to determine how the user chooses an information channel for a particular task. Rosenberg (15) interviewed 96 research and non-research personnel and asked them to rank eight information gathering methods according to ease of use and amount of information expected, and to indicate which method they would use to solve three hypothetical questions. The study revealed no relationship between performance ranking and the amount of information anticipated, e.g. the use of an on-site library headed the preference list, but came seventh in the list of methods arranged in anticipated value. A further study (16) showed that the quality of an information channel had no bearing on frequency of use. The over-riding factor in both these studies pointed to accessibility and ease of use. Back (17) showed that scientists working in a new field had broader information needs and different information uses than those working in well-documented fields. His surveys also showed that there was no difference in the information seeking habits of the Ph.D. and non-Ph.D. scientists in his sample. This last observation was borne out by another large study of chemists and physicists (18).

These studies on information needs and uses (13-18) reveal very clear guidelines for any systems designer:

(a) The objectives of the system must be carefully established so that the user knows what type of questions can be answered and which alternative channels he needs to consider.

(b) Availability and ease of use will govern the overall response to a new system, however elegant.

(c) A study of the user population is essential to establish patterns of information use and how these can be effectively met.

1.3 Information Services Available to Scientists

In designing a system based on the use of one information channel, it is important to consider the range of information services that are available to an individual scientist. This helps to build a clear picture of the objectives of the information channel and its inter-relationship with other channels. There are three distinct but interdependent information distribution systems basic to the scientific community today:

(a) The traditional primary publications, many of which are still proceedings of meetings.

(b) Libraries which collect, organise, store and make available the accumulations of primary publications.

(c) Secondary services, each of which provides, for a specified range of primary publications, subject-oriented access to newly available information and/or subject-oriented access to the information store (19).

These secondary sources act as reference retrieval devices, enabling a user to find specific relevant articles in the primary journals. The development and success of Chemical Abstracts journals stresses the importance of these publications. Chemical Abstracts compacts into a highly useable form the findings reported from a wide variety of sources, and most chemists rely on it for many of their requirements, stressing the importance of its completeness and accuracy.

However, the two main information uses for Chemical Abstracts are mutually exclusive:

(a) The current-awareness function where the abstract requires to be available quickly and have maximum readability.

(b) The retrospective searching function in response to specific queries, where the abstract should contain maximum information.

There is little doubt that Chemical Abstracts is the best overall retrospective searching source available to the chemist. Its abstracts are particularly useful when the reader knows exactly what he is looking for; but as a current-awareness tool it is far from ideal. Over the past twenty years much research and development in the information field has been devoted to the problem of current-awareness and the selective dissemination of information. Publications such as Chemical Titles were introduced by Chemical Abstracts Service to fill the time gap between the primary publication and the full abstract (20). However, perhaps the most significant step forward was the application of computer technology, both in the preparation of current awareness publications such as Chemical Titles and later to the search and retrieval of selected aspects of such information.

Services based on "machine-readable data bases" are now widespread. The 1972 survey carried out by Williams and Stewart in the United States (21) showed 56 major processing centres of which 33 were offering services to the general scientific population. In the UK the largest centre is the United Kingdom Chemical Information Service (UKCIS) based in Nottingham. It primarily runs a number of services based on Chemical Abstracts Service products (22); mainly for academic

scientists and those engaged in research in smaller companies.

Many of the larger industrial concerns tend to buy copies of the data bases and process the company's requirements themselves. The advantages are confidentiality, flexibility, the lower costs for large user populations and the option to use the same computer programs to search both in-house and external information. For example, ICI's ASSASSIN system (23) where inputs from many sources, internal and external, are merged into an integrated information system which provides selective dissemination of information, current-awareness bulletins and retrospective searches. Nearly all available computerised information services are based on a textual descriptor, however short, of the primary publication. They provide a fast and effective route by which a scientist can gain access to a document, but he still has to manually extract any relevant data contained in that document.

1.4 Availability and Importance of Data

The location of a specific data item using literature-based information systems can prove a long and tedious business. However, it is a common problem - a recent ASLIB publication on data (24) showed that 70% of the chemists questioned needed property data. In addition, a survey by the Journal of Industrial Research (25) showed that 55% of their respondents looked up property data at least once a week. The ASLIB survey pointed to two main sources of data retrieval - handbooks, particularly The Handbook of Chemistry and Physics, together with Chemical Abstracts. The latter is surprising as it does not usually contain the numerical values of the properties measured. It does, however, demonstrate the reliance most research workers prefer to place on old, well-tried sources despite today's trend towards more and more specialised data activities (26-28). This is borne out by a recent study on thermodynamic data (29) which showed an apparent indifference by research workers to the various data projects available in the field. Edmonds attributed this to the proliferation of these services, and a lack of knowledge of their activities.

One of the more successful projects in this country has been the Cambridge Crystallographic Data Centre (30,31). This concerns itself with the retrieval, evaluation, synthesis and dissemination of structural data obtained by diffraction methods. The centre provides both the traditional handbook and the more flexible computerised retrieval systems, thus managing to cater for all types of user.

In this project, a series of pharmaceutical research teams were being examined. Their primary target was the chemical synthesis of potential drugs - the matching of synthetic effort against known biological targets. In this highly competitive area of research, the concept of confidentiality was of paramount importance. In addition, because of high research costs, the need to exploit information already known to the company was essential. There had been a gradual development of specialised information groups looking at the data generated from the various research activities (32). Data Services Section is one such group looking after data generated within the chemistry, biology, drug evaluation and field trial departments.

This study centred around the Section's role in the exploitation of chemical information generated within the company. It was a data retrieval service, and as such, had to concern itself with the structures of the molecules themselves.

1.5 Developments in Computerised Chemical Structure Handling

Computer technology is widely used in information storage and retrieval since it can handle large volumes of textual data quickly and effectively, but can it cope with the special problems associated with the handling of chemical structural information? Indeed, what are the special problems? The 2-dimensional graphical

diagram is considered to be the most explicit representation of a compound - it provides instant visual communication of the main structural features and their relationships. It, however, cannot be communicated verbally. Chemists, therefore, developed nomenclature, using "family names" wherever possible so that they could interconvert between the structural diagram and the name intuitively. The rapid growth of chemistry in the last half of the nineteenth century meant that the number of known compounds eventually exceeded the memory of the average chemist. This, and the need for international co-operation led to the development of systematic nomenclature.

At the Geneva convention in 1892, international rules were proposed. A very important character of these rules was that they were intended to give one official name for one organic compound - changing the role of nomenclature from solely verbal communication to include registration. This enormously increased the complexity of the names and indeed many of the Geneva names failed to gain acceptance amongst chemists for this reason.

As the number of known compounds increased so chemists were forced to turn more and more to secondary journals - Chemical Abstracts gaining the most importance. Systematic nomenclature and molecular formulae became the two chemical entry points into the literature. Chemical Abstracts Service have developed their nomenclature along the same paths as the original Geneva rules - one name, one molecule - the problem of registration being the most critical. Over the years there have been a number of amendments to Chemical Abstracts index names. Firstly, "trivial" or "author" names are converted into more "systematic" names which are more descriptive of the total molecular structure. Secondly, there has been a unification and simplification of naming principles for all chemical substances (33). These amendments and the complexity of the systematic nomenclature mean that the average bench chemist will use the molecular formula index (however imprecise) in preference to a name index.

However, to the chemist, the 2-dimensional structure diagram does not only identify the chemical compound, it also conveys relationships concerning functions, reactivity and other properties. Today a chemist in pharmaceutical research is called upon to seek more and more relationships between classes of compounds. Traditionally, he visually examines the structural diagrams and he may then group those with common attributes under a generic structure. This is becoming more and more difficult as the number of known compounds rapidly increases. Some chemists have recognised that the computer offers a further dimension in the identification, correlation and search of chemical structures and have exploited this in such applications as drug design and synthetic pathway generation (34).

The introduction of computer techniques led to the development of new languages for structure manipulation, particularly in the 1960's (35,36). These languages were designed to produce unambiguous representations which fully described the structure of the molecule for later manipulation. These fall into two classes:

(a) Linear notations using compact and concise strings of symbols. For example, the Wiswesser Line Notation (WLN) (37), the IUPAC notation and the Hayward Notation (34).

(b) Connection tables which explicitly record each atom, and often each bond. For example, the Chemical Abstracts connection table (34).

Since the introduction of these languages, there has been a great deal of activity in the development of computer techniques for searching and manipulating them. Many of the techniques used for the manipulation of Wiswesser Line Notation were pioneered by ICI. Hyde and Matthews developed computer programs (the CROSSBOW system) to generate a connectivity matrix using as input an unmodified WLN (38). This showed that the notation effectively described the chemistry and the topology of the molecule. Having derived the connectivity matrix they went on to generate chemical fragments and two-dimensional structure diagrams (39,40). These studies pointed to the possibilities of a chemical information system based on WLN which also offered all the manipulative capabilities of a connection table system. In addition, it meant that the output from the system could be the two-dimensional structure diagram, avoiding the necessity for the user to learn WLN.

More recent studies have been aimed at converting structures automatically to WLN (41). The computer programs at the National Institutes of Health (NIH) produce canonical notations for monocyclic, benzene-containing or acyclic compounds. A comprehensive solution to this problem has yet to be found.

However, sufficient basic techniques for the computer handling of chemical structures have been developed, and it remains for these techniques to be exploited within chemical information systems.

There are only four publically-available services for chemical structure and substructure searching. In 1965, Chemical Abstracts Service initiated the Compound Registry System, which now encompasses over two million structures in the form of connection tables. In 1967, International Documentation in Chemistry (IDC) was founded in West Germany, utilising a combination of fragment code and connection table. The Institute for Scientific Information, in 1968, introduced its Index Chemicus Registry System (ICRS), based on WLN (44). In the patent area, Derwent have for many years produced a series of machine-readable data bases using fragment codes (86).

Most of the development in chemical structure searching by computer has been carried out in the larger industrial companies, e.g. the Dow Chemical Company (42,43). Such facilities are designed to provide two main functions: to register compounds made by the Company such that all data pertaining to that compound is found under one unique key, and to retrieve individual compounds or generic classes of compounds, already known to the company.

The company may dictate that compounds are registered in a particular place, but the subsequent use made of the system by the chemists will depend on its usefulness.

This study is attempting to develop a chemical structure handling service based on user needs and to modify the system according to usage and requirements.

1.6 Identifying User Needs

In 1971 there was a large growth in the literature on the subject of information needs and users. Many of these have been reviewed by Lin and Garvey (45).

From some of these studies they observed that information needs can be stimulated simply by a knowledge of what facilities and/or materials are available. For example, in the study carried out by the School of Library Science at Syracuse, users tried their novel retrieval system for psychological abstracts (SUPARS) simply because it became available.

Lipetz (46) stated the following three objectives for any study of information needs:

(a) Explanation of observed phenomena of information use or expressed need.

(b) Prediction.

(c) Control and improvement.

To fulfill these objectives he maintained that the activities must be carried out which fully describe the existing facilities, define user needs and develop a theory and method of on-going work.

In general, there have been five methods used in the study of information users and their needs:

(a) Questionnaires.
(b) Interviews.
(c) Diary methods, systematic self-observation by the user.
(d) Observation by person studying the user.
(e) Analysis of existing data.

The principal method for collecting data regarding information transfer appears to be questionnaire followed up by an optional interview. Some have preferred the less personal approach and have studied the existing data alone, e.g. a count of citations, analysis of loan records.

Woods (47) reported over 100 user studies in the period 1966-1970. Many of these studies dealt with very local issues which cannot be extrapolated to the wider case. Barnes (48) examined the common ground between various information user studies. He found that whilst the results were not inexplicably contradictory, differences in principle and method often make it difficult to demonstrate close agreement.

A number of past surveys do, however, have some relevant results to this study. Lancaster (49) evaluated the operating efficiency of Medlars. He measured the frequency of inadequate user-system interaction in the submission of search requests. In 29% of the searches studied, the user had stated a request narrower than his actual interest, and in 40% his interests were represented too broadly. This highlighted the need for good human communications if a remote computerised system such as this was to be successful.

Rubinstein and Schulz (50) evaluated the use of a custom-built biological literature search service, over a three-year period. The information needs of biomedical researchers were observed by studying the patterns of the repeated use of the service, BIOSIS. They found that use grew with time, as the biomedical researchers accepted the service.

Dubinskaya (51) studied by questionnaire the information practices of chemists in six research and planning groups in USSR. Like many studies, his questionnaire had a poor response, some 245 out of 540. His conclusions were, therefore, vague since 50% of his sample had failed to complete the questionnaire. This appeared to be a common problem with this form of approach, and is why an alternative study method was sought.

It was felt that any method involving participation by the user was undesirable. Delicate communication links had been established, but it was felt that these would not withstand any pressure. Techniques involving questionnaires, interviews or diary methods were, therefore, dismissed. Studies using observation techniques usually were considering library use. For example, Slater (52) studied users in some 14 industrial libraries by observation. Her study set out to find

the major and minor user groups and their habits, e.g. how frequently they used the libraries. However, studies involving users' behaviour are much more difficult to conduct by observation techniques since a detailed study is involved.

For the detailed type of survey being carried out on a known population, analysis of both existing and new measurements was chosen. Firstly, it was necessary to define the measurements required to survey the users, their needs and the performance of the system.

The expression "satisfying a requester's information need" is often used, but its meaning is rather obscure. O'Connor (53) discusses three possible meanings, but Lancaster probably provides a better definition:

(a) As many answers as possible which answer the question.

(b) As few answers as possible that do not.

These terms, frequently called recall and precision, are the most common measurements for the effectiveness of an information retrieval system. They are widely used by Cleverdon (54) and others, and have formed the basis of many mathematical models (for example, that of Brisner (55)). Recall and precision are, however, difficult to objectively quantify since the opinion of the enquirer is always required. It was therefore decided to use a chemist's performance as a measure of the success of the information system. If the information was conveniently obtained and easily digested, a frequent user of the system may show good performance figures in his research function. A number of measurable parameters were, therefore, required to define performance.

Farradane (56) suggested that a well-designed user-oriented evaluation should provide information on the following:

(a) To what extent is the system meeting user needs.

(b) The reasons for failure to meet those needs.

(c) Cost-effectiveness of searches made by user's against searches by intermediaries.

(d) Can improvements in performance be made by basic changes?

(e) Can cost savings be made without impairing performance?

(f) Effect of any change on the usage.

A number of these concepts were relevant to this study, and have formed the basis of the chosen criteria:

(a) To quantify the usage of the chemical information services provided at the beginning of the study.

(b) To measure the change in usage with any improvements in the system or with user acceptance.

(c) To improve the facilities offered by the system to meet user need.

(d) To improve the computer performance of the system to reflect cost-effectiveness.

(e) To examine the relationship, if any, between user's performance and his use of the system.

(f) To move the system towards on-line facilities and to examine these effects on users - both information scientists and the originating enquirer.

At the start of the study, there was few published papers on on-line applications in the chemical structure retrieval field. The most substantial piece of work was that carried out at the National Institute of Health, Bethesda by Feldmann and Heller (57-59). They

developed a collection of computer programs to do chemical information retrieval on a graphics terminal. The medium of interaction was a structural diagram and could therefore be used directly by the chemist. In interacting with the computer, the chemist could graphically specify two-dimensional structures as queries and view structures as search results. The system was developed in an academic environment and its performance or acceptability for use by chemists has never been established. In a study using a functioning research team, a more gradual evolution of an on-line system was required. Firstly, by developing an on-line retrieval system for use by information scientists and so improving the performance of the services to the users. Once this had been developed, tested and evaluated, the move to more sophisticated structure retrieval for the chemist himself could be planned.

Since the data bases available for searching were based on Wiswesser Line Notation (WLN), the intermediate system could use WLN as the interface language. In the sense the input and output facilities were similar to text information retrieval techniques and experience could be gained from such systems.

The most promising system in the UK available at the beginning of the study was the Culham RIOT II system (60,61). The Culham Library had been carrying out experimental on-line retrieval studies since 1968. There were two main purposes: to develop a cost-effective and robust system, and to use this to access the potential benefits that interactive retrieval could bring to the user community. A visit was made to the Culham Laboratory in August 1972. The system available at that time was based on video screens and an ICL System 4-70 computer, operating on a data base of 25,000 to 30,000 references. The system had been designed to allow the user to simply and easily make use of the system. At all stages in the search system he could ask for information by typing in the directive "HELP". The user was also helped in query formulation by filling in a matrix which showed him how his Boolean logic units were linked. A tab key ensured the user jumped from box to box. When the user asked for a concept, the frequency of occurrence of that concept and the five most recent references were shown. When the total search was instigated, references were shown in blocks of ten, as this was found to prevent the user sitting in front of the screen for any length of time. The design of this on-line retrieval system has taken into account the user needs and requirements, and was aimed at the uninitiated user.

Fishenden (62) expressed concern in 1965 that information services were not developing correctly. He felt that scientists were still human beings and their personal likes and dislikes had to be taken into account. It is no good creating an information system that is 100% efficient, but unacceptable to users. These concepts become more important when on-line systems are developed. The information scientist no longer intervenes, and therefore cannot bend the system to meet unthoughtof user needs.

In designing on-line systems, considerable thought must be given to the monitoring and evaluation stages, e.g. language design, responsiveness to user, error diagnosis, use of system resources. Mittman and Dominck (63) suggested the following tools for the effective software monitoring of on-line systems:

(a) On-line log generator for recording both length and time of occurrence of each significant event.

(b) Information management system permitting storage and retrieval of all logged variables.

These comments were based on their experiences with RIQS - Remote Information Query System. At ICI such variables would be measured by the existing on-line capabilities and no new techniques had

to be developed. A number of studies have been carried out on the costs and effectiveness of information systems, but these tend to deal with specific systems and were not generally applicable. For example, the case study by Standera (64).

There have been very few studies on the user reaction to an on-line system, probably the most detailed is the evaluation of the NASA/RECON system by Meister and Sullivan (65). They used two basic methods to determine acceptability and usability: frequency of system usage, and personal opinion of user population. The major source of user satisfaction with the RECON system was the speed with which a search could be initiated and completed. This was one area of interest - as questions are answered more quickly, will usage increase?

To date, the most significant user evaluations have been carried out by Garvey (66-72). His objectives were to examine the variations in user behaviour and to analyse the pattern of information need and use so that the information scientists could improve information services to the scientific population. Garvey concentrated on research scientists, and as such the population was similar to the group at ICI Pharmaceuticals Division.

Garvey first examined the way in which a scientist conducts a research project, and found that the typical scientist appears to progress from stage to stage in a rational manner. An examination of the scientific activities involved showed that there were different information needs associated with the different stages of scientific work. Hence, if information services were informed of the stage of research for which information is being sought, Garvey felt that they might be able to provide information more efficiently.

Garvey also examined the individual differences among the 2030 scientists in his sample. He found that not only does an individual scientist's behaviour, information needs and media-use change as he progresses from stage to stage of his scientific work, but such needs and behaviour vary from scientist to scientist. He isolated the following factors:

(a) Physical scientists make more use of meeting presentations, technical reports and journals than do social scientists.

(b) Basic scientists have greater information needs to aid in the perception or definition of problems and to enable full interpretation of collected data, and to relate work to on-going work in the area, than do the equivalent applied scientists.

(c) Experienced scientists made more use of colleagues and journals than their newly graduated fellow workers.

(d) Scientists who change research topics had more frequent need of information methods.

He found them to be inter-related within this complex picture and identified the need for dynamic information systems to cater for the ever-changing and individual needs of research scientists.

Extrapolating from these studies by Garvey, this study expected to find some individual characteristics in the user population, and had therefore to define a number of groupings which might adequately describe the population. Two of Garvey's factors were not applicable - this study dealing only with pure research scientists. However, they may be experienced or inexperienced (as defined by the length of service in the Company) and they may have spent varying times on research projects.

The environment at the start of the study will now be discussed in more detail, in particular:

(a) The user population and the working environment at ICI Pharmaceuticals Division.

(b) The systems facilities available at the start of the study.

1.7 The User Environment

1.7.1 The user group

The Research Department at ICI Pharmaceuticals Division is a large research team of biologists, chemists and biochemists. The principal user group being considered here was the Chemistry Department, supporting some 80 chemists. The number of personnel was fairly constant; of the 80 chemists, 54 have been in the company more than five years, 20 from one to five years, and 8 less than one year.

The chemists were divided into nine sections (1 to 8 and 10). The distribution of the chemists in these sections is given in Figure 2.

Figure 2 - Distribution of Chemists in the Various Sections

Section	No. of Chemists
1	8
2	8
3	10
4	8
5	10
6	10
7	7
8	11
10	8
TOTAL	80

There were, therefore nine section leaders. Chemists of merit were awarded the position of "Senior Scientist" and there were six of these in the department. Of the remaining chemists, 49 were Ph.D. qualified and 17 were internal promotions from graduate status. Chemists moved between the various sections. In this user group, 56 had been in the same section more than two years, 16 from one to two years, and 8 less than a year. Each section dealt with one or two areas of biological activity, e.g. Central Nervous System drugs or Virology, and was responsible for supplying compounds for a series of biological tests. Each chemist, working in a multi-disciplinary project team, designed drugs for a specialised area.

The chemical sections were generally reorganised each year - from 5 to 15 are involved in each move. Graduate assistants moved freely amongst chemists in the section, and each chemist had, on average, two graduate assistants.

1.7.2 Need for chemical information

In their drug research the chemists may have required structural information when examining:

(a) The availability of samples of compounds for use in biological testing or for use as intermediates in a reaction. They may be interested in compounds available within the Company (access within a week) or compounds which are commercially available (access within a month).

(b) Novelty checking either at the start of a project or when a lead compound was found. The chemist may be interested in one compound or a class of compounds, and may be interested in compounds available within the Division, the Company, the general literature or the patent literature.

(c) The ability to correlate chemical structures with biological activity.

(d) Methods for making compounds.

1.7.3 Divisional information services

Information services within the Division were carried out by two units:

(a) Industrial Property Department covering patents, trademarks, reports and the general literature.

(b) Data Services Section covering internally-generated research data, its collection, storage and exploitation.

Data Services Section was responsible for chemical structural search facilities for the Division and aided the other Divisions in this area.

1.8 Systems Available at the Start of the Study

Several systems were available at the start of the study:

1.8.1 Divisional registration and search

Prior to September 1972, Pharmaceuticals Division registered compounds independently of the rest of the Company. Manual card indexes were maintained showing the structural diagram, systematic name, molecular formula, reference numbers, sample sources and test information. Two indexes were maintained:

(a) In reference number order.

(b) In molecular formula order.

Having established a compound's novelty in the molecular formula index, a chemist would submit his compound to an information scientist. The information scientist would be responsible for ensuring that the compound with the appropriate reference number would go through for biological testing.

In the registration process, the information scientist would check the molecular formula for accuracy and again look up the manual molecular formula index. If he could find no previous record of the compound, he would assign the next sequential reference number, code the compound in WLN and the fragment code, and pass the necessary information forward for biological screening.

The WLN/molecular formula record so produced was entered into a computer system. An automatic check on the WLN/MF was made, and if there was any discrepancy, the compound was rejected for correction. The computer file of compounds was held on magnetic tape in reference number order and hence could not be used for novelty checking (to do this, any incoming WLN would have to be checked against all others on the file, one at a time - a very time-consuming process). The fragment code was recorded in a punch card. This was overprinted with the structure at the same time as the structure cards were prepared for the manual indexes. The fragment code was also read to a second computer file. There were several disadvantages to this system:

(a) There was the possibility of duplication, i.e. the same structure being given two separate reference numbers. The manual molecular formula index was the only checkpoint - cards could be missing or misfiled, or it was possible to misinterpret structural diagrams when scanning large numbers of entries. Also, the information scientist tended to rely on the chemist's novelty check when assigning the reference numbers.

(b) The registration process was time-consuming, the compounds were coded and drawn several times. However, registration had to be carried out as quickly as possible so that the necessary documentation could be put into the indexes.

(c) The fragment code, the main search tool, was subject to error. This could be caused by misinterpretation of the rules, by miscoding or by mispunching, and was not checked automatically.

The system, however, was successful, its success largely due to the excellent memory of the information scientist in charge of registration.

The substructure search service largely depended on the manually assigned fragment code. The punch cards were sorted using a punch card sorting machine. KWIC indexes were produced from the WLN/MF file and these were used to answer specific questions.

It was found that the substructure search system was limited since:

(a) Sorting 80,000 punch cards was time-consuming.

(b) Any search hits were returned to the chemist in the form of the punch cards (the structures were overprinted). The chemist was only allowed to keep the cards for a limited time - otherwise, other searches could be held up or incomplete.

(c) A number of searches were too specific to be effectively carried out with a fragment code and, hence, labour was required in editing.

1.8.2 Company compound registration and search

The Company compound centre collected together compound information from all the divisions. This was maintained on a large magnetic tape file (150,000 records) on the Company's IBM 370 machine, and was based on the research carried out by Hyde and Thomson (38-40). Several files were maintained:

(a) WLN file containing all Divisional numbers. Other files were only referenced by the CR number (the Company Registry number) and had to be related to this file.

(b) Connection table file. These were generated from the WLN on registration and stored on magnetic tape.

(c) Fragment file. An open-ended fragment code (> 5,000 separate fragments) was automatically generated from the connection table on registration. The file was inverted for searching.

The registration process was expensive and the search system was difficult to interpret - about 4,000 of the fragments had a posting of less than 10 and generic fragments were often not available. The string search program was based on normal text search and was often inadequate. 95% of the searches were carried out manually using a KWIC index and the structures then generated from the stored connection tables. The company service was used by Pharmaceuticals Division chemists. They were becoming dissatisfied with computer services as a number of constraints were imposed on them:

(a) The service was expensive - limited output on standard computer stationary.

(b) There was a long turnaround time of up to 4 days. Manual searching of the KWIC Index could take up to a day.

(c) The information was often out-of-date - the search files were only updated every six months because of cost factors.

By the end of 1970, the number of chemists at Alderley Park had risen to about 70. The demands on the Divisional manual and punch card services were large and chemists had little support for application of computers in this area.

1.8.3 Research carried out by the company

Following the research carried out in the late 1960's (38-40), the company had supported a research program into the application of computerised structure handling techniques to chemical information.

Experience with the Company Compound Centre led to the development of the following:

(a) Mark II versions of the CROSSBOW connection table and display programs. These were designed to overcome some of the problems of the Mark I version; for example, to facilitate conversion to the Chemical Abstracts Connection Table (73).

(b) An atom-by-atom search program. This was written to allow detailed interpretation of the Mark II connection table to take place. The program allowed up to 8 nodes (or atom symbols) to be specified and a simple pathway (linear, one-branch, two-branch) between them. It then generated a mini-connection table which was matched against total connection tables of the compounds to be searched.

(c) A program to convert standard WLN notation to the expanded notation used by CROSSBOW programs. Hyde and Thomson (38) had recognised the extra complications which arose out of use of the multiplier and contraction rules in WLN (37), and had developed the CROSSBOW system to use notations coded without them. This did in fact mean that ICI was at variance with the rest of the WLN-using population, and the program was written such that externally-generated data bases could be used with the CROSSBOW system.

These programs, together with existing facilities were used as the building blocks of a new system to meet user needs.

1.9 Outline of the Study

The objective of the study was to design and develop a computerised chemical structure handling service to meet the chemists needs and requirements for an internal information handling facility. This development had to take into account:

(a) Existing data bases and their use.

(b) The divisions decision to change computers and to develop the Company Chemical Data Bank on the Burroughs 3500 computer, with its particular capability of handling on-line systems.

(c) The need to develop services to meet the continuing needs and requirements of Research Department.

The ability of the system to meet the user requirements was seen as of paramount importance. Two sets of users can be identified:

(a) Information scientists using the system as an interface for members of Research Department.

(b) Chemists in the Research Department seeking information to aid them in their research.

Both sets of users were studied. The information scientists were monitored constantly and any changes to meet their requirements acted on as soon as possible. The chemists, however, represented a much larger group and could not be subjected to the same interrogation techniques. Their response to the system has been studied indirectly - observations being made over the two years and the results analysed at the end of the study. The technical development of the system and the user studies were carried out concurrently, but will be reported separately.

Chapter 2 - Chemical Data Base Design

2.1 Introduction

The introduction of the Burroughs 3500 as the Divisional computer, with its particular suitability to on-line direct-access working, introduced a new philosophy into the development of chemical information systems within Pharmaceuticals Division. Consequently, an investigation was set up to look at new systems for compound registration and search, based on the new computer facilities, and aimed at overcoming the deficiencies and limitations of existing systems. At the same time, the Division decided to amalgamate the chemical registration and search facilities of the Head Office Company Compound Centre with those of Pharmaceuticals Division, Data Services Section.

As initially set up, the Company Compound Centre only registered compounds of complete structure, whereas Pharmaceuticals Division had to register all substances submitted for biological test. The overlap, however, between the two centres was large, and a lot was to be gained if it was possible to define a common system for registration. The first stage was therefore to design a Company data base suitable for all the registry and search functions to be carried out, taking into account data bases and services already in existence, surveying the contents of the files, etc.

2.2 Available Company Data Bases

2.2.1 Pharmaceuticals Division compounds

There were five files of chemical data on Pharmaceuticals (M) compounds held within the Division (see Fig. 3).

Figure 3 - Chemical Files Held in Pharmaceuticals Division

	MAIN CONTENT	PURPOSE	MEDIUM	ORDERED BY	USED FOR	
					REG.	SEARCH
1	ICI and M nos., MF, structural program test submissions.	A complete record of all test sub-missions.	Manual cards.	M no.	✓	
2	ICI and M nos., MF, structural diagram.	To allow chemists to perform novelty checks.	Manual cards.	MF	✓	
3	ICI and M nos., MF, WLN.	Computerised record of compounds.	Magnetic tape, ICL 1902A.	M no.		✓
4	M no., fragment code.	Searched using statistical sorter for substructure classes.	Punch cards.	Essent-ially unorder-ed.		✓
5	M no., fragment code.	Searched by computer for substructure classes.	Magnetic tape, ICL 1902A.	M no.		✓

MF = Molecular formula.
WLN = Wiswesser Line-Formula Notation.
M No. = Pharmaceuticals Division no.
ICI No. = ICI Central Registry no., also referred to as CR no.

The composition of these files and the indexing procedures used reflected the Divisional need to relate chemical and biological information.

Within the Division, registration was carried out manually by checking a potentially new M compound for novelty in the manual MF index, assigning an M number, working out the WLN and fragment code, creating records for the appropriate files and updating these files.

The compound was then passed onto the Company Compound Centre where a separate registration was performed using manual and automatic novelty checks by WLN. In the updating of Divisional and Company WLN files, an MF was automatically generated from the WLN and compared with the input MF previously calculated manually: this acted as a partial check of both WLN and MF. The Divisional and Company WLN files were instituted to provide a unique and unambiguous representation of structures in a form suitable for automatic processing whether for registration or substructure search purposes.

Within the Division, it was not possible to automate directly the manual novelty check of the MF index used in registration. This was because the MF was ambiguous, and the final stage of the check depended on a visual comparison of structural diagrams. A WLN file was thus essential for an automatic registration system. The following criticisms could be levelled against the Divisional system:

(a) Duplication of the registration function by the Division and the Company CROSSBOW Centre was wasteful - 70% of the Company file was made up of Pharmaceuticals Division compounds.

(b) The maintenance of five separate Divisional files for chemical information was complicated and unnecessary. In particular, the updating of the manual MF index was very laborious.

(c) The manual MF novelty check in Divisional registration was prone to human error, whether due to the wrong arrangement of entries from a previous update, or due to looking for the new compound in the wrong place.

(d) The use of two independent structure codes (WLN and fragment code) both requiring manual encoding and their own file maintenance, was unnecessary. The WLN contains implicitly all the information recorded in the fragment code and an equivalent to the latter could be automatically generated from the former.

(e) The operation of the Divisional registration procedure on a daily batch basis caused a delay of some hours between the submission of a compound by the chemist, and the confirmation by the indexer that the compound was novel and the provisional M number valid. If the compound was found not to be novel, the chemist had to be contacted again to decide on further action.

(f) The present location of the manual indexes and the registration entry point was not central to the chemical laboratories.

(g) Compounds were not accepted for Company registration unless they could be encoded into WLN by a certain well-established set of rules.

Thus, compounds of only partially known structure, complexes, polymers, polypeptides, etc., were not assigned ICI numbers, since WLN rules for such compounds were only tentative or non-existant. Within the Division, such compounds were assigned an M number, and were entered in some way on the WLN file.

2.2.2 Company Compound Centre

A record of all chemical compound information generated within the Company was held by the Company Compound Centre on the IBM 360/65. Registration of the Company compounds took place on request from the Divisions, and four large files were maintained (see Figure 4).

Figure 4 - Chemical Files Maintained by Company Compound Centre

	MAIN CONTENT	PURPOSE	MEDIUM	ORDERED BY	USED FOR: REG.	USED FOR: SRCH.
1	ICI no., WLN, MF and Divisional nos.	To provide a complete record of all compounds requested by the Company.	Magnetic tape, IBM 360/65.	(1) ICI no. (2) WLN	✓	✓
2	ICI no. Mark I connection table.	Used to generate structure displays.	Magnetic tape, IBM 360/65.	ICI no.		✓
3	ICI no., fragment file (based on Mark I connection table).	Used for substructure searching.	Disk, IBM 360/65.	Inverted file.		✓

MF = Molecular formula.
WLN = Wiswesser Line-Formula Notation.
ICI No. = Company registry number.
Mark I connection table = See full description by Hyde & Thomson (38).

This Company file contained information from the following collections (the letter in brackets is the Divisional code):

(a) Pharmaceuticals Division (M) - about 72,000 compounds tested within the Division. The Company file stored information on the compounds having definite structures, but it was not directly compatible with the Divisional file. Different indexing conventions were used to record salts, stereochemistry, etc. reflecting the need to relate information from various Divisions.

(b) Plant Protection Ltd. (R) - about 50,000 tested or purchased by Jealott's Hill Research Station.

(c) Organics Division (SC) - about 48,000 compounds in their specimen collection.

(d) Organics Division (A) - about 60,000 compounds, mostly indexed from Company reports, works processes and about 26,000 indexed from commercial suppliers catalogues.

(e) Head Office (Z) - compounds indexed by Head Office from Company Reports - about 3,000 compounds.

(f) Research Station, Brixham (B) - about 600 compounds tested by this research unit.

The total number of unique compounds on the Company file was 165,000, each of these records being assigned a Company reference number (CR number). A record may have any number of reference numbers from any of the Divisions assigned to it, and no attempt was made to prevent duplication within Divisions. Also, the rules governing the allocation of CR numbers were not stringently applied and correlation between compounds essentially similar was often difficult.

However, this was the only computerised record of compounds outside the M collection.

2.3 Requirements of a Company Chemical Data Bank

A Company Chemical Data Bank was to be set up on the Burroughs 3500 to carry out the following functions:

(a) Fast registration of new compounds into the Pharmaceuticals Division collection for biological testing.

(b) Registration of all new compounds into the Company compound file.

(c) Inter-relation of compounds indexed in the various Divisions.

(d) Substructure searching on the Company collection or on Divisional collections.

(e) Specific compound location within the Company data base.

(f) Inter-relationships with related data bases, e.g. biological test results, sample data.

(g) Ability to make comparisons with externally available data bases, e.g. commercial catalogues, literature information.

2.4 On-Line Considerations

The B3500 was particularly suitable for on-line applications, and the commercial department had already made available an on-line ordering and invoicing system. The system provides all the basic software for other applications to be developed on-line. A total on-line chemical data base would have had the following advantages:

(a) Up-to-dateness: compounds would be available in the system on allocation of the reference number.

(b) Ease of use: random access techniques allowing fast retrieval, even in a large data base.

(c) Automatic novelty checking using a Visual Display Unit (VDU) terminal

(d) Immediate correction of input errors.

(e) A compound's performance in the search system could be evaluated, and appropriate action taken on input. For example, if a compound failed at the connection table generation stage, the manually generated connection table could be generated and added to the file.

However, there were a number of constraints:

(a) Usage - economic and efficient use of an on-line terminal is closely related to the number of transactions/day and the complexity of these transactions.

(b) Size - limits on disc availability dictated that the data base needed to be as small as possible both in terms of the number of records present and the size of each record.

(c) Response time - for maximum advantage, the response time had to be as short as possible. The length of time to find any one record using a given key would depend on the size of the file and the extent and type of file organisation.

(d) Variable length records and variable length keys were not possible.

Bearing these constraints in mind, the Divisional requirements were judged on the following criteria:

(a) Size of the collection using the following guidelines:

$<$10,000 = Small
10,000-40,000 = Medium
$>$40,000 = Large

(b) Type of growth to the collection, i.e. whether daily additions being made or occasional batch additions. The rate of daily growth was also important.

(c) Overlap between collections, i.e. movement of compounds from one Division to a second.

(d) Use of the collection, i.e. were the compounds being currently submitted for biological evaluation?

In the light of these conditions, the data was divided into the following:

(a) Constantly growing, large collections with some overlap and current biological evaluation programmes:

An on-line registration system, the main Company Data Base, based on a random-access disc data base was designed to cope with data from Pharmaceuticals Division (M collection), Organics Division (SC collection), and Plant Protection Division (R collection).

(b) Other collections where growth was intermittent and the data base was small or not attached to a current testing programme.

A batch registration system based on a serial magnetic tape data bank, the additional Company Data Base was designed to cope with data from Brixham (B collection), Head Office (Z collection), Australia and New Zealand (I collection), and the Fine Chemicals Service source collection (A collection).

Figure 5 illustrates the selection of data for the two registration systems:

Figure 5 - Proposed File Usage

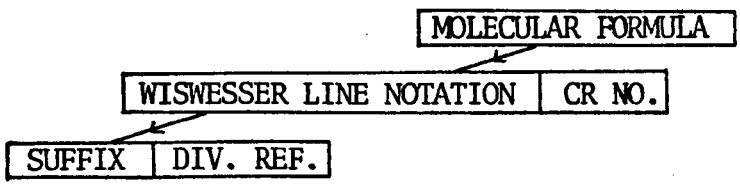
	COLLECTION	SIZE	GROWTH/DAY
ON-LINE DISC ↑	M	85,000	25
	R	50,000	24
	SC	49,000	6
BATCH TYPE ↓	A	67,000	Variable, from 0 to 200
	Z	1,800	4
	B	1,000	0

2.5 The Main Chemical Data Base

2.5.1 Basic design

The fundamental concept of the random-access data base is a multi-level compound description as shown in Figure 6.

Figure 6 - Relationship of the Data Elements



A Company number (the CR number) is allocated to each chemical species as defined by a unique WLN. There may be more than one WLN for each molecular formula and there may be more than one suffix for any Wiswesser Notation, where the suffix is a modification to the basic species, e.g. salt, stereochemistry:

DIV. REF. = NOTATION + SUFFIX

where the suffixes are denoted according to the indexing conventions of

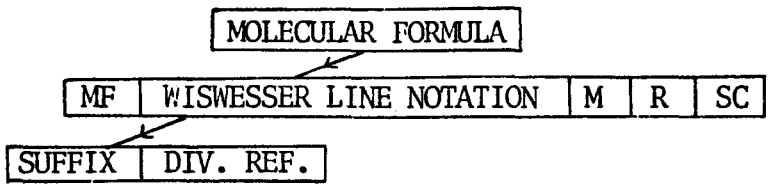
the Division concerned. There is only one entry for each Divisional reference. There may be one or more Divisional references for each CR number, e.g. "Hibitane" where a large number of salts have been prepared for the same basic molecule, and the compounds have been tested extensively in the various Divisions (see Figure 7).

Figure 7 - References for Hibitane

				Molecular Formula C2OH18N9-CL2
CR No.	M No.	SC No.	R No.	WLN
238979	10040	10661	15599	GR DMYUM&MYUM&M6MYUM&YUM&MR DG
				SUFFIX
		16389	17548	PALMITATE
		16490	16834	STEARATE
	11976		57614	CU COMPLEX

95% of all compounds on the file have no variation on the basic species. To make access to the file more efficient, the original multi-level concept was modified so that the main record holding the WLN and the CR number may also hold one M, R and SC number relating to the basic network (see Figure 8). In addition, it is usual to require both WLN and molecular formula together in response to a question. Since there could be a large number of WLN's for any molecular formula, it was decided to hold the MF on the main record, i.e. one MF for each WLN.

Figure 8 - Data Structure on the Chemical Data Base



2.5.2 Accessing the data base

To make maximum use of the Company data base, access would be required through each of the following keys: reference numbers - M, R, SC and CR; molecular formula; and Wiswesser Line Notation.

To cater for immediate access using any of these keys, the data base must be held on random-access disc, and the files must be organised such that efficient access can be made to any record on the data base using any of these six keys.

Each Divisional number, M, R or SC, can be present on the data base once only, i.e. the information for that compound is present in a known area on disc. Similarly, the CR number represents a given Wiswesser Notation, and this will occupy a known area on disc.

To access data under a known reference number, indexes are required to relate each reference number with the position of its relevant data on the disc. The reference numbers cover the following ranges:

M	1 - 92,000
R	1 - 75,000 with some nos. yet to be allocated
SC	1 - 52,000
CR	200,001 - 325,000

All Divisional collections are growing rapidly and hence the provisions were made as shown in Figure 9.

Figure 9 - Division Reference Indexes

M/CR INDEX

M
1 - 200,000
CR
200,001 - 399,999

S/R INDEX

S
1 - 100,000
R
100,001 - 199,999

Notes:

(a) Up to 200,000 M numbers were created for, since nearly 100,000 has been reached and the collection was growing rapidly.

(b) Up to 100,000 R and SC numbers were allowed for, although R numbers could expand beyond this by adding to the end of the index.

(c) To look up an R number the program adds 100,000 to the number input.

Every reference number has an entry on file, whether present or not. The entry takes the form of a 6-digit number indicating the position of the main record on the disc, and a marker indicating where the record is present or not; an "*" indicates the presence of the compound on file.

Look-up for a reference number is very simple. For example, what is the structure of R465? This number is input to the program: the R tells it to examine the S/R index and to add 100,000 giving 100,465. The 100,465th record of the S/R index is then accessed giving the following information:

095432*

i.e. the compound is present and starts at position 095432 on the main file. In this way, any reference number may be accessed very quickly.

The molecular formula is not a complete description of the structure, but summarises the atomic content of the molecule. The distribution of structures amongst molecular formulae is very uneven. Some formula, which may include combinations of the more common elements or unusual elements, represent a single structure yet other formulae represent large numbers of structures. Appendix I shows the molecular formula population in a file of 125,000 compounds (74).

A member of the chemist user population knows no WLN and would use molecular formula as a main key to find a given compound. Hence, if this data base is to be used by the chemist himself, it must contain a molecular formula index.

The length of the molecular formula varies, but it is almost always complete or unique by 18 characters. It was decided, therefore, to store the molecular formula key as a fixed length record of 18 characters, any remaining symbols being ignored.

All compounds in the file must have an entry in the molecular formula index. Each molecular formula index record contains the following information: molecular formula, 18 characters using the Richte ordering convention; position on the disk of the first compound with that molecular formula; and the position on the index of the next molecular formula. On updating, any new molecular formulae are added to the end of the file and the pointers used to maintain the correct sequence. Re-organisation could then be carried out when the file became uneconomic.

The main WLN record will contain a pointer to the next record on the disc with the same molecular formula. All compounds with the same molecular formula can, therefore, be found by tracing the pointers from the first record to the last where the pointer will be set to zero.

It must be possible to look for the presence of any molecular formula on the file quickly. It is not efficient to scan the whole molecular formula index (about 70,000 entries) each time. Hence, index tables have been set up to allow selective entry into the molecular formula index.

The first table, the coarse table, holds the formula of every 7,000th entry on the index. These ten entries are held in alphabetic order and each points to the next lookup table, the fine-table. There are ten fine tables, one for each entry in the coarse table and the molecular formula in the coarse table shows the first entry in the fine table. The fine table holds every 100th molecular formula and its position on the molecular formula index file (70 entries in all).

Using these tables, any molecular formula can be found in a maximum of 270 comparisons. For example, is C6H17ON on file? The coarse table is examined (see Figure 10).

Figure 10 - Part of the Molecular Formula Index Coarse Table

MOLECULAR FORMULA	FINE TABLE
C3H8ON2	2
C5H18O	3
C8H10ON2-CL2	4

Since C5H18O is the first entry in fine table 3 and C8H10ON2-CL2 is the first entry in fine table 4, C6H17ON can be expected to be in fine table 3. This is now examined (see Figure 11).

Figure 11 - Part of a Molecular Formula Index Fine Table 3

MOLECULAR FORMULA	POSITION ON INDEX
C5H18O	010010
⋮	
C6H15O	010650
C6H18O	010875

Here, C6H15O is the last entry in the file to have a value of less than C6H17ON. The program then moves to the molecular formula index, starting 010650. It scans the entries matching against C6H17ON until a match is found or the record is higher than C6H17ON, in which case the record is declared "not on file".

New molecular formulae are being added all the time. These are added physically to the end of the file, and pointers set on the entry logically before on the molecular formula file and the new entry points to the molecular formula immediately afterwards. The coarse and fine tables are not changed.

Using these files, any molecular formula can be found quickly and efficiently. Update is simple since the records are logically connected rather than physically adjacent.

The Wiswesser Line Notation, WLN, is the complete description of the molecule and its length varies from 2 characters to over 150 characters, the average length being 19 characters.

To allow entry using WLN meant that some form of key was required. For a record to be a simple key, it must have fixed length, must be unique, and must be ordered alphanumerically. To make a WLN a

fixed length record of say 144 characters was impractical. A large amount of disc would be required and most would be taken up with spaces. Matching on 144 characters would be inefficient and slow, and again much of the matching would be space against space.

It was therefore decided to break the WLN into a variable number of fixed length blocks, and only the first block would be present on the main file and be used as a key. An analysis was carried out considering the following three points:

(a) The average length of notation was 19 characters - the block size should be as near to this as possible to avoid wasted space on the disc.

(b) The block size should be selected such that as many compounds as possible are unique within one block, i.e. the number of "synonyms" should be kept to a minimum.

(c) The number of blocks should be kept to a minimum to avoid a large number of disc accesses to find the complete record, i.e. as many compounds as possible should be unique within one block.

The results led to the selection of a 24-character block. A summary of the results is given in Figure 12.

Figure 12 - Summary of Results on WLN as Key Using 16, 24 and 32-Character Records

LENGTH OF BLOCK	% COMPLETE	NO. OF SYNONYMS	NO. OF BLOCKS REQUD. TO COMPLETE THE WLN		
			1	2	>2
16 characters	40	7	40%	50%	10%
24 characters	80	1.5	80%	19%	1%
32 characters	90	1	90%	9%	1%

Hence, the main chemical file is divided into two areas (see Figure 13).

Figure 13 - Information on Main Chemical Files

MAIN FILE (WW)

Reference numbers.
CR number.
First 24 ch. of WLN.
MF.
Pointers to other MF records.
Pointers to overflow record.
Pointers to other WW records.
Synonym count to show no. of WLN's with same first 24 ch.
WLN length.

OVERFLOW FILE (WO)

24 ch. of WLN.
Pointer to next overflow record.

OR

Suffix.
or
References to suffix.

The main file, the WW file, is logically ordered in Wiswesser order by the use of pointers to the next and previous records in the chain. New records are physically placed at the end of the file and logically related to their position in the file.

As in the molecular formula index, coarse and fine tables are used to enter the file by any WLN. There are 20 entries in the coarse

table, and therefore 20 fine tables.

Suppose it is necessary to find some information on "Hibitane" and, therefore, to know its reference number at Plant Protection Division, i.e. the R number. The WLN for "Hibitane" is first input - GR DMYUM&MYUM&M6MYUM&YUM&MR DG. The program first counts the number of symbols - 30 characters. It examines the coarse table for GR DMYUM&MYUM&M6MYUM&YUM, i.e. the first 24 characters. It finds the next lowest WLN entry and goes to the appropriate fine table. It now finds the next lowest WLN and an address on the main file. It now follows the "next WLN" pointers on the main record until it finds a match. Here it finds there are two synonyms and examines the first. To avoid unnecessary looking up on the overflow file, it first compares the WLN length on the file with 30 (for Hibitane). They are not equal so it proceeds to the second synonym where it finds an equivalence on length. The program now follows the pointer to the overflow record and finds the next 24 characters of WLN. A match is found and the main record will give the reference numbers, etc. for Hibitane. However, the overflow pointer points to more overflow records.

The WLN length has been satisfied: therefore, the remaining overflow records hold the suffix data. The first record, and maybe the second, will contain textual information on the nature of the suffix. This is followed by a "reference number record" holding any reference numbers pertaining to the previous suffix. Up to 8 reference numbers may be stored in the 24 character area allowed ensuring maximum utilisation of the disc. The program translates these codes back to their Divisional meaning before the user receives an answer.

The storage of the Hibitane information (as shown in Figure 7) on the disk is shown in Figure 14.

Figure 14 - Storage of Hibitane Information on the Chemical Data Base

MAIN FILE

CR No.	M NO.	SC No.	R No.
238979	010040	10661	15599

WLN	Pointer to overflow
GR DMYUM&MYUM&M6MYUM&YUM	000001

Synonym count	WLN length	Next WLN pointer	Previous WLN pointer	Next MF pointer	Previous MF pointer
002	030	010050	010010	000000	056210

Notes:

(a) The M number is held as a 6-digit number and the SC and R numbers as 5-digit numbers.

(b) The remainder of the record is held in overflow area 1.

(c) There are two compounds with this 24-character WLN. It is the last record in the MF block, i.e. next MF pointer = zeros.

OVERFLOW FILE

1.	&MR DG	000002
2.	PALMITATE	000003
3.	716389 | 517548	000004

4.	STEARATE			000005
5.	716400	516834		000000

Notes:

There are two suffixes:

(a) Palmitate with SC16389 and R17548.

(b) Stearate with SC16400 and R16834.

This overflow system had two main advantages. It allowed for the duplicate records already on file, i.e. two reference numbers in the same Division for the same WLN. If this occurred on the main record, a suffix of spaces could be held. If the duplicate was on the suffix up to 8 reference numbers could be held. Additional information could be held on the overflow file, e.g. non-generatable connections tables or 2-D displays.

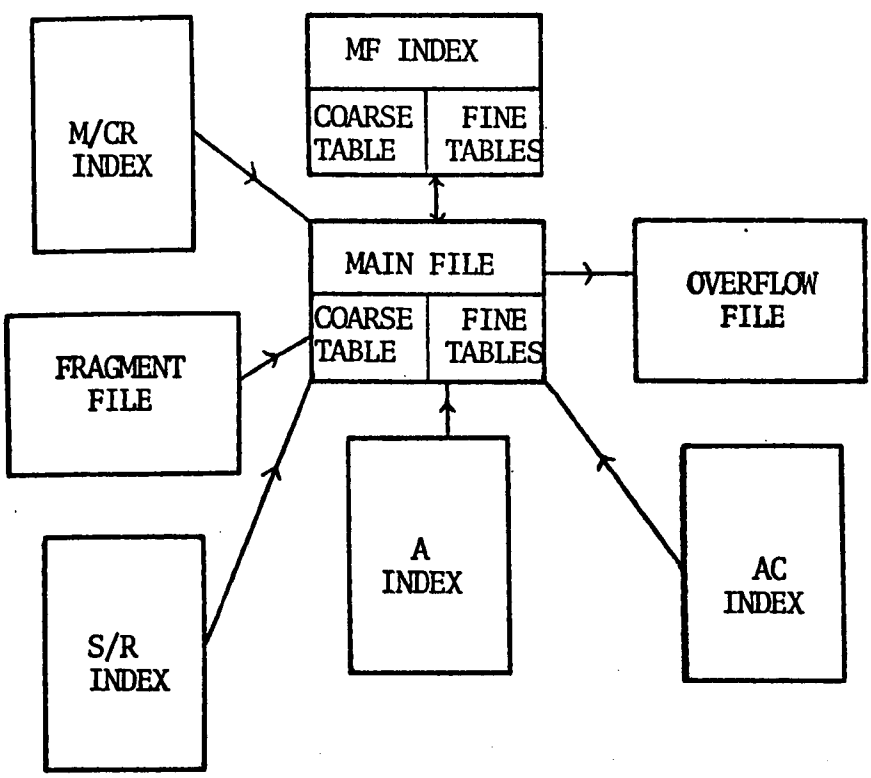
2.5.3 Format of the main chemical data base

Figure 15 shows a schematic representation of the total chemical data base.

The fragment file is used for substructure search. Here, all records in the file must be searched, hence organisation of the fragment file is not important. It is held in CR order with one entry for each CR number - the CR number is not stored on the record, but the actual address of the main record on the Wiswesser file. One fragment (bit 152) is used to indicate whether the record is live (present) or dead (not present/deleted). Any hits from the fragment search are used to access the Wiswesser file so that further searching can be carried out. The efficiency of this two-stage search lies in the fact that few hits are obtained from a fragment search (less than 10% of the total WLN's present). It is faster for this two-stage examination of the data base than for all the data to be searched on one pass of the file, since less read accesses are required.

The A index relates the additional chemical file to the main Company file (see Section 2.6).

Figure 15 - Schematic Representation of the Files in the Chemical Data Base



The AC index relates the commercial availability file to the main Company file (see Section 2.7).

The main chemical data is permanently resident in 30 million bytes of fixed head per track disc on the B3500 (now B4700). It is accessed many times a day for both specific compound enquiry and sub-structure search in both on-line and batch modes (see Chapter 3).

2.6 The Additional Chemical File

Having extracted the most used and most valuable parts of the Company Compound Centre data base, there was left a smaller amount of material requiring less and slower access. The decision had to be made as to whether to devise a second, less sophisticated batch system for the remainder of the data or to rely on purely manual systems. The following collections were to be considered (see Figure 3):

(a) Brixham (B) Collection - Information officers at Brixham expressed little interest in their information being recorded. In addition, the compounds had largely come from other Divisions, and therefore did not represent an additional source of compound information. The numbers involved, less than 1,000, meant that manual systems were still feasible, and hence the B collection was removed from the Company's computerised files.

(b) Head Office Reports (Z) Collection - The Z collection represents compounds found in reports and indexed by Head Office. However, these reports were also found to be part of the Organics Division A collection and were thus doubly indexed. For this reason, the Z collection was also removed from the computer files.

(c) Organics Division (A) Collection - The 67,000 compounds here fall into three main classes:

(1) Commercially available compounds indexed from ten major catalogue suppliers - about 30,000 compounds.

(2) Organics Division compounds reported in Works Processes - about 4,000 compounds.

(3) Compounds mentioned in Company Reports or selected literature references - about 40,000 compounds.

There is some overlap between the various classes, e.g. a compound may be involved in a Works Process and referred to in a Company Report. The first class of compounds were not truly Company compounds, and were, therefore, also transferred to a commercially available file (see Section 2.7).

The A collection compound information was written to a magnetic tape file. Each record containing A number, fragments for substructure search, WLN including suffix, if one present, and molecular formulae. All fields were initially held as fixed length and the tape file, maintained in A number order, was updated and searched in batch mode.

After about one year of searching the Company and additional chemical files, it was obvious that some form of overlap indication was required. Users were interrogating the main Company file and the additional chemical file and follow-up work was being carried out by the Section on the same compounds from both sets of data. To find the amount of overlap, the two files were matched on WLN (not suffix) and the appropriate CR numbers found and written to the magnetic tape. On searching the additional compounds file, it was therefore possible to exclude compounds which were/were not in the main Company file.

Most searches, however, were being carried out primarily on the main Company file, and some indication of the A number, if available, was necessary. This, without searching the tape file, would give the chemist an extra access point for information on the compound, e.g. a recipe from a Works Process. To do this, an additional file was added

to the chemical data base detailing the A/CR equivalent. It was similar in concept to the M/CR and S/CR indexes, i.e. every CR number has an entry on file, whether present or not. The entry takes the form of:

(a) A 6-digit number representing the A equivalent. The first digit is 8 or 9 if the compound is part of the commercially available file.

(b) An "*" if there is more than one A number for the CR number.

The latter is possible since the A number represents a WLN + suffix and the CR number, the WLN. The "*" is only an indication of the availability of other numbers, these are checked by Organics Division when a request for an A number is made. There are 160,000 records allowed for on the index, and the key is the CR number minus 200,000.

This file is accessed at the end of a substructure searched on the Company collection. Any A number found is printed on the structure card along with an "*" if more than one number, and a "C" if commercially available. After a year of searching, the position of the A collection was reviewed. If usage had been higher, the data would have been transferred to the main data base or written to disc pack. However, the total data base was searched only 32 times in 1973 (as opposed to 450 searches on the commercial availability files, and 990 on the Company data base). Its main use was as a source of additional information on compounds in the Company files, and hence, the A index provided this function adequately.

There was some debate as to whether the A collection should be maintained, but the costs of maintenance are low (updates being carried out every six months), the magnetic tape search file adequate and the data potentially of value. It has, therefore, been continued as a batch magnetic tape system.

2.7 Commercial Availability Files

The Company has access to files listing commercially available compounds.

2.7.1 The A collection

Organics Division maintain a punch card index to the compounds in the A collection which are commercially available, and giving their sources. The sources are merely recorded as the catalogue name, and are, therefore, of little value. To find which catalogues a compound with an A number may be purchased from, and its reference number and price, Organics Division must be consulted. Their data is kept in a card index and searched manually.

The punch card file was therefore used to select compounds from the A collection which were commercially available. Some 28,500 compounds were withdrawn and written to an "A commercial file". The selection was carried out on the A search file, and the following information was therefore available for searching: the A number; the CR number, if in Company data base; fragments, WLN, suffix (all held as one field); and molecular formula.

A commercial availability file was set-up, and it was possible to exclude compounds which were in the Company Data Base to avoid double selection. The system was, however, limited as:

(a) There was no price information on the data base.

(b) The compound could not be directly ordered, details had first to be obtained from Organics Division.

(c) It was not always up-to-date, time being necessary to manually access new catalogue information. This was done at Organics Division and was, therefore, outside our control.

2.7.2 The Aldrich chemical files

The Aldrich Chemical Company produces a magnetic tape of all compounds available commercially through them. Most of these 19,000 compounds are available in their catalogues, and the remainder are new additions (available at a fixed price). In addition, Aldrich agreed to alert ICI of all compounds newly produced six months before they become available in the catalogues.

The records on the magnetic tape contain the catalogue reference number and the WLN. The latter was not directly comparable with the Company files. A program had been written to cater for this problem (75). ICI does not use the multiplier and contraction rules as given in the Smith manual (37). Most other companies did begin by using these rules, but there has been a gradual swing to the ICI system.

The program which reformats the notation does not always produce the same notation as would be produced if the compound had been coded without the rules. The difference lies in order of symbols rather than their content, and substructure searches would produce the same answers.

There was not enough information on the tape to produce an effective search system. The Aldrich Chemical Company was approached, and a second file was obtained. This detailed the catalogue reference number, the chemical name and the molecular formula as given in the catalogue. They had no way of giving us a computer tape of prices. The price was, therefore, transferred from the catalogue on to punched cards.

Enough information was therefore available to design an efficient search system giving:

(a) Full substructure search facilities.

(b) Enough information after searching for the compound to be ordered from Aldrich.

(c) A price per gram so that price limits on searches can be carried out.

The data base was in two parts:

(a) Substructure search tape detailing catalogue number, CR number (if present in main Company file), WLN containing the suffix, fragments for substructure search and molecular formula.

(b) Additional data file stored on magnetic tape, but loaded onto disc when required. This detailed the molecular formula, catalogue name, actual price and price per gram. It was used most often as a data file for printing or could be searched in conjunction with the substructure tape to obtain compounds of a certain type with a given price limit.

2.7.3 Detailed indexes of small catalogues

Jealott's Hill Technical Information Unit began to index in the Wiswesser Line Notation various small, but potentially useful catalogues. These were maintained on punch cards and were searched by use of printed lists. The information recorded included catalogue number, WLN containing the suffix, molecular formula, catalogue name and price. It was felt that such data would complement commercially available data already maintained by Pharmaceuticals Division.

2.7.4 An integrated commercially available index

After a year of searching the existing commercially available files, a number of points arose:

(a) It was time-consuming to search both the commercial A file, and the Aldrich files. However, this was the only way in which full coverage could be guaranteed. Some inconvenience was caused by the same compound appearing in both sets of answers. It was usual in fact to restrict questions to one data file.

(b) On a Company search, the Aldrich availability was not known.

(c) Searching of only the commercial A collection was not making use of our knowledge of commercial availability.

(d) Other commercial information was becoming available, e.g. that produced by Jealott's Hill.

It was therefore decided that an integrated index of commercially available compounds should be set-up - the available chemicals (AC) index.

The design of such a data base was now considered. Disc packs were now available on the Burroughs 4700, and it was decided to use this medium for this index. It had several advantages over the tape systems, e.g. fast access times, less prone to handling problems and capable of random access, yet was not as expensive as the head per track disc used for the on-line Company file. For convenience, master files for the various catalogues could be maintained on magnetic tape, and merged and loaded on the disc pack when necessary. This was feasible since updates of the data are not frequent (a maximum of one every three months) and copies on magnetic tape were required, in case the system failed.

On-line access was not anticipated, hence the added costs of head per track disc could not be justified.

There are three potential search items on such an index:

(a) The chemical information: fragments, WLN and molecular formula.

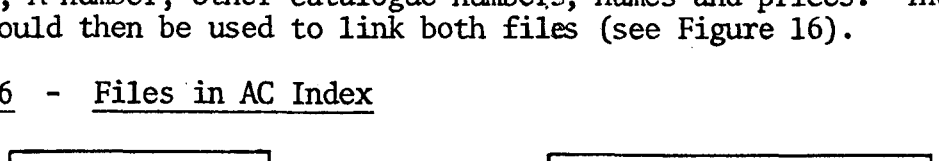
(b) The price per gram.

(c) The presence in the main chemical data base.

The last is a yes/no option and can be stored on the search file as a fragment (no. 149) in the fragment screen. The actual CR number can then be recorded in the additional details file. The price per gram can therefore be stored in the positions 1-6 (usually occupied by the CR number) in the standard search file.

As in the main chemical data base, similar compounds need to be related - the AC number, therefore, refers to the molecular species only and there may be a large number of compounds for any one species. The data required on the additional data file is, therefore, CR number, suffixes, A number, other catalogue numbers, names and prices. The AC number could then be used to link both files (see Figure 16).

Figure 16 - Files in AC Index

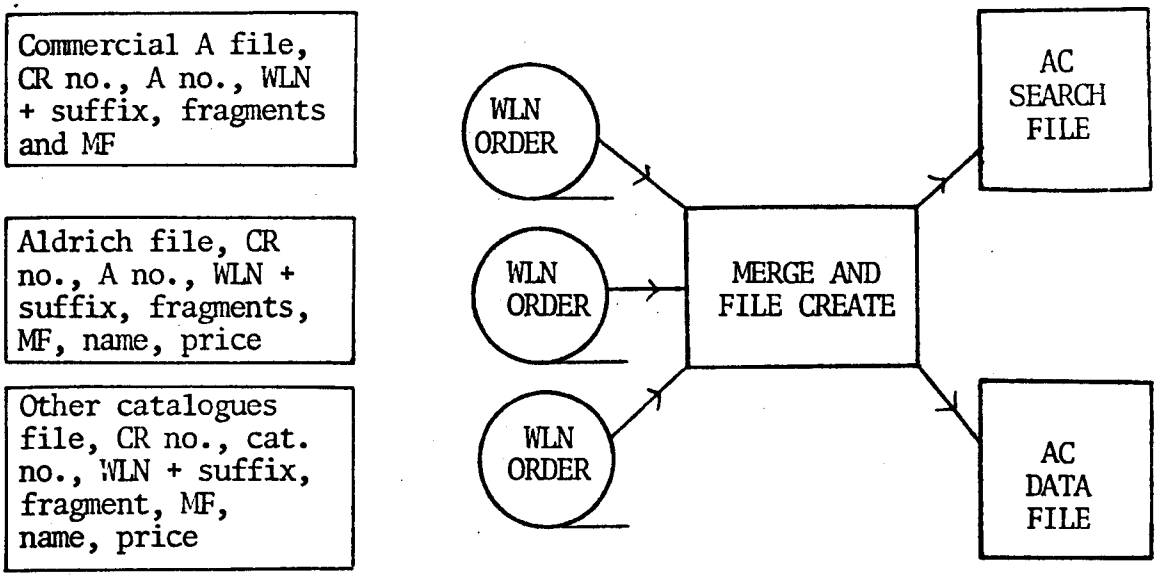


There may be any number of additional data elements for a record on the search file, and some effective way of linking these two files together is required.

The files are created by the merging of existing files, presorted into suffix within WLN order (Figure 17).

Compounds with like WLN are brought together, and the WLN, MF and fragments are written to the search file. The additional data for that one compound is then packed into the additional data file. The data is stored as variable length fields in fixed length records (each one disc segment in length). The address of the first data element is written to the search file as the AC number and the lowest price per gram found is stored in the remaining data field.

Figure 17 - Creating the AC Index



Each field on the additional data record is separated by a unique character (FA-FE) and the end of record is defined by the character (FF). The next available AC number is the next element on the additional data record. An example of a record is shown in Figure 18.

Figure 18 - Part of the AC Index

PRICE AC NO.

100000	000001	Fragments	WLN	MF
050000	000006	Fragments	WLN	MF
200000	000007	Fragments	WLN	MF
SEARCH FILE				

1	CR No.	FA	Suffix	FB	A Nos.	FC	Cat. No.
2	Cat. Name	FD	Price	FE	Suffix		
3		FB	A No.	FC	Cat. No.	Cat. Name	
4	FD	Price	FE	Cat. No.	Cat. Name	FD	
5	Price	FE	FF				
6	CR No.	FA	A No.	FC	FF		
7	A No.	FC	Suffix	FB	A No.	FC	FF
DATA FILE							

This index has a number of operational advantages:

(a) The search file is in WLN order, hence, printed lists in WLN order (for specific compound enquiries) can readily be obtained.

(b) Information on compounds of the same species but different salts or purities is brought together.

(c) The AC number is the actual key to access the data file at the end of the substructure search.

(d) The actual keys are in ascending order when examining the data file, thus minimising disc head movement.

(e) The system is flexible allowing for any number of catalogues to be introduced.

2.8 Literature Files

Literature based information services are carried out by the Literature Services Section of Pharmaceuticals Division and are divorced from the functions of Data Services Section. Many of the services based on commercially available literature services are run by ICI Head Office for all the interested Divisions. These include services such as Chemical-Biological Activities (CBAC) and Chemical Abstracts Condensates (CAC), and are text-based information retrieval activities.

There is only one structure-based literature search file - the Index Chemistry Registry System™ (ICRS) produced by the Institute of Scientific Information, and as such is now the responsibility of Data Services Section. The ICRS data base is the only literature search file using the WLN which is commercially produced. The data base is produced monthly by ISI, primarily as a current awareness service. It reports the new compounds mentioned in the literature during the previous month (about 110 journals are scanned). ISI also produce their own search package (RADICAL II) for use on IBM 360/370 computers (76). The current awareness service had been run by Head Office for a number of years. Use by the various Divisions was low - about 100 profiles (60% Organics Division, 30% Plant Protection Division and 10% Pharmaceuticals Division), and the annual cost (including systems maintenance) was around £4,000. The low usage was thought due to the poor search facilities in RADICAL II, lack of adequate structure representation on output and lack of experience of WLN by the personnel involved.

Data Services accepted responsibility for the ICRS data base, with the prime objective to develop a fast and efficient system to cater for the chemists needs in compound searching from the literature. It was felt that these would be in three main areas:

(a) SDI - the use of the CROSSBOW programs should improve the current awareness service. The main benefits would be that searches would be coded by people already working in the area, a greater depth of search specification was possible and structural diagrams could be printed on output.

(b) Retrospective search - chemists require some form of compound search in the past literature. An 18-month sample was thought ideal since it was a manageable size (about a quarter of a million records), represented the most up-to-date information, and filled the gap until the Chemical Abstracts formulae indexes were available.

(c) Reaction indexing - a data base of novel reactions could be abstracted from the ICRS files and form the basis of a reaction indexing system.

2.8.1 Format of the ICRS data base from ISI

The monthly data base consists of a multi-file magnetic tape in IBM format. It contains two files: the compound file and the bibliographic file. The compound file records the WLN and molecular formula for each compound considered novel in the journal article. Each journal reference is given an abstract number and each compound is given a compound number within that abstract. The WLN used complied with the standards recognised by the Chemical Notation Association (CNA) and is not directly compatible with those present in the Company generated files. The molecular formula is also of a different format. Here the field is laid out in twelve sub-fields of five characters each. The first five positions are always carbon; the second five positions are always hydrogen, and the other elements follow in alphabetical sequence. Each record is 210 characters long.

The bibliographic file records the source of the journal articles. There may be any number of records (each 90 characters in length) for each bibliographic record, each with a separate card code in column 10 (A - author's name, G - journal reference, H - organisation address, I - instrumental data alert codes, N - use profile subject, S - subject, T - article title). There may be up to 70 characters of data on each code, and if more than 70 characters of data is present it overflows onto further records. The abstract number is present in the first 6 characters of the record.

Any proposed system was required to:

(a) Provide fast and efficient substructure searches on a current awareness basis and on a retrospective basis (for the past 18 months).

(b) Provide 2-dimensional structural output for these search services, the structures being presented along with the bibliographic data.

(c) Provide textual searches on the bibliographic data on a monthly or a retrospective basis.

(d) Allow for the preparation of structural subfiles based on various parameters, e.g. biological activity, new reactions.

I felt that with a small amount of systems design we could develop an ICRS system more suitable to our needs. The following criticisms were levelled at the facilities offered by ISI:

(a) Lacked adequate screening facilities for substructure searches on large data bases. This is reasonable since the facilities are geared to current awareness only.

(b) No structural output.

(c) Bibliographic files were large and expensive - records were space filled to fixed length records. Any number of these records may be present for each abstract.

We therefore set up two files.

2.8.2 Substructure search file

Firstly, the WLN and molecular formula were processed into ICI format:

(a) The molecular formula was contracted to a continuous record in Richter format (maximum 18 characters).

(b) The multipliers and methyl contractions were removed from the WLN (7S

(c) All salts are treated as free acid or base, and any additional information placed in the suffix (preceded by ▽&&).

The compounds were then submitted to the fragmentation generation program and the fragment screen set up. The search file holds the following information, as shown in Figure 19.

Figure 19 - Contents of ICRS Search File

FIELD	LENGTH
Abstract no.	6 digits
Compound no.	3 characters
Fragment screen	152 bits
WLN field length	3 digits
(WLN + Suffix) field length	3 digits
WLN + Suffix	144 characters
Molecular formula	18 characters

The total record length was 190 characters. The substructure search files were held on disc pack with security copies on magnetic tape.

2.8.3 Bibliographic file

The bibliographic file had to perform two functions:

(a) Provide bibliographic data for printing on structure cards such that the chemist could find the original references.

(b) Provide text searching on data such as subject fields, titles, instrument data.

There is one variable length record for each abstract number. This record is separated into a number of variable length fields, each separated by an "@" sign followed by the next card code. The exception

is the instrument data - this is simply the presence/absence of the use of a given technique, each represented as an alphabetic code. Since this was a common search item and had no display value, it was decided to set up a bit screen, bit 1 = presence of A, etc. The file, therefore, has the format given in Figure 20.

Figure 20 - Format of ICRS Bibliographic File

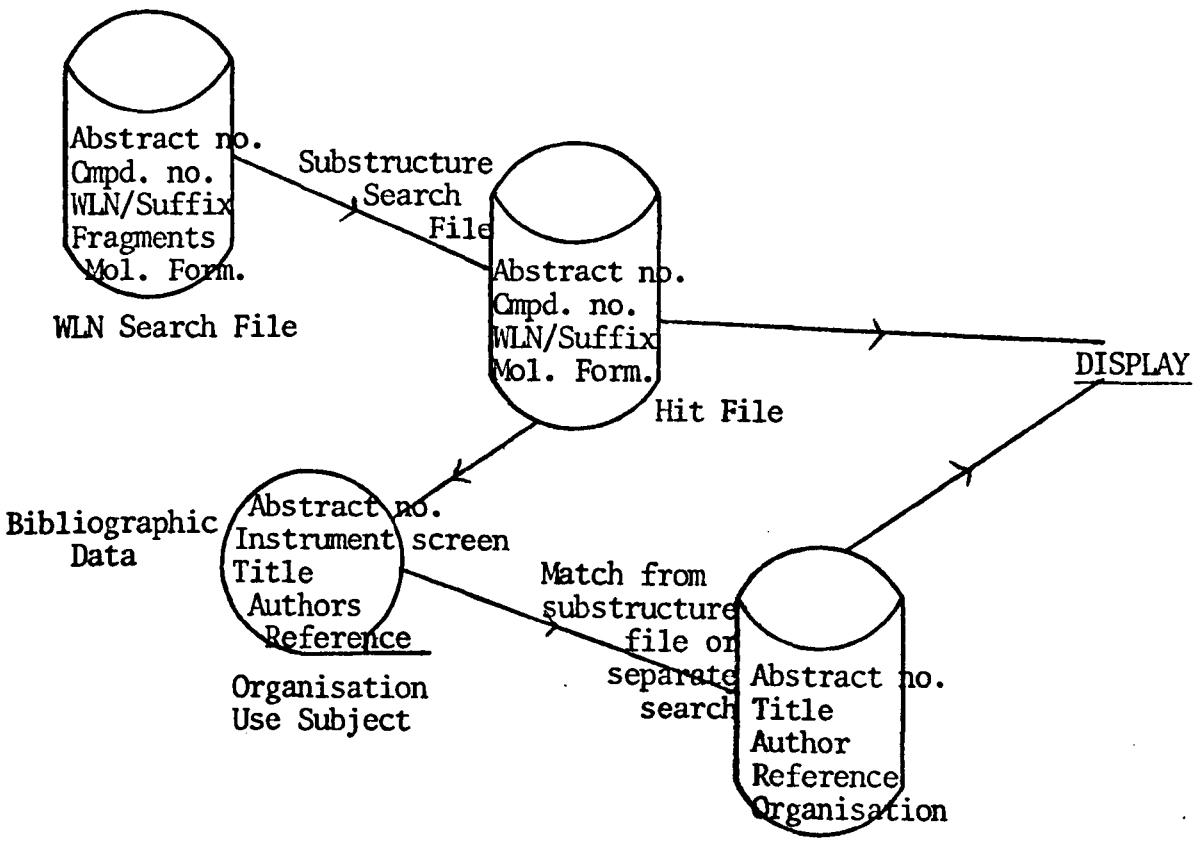
POSITION (DIGITS)	FIELD
1 - 6	Reference number of abstract.
7 - 13	Bit screen for I data.
14 onwards	Data.

The data is recorded in the following order: title (T), authors (A), journal (G), organisation (H), use (N) and subject (S). The first four fields only are used when printing the structure cards, hence the remainder of the record can be ignored. Of the remaining two fields, the use profile is more likely to be searched than the subject, and hence, is coded first. This indicates any field of biological activity with which the compound might be associated.

All data on A, S and T records are joined together to give one field for each code: A fields separated by commas, S fields separated by semi-colons, and T fields separated by spaces. Hence, there is only one field for each code letter, making searching and presentation simple.

When the file was set up, some amendments were necessary to cater for differences in representations on the IBM and the Burroughs, i.e. all occurrences of "%" are replaced by "(", and all occurrences of "¢" are replaced by ")". The file is held on magnetic tape, in abstract number order, and can be accessed separately or after a substructure search. The inter-relationship of the two data files is shown in Figure 21.

Figure 21 - Diagrammatic Representation of ICRS Data Base



2.9 Specialist Files

2.9.1 The Hansch data base

Hansch and his colleagues at Pomona College produce a magnetic tape twice yearly, detailing compounds known to have partition coefficient data. Such information is useful for chemists using the Hansch technique (83) for structure/activity correlation. If a chemist cannot find information on specific compounds, he can often estimate the partition coefficient value from values for related compounds, hence the need for substructure search services.

The data recorded on the magnetic tape is given in Figure 22.

Figure 22 - Contents of Hansch Data Base

WLN (not compatible with ICI).
MF (not compatible with ICI).
Name.
Solvent.
Journal reference.
Footnote.
Partition coefficient value.
PKa values.
Sigma values.

The tape is prepared as an image of the print files which would be used to produce the printed form of the index on the IBM 360/370 series. The main file is repeated a number of times in molecular formula order, WLN order and in the form of a KWIC index on WLN. We isolated the WLN order file onto disc and selected the following information:

WLN (may take up more than one print line, and is processed to obtain a complete WLN record).
MF.
Name (may take up more than one print line, and is processed to obtain a complete name record).
Number indicating solvent used (refers to solvent index).
Number indicating journal reference (refers to journal index).
Number indicating footnote (refers to footnote index).
Note indicating whether value measured or calculated.
Partition coefficient value in solvent used.
Partition coefficient value in octanol.

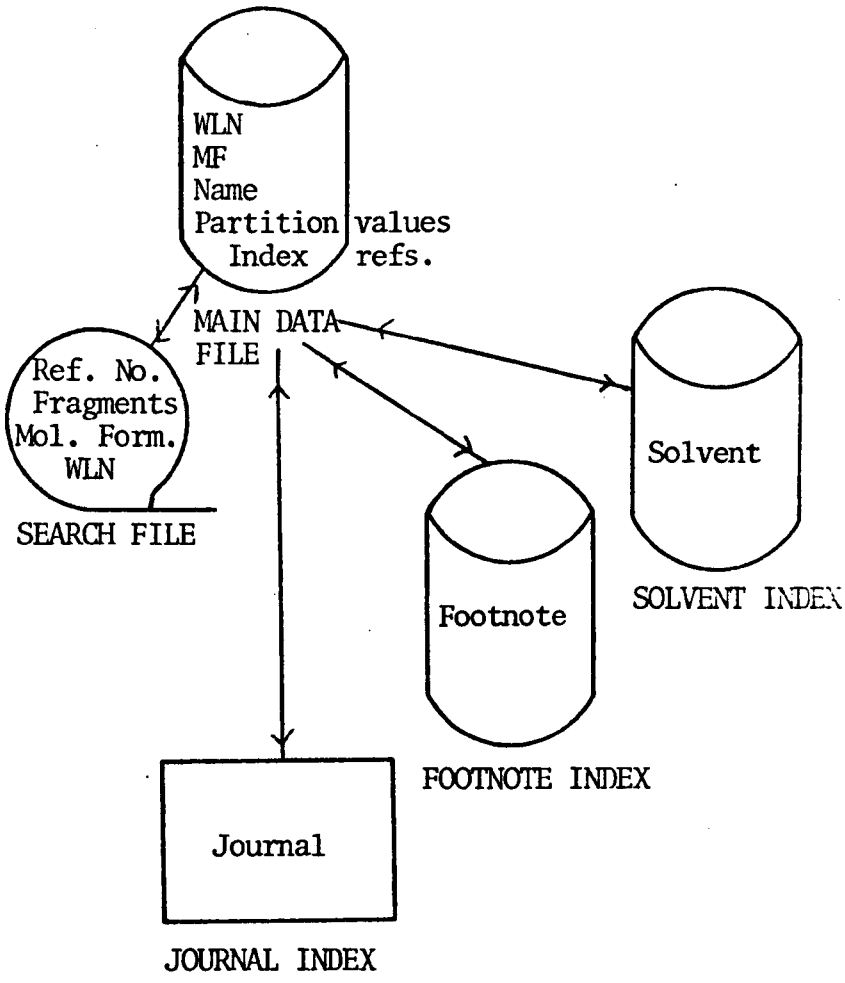
The PKa values are appended to the end of the name record, enclosed by slash marks. The sigma values were on a separate file which was removed and stored on disc. The solvent index, journal index and the footnote index were converted to random access files such that the numbers on the main files could be used as actual keys to the index files.

The WLN search file was produced by isolation of the WLN and MF from the main data file, and the allocation of a unique number to each record. This unique number is used as the random access key to access the main data file for information after a substructure search has been carried out. The WLN search file is set up by generation of the fragment code and holds the following information:

Hansch reference number.
Fragments.
WLN.
Molecular formula.

A diagrammatic representation of the Hansch data base is given in Figure 23.

Figure 23 - The Hansch Data Base



After substructure searching, the reference number on the search file is used to access the remainder of the information on the main data file. Here, the references to each of the files are used as actual keys to find data relating to the reference numbers.

Hence, after a substructure search has been carried out, the following information is presented:

Chemical information - WLN, MF, name and structural diagram.
Partition values, and an indication whether measured or calculated.
The originating reference, authors name and journal source.
The solvent used, e.g. octanol or diethyl ether.
Any footnote clarifying the data, e.g. equation used in the calculation.

In addition, the data file may be searched independently of the search file, e.g. find all compounds tested in octanol and diethyl ether.

2.10 Processing Other Data Bases

The use of random access techniques with the Company Chemical Data Base has involved special design criteria. The advantages of the sophisticated file structures became apparent when techniques were developed for exploitation of the data base.

A simpler approach was taken with data bases from external sources. The files are simple and flexible enough to cater for any new supply of compound information which may be available. In this way we hope to keep pace with the increasing and changing demands of users.

For a potential application we would set up at least two separate files:

(a) Compound file, detailing the chemical entities involved.

(b) Additional data files, holding textual or value information which can be searched in conjunction with the chemical information.

In this way, the same systems can be used to interrogate chemical information from any source.

Chapter 3 - Techniques For Use With The Chemical Data Base

3.1 Introduction

A set of programs was required to access the "Chemical Data Base" and to carry out the various job functions for which the data base was designed. On-line techniques had to be evaluated and their impact on the two main job functions monitored. A gradual movement to on-line working was considered vital since at all times cost-effectiveness was paramount. The full development of the facilities discussed occurred over three years and was closely associated with the results of the survey on usage.

3.2 Compound Registration - Outline of Facilities

The "Chemical Data Base" was set up on the B3500 to carry out the following registry functions: on-line registry of new compounds into Pharmaceuticals Division for biological testing, registry of all new compounds into the Company compound centre, and the inter-relation of compounds indexed in the various Divisions. Initially, a totally on-line registry system was envisaged only for Pharmaceuticals Division where fast response times had to be achieved, other Divisions using batch facilities. This had a number of advantages:

(a) The data base was always up-to-date with Divisional information and was immediately accessible for registration and search.

(b) The manual novelty checking procedures could be replaced by automatic methods.

(c) Immediate correction of input errors was possible.

(d) Batch registration for the other Divisions prevented their abuse of the system, all data being first checked by Pharmaceuticals Division.

This latter point was important since for maximum effectiveness of a registry system, like compounds must always be indexed in the same way. The first stage in the development of the registry system was the batch update facilities for information from all Divisions. The move to full on-line working was seen in three phases: on-line enquiry followed by batch registration; on-line enquiry, data validation and batch update finally leading to on-line enquiry, data validation and update. The on-line operation on the Burroughs 3500/4700 was never reliable because of machine breakdown and software inadequacies. Therefore, after some experience it was decided not to implement the total on-line registry. To maintain file integrity it was felt that on-line enquiry and data validation, followed by a nightly batch update was the best possible solution. On-line facilities were, however, extended to all Divisions since adequate control could be maintained by monitoring the batch update and restricting the transactions possible on the on-line system.

3.3 Compound Registration - Batch Facilities

The first batch system was written such that the transition to on-line working would be as simple as possible. To fit in with Divisional standards two programs were written: the user routines interfacing with the terminals and the file handling routines processing the "Chemical Data Base".

Since the data base had been designed using the FORTE package, Burroughs file organisation software (77), the file handling routines were merely a collection of FORTE statements necessary to carry out the various functions. The user routines, on the other hand, contained all the computer codes necessary to carry out the registry job functions. This user program was in five parts:

(a) Validation - all input data was on punched cards and routines were written to ensure against punching and transcription errors and to ensure the data was in the correct format.

(b) WLN/molecular formula check - errors in the translation of the two-dimensional structure diagram into WLN or molecular formula can be detected by generating a molecular formula from the WLN and comparing the generated value with that input. Mistakes in either of the two data elements can often be found in this way.

(c) Novelty checking - new compounds being entered must first be examined since the data may already be known. The WLN is compared with those already in existence, and, if present, rejected.

(d) Update - all new information has to be added to the files.

(e) Report production - all validation errors, WLN/molecular formula check messages and data already present on file must be reported in an error listing and all data successfully added to the file reported in an update listing.

The following types of transactions were catered for:

(a) Addition of new compounds as defined by unique WLNs.

(b) Addition of further Divisional references for an existing WLN.

(c) Addition of salt or stereochemical descriptors (suffixes) to existing WLNs.

(d) Deletion of total records, suffix only descriptors or reference numbers.

The type of transaction required was specified on the input punched card, and the program carried out the necessary functions.

The batch registration system worked well when first introduced, but was often unnecessarily laborious because of the large number of checks built in. Two major alterations were made to speed up the processes:

(a) Company Registry Numbers (CR numbers) were automatically allocated for unique WLNs. This removed the need for manual records on which CR numbers had been allocated or recently deleted. Manual methods can become tedious when a number of sources are all allocating CR numbers.

(b) The flow of compounds between the various Divisions was steadily increasing and hence the need to add additional Divisional numbers to existing WLNs was increasing. This facility was improved by adding a further type of transaction which linked two Divisional numbers. Previously, additional reference numbers had to be related back to the CR number.

In addition to registration programs, it was also necessary to prepare a batch enquiry routine so that selective prints of the "Chemical Data Base" could be obtained. The program written catered for specific compounds or ranges of compounds using reference numbers or WLN as entry points.

The program was used when there was some doubt as to whether a record was already on file. It was not entirely functional since turn-around was slow (often overnight or longer). For this reason, the program was mainly used to periodically prepare retrospective lists. These would then be used by the various information officers to look up specific pieces of information.

3.4 Compound Registration - On-Line Facilities

3.4.1 On-line enquiry

To aid novelty checking and to reduce the amount of clerical effort required for the nightly update, an on-line enquiry facility was urgently required. Interrogation was to be through a VDU terminal with a "teletype" attachment for hard copy, and the program was to be simple to use and require the minimum of operator intervention. The design of the "Chemical Data Base" allowed for on-line access through the following keys: molecular formula, Wiswesser Line Notation (WLN), CR number, M number, SC number and R number. One enquiry format was thought easiest to use and consisted of:

(a) The routine identifier (%401 for all routines). This allows the main on-line system to find the chemical data base routines rather than, say, commercial or biological routines which can be accessed by the same system.

(b) The enquiry type - MF for molecular formula, WW for WLN, CR for CR number and DR for M, R or SC numbers.

(c) The data to be found, i.e. molecular formula, WLN or reference numbers.

There are two basic replies for each type of enquiry. Firstly, if the data is not present, the input request is returned with the appropriate error message:

MF
WW
CR
DR } NOT ON FILE

If the data is present, corresponding answers will be obtained on the screen. For a molecular formula enquiry, there may be any number of answers. The information on each compound is restricted to its WLN and the main references. If further information is required about specific compounds in the list, then the CR number or WLN enquiry will supply them. The amount of data possible on a screen is limited and only the first three potential answers are displayed. The next three notations can be examined by retransmitting the screen, i.e. "pages" can be turned. This process can be time-consuming if there are a large number of entries for a given molecular formula (74). To reduce the amount of processing required to deal with one enquiry, the molecular formula routine was later modified. Only the first twelve compounds are now accessed and prepared for display and the total number of entries indicated. To do this it has been necessary to add a count of the number of entries on the molecular formula index.

In such cases, a WLN enquiry is more fruitful since it results in all the information about one compound being displayed. If the WLN is present, the reply is the WLN, MF, main references and the number of suffixes on file. The screen is filled with the first three suffixes together with the appropriate references. "Pages" may be turned as in the molecular formula enquiry. Again, this may be time-consuming if the WLN and suffix is known. A suffix search may then be carried out where WLN and suffix are both input. A suffix enquiry produces pertinent information on molecular formula, WLN, suffix and suffix reference numbers. To aid novelty checking, the molecular formula can be input in addition to the WLN, and this will be checked and verified by the MF/WLN check routine. The check is carried out before entering the normal enquiry function of the routine, and the enquiry will not proceed if the check fails. Appropriate error messages are then displayed.

A Company reference number enquiry results in all the information about the specific CR No. being listed whilst the Divisional reference number enquiry produces information specific to that number.

These routines have been extensively used in the day-to-day working of the section and have proved very successful, both in novelty checking and in specific compound location.

3.4.2 On-line data validation

A simple extension of the on-line enquiry routine was the ability to input data after a novelty check, and to store the data ready for update at a later date. These modifications were felt necessary since:

(a) Novelty checking and validation would be done by the Divisions concerned and could therefore be corrected by them.

(b) The time lag due to post and punching would be reduced.

(c) There would be less chance of error by transcription or punching.

One new enquiry type (RG) was developed. This was very similar to the Wiswesser Notation enquiry insomuch as a WLN and MF are input, but the enquiry is made with registration in mind. The WLN/MF check is first carried out - if the input fails the appropriate error message is produced and no further action is taken. If the check is passed, the WLN is then searched for on the data base. If not found, the user is presented with a new compound registration form with the necessary WLN and MF already filled in (see Figure 24). If the WLN is on file, the user is presented with the cross-reference or suffix amendment form containing all the known information - CR number, WLN, and MF (see Figure 25). Both forms are easily filled in, the cursor jumping from place to place where input is required. The screens are then retransmitted and enter the new "Registration Routines" (%402). These routines receive the information, vet the requests and store the data on a temporary file for later processing in batch mode. The routines are similar to those in the batch update routine except the data is not updated on the file.

There are also a number of housekeeping routines to enable users to keep track of progress on a given day (WF routines). For example, it is possible to list all the Divisional reference numbers that have been input since the last update. Alternatively, all information input on a given reference number on the temporary file may be retrieved or deleted.

Using these routines, information officers at remote locations can enter data for immediate checking and later updating. The one exception is total record deletion. Since here it is possible for one Division to delete information from another Division without their knowledge, it was decided that these transactions should be carefully monitored. They can, therefore, only be carried out in batch mode. A program was designed to process the registration data accumulated through the on-line system and reformat it into a form suitable for use with the batch update programs.

3.5 Substructure Search - Outline of Facilities

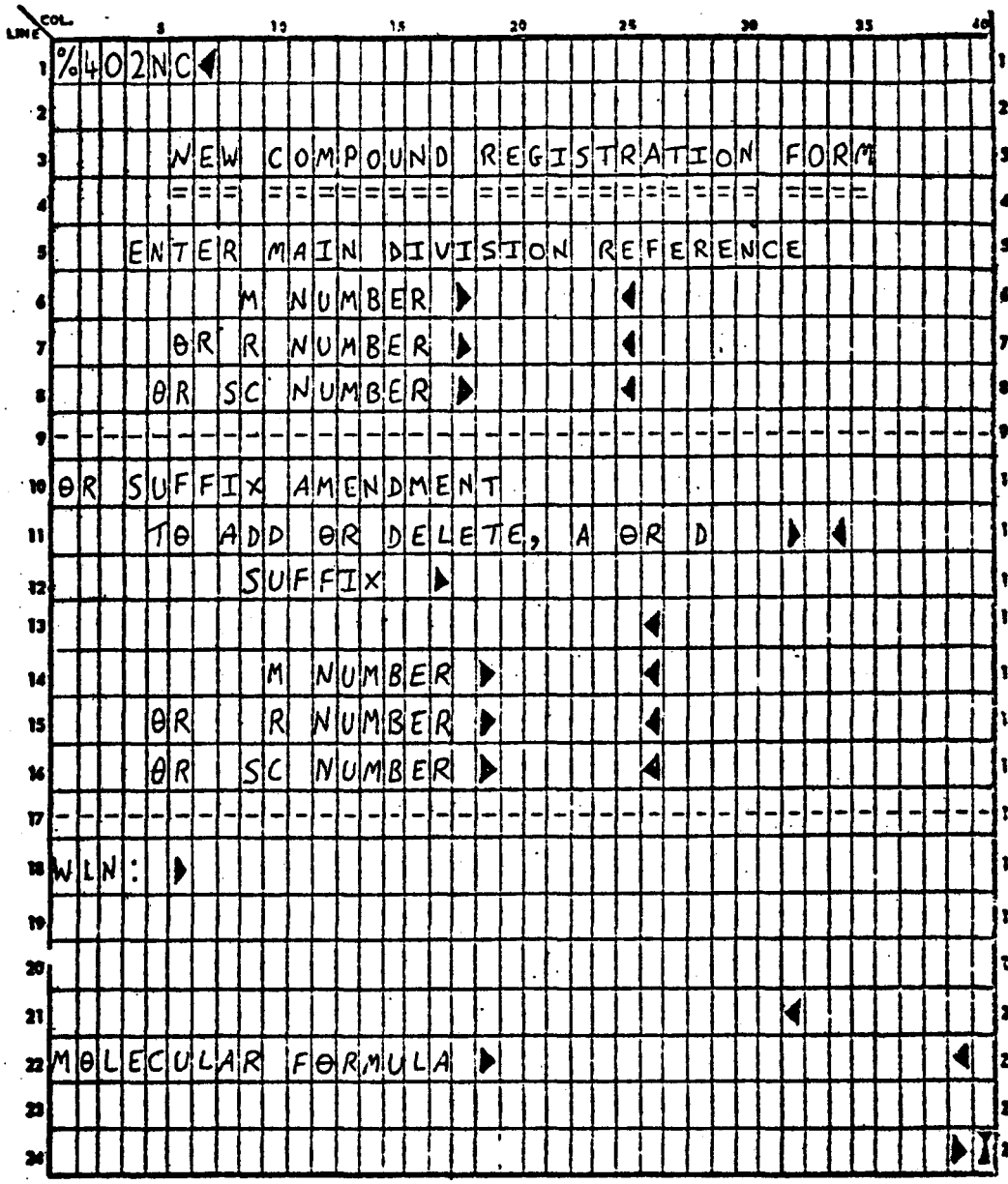
3.5.1 Developing a search system

With the amalgamation of the Company Compound Centre into Data Services Section, there was an urgent need for a substructure search system capable of carrying out fast and efficient searches both on Company and Divisional data. In addition, extensions and modifications to existing substructure search services would be made easier if a system for the manipulation of structures could be obtained which was independent of a data bank. This suggests that one CROSSBOW search system should be developed, and this should meet as many of the Company's structural information needs as possible.

The Mark I CROSSBOW system (73, 74) was not considered suitable for future development since:

SHEFFIELD UNIVERSITY LIBRARY

Figure 24 - New Compound Registration Form



(a) Many of the programs were written in PL/1 and would require extensive translation for use on the Burroughs 3500.

(b) Housekeeping was expensive. Three files were maintained: WLN's and reference numbers, connection tables and an inverted file of fragments. These files were large and updates costly.

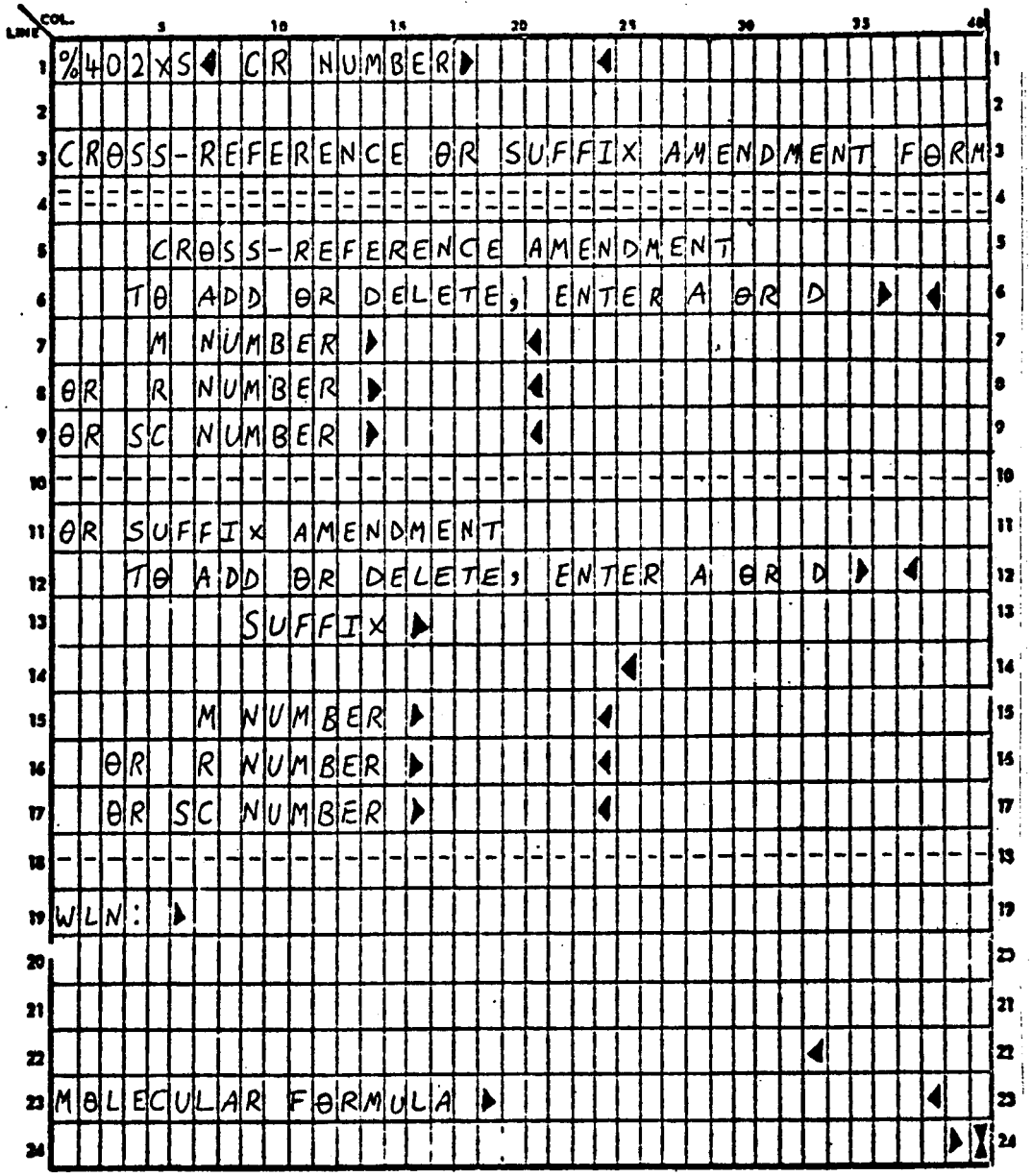
(c) The fragment code was difficult to use - there were some 5,000 separate fragments, over 4,000 with a posting of less than 10. In contrast, the manual fragment code used by Pharmaceuticals Division had 312 fragments.

(d) There was no atom-by-atom search for use on the connection table - all substructures had to be defined using WLN making search formulation difficult. This is supported by other groups working with WLN as the sole search mechanism (78).

(e) Structures were printed on continuous stationery, several to a page. This made the editing of false drops cumbersome.

The Mark II programs had been written to overcome some of the operational problems encountered with the Mark I programs. In particular:

Figure 25 - Cross-Reference or Suffix Amendment Form



(a) The programs were written in COBOL and readily convertible between computer manufacturers.

(b) The connection table was simplified and the WLN to connectivity routines improved. Generation of connection tables was no longer the costly business it was.

(c) An atom-by-atom search program had been developed to perform detailed network searches.

The main problem was, therefore, the fragment code. Manually generated fragment codes had played an important role in the development of chemical substructure search systems (24), the one developed by ICI having been used for twenty years. However, fragment codes had traditionally been used as a stand-alone search tool. An evaluation was therefore required as to their effectiveness in a substructure search system where WLN string search and connectivity table atom-by-atom search were also possible. An exploratory system was set-up on the ICL 1902A where the Divisional fragment and WLN files were held. Firstly, a number of searches were run against the files both containing 75,000 compounds:

(a) Fragment file: 312 attributes stored as a bit matrix.
(b) WLN file: WLN/MF stored as fixed length records.

Search times for a batch of five questions varying in complexity were: fragment file: 25-30 minutes; and WLN file: 40-180 minutes. The WLN file did have some advantage in that it could be used by other CROSSBOW programs to produce the 2-D structure display on the connection table for further detailed searching.

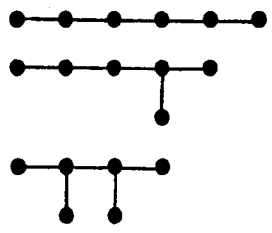
The initial substructure search system set up, therefore, relied on a first initial search on the fragment code, selection of the appropriate WLN records for hits from the fragment search, optional further WLN searching, connection table generation for WLNs selected as hits, optional atom-by-atom searching of the connection table, and finally generation of the two-dimensional structure record. The search logic used at each stage was fairly simple. Normal Boolean logic (AND, OR, NOT) was used for the bit fragment search and an existing program was used. The standard ICL package, FIND II, was used for the WLN searching (79). The FIND II package is designed for the string searching of textual data. It can perform AND, OR and NOT logic on single characters or groups of characters, and will search through a record looking at each character one at a time. The atom-by-atom search program handled one question at a time and performed an iterative search on the connection table network.

Up to 8 nodes could be searched at one time, and there could be up to 10 alternative connection table units for each node. The types of network possible were:

Linear

One-branch

Two-branch



where the branch could be only one node in length. When the search system was first set up most questions were completely answered at the fragment level. Access to the remainder of the system being for structure display only. As the information officers became used to the multi-level approach, so the use of WLN string searching and atom-by-atom searching developed (80).

Experience showed that by adding a number of additional features to the string search program, greater specificity of search could be obtained and considerable savings in processor time could be achieved. A new WLN string search program was designed which took advantage of standard WLN features. The following logic was added to the standard string search logic found in text searching programs such as FIND II:

(a) Followed-by logic - this ensures that one character string follows another in the notation. This is useful where substituent patterns on a given ring system are required. For example:

AND, 'T6NJ' - /∇BQ/ , /∇CQ/. END

where the substituent pattern (∇BQ or ∇CQ) is attached to the ring 'T6NJ' and must therefore follow it in the notation.

(b) Start of notation only - this ensures that the string of characters stated must appear at the beginning of the notation. This can save processing time since some strings can only start the notation. A second string would then be used to locate the grouping in the molecule. For example:

OR,$'QV8' , '8VQ' . END

The Q symbol is terminal, hence the grouping 'QV8' can only start the notation whilst '8VQ' can appear anywhere in the notation.

(c) No-space character - the '*' sign does not normally appear in the notation and its presence in a question was used to indicate 'any character but not space'. Because of the special meaning of a preceding space, the character could be used simply to distinguish between locants and other characters. For example:

AND, '*B'. END

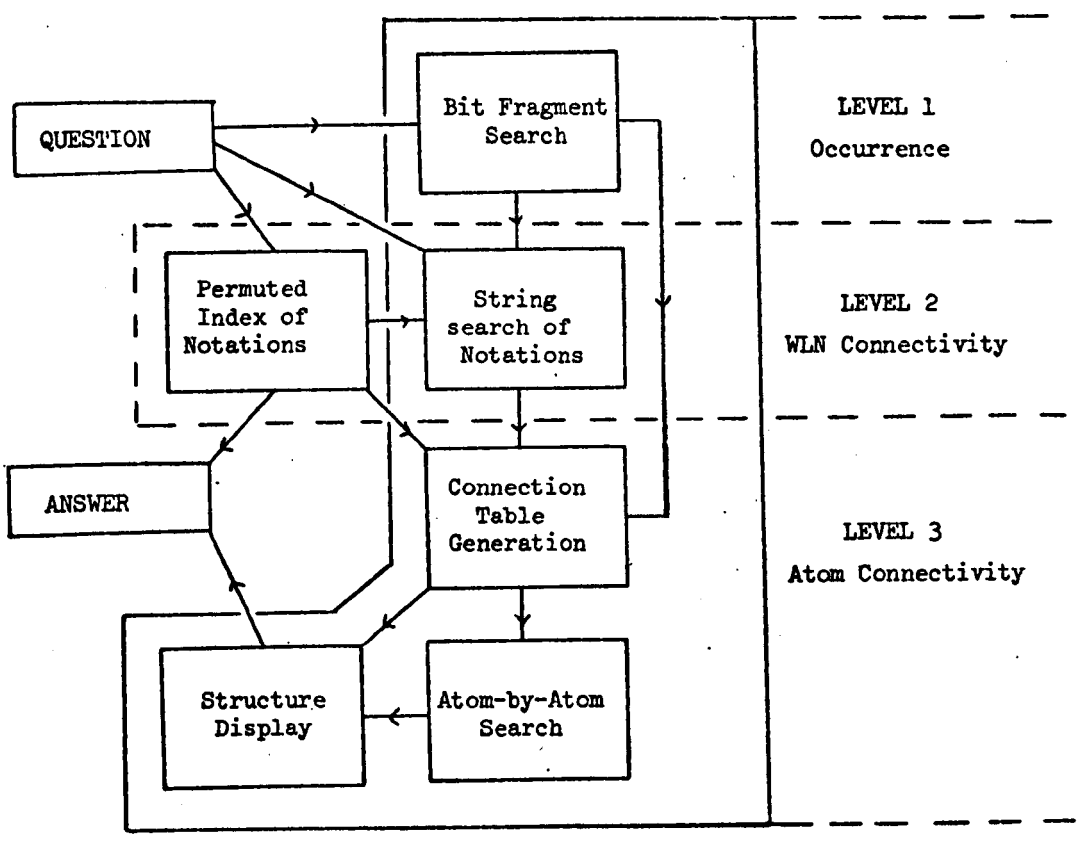
The search is for boron compounds and therefore needs to exclude the occurrence of the locant ' ∇B'.

(d) Notation length - the total notation field length used was 144 characters, but the average length of the notation was only 19 characters. To save time in searching unnecessary characters, the program was made first to scan the notation for the occurrence of two spaces. This marked the end of the notation and only preceding characters were searched.

The use of a purpose built string search program instead of FIND II was very successful. Coding was simpler, more questions could be answered more exactly by string search and the search times were considerably reduced.

The substructure search system (Figure 26) seemed to work well. It was able to efficiently answer 95% of the questions put to it, but the multi-level concept took time to develop (80). The efficiency of the system depended on the interaction of the three levels. The most expensive iterative procedures were only used on small numbers of compounds (usually less than 3,000) and only when the problem had not been solved by cheaper search levels. Conversely, use of one of the simpler levels in areas where it was not suitable could lead to long search times and the possibility of relevant answers being missed.

Figure 26 - Multi-Level Search System



Following the use of this system on the ICL 1902A, it was decided that the larger Company substructure search system should be based on this type of multi-level concept. However, it was not possible to convert the system directly. The fragment coding was only available for compounds in the Pharmaceuticals Division collection - it was not feasible to produce the manually assigned fragment code for the other Divisions. It was, therefore, necessary to design a suitable fragment code which could be generated automatically from the WLN. If this was possible, then the multi-level substructure search system could be used with any WLN based data base.

3.6 Substructure Search - Generating a Fragment Screen Automatically From The WLN

Experience had shown that the fixed fragment approach (as in the 1900 program) was more suitable for substructure search than the open-ended approach (as in the original Mark I fragmentation program on the IBM computer):

(a) A fixed fragment code was easier to interpret on searching - the choice of fragments being from a smaller, better defined set.

(b) Searching was more efficient - a fixed code could be stored in bit form and searched by simple Boolean logic.

As the fragment code was being designed to operate efficiently within the multi-level search concept, a number of criteria had to be satisfied:

(a) Searching was to be fast and efficient since it was to be the initial screen and searched for all records on the file.

(b) Storage should be as small as possible to minimise the overheads when searching the chemical data base.

(c) It should be readily generated from the WLN - this would give some flexibility if a decision to change some of the fragments was made.

(d) The fragment code should be generateable for at least 99% of the file - 1% of the file was the maximum that could be handled manually.

(e) It should complement the searching facilities already available in the WLN string search facilities.

Two points were extensively explored:

(a) What type of information could be readily obtained by processing the WLN?

(b) From the manually assigned fragment code, which fragments would complement WLN string search and which overlapped with it?

The WLN was first examined and 62 fragments isolated which could be readily obtained from WLN (see Figure 27). After analysing these fragments and the manual fragment code, it became apparent that fragments selected would depend upon four main functions:

(a) Distribution: The more frequently occurring elements - carbon oxygen, nitrogen and sulphur - were to be represented by a number of fragments whilst the less frequently occurring elements were to be grouped together. The distribution of the fragments was to be designed, where possible, to give good screenout prior to further levels of substructure search (see Appendix II - Distribution of Fragments).

(b) Ease of characterisation: Only fragments well defined by the notation were to be isolated. It was necessary to decode the ring information within the notation to establish the characteristics of the individual rings. Less processing was required for the acyclic parts of the molecule - branch groups were to be merely cited as present whilst linear groups were to be further characterised by their position as ring substituents or not.

Figure 27 - Fragments Which Are Easily Obtained From WLN

1. Metal atom.
2. Methyl group.
3. Alkyl chain with 2-9 carbons.
4. Alkyl chain with 10 or more carbons.
5. Y branching carbon.
6. X branching carbon.
7. Generic halogen.
8. 1 chlorine group.
9. More than 1 chlorine group.
10. Bromine.
11. Fluorine.
12. Iodine.
13. $-NO_2$ group.
14. N atom bonded to more than 3 atoms.
15. Trivalent nitrogen.
16. 1 -NH- group.
17. More than 1 -NH- group.
18. 1 -OH group.
19. More than 1 -OH group.
20. 1 -S group.
21. More than 1 -S- group.
22. 1 carbonyl group.
23. More than 1 carbonyl.
24. 1 linking -O- atom.
25. More than 1 linking O atom.
26. 1 primary amine.
27. More than 1 primary amine.
28. Single benzene ring.
29. More than 1 benzene ring.
30. Phosphorus.
31. Terminal dioxo- group (not NO_2).
32. Double bond.
33. More than one double bond.
34. Triple bond.
35. Simple spiro rings.
36. Single carbo ring system.
37. Single hetero ring system.
38. Bicyclic - carbo ring system.
39. Bicyclic - hetero ring system.
40. Poly (>2) cyclic - carbo ring system.
41. Poly (>2) cyclic - hetero ring system.
42. Rings other than 5-6 membered.
43-58 Consider presence of certain heteroatoms (O, N, S, carbonyl) in a ring system and their duplication in the same ring system or in other ring systems:
    (a) single occurrence in 1 ring system;
    (b) duplication in 1 ring system:
    (c) duplicated (a) in more than 1 ring system;
    (d) duplicated (b) in more than 1 ring system.
59. Other hetero atoms in a ring system.
60. Spiro carbon indicator.
61. True bridge indicator.
62. More than 1 multicyclic point.

(c) Relationship to other search techniques

The fragments were to be generic, especially for ring systems, and designed to be used as screens before detailed string or atom-by-atom searches. They were not designed to be used as the sole search technique, nor to give a total representation of the molecule.

(d) Experience

One of the requirements of the system was simple questions formulation. This was established by an analysis of questions asked to similar systems over the years.

Using these criteria, various classes of fragments were established as shown in Figure 28.

A Wiswesser ring notation fully describes the whole ring system without an explicit definition of each individual ring. However, many of the cyclic fragments required for screening relate to individual rings rather than the whole system, and so the ring notation must be decoded to identify the nature of the component rings. A subroutine has been written for the purpose, and successfully decodes the notations for nearly all ortho- and peri-fused, bridged, and spiro ring systems. It excludes systems with a branched locant path and those coded with the ring-of-rings contraction: these and some other rarer unusual conditions are recognised and signalled.

Figure 28 - Type of Fragment Generated

TITLE	DEFINITION	NOTES
All parts of the molecule.	May be present anywhere in the molecule.	Used when it would be difficult to separate the classes further, or where the fragment is very rare.
Acyclic parts:		
(1) Non-cyclic parts.	Items of low frequency or branch units found anywhere outside ring systems.	
(2) Non-substituent groups.	Groups with no direct ring attachment.	Essentially the same groups.
(3) Substituent groups.	Groups directly attached to a ring.	
Cyclic parts:		
(1) Heteroatom content.	Number of atoms of specified types within individual rings.	Cannot be obtained directly from the WLN. Some prior ring analysis required.
(2) Ring size and type.	Details size and general type, e.g. carbocyclic 6.	
(3) No. of hetero-atoms per ring.	Number of heteroatoms of any type in a given ring.	
(4) Fusion type.	Details whether ring fused or not and if so, what type of fusion, e.g. carbo/hetero.	
(5) Ring link types.	Simple screens for given ring types, e.g. bridge, spiro.	
(6) Ring counts.	Details total number of simple ring conditions.	
General molecule types.	Molecules having no other characterisation in the code, e.g. polymers.	Non-chemical.
Company file types.		

In fact, the subroutine generates a full connection table for the ring system and extracts from it the details relating to each component ring. The potential applicability of the routine, therefore, extends beyond fragment generation to notation checking and connectivity generation prior to atom-by-atom search and structure display.

The first version of the program, incorporating the ring analysis routine, generated the 122 fragments given in Appendix III. After 18 months experience with these fragments, a number of minor extensions were made making a total of 152 fragments (for the full list see Appendix II). The extensions included:

(a) Better definition of the sulphur atom. The class -C=S was separated from the remainder of the S groups. This group was the most common, widely-used and separate fragments thought beneficial.

(b) The -C(=O)-OH (acid) and -C(=O)-O- (ester) groups were removed from the general carbonyl class. These again are commonly occuring, widely-used and their separation thought beneficial.

(c) The presence of an exocyclic double bond in a ring was detected. This allowed searches for the group: X=C

to be carried out, where X is not necessarily O.

(d) A new group was introduced giving the total number of heteroatoms in the ring. This group was present in the manual fragment code, but had been excluded in the original specification. It had proved an ommission, and was included at this stage.

(e) The bi-linkage fragment was included, i.e. a direct link between two ring systems. A number of searches had been processed where this was thought useful; hence, it was included at this stage.

(f) The three-branch carbon atom was broken down into two fragments since the group had a rather large distribution. Two fragments were therefore assigned:

(i) occurrence of one Y branch;
(ii) occurrence of more than one Y branch.

The list of fragments together with their occurrence in the Company File is given in Appendix II.

To test the new fragment code, it was decided that an equivalent file should be set up and its performance compared with the system already available on the 1900. Since the fragments were to be generated from the WLN, it was decided to hold both fragments and WLN on the same record and search them together. A new program was therefore required to perform bit and string search on the same record. The fragment search was performed initially - the logic available being identical to that in the manual fragment search program. Any hits from the fragment search were then passed through to string search - the logic available being identical to that in existing string search program. The bit and string routines were therefore written into one program. No other programming changes were necessary.

The system was simple and worked well in test runs. Transfer to the Burroughs 3500 was scheduled before any parallel runs could take place, hence no direct comparison data is available.

3.7 Substructure Search - Batch System for the Company

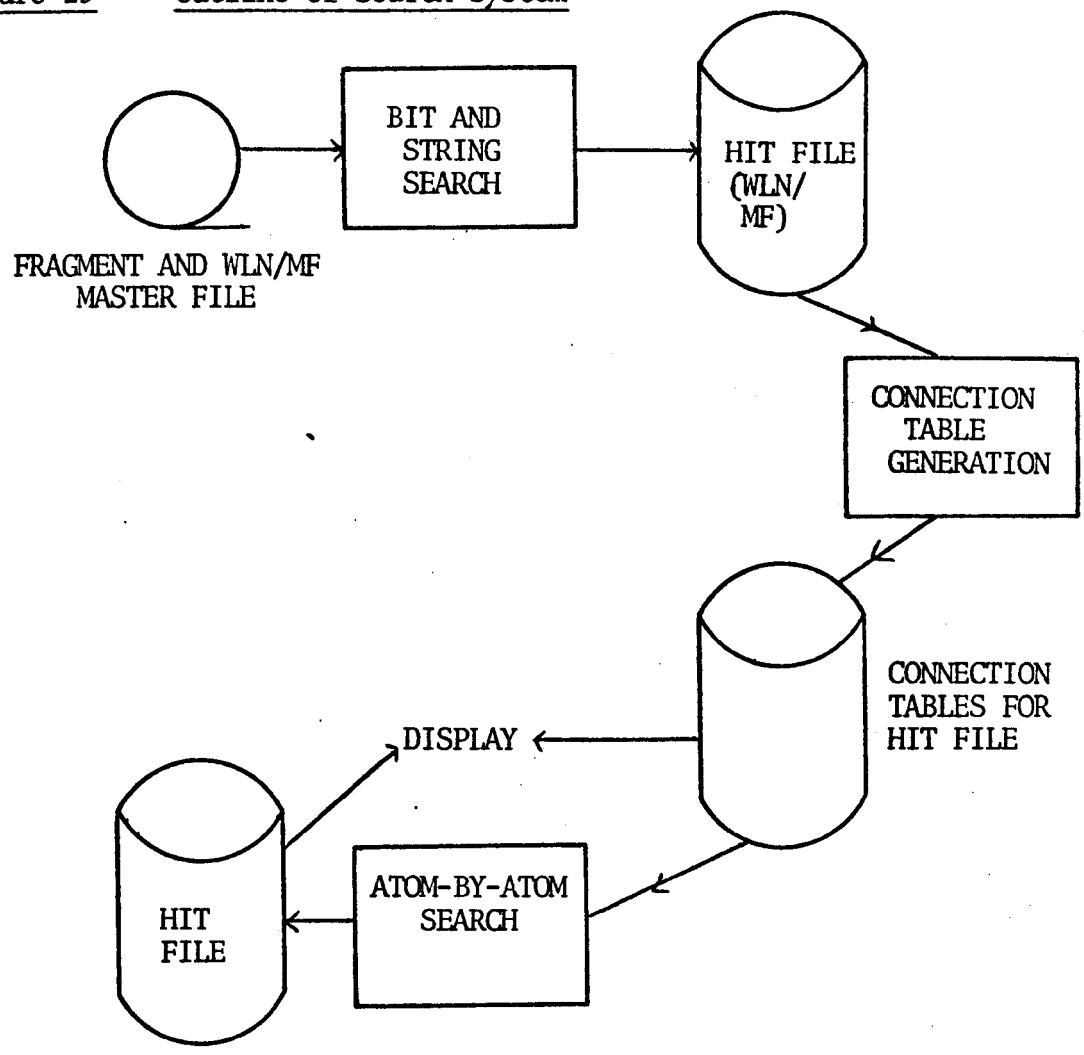
The initial system for Company data on the Burroughs 3500 was to be as near as possible to the multi-level system designed for the Pharmaceuticals Division data on the ICL 1902A. Where possible, programs were directly converted or substitutes found. An outline of the system is given in Figure 29.

The search system worked, but had a number of disadvantages:

(a) A fragment-WLN was generated on magnetic tape for each unique WLN on the 'Chemical Data Base' (as depicted by a unique CR number). Additional data, such as other reference numbers, suffix information, had to be recovered from the 'Chemical Data Base' prior to printing. This was fairly time-consuming and required that the disc-based 'Chemical Data Base' be available when searches were carried out.

(b) The bit and string search routines had been adapted from existing B3500 programs for text searching. They were, therefore, not ideal, questions often being difficult to phrase. Problems were very similar to those found with FIND II in the existing system.

Figure 29 - Outline of Search System



(c) Much editing was often required on output - the atom-by-atom search was not always explicit enough, and the structure display program sometimes poor. This was more noticeable on the Company file since:.

(i) the file was larger, > 120,000 records (as opposed to 85,000 records);

(ii) compounds often had to be hand-drawn, stencils only being available for the M collection;

(iii) the number of searches was increasing;

(iv) update was only carried out every 6 weeks - the fragment file taking 6 hours to be re-generated. A complete update was necessary since there was no way of telling whether CR numbers were being deleted, amended or additional information being added. The search file could, therefore, be up to 6 weeks behind the on-line chemical databases;

(v) only one atom-by-atom search could be carried out at a time - hence, much data manipulation was required to allow a multi-question bit and string search to be processed.

It was realized that improvements to the system would be necessary if we were going to cope with the expected increase in work load. The new chemical search system took 18 months to develop - slight improvements being introduced whenever possible, and advantage being taken of on-line techniques.

3.8 Substructure Search - Batch/On-Line System For The Company

The 'Chemical Data Base' was available on disc and accessed through the on-line system for chemical registration. It was therefore

available for on-line substructure searching. The advantages of on-line access were examined and thought to be:

(a) Immediate parameter validation.

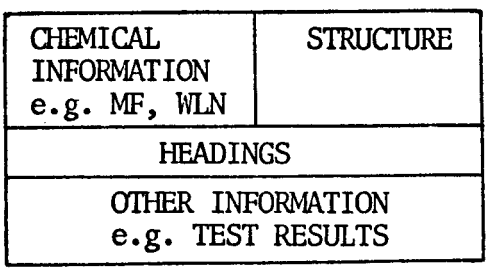
(b) String search formulation - several attempts at search coding could be made - the answer set for each counted and the main run performed on the best formulation. This should exclude the over-specification of searches (and subsequent loss of answers) and the general questions leading to very large answer sets.

(c) Elimination of unnecessary atom-by-atom searches - low answer sets would be printed after the string search phase.

(d) Urgent searches could be handled within an hour - reference numbers being printed on the terminal.

To implement on-line substructure searching it was necessary to add a disc fragment file to the chemical data base and to design the bit and string search to work both in batch and on-line modes. The system was only to be partially on-line and therefore a second requirement of the system was that it should produce a minimum number of batch runs. The whole system was examined and common file layouts identified. These resulted in a common structural card format (see Figure 30).

Figure 30 - Common Structural Card Output From All Searches



Any search, batch or on-line must therefore produce the following standard files:

(a) "A" file - output from any bit and string search program. Details the main reference number, molecular formula, WLN, suffix, additional reference and the search tag. This file is input to the connection table generation program and also used to produce the chemical information on the resulting structure card.

(b) Structure file - output from structure display program. Details reference number, structure and search tag.

(c) "B" file - may come from a variety of sources, e.g. test results from biological search, bibliographic details from ICRS search, further chemical information for company searches. Used solely for printing on the bottom of the structure cards.

To minimise batch runs following on-line search it was also required that "A" and "B" files could be merged prior to batch processing and that a number of atom-by-atom searches could be executed at the same time. All these features were considered in the design of the new system, but major changes were required in the bit and string search program, and in the atom-by-atom search program.

3.8.1 New bit and string search programs

There are three programs containing the same basic search routines:

(a) An on-line program, where the parameters are entered on a VDU. Bit and string searches are carried out immediately for the company data base. Parameters for other data bases can be validated and written to a work file for later execution.

(b) A batch version of the on-line program for use when the on-line system is not available or when large run times were anticipated.

(c) Batch program for searching data bases maintained on tape or disc pack.

The purpose of the programs is to search the various chemical data bases for compounds as defined by fragment code, WLN, molecular formula and/or reference number parameter and to produce the necessary "A" file. A "B" file is also produced when a search has been carried out on the total "Chemical Data Base". Here, the CR number is the main key and there may be a variable number of suffixes and related reference numbers - these are written to the "B" file.

The variety and type of parameters will vary with each request, but any combination of Boolean logic is allowed in a single enquiry. The WLN string search facilities have again been extended over and above what could be expected in a text searching package. Additional facilities include followed by logic, ignore logic, the specification of various characters in one position in a string, the "any character but not space" feature and the ability to limit a string to the start of the search area. The only facility not available in the 1900 system is ignore logic, e.g. the ability to find "AB" but not when it forms part of "ABC". This has been added because of the particular problems with NOT logic. In a chemical molecule it is difficult to imagine what could be present other than the substructures requested - NOT logic should be used only in very specific circumstances. Ignore logic overcomes many of these problems since only the immediate environment of the substructure is being considered. The molecular formula logic has been inserted and it is merely necessary to specify the atom required and the minimum and maximum number of times it should occur if they are known.

There are seven operation types in the on-line program:

(a) Normal run - Searches will be carried out on the input and output files specified. If there is no output files specified, the searches will be merely counted.

(b) Follow-up WLN search - A further WLN or MF search can be on the existing "A" file. This allows for further re-definition of a question or for more than one search to be carried out on one initial broad search.

(c) Parameter validation - Parameters are validated and written to a work file for later processing.

(d) Parameters input - Searches the chemical data base using parameters already input to a work file.

(e) Error report - If an error on the data base is detected during searching, it is written to the "B" file using a given flag. This message can then be re-called from the "B" file using this routine.

(f) Print routine - Accesses the "A" file and displays 80 references on the screen. Either the first 80 may be selected or a given number may be skipped. In this way, results in reference number form can quickly be obtained from a previous search.

(g) Merge routine - Up to five "A" files and five "B" files may be merged to give two master files for later processing. Either the whole five may be merged or individual searches may be selected from each input file.

The output, therefore, from a session on the search system should be two files - one containing "A" data and one containing "B" data for the searches carried out on the Company Data Base. In addition, a number of work files containing search parameters may be set up for later batch processing on other data bases.

3.8.2 New atom-by-atom search facilities

A new atom-by-atom search program was designed which would:

(a) Cater for more than one atom-by-atom search per run.

(b) Allow better specification of atoms in a network using ring or terminal information.

(c) Make better use of screening facilities available in the connection table.

(d) Cover a larger range of networks.

In addition, modifications and corrections were made to the connection table generation program such that the percentage coverage of input WLN's converted was much higher. This eliminated much of the editing usually required at the end of a search since compounds with no connection tables are automatically accepted as hits.

The new programs allow any number of searches to be carried out in the same run. All the requests are read in from cards and transferred to a work file which is sorted into the same order as the input connection table file. The program reads down the connection table file executing the appropriate atom-by-atom searches. The atom-by-atom search itself has been extended to cover from 3 to 10 nodes and for any node the following information may be requested:

(a) Atom type - up to 10 alternative connection table units may be specified.

(b) Position in a ring - nodes may be classified as cyclic, acyclic, either cyclic or acyclic but not in the ring type specified, or cyclic but not in the ring type specified. The ring atoms may be further specified by ring size, ring type, e.g. carbocyclic with one fusion or may be related to a specific ring content. In addition, they may be classed as fusion, non-fusion, spiro or bridge atoms.

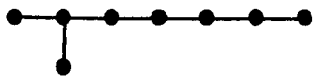
(c) Whether terminal or linking. The allocation of unit symbols in the connection table was based on the premise that a symbol should cover as many atom/bond configurations as were unambiguous. Thus, -O- and -OH are represented by Q. Hence, -OH groups only, may be found in an atom-by-atom search by specifying the Q node as terminal.

All these facilities improve the specificity of an atom-by-atom search question, and hence reduce the number of false drops, i.e. improve precision. The network descriptor links the node atoms together and the following types may be searched for:

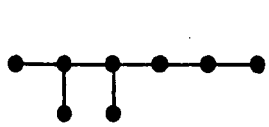
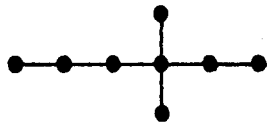
(a) Linear

e.g. 

(b) One-branch

e.g. 

(c) Two-branches

e.g.  or 

In addition, the program has been made more efficient by improving the facilities for "pre-screening". The program, prior to carrying out the detailed atom match, can carry out various screening processes on the connection table. These make use of two features:

(a) Detailing of ring information.

(b) Precise classification of chemical units.

Only possible hits enter the detailed atom-by-atom matching processes.

3.8.3 Structure display

Again, many improvements were made to the coverage of compounds correctly displayed. This was largely done by monitoring the performance of existing programs and suggesting modifications, in the light of occurrence, etc. The structure display function was split into two parts:

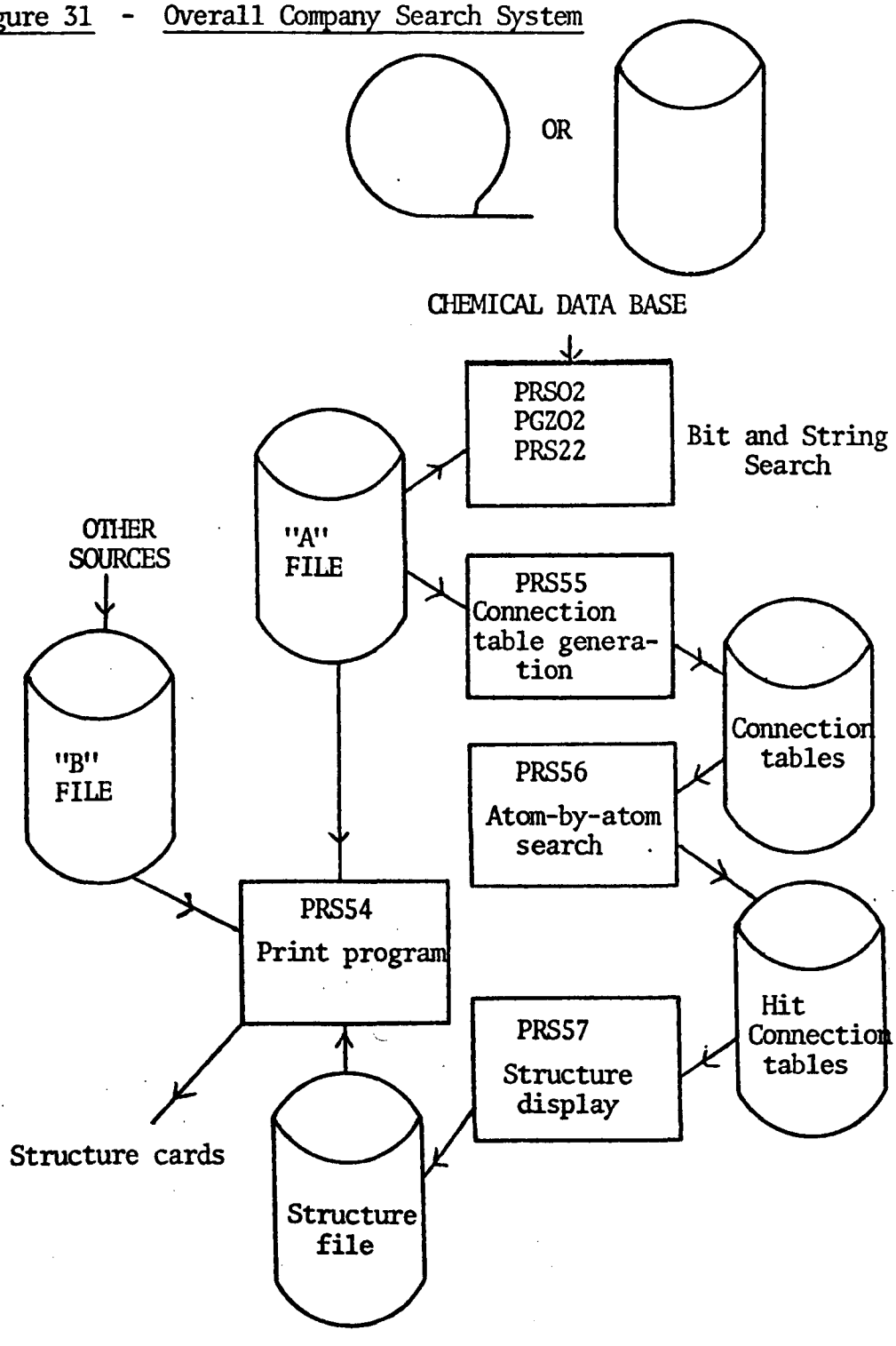
(a) Structure generation from the connection table.

(b) Printing together with associated data. The program sorts each of the three files ("A", "B" and structure) into enquiry number within tag, and prints the cards together with any headings specified on cards.

This separation gave a flexible system, enabling structures from various sources to be printed at the same time. A detailed systems flowchart is given in Figure 31.

The system depends on the maintenance of few programs, is very flexible and fast to run. It is, in fact, a firm building block for the development of future search services around chemical information.

Figure 31 - Overall Company Search System



Chapter 4 - Systems Design and the Response of Information Scientists

4.1 Introduction

Data Services Section operates an information service to members of Research Department. The service is based largely on data generated from within Pharmaceuticals Division. Members of the Chemistry Department are offered three main facilities centring around:

(a) The maintenance of a Company chemical data bank and the registration of new compounds into the collection. The registration service is run by information scientists, although the chemists are expected to have carried out a novelty check prior to registration.

(b) Location of classes of compound by substructure search. These may be carried out on the Company compound files or on files detailing commercially available compounds, and is always performed by information scientists.

(c) Location of samples for specific compounds - either from within the Company or the availability from commercial suppliers. When a chemist cannot locate the compound himself, he will approach the information scientist offering the service.

Each of these services has been influenced to some extent by the advent of computer systems - whether batch or on-line. The computer system has been aimed at helping the services run more efficiently and effectively. This part of the study looks at the response of the information personnel to the computer system, and the ways in which the computer system has been modified to meet their needs.

4.2 Registration Systems

4.2.1 Company registration

Since Pharmaceuticals Division was producing more than 60% of the work of the Company Compound Centre, it was decided to amalgamate the two functions to save duplication of effort. Pharmaceuticals Division took over responsibility for Company-wide compound registration in September 1972. This more or less coincided with the loss of both information scientists from the registration service. Since neither of the available registration systems were suitable to meet the two differing needs, a new system was designed to:

(a) Reduce the amount of work involved in the registration of one compound.

(b) Reduce the chance of duplication.

(c) Cover the registration and search needs of the three main chemical-producing Divisions.

A new "Chemical Data Base" was designed, based on the Wiswesser Line Notation, to allow the three Divisions to register compounds simply and to provide a data base for single compound and substructure searching. (For full details of the design of the chemical data base, see Chapter 2, and for details of the facilities developed around the data base, see Chapter 3.)

4.2.2 Batch registration

For the first twelve months of use, access to the chemical data base was in batch mode only. New compounds were updated each night and specific compound enquiries could be carried out by a batch program on demand. The update routines checked that the WLN/MF record was correct, that the WLN was novel for a new entry, or that a reference from the same Division was not already assigned to a given WLN. All the appropriate flags and pointers were set up such that enquiries can be made through WLN or reference number. Initially there was little confidence in the computer data base. The registration routines were

carried on as before. The nightly update run was used as a final check rather than as part of the registration process. The batch enquiry program was little used. Operationally, it was not possible to guarantee its execution within a given period of time, hence, it was difficult to fit it in with work schedules. To help with this problem, printed indexes of the data base were produced in WLN order and in reference number order for each of the Divisions. These were widely used instead of the batch enquiry run, but they were quickly out of date and did not give a fully comprehensive answer to many questions.

At Pharmaceuticals Division trust still remained in the manual molecular formula indexes. Novelty needed to be established as soon as possible, and this seemed to be the biggest setback against a computer-based system. On-line enquiry was a must. This was not so at Jealott's Hill, where a two-week cycle was used. Novelty checking had to be complete before the fortnightly compounds list was produced. A system was set up around the production of this compounds list. This proved successful and Jealott's Hill relied heavily on the computer enquiry and update procedures. However, it involved our registration unit in a great deal of clerical effort, and from our point of view was wasteful in resources.

4.2.3 On-line enquiry

The ability to enquire on the status of the chemical data base through a VDU terminal became available in September, 1973. Interrogation could take place through molecular formula, WLN, CR number or any of the Divisional reference numbers (for full details, see Chapter 3).

Training sessions were held in August, information scientists being trained separately from information assistants. The immediate response of the information scientist was to treat it as a challenge, and find ways in which it could produce the wrong answer. The training sessions developed into games; this, however, ensured the programs were fully tested. It seemed important that the system was complete and error free before the information scientist was willing to use it. On training, the information assistants were more concerned with the VDU itself and the physical handling of it. They appeared little interested in the processes being carried out and required on-the-job training in the applications.

When the system went live a slight fault was found. On a molecular formula enquiry it was only possible to examine three (one screen) of the structures in the class. Although a slight fault in a now little-used routine, this limited the systems acceptability and usefulness. Novelty checking took place through the on-line system when convenient. However, there seemed to be no desire to use the facility and the molecular formula index was still often used. Input was examined in batches of 12 enquiries using the molecular formula enquiry routines, i.e. staff were duplicating how they would use the manual index. Usually, around five compounds required coding and use of the WLN enquiry i.e. having greater than three compounds per molecular formula. The information scientist considered the terminal novelty check dangerous - the possibility of mistyping could produce the message "MF NOT ON FILE", there was no check on the WLN code, etc. The nightly update was still considered the final validation for novelty.

For jobs other than registration, the use of the terminal by an information scientist became a prestige job, and all possible accesses to the chemical data base were carried out through the terminal. Often these jobs were more suitable for a batch program having large numbers of accesses and not requiring a fast turnaround. To prevent unnecessary time-wasting and to monitor the use of the terminal, users were asked to log in and out. After about six weeks, usage became normal. Applications using the terminal are given in Figure 32.

Figure 32 - Usage of Terminal For On-Line Chemical Enquiries

MONTH	USAGE	NOTES
September		Not monitored.
October	Registration - 3 sessions per day. Other registration or corrections to file - 1 session per day. General enquiries by Registration Unit - 3 enquiries per day.	Each session lasted about 20 mins. Usually about 30 mins. in length. Usually done one enquiry at a time.
November	Registration - 2 sessions per day. Other registration or corrections - 1 session per week. General enquiries - 4 per day. Other enquiries - about 3 a week.	Each session lasting about 15 mins. By people in Registration Unit. By information assistant in charge of sample location.
December	Registration - 2 sessions per day. Other registration or corrections - 1 session per week. General enquiries - 2 per day. Other enquiries - 6 per day.	Sample location done by technical officer.

Information assistants were very slow to begin using the terminal. Much advice and training was necessary, even though the applications were simpler, usually reference number enquiries.

4.2.4 Design amendments

A number of design amendments were made after the first month's operations. Response was not instant - it was found desirable to input the next enquiry whilst awaiting a reply. To do this, two changes were made to the VDU itself:

(a) After a message was received, the VDU normally set itself ready to receive another. To allow time for the user to examine the screen, the VDU was altered so that the user could have the option of either inputting a second enquiry or to receive any further answers.

(b) The VDU normally "buzzed" when the computer wished to send a message or a response. This was silenced allowing the user to call messages when required.

A systems change was also made at this point. Messages "MF NOT ON FILE" or "WLN NOT ON FILE" were not accompanied by which MF or which WLN was input. Hence, uncertainty could arise as to whether the MF or WLN was input correctly, or in fact, which enquiries the message applied

to. This was altered to show the input question as well as the output message. At the same time, the system was corrected so that any number of WLN's could be examined using a molecular formula enquiry. Use was again monitored. There was a great deal of reluctance to use the WLN enquiry facilities. Information scientists seemed prepared to examine large numbers of alternative WLN's in a molecular formula enquiry rather than use the WLN facility. They seemed concerned over the possibility of error.

After the second month of operation, two more systems changes were considered. If there are more than twelve alternative WLN answers on a molecular formula enquiry, the computer should stop following the chains at the thirteenth entry, displaying a message " > 12 MFS ON FILE", and processing the first twelve answers for display. Normally it would find all the alternative answers to an MF enquiry before displaying the first three answers. Hence, the system does the most work for the answer of smallest value. A limit of twelve reduces this overhead and twelve appeared to be the outside limit for scanning. The second change was to include the WLN/MF check facility in the enquiry routine. Here, the information scientist enters both the WLN and the molecular formula. The system first checks the two input notations, and rejects the query if there is any discrepancy. If they agree and are valid, the program performs the WLN enquiry.

These facilities became operational at the beginning of January 1974. The response time was slow since the WLN/MF routines had to be heavily overlaid to become part of the on-line system. Nevertheless the system became widely acceptable to the information scientists. They now had sufficient confidence in the system to pass compounds through for biological testing before the overnight update took place.

4.2.5. On-line registration

Although the Company registration system had been originally designed for the on-line registration of compounds, experience had indicated that it would not be wise to allow each Division to update the files on-line. The WLN/MF check program was not comprehensive enough to check many common fault types and the disc had not been very reliable. The compromise solution was to allow the on-line enquiry program to transfer the WLN/MF/reference number to a work file. This work file would be accessed by the on-line program during the day, and would be used to update the file each evening. At the end of the update, a hard copy printout would be available, and the on-line input could be checked. Only in this way could the necessary checks on accuracy be made.

On-line registration (with batch update) appeared satisfactory to the information scientists. It saved them effort as they have only one input - the WLN/MF/reference number through the VDU. The on-line enquiry facility answers that data they are inputting is novel, and the input each day is sufficiently small to ensure that the same compound is not input twice. The on-line registration program does, in fact, check that any one set of information on any one reference number is input in one day. (Full details of the on-line registration program is given in Chapter 3.)

The system became operational in Pharmaceuticals Division in March, 1974, at Jealott's Hill in April, 1974, and at Organics Division in August, 1974. Pharmaceuticals Division is now only directly responsible for the nightly update and subsequent checks on the accuracy of the data. The response from the various Divisional information officers has been very encouraging. They feel they are getting equal access to the data base without problems of update, maintenance, etc. Very little use of the manual molecular formula indexes is being made by information officers. However, they are still being maintained for use by the chemists. Until the on-line system can be directly accessed by chemists, full value of computer technology will not be obtained.

At Pharmaceuticals Division, we have been able to cope with large increases in work load with an effective loss of two information scientists. Hence the policy to amalgamate Pharmaceuticals Division registration and the Company Compound Centre into one unit with more sophisticated computer facilities seems justified.

4.3 Compound Search Services

Prior to the introduction of the three-level substructure search facilities, there were two main search techniques in use in the Company:

(a) Fragment codes - the manually assigned fragment code could either be searched by computer or using a punch card sorting machine.

(b) Wiswesser line notation - the WLN could be searched by computer, but was mainly searched manually using KWIC indexes.

Information scientists who were to use the new system had experience using either the fragment code or the KWIC index. An analysis was made to see how the information scientists adapted to using more than one technique and how the difference in backgrounds affected the use of the various levels. The analysis was carried out over the first 18 months the multi-level search system was in operation. (The multi-level search system is described in Chapter 3.) During 1973, various experimental on-line search facilities were introduced. The response of the information scientist was monitored.

4.3.1 A batch search system

Users questions were divided into three batches, each holding 100 questions. For a detailed breakdown of the batches, see Figure 33.

Figure 33 - A Breakdown of Substructure Search Questions Into Batches

BATCH 1	Two to six months after the start of the multi-level substructure search service. Only two computer techniques were available: (a) fragmentation using the manually-generated fragment code, and (b) WLN string search. KWIC indexes were available for manual interpretation.
BATCH 2	Eight to twelve months after the start. The two computer techniques, fragmentation and WLN string search, were firmly established, and the third, atom-by-atom search, was just introduced.
BATCH 3	Fourteen to eighteen months after the start. All three techniques, fragmentation, WLN/MF string search and atom-by-atom search, were available with enhanced facilities.

The usage of each technique within each batch was first established independently of information scientists. The results are given in Figure 34.

In 1965 all the substructure search questions in Pharmaceuticals Division were answered using a fragment code and a 335 statistical sorter. Today, computerised techniques based on notations and connection tables are available, what effect has this had on the use of the established methods? The total usage of each technique over the three batches is given in Figure 35.

Figure 34 - Usage of Each Search Technique - Details For Each Batch

BATCH 1

Total no. of questions = 100
No. of times each technique used:

Fragmentation	72
String	40
KWIC	14

} Computer methods 8[illegible]

One technique only		Two techniques	
Fragmentation	46	Fragment/String	26
String	14		
TOTAL	60		

BATCH 2

Total no. of questions = 100
No. of times each technique used:

Fragmentation	75
String	52
Atom-by-atom	17
KWIC	7

} Computer methods 9[illegible]

One technique		Two techniques		Three techniques	
Fragment	25	Frag/String	34	Frag/String/A-by-A	0
String	15	Frag/A-by-A	16		
Atom-by-atom	0	String/A-by-A	3		
TOTAL	40	TOTAL	53		

BATCH 3

Total no. of questions = 100
Total no. of times each technique used:

Fragmentation	92
String	58
Atom-by-atom	42
KWIC	4

} 96 by computer

One technique		Two techniques		Three techniques	
Fragment	9	Frag/String	41	Frag/String/A-by-A	13
String	4	Frag/A-by-A	39		
Atom-by-atom	0	String/A-by-A	0		
TOTAL	13	TOTAL	70		

Figure 35 - Usage of Search Technique - Summarised Over All Batches

TECHNIQUE	BATCH 1	BATCH 2	BATCH 3
Fragment	72	75	93
String	40	52	58
Atom-by-Atom	*	17	42
KWIC	14	7	4

* Not available.

The usage of the manual method has gone down, but the total usage of each of the computer techniques has increased. The introduction of sophisticated techniques makes the questions easier to answer but old established techniques play their part either as screens or to answer those questions within their power. It is also interesting to note that the introduction of a second connectivity technique, atom-by-atom search does not inhibit the usage of string search.

In 1965, all questions were answered by one technique - a fragment code. The amount of editing required before the chemist received the output varied from question to question. Today, three levels are available, but to what extent are the questions being answered by a combination of the levels? The use of multiple techniques is given in Figure 36.

Figure 36 - Use of Multi-Level Techniques

NO. OF TECHNIQUES	BATCH 1	BATCH 2	BATCH 3
One technique	60	40	13
Two techniques	26	53	70
Three techniques	*	0	13

* No result possible.

Use of the single level approach fell with time and the use of two or even three levels became established. Analysing searches carried out by individual information scientists led to one important conclusion. Individuals tended to continue using the technique best known wherever possible. Use of other techniques began for searches which were not easily undertaken by the technique best known to them. Experience showed, however, that the multi-level technique was easier to handle for a larger number of searches, and so the usage increased.

Perhaps the most important result from this part of the study was the fact that experienced WLN coders did not make the best users of the system. They tended to be strongly attracted to WLN string search and use it frequently to the exclusion of atom-by-atom search. They often made cumbersome string searches at the expense of efficiency in computer time and sometimes left out possible strings leading to less than 100% recall.

4.3.2 Effect of on-line systems on search methods

During 1973, various experimental on-line facilities were introduced for bit and string search. The use of the various techniques was monitored, particular attention being paid to the use/non-use of atom-by-atom search. In July the fragment search was available on-line and in August the on-line string search was introduced. At that stage, it was not envisaged that atom-by-atom search would be put on-line because of the large programs involved. Figure 37 shows the use made of the batch atom-by-atom search program whilst on-line and string searches were introduced.

The introduction of on-line fragment search had little effect on the search method, but the use of atom-by-atom search dropped markedly when on-line string search was introduced. Analysis of the questions themselves showed that the string searches were becoming more complex. Two factors could be influencing the low use of atom-by-atom search:

(a) Knowing the number of answers after string search and discounting the use of atom-by-atom because of the small numbers involved.

(b) The desire to complete the job through an on-line system rather than waiting overnight not knowing whether a mistake had been made.

Figure 37 - Results For Use of Atom-by-Atom Search in Partially On-Line Systems

MONTH	NO. QUESTIONS		% OF TOTAL QUESTIONS ANSWERED BY A-BY-A
April	6	Atom-by-atom	50%
	6	Non atom-by-atom	
May	8	Atom-by-atom	30%
	19	Non atom-by-atom	
June	13	Atom-by-atom	40%
	18	Non atom-by-atom	
ON-LINE FRAGMENT SEARCH INTRODUCED			
July	15	Atom-by-atom	45%
	20	Non atom-by-atom	
ON-LINE STRING SEARCH INTRODUCED			
August	4	Atom-by-atom	20%
	16	Non atom-by-atom	
September	8	Atom-by-atom	18%
	36	Non atom-by-atom	
October (1st half)	9	Atom-by-atom	22%
	29	Non atom-by-atom	
ON-LINE PROCEDURES ABANDONED ON 18 OCTOBER			
October (2nd half)	7	Atom-by-atom	18%
	30	Non atom-by-atom	
November	32	Atom-by-atom	35%
	62	Non atom-by-atom	
December	18	Atom-by-atom	35%
	36	Non atom-by-atom	

To test these two points, the on-line procedures were abandoned in mid-October, and the batch system was modified to enable the information scientists to know the number of hits after string search. The use of atom-by-atom rose slowly and finally settled just below the figures for the start of the year. From this it was concluded that about 35% of all questions should be carried out by atom-by-atom search in an on-line environment. To prevent recurrence of the previous use of string search, it was decided to include atom-by-atom search in the on-line system. The search would then in effect be completed by the information scientist and accepted by the computer system before the information scientist left the terminal. This is a longer-term development since the connection table generation program and the atom-by-atom search program have to be re-written.

4.4 Location of Samples

The sample request service was run by an information assistant used to manual systems. The on-line chemical enquiry system should have provided her with a fast method for finding out whether a compound with a given reference number had a reference number in another Division. This is an important part of her job, although the information was already

recorded in the manual indexes. The information assistant was very sceptical about using the terminal although she was located in the same office. It was rarely used (12 enquiries per month) and only when the information could not be found elsewhere. Because it was always used as a last resort, it gave little positive feedback to its usefulness. Owing to holidays, the information assistant was replaced by an information scientist for one month. The use of the terminal was found to increase to 120 enquiries per month, even though the work load should have been lessened because of the holiday period. About 50% of these enquiries were based on WLN and outside the scope of the information assistant, but the information scientist tended to make better use of the terminal in all applications.

4.5 Conclusions

Both the specific compound enquiry procedures and the substructure search procedures have been substantially modified in response to the use by information scientists. Factors of importance were:

(a) People would only accept a new system if it offered considerable advantages over current practices. It was not possible merely to introduce facilities which were better from the systems considerations if they did not bring some advantages to the information scientists at the same time. In some ways it is easier to introduce a computerised system when the alternative methods of working no longer exist. Constant training was necessary to show people the advantages of the system in day-to-day working.

(b) It is easy to misuse on-line systems. The introduction of on-line facilities needs careful monitoring, particularly if users have no knowledge of the consequence of their actions inside the computer. The need to introduce some systems security to prevent people using unnecessary computer time was soon realised.

(c) Information scientists tended to use the terminal more readily than information assistants, regarding it as a prestige piece of equipment. Once established, however, an information assistant used the terminal effectively.

Chapter 5 - Use of the Search Facilities by Chemists

5.1 Introduction

Data Services Section had for some years been offering chemical search services to Research Department, and hence, at the beginning of this study had built up a pattern of interaction with individuals. This part of the study monitored the use of the various services over the two-year period when new computer facilities were being developed and changes in methodology were taking place.

There were three main categories of services monitored:

(a) Location of samples for specific compounds - either from within the Company or from commercial suppliers. Chemists may request information on the availability of samples after reading the literature, by planning possible synthetic routes, after analysing the results of a substructure search, etc.

(b) The maintenance of the Company chemical data bank and the registration of new compounds into this collection.

(c) The execution of substructure searches for classes of compounds, either in the Company files or from commercially available sources.

In the first two services - sample location and compound registration - one might not expect any dramatic change in use since the computer systems only affected the way in which the job was carried out not the end product. Studies like those of Rosenberg (15) and Allen (16) pointed to the fact that within their user population, the quality of the service had no bearing on the usage, the over-riding factors were accessibility and ease of use.

However, the way in which the substructure search service had developed had changed the whole concept of the service and hopefully this would be reflected in the growth in use. Another factor being considered here was the overlap between the various services. Traditionally they had been managed as separate entities with little information flow between the various information scientists. The relationship between the various services needed to be established if chemists were to get the best possible service with the limited labour resources.

5.2 Sample Information

5.2.1 Introduction

Requests for samples were handled by an information assistant who received requests by personal contact, telephone or by post. She handled about 7,000 samples each year, locating actual samples of compounds available within the Company or finding the best supplier, price, etc. of compounds available from commercial suppliers. The latter information enabled chemists to order the necessary chemicals through the Research Stores.

The information assistant recorded all requests in a log book. This detailed the user (the chemist himself or one of his assistants) the type and amount of sample requested, and the information supplied, i.e. actual sample or supplier information. About 20 requests were logged on one page, and the date was stamped at the top of each page.

5.2.2 Methodology

Information was abstracted for the years 1972 and 1973. As all information on other services was recorded against chemists, "staff lists" were used to relate experimental officers to the appropriate chemists.

The information recorded was: date (approximate), chemist, section, number of compounds on which information was requested/visit, and the number of samples obtained/visit.

The data was transferred to punch cards, one visit per card. The cards were then sorted and printed on the B4700 using a standard reporting package. Two lists were obtained for each year:

(a) Section order - further subdivided by chemist and then by date.

(b) Month order - further subdivided by section and then by chemist.

This raw data was examined and a secondary record set up showing: year, chemist number, section, total number of visits, total number of compounds requested and the total number of samples obtained.

Prints were prepared: in chemist order, in descending order of number of requests, in descending order of number of compounds requested and in descending order of number of results obtained.

From the primary data prints, it was possible to establish the total usage by section and the usage by month. The results are shown in Figure 38 and Figure 39.

Figure 38 - Use of Sample Location Service by Section

SECTION	1972				1973			
	NO.	VISITS	REQUESTS	SAMPLES	NO.	VISITS	REQUESTS	SAMPLES
1	6	112	375	133	7	104	411	216
2	7	190	454	189	8	163	400	225
3	7	208	812	393	11	262	943	490
4	9	111	265	159	10	142	304	131
5	8	231	813	447	10	210	645	322
6	4	42	97	36	6	46	104	46
7	6	74	279	101	7	59	185	57
8	8	104	190	90	12	237	860	378
10	10	166	383	198	10	102	226	87

Figure 39 - Use of Sample Service by Month

MONTH	1972			1973		
	VISITS	REQUESTS	SAMPLES	VISITS	REQUESTS	SAMPLES
Jan.	102	294	120	142	346	128
Feb.	123	353	151	121	363	152
Mar.	144	424	207	116	361	138
Apr.	120	292	187	110	218	92
May	109	260	111	111	277	135
June	111	242	120	124	443	210
July	88	225	93	87	323	200
Aug.	108	298	150	131	298	158
Sep.	102	382	166	80	162	113
Oct.	106	222	160	136	524	197
Nov.	78	211	91	150	477	227
Dec.	96	288	97	77	310	183

From the print of requests by chemists, it was possible to deduce the following facts:

(a) Sixty-three chemists used the service in 1972 and 61 chemists in 1973. The results divided by section are given in Figure 40.

(b) Of the ten chemists using the service in 1972 and not in 1973, five had moved to other departments. Of the eight chemists using the system in 1973 and not in 1972, six were either newly promoted or new entries.

(c) Of the chemists using the system in 1972 and 1973:

15 increased their number of requests
21 decreased their number of requests
17 made about the same number of requests
21 increased the number of compounds requested
21 decreased the number of compounds requested
11 requested about the same number of compounds
18 chemists had more compounds located
20 chemists had less compounds located
15 chemists had about the same number of compounds located

Figure 40 - Use of Sample Location Service by Section (Summary)

SECTION	1972	1973
1	6	7
2	7	8
3	7	11
4	9	10
5	8	10
6	4	6
7	6	7
8	8	12
10	10	10

5.2.3 Discussion

Sample requests seemed to come in steadily. The usage of the service by month seemed very stable - about 100 requests each month. The number of transactions carried out in 1972 and 1973 were about the same. However, the number of compounds requested increased by 16% and the number of compounds found increased by 20%. Usage in most sections was fairly stable. It did, however, drop considerably in four sections (2, 5, 7 and 10). There was one large increase in visits by section 8 and this was accompanied by a very large increase in the number of compounds obtained.

The number of individual chemists using the service remained fairly constant. About 30% increased their usage of the service, 30% decreased their use and the remaining 30% asking about the same number of questions. 45% of the chemists used the service rarely and only 15% of the chemists were very high users. Of the high users, five were from Section 3 and three were from Section 4. At least one chemist in each section was a poor user.

5.2.4 Conclusions

The use of the sample location service seemed very stable over the two years. It was a manual system and had been run by the same information assistant for about ten years. There have been few changes in the way the service was run, and hence any changes in usage must be caused by external factors.

The sample location service could be considered a control and could be used to measure other fluctuations in the system or user population.

5.3 Sample Location and Substructure Search

5.3.1 Introduction

There was no way of analysing from the logbook records of sample requests which compounds were being requested as a result of a substructure search being previously carried out. To do this, the information assistant was asked to keep a special record of those requests known to have come from a substructure search - usually the chemist would present her with the structure cards. This data was being recorded during the period December 1973 and January 1974.

5.3.2 Methodology

The information assistant recorded the chemist, the search identifier allocated when the substructure search was carried out (if known), the date, the number of compounds for which information was requested. There were thirty-two such requests in the two-month period. These do not necessarily represent all the sample requests following up from substructure searches, only those directly related.

The sample location record was supplemented by data from the substructure search log - showing the date the search was carried out and the number of compounds originally obtained. The type of information requested has been summarised to show:

(a) Samples from Company collections required for chemistry.

(b) Samples from Company collections required for biological testing.

(c) Availability of compounds.

The results are given in Figure 41.

5.3.3 Discussion

At least 20% of all sample requests appear to be generated from substructure searches, the majority of these being searches on the Company files. About 30% of the requests were to enter compounds for biological testing. This indicates the use of the substructure search services to find compounds required for test. A further 50% were requests for samples for use in chemical reactions, and the remaining 20% were for sources of the compound for future reference, e.g. availability, price.

If a chemist required compounds for testing, on average he requested the samples two to four days after the substructure search had been carried out. The decision to test the compound involves the chemist in little work; he therefore responded quickly with large numbers of compounds.

If a chemist required compounds for synthesis, the samples were often requested months after the search had been carried out. They seem to be requested as and when the preparative effort was available.

5.3.4 Conclusions

Sample requests as a result of a substructure search on the Company files accounted for at least 20% of the service. Compounds for test were usually requested soon after a search had been carried out, whilst requests for compounds for chemistry were usually processed in small numbers. They could be requested months after the original search was carried out. This last point reinforced the systems decision to date the structure cards at the time of the substructure searching. This was particularly important when biological and sample information was printed on the cards. A second feature was noted. If chemists were still working with information from searches carried out months ago, a system of substructure search update on a profile run on the Company files was necessary, i.e. chemists could be kept informed of all compounds new to the collection in their field of interest.

Figure 41 - Sample Requests After Substructure Searches

DATE OF SAMPLE REQUEST	CHEMIST	SECTION	DATE OF SEARCH	NUMBER RETRIEVED	NUMBER REQUESTS	TYPE OF INFORMATION
04.12.73	83	8	26.10.73	251	30	A/B
05.12.73	85	8	13.11.73	21	3	B
05.12.73	34	3	29.11.73	58	6	A
06.12.73	11	1	04.12.73	2	2	B
06.12.73	34	3	29.11.73	124	4	A
06.12.73	34	3	26.11.73	168	2	B
07.12.73	85	8	05.12.73	32	14	B/C
10.12.73	84	8	17.08.73	35	11	B/C
10.12.73	85	8	05.12.73	73	15	B/C
10.12.73	85	8	06.12.73	12	12	B/C
10.12.73	84	8	04.12.73	68	23	B/C
12.12.73	38	3	30.11.73	18	4	B/C
14.12.73	30	3	04.12.73	9	9	A
17.12.73	85	8	06.12.73	12	4	B
21.12.73	43	4	26.06.73	22	1	C
08.01.74	34	3	24.07.73	61	1	B
10.01.74	26	2	08.01.74	18	2	A
10.01.74	26	2	08.01.74	20	14	A
11.01.74	30	3	02.01.74	101	10	B
17.01.74	69	6	05.01.74	28	5	A
22.01.74	22	2	16.10.73	21	15	B
23.01.74	54	5	17.10.73	15	15	C
24.01.74	85	8	05.12.73	32	4	B
25.01.74	85	8	16.01.74	18	2	B
26.01.74	44	4	03.05.73	558	6	B
28.01.74	54	5	21.01.74	32	33	C
29.01.74	22	2	23.01.74	27	3	A
05.02.74	85	8	01.02.74	18	10	A
06.02.74	71	7	19.11.73	33	12	B

Notes on type of information:

A = Sample required for test.
B = Sample required for chemistry.
C = Information on availability required.

5.4 Compound Preparation

5.4.1 Introduction

Compounds prepared by chemists must all be registered and allocated an M number before they can be submitted for biological test. The M number was allocated when the compound was registered onto the chemical data base, but the chemical data base did not record information on the chemist who prepared the compound. This information was only recorded in two manual indexes:

(a) The card index maintained in M number order.

(b) The manually produced compounds list. This was produced fortnightly by Data Services Section. It recorded the structure of each new compound made, along with the initials of the chemist and any biological tests for which the compounds had been submitted.

5.4.2 Methodology

The compounds lists were scanned for 1972 and 1973, and the following information was recorded: date, chemists initials, section, M reference number and main biological tests. This data was then transferred to punch cards and sorted and printed on the B4700 using the standard reporting package.

The raw material was first sorted and printed by year into:

(a) Section order (further sub-divided by chemist and then by date).

(b) Date order (further sub-divided by section and then by chemist).

From these initial lists a second set of data was prepared detailing: year, date, chemist, total number of compounds prepared, section and main biological test areas.

From the data prints, it was possible to establish the total usage by section and by month. The results are shown in Figure 42 and Figure 43.

Figure 42 - Breakdown of Compounds Prepared By Section

SECTION	1972		1973	
	NO. IN SECTION	COMPOUNDS MADE	NO. IN SECTION	COMPOUNDS MADE
1	8	263	8	397
2	8	540	8	582
3	8	640	10	1118
4	8	358	8	666
5	8	605	10	627
6	10	110	10	191
7	8	175	7	370
8	10	536	11	696
10	8	390	8	429
TOTALS	72	3617	80	5076
Average/chemist		50 cpds/yr.		63 cpds/yr.

Figure 43 - Breakdown of Compounds Prepared by Month

MONTH	1972	1973
January	298	367
February	310	434
March	355	458
April	314	366
May	348	457
June	274	365
July	312	458
August	322	409
September	359	355
October	382	461
November	406	467
December	330	461
TOTALS	3628	5058
Average/month	302	421

Note: the low value for September 1973 was caused by a major laboratory removal.

From the results it was possible to deduce the following facts:

(a) There was a 39% increase in the number of compounds prepared in 1973, with only a 5% increase in staff.

(b) On average, a chemist's output rose from fifty compounds a year in 1972 to sixty-three compounds a year in 1973. This is not a good estimate since: three section leaders prepared no compounds; seven members of section six have a very low output since they work in the polypeptide area.

Allowing for these facts by making the polypeptide team equivalent to one synthetic chemist, and hence subtracting nine from the total in each year, the chemist's output rose from fifty-seven compounds to seventy-one compounds, i.e. a 25% increase in output.

(c) The output of compounds/month rose slowly from the beginning of 1972 to the end of 1973, allowing for fluctuations due to holidays. The average output for the first quarter in 1972 was 321 and the average output for the last quarter in 1973 was 463 - an increase of 44%.

5.4.3 Discussion

The preparation of compounds increased steadily from the beginning of 1972 until the end of 1973. The average productivity per chemist increasing from fifty-seven compounds in 1972 to seventy-one compounds in 1973. The largest increases were in Sections 7, 4, 3 and 6. Productivity in Sections 2, 5 and 10 showed only a slight increase.

5.4.4 Conclusions

A chemist's ability to prepare novel compounds appeared to have increased gradually over the two-year period - some areas being more affected than others.

5.5 Substructure Searching

5.5.1 Introduction

The information scientist maintained a logbook of all substructure searches carried out indicating the user, the search request and the type and size of output he received. This logbook was analysed to find the volume and type of use of the service.

On 1 October, 1972 saw the amalgamation of the Company Compound Centre with Data Services Section. From that date onwards, the Section was responsible for searches on Company data for all Company uses.

5.5.2 Methodology

The "Search Book" was scanned for 1972 and 1973 and the following information was recorded: date, chemist, section, search tag (to identify the search type), number of compounds output. This data was then transferred to punch cards, then sorted and printed on the B4700 using the standard report package.

The raw material was first sorted and printed by year into:

(a) Section order (further subdivided by chemist and then by date).

(b) Month order (further subdivided by section and then by chemist).

From these initial lists a second set of data was prepared detailing: year, chemist, section, total number of searches requested and date. From these prints it was possible to establish the total usage by Section and by month. The results are shown in Figure 44 and Figure 45.

Figure 44 - Breakdown of Substructure Search Use by Section

SECTION	1972		1973	
	NO. OF USERS	NO. OF SEARCHES	NO. OF USERS	NO. OF SEARCHES
1	5	41	5	16
2	6	16	6*	24
3	6*	74	10*	138
4	7*	27	7*	20
5	8*	79	6	44
6	2	10	4*	14
7	4	9	6	18
8	7*	17	9	101
10	7*	33	4*	10
TOTALS	52	306	57	385
Average use		5.88		6.75

*including Section Leader

Table 45 - Breakdown of Substructure Search Use by Month

MONTH	1972			1973		
	TOTAL	PHARMS. DIV.	CHEM. DEPT.	TOTAL	PHARMS. DIV.	CHEM. DEPT.
January	36	36	30	44	31	30
February	28	28	27	55	37	35
March	55	55	46	66	32	29
April	19	19	18	39	12	11
May	38	38	36	56	35	30
June	32	32	29	74	31	31
July	42	42	38	51	35	29
August	34	34	28	57	20	19
September	26	26	25	64	44	28
October	49	27	23	103	76	60
November	66	28	25	96	70	58
December	43	25	24	58	44	29
TOTALS	470	344	304	763	467	389
Average/ month	39	29	25	64	39	32

The usage by Departments other than Chemistry Department is given in Figure 46.

Figure 46 - Use of Substructure Search Service by Departments Other Than Chemistry Department

Other Internal Departments:

DEPARTMENT	1972		1973	
	NO. OF SEARCHES	NO. OF USERS	NO. OF SEARCHES	NO. OF USERS
Biology	8	3	1	1
Biochemistry	8	5		
Data Services	19	5	19	5
Physical Methods			54	1
Process Dev.	2	2		
Medical Services	1	1		
Safety of Meds.			1	1

External Users:

ORGANISATION	1972	1973
	NO. OF SEARCHES	NO. OF SEARCHES
Plant Protection	72	220
Organics Division	16	60
ICI United States	6	18
Corporate Labs.	5	4
Plastics Division	1	
ICI Brixham		1
Bexford Ltd.	1	

From these results it was possible to deduce the following facts:

(a) Although chemical project teams were reorganised at the start of each year, this seemed to have little effect on the number of searches requested.

(b) Use by Section Leader did not affect use by other members of the Section. Five Section Leaders used the service each year.

(c) In sections with the highest usage each year (Section 5 in 1972 and Section 3 in 1973), all members used the service. Both areas of research had produced a large number of Company drugs, and much of the research was centred around the improvement of available drugs.

(d) Both Section 7 and Section 6 had a low usage and were used by few members of the Sections. Both areas involved work on compounds not adequately covered in the system.

(e) The number of searches carried out by the speculative chemistry project (Section 4) was lower than expected. This could have been a reflection of their interest in synthesis rather than sample availability.

(f) The large usage by chemists at Plant Protection Division could be a reflection of the type of work carried out - mainly general chemistry.

(g) There were significant additions to Research Department and Chemistry Department on 1 October, 1973. Six chemists were either recruited or transferred. Of the sixty searches in that month, twenty-seven were from the new chemists. However, of the fifty-eight searches in November, only nine were from the new chemists.

(h) Of all the chemists who joined Chemistry Department in 1973, eight were using the system and six were not.

5.5.3 Comparisons of usage in 1972 and 1973

Of the chemists not using the system in 1972, seven were using the system in 1973 and twelve were not. Of the chemists using the system in 1972, eight were not using the system in 1973. Of the chemists joining Research Department in 1973, eight were using the system and six were not. Of the chemists changing section, five were making some use of the system, seven were making more use of the system, and four were making less use. Of all chemists in Research Department, twenty-two were making the same use, nineteen were making more use and fifteen were making less use.

The total searches for Chemistry Department increased from 304 in 1972 to 385 in 1973. Three sections requested less searches, three requested about the same, and three sections requested more. An analysis of all sections is given in Figure 47.

5.5.4 Discussion

The use of the substructure search service by members of Chemistry Department increased by 28% in 1973 with three sections having large increases. More chemists used the service (71% of the total population in 1973, as against 68% in 1972), each chemist asking more questions. The increase was gradual throughout the two years - the average usage for the first quarter of 1972 was thirty-four, and for the last quarter of 1972 was forty-nine, i.e. an increase of 44%.

There was a marked alteration in the service around the chemists laboratory move in September. The figures for August and September are low, and those for October and November very high.

5.5.5 Conclusions

The use of the substructure search service was increasing, more questions being asked by more chemists.

Figure 47 - Analysis of Substructure Searches by Section (Summary)

SECTION	YEAR	NO. OF SEARCHES	NOTES
1	72	41	Large decrease.
	73	16	
2	72	17	Slight increase.
	73	24	
3	72	74	Large increase.
	73	138	
4	72	27	Slight decrease.
	73	20	
5	72	80	Large decrease.
	73	44	
6	72	15	
	73	14	
7	72	9	Large increase.
	73	18	
8	72	17	Very large increase.
	73	101	
10	72	32	Section leader changed.
	73	10	Large decrease.

5.6 Analysis of All Services For 1972 and 1973

5.6.1 Use of services

There had been a significant increase in demand for all three services performed by Data Services Section:

(a) Sample Request Services - The number of visits made by chemists remained about the same although the number of samples requested increased by 16% and the number of samples obtained increased by 20%.

(b) Compound Registry Services - A 25% increase in compound preparation was achieved with a 5% increase in the staff within Chemistry Department. Hence, there was a 34% increase in the compounds registered into the collection, the remaining increase being compounds requested from the outside sources, mainly Jealott's Hill and Organics Division.

(c) Substructure Search Services - The increase in use of the substructure search service was 28%.

5.6.2 Maintaining the services

There has been only a small increase in staff to deal with these large increases in use. The problem has largely been solved by the provision of sophisticated computer systems. These are used by the Department, and although members of Chemistry Department have been encouraged to use them, there has been little response.

The on-line facilities enable fast responses to be made on up-to-date data. In addition, the structure cards from the substructure search service enable more efficient follow-up requests, e.g. structure and all reference numbers are available for sample location, biological results available for test submission, etc.

5.6.3 Overall conclusions

There have been large increases in the demands made to Data Services Section over the two-year period 1972-1973. These had effectively been met by the introduction of both on-line and batch computer systems used by the information scientists within the Section.

Chapter 6 - Analysis of User Study

6.1 Introduction

Part of the purpose of the study was to look at the relationships, if any, between usage of the various information services provided by Data Services Section, and the performance and background of each chemist. It was therefore necessary to define distinguishing features in the performance of a chemist. A chemist's role in Research Department was to design and synthesise potential drugs. Performance could therefore be crudely measured in terms of the numbers of active compounds produced. This, however, was too difficult to measure so that comparisons could be made between the individual chemists. Activity in a test could vary from 0% to 90%, depending on the type of test and the ratio selected by the biologist. A better measure might therefore have been the difference in the success of a chemist in finding activity in a particular test for which he was designing drugs against the average activity found in all compounds submitted to that test. However, this measure would only be appropriate in those tests which obtained compounds from a wide variety of sources. Since there was no well-defined way to characterise chemists' performance, it was decided to record a large number of contributing variables which together might give some measurement of performance and which could be meaningfully compared between chemists.

The sample population was eighty, and a total of thirty variables were measured for each chemist. The main part of the survey was carried out on data collected for the year 1973. A number of measures were also recorded for 1972, and the difference in values between the two years calculated. This gave a number of use and performance differences for the two years.

Some of the variables used have common sub-variables. For example, the success rate in sample location is the ratio of the number of compounds requested to the number of compounds obtained. The variables therefore contain both relative and absolute measures. The correlations observed between relative factors have an obvious source.

The survey was in three parts:

(a) Performance appraisal - examination of the various parameters measured and the relationships between them.

(b) Relationships between the use made of the services and chemists' performance.

(c) Relationships between use made of the service and chemists' background.

6.2 Use, Performance and Background Measures

The following variables were measured for each chemist for the year 1973:

(a) Use of the substructure search system: measured in terms of the number of substructure searches requested.

(b) Use of the sample location service: subdivided into five measures:

1. Number of visits made to the sample location service.
2. Number of samples requested.
3. Number of samples obtained.
4. Success rate in sample location, i.e. $\frac{\text{no. samples obtained}}{\text{no. samples requested}} \times 100\%$
5. Number of compounds requested per visit, i.e. $\frac{\text{no. samples requested}}{\text{no. of visits made}}$

(c) Productivity: subdivided into eight measures:

1. Number of compounds prepared.
2. Proportion of compounds made by novel reactions,

   i.e. $\frac{\text{no. of compounds made by novel reactions}}{\text{total no. of compounds made}} \times 100\%$

3. The proportion of compounds tested but not made,

   i.e. $\frac{\text{total no. compounds tested - no. compounds made}}{\text{total no. of compounds tested}}$

4. The number of biological tests requested, on average for each compound,

   i.e. $\frac{\text{total no. of results}}{\text{total no. compounds tested}}$

5. The success in obtaining biological activity in compounds made by him,

   i.e. $\frac{\text{no. of active compounds made}}{\text{total no. of compounds made}} \times 100\%$

6. The success in finding biological activity in compounds obtained for test,

   i.e. $\frac{\text{no. of active compounds acquired}}{\text{total no. of compounds acquired}} \times 100\%$

7. Success in finding biological activity in any compounds,

   i.e. $\frac{\text{number of active compounds}}{\text{number of tested compounds}} \times 100\%$

8. The proportion of compounds tested only in the chemists specialised area,

   i.e. $\frac{\text{no. of compounds tested in given test}}{\text{total compounds made}} \times 100\%$

In addition, the following variables were measured for 1972 and compared with those obtained for 1973 to give a relative increase or decrease:

(a) Difference in use of substructure search service.

(b) Difference in numbers of visits to the sample location service.

(c) Difference in numbers of compounds requested from sample location service.

(d) Difference in numbers of samples obtained from sample location service.

(e) Difference in numbers of compounds made.

In addition, the normal biological activity ratios were established for each test during 1973. These normal values were then compared with the chemist's individual performance in each test, to give differences from the average rate of success in obtaining biological activity. These measurements were taken for the compounds made by him, the compounds acquired by him, and lastly for all the compounds tested by him.

The chemist's administrative unit - the section - was taken

as a guide to his therapeutic area. Each section contained one or two main biological areas and these were usually related in some way.

Four indications of background were recorded:

(a) Status - section leader, senior scientist, Ph.D. chemist or promoted chemist.

(b) Length of time on present project - measured by length of time in same section.

(c) Length of time in Research Department indicating the age and experience of the chemist.

(d) Role of chemist in biological testing. A lead chemist was responsible for ensuring compounds of the right type were tested in a given testing programme.

In addition, a number of other variables were measured as they were thought possible influencing pointers:

(a) The number of publications in external journals - a measure of novel work areas, mainly in chemical synthesis, and his outside esteem.

(b) Assistance in experimental work - the number of experimental officers and laboratory assistants allocated.

(c) Use of computerised literature service for chemical information. Access was only available on the ICRS system, and use of this service was recorded.

6.3 Methodology

6.3.1 Data collection

The information on each chemist was gathered from various sources and recorded against the chemist's name and number on index cards. This information was collected from four main sources.

Firstly background information - section, status, time in section and time in company - was collected from staff lists. Lead chemists were identified from testing programme sheets maintained by Data Services Section.

The 1973 figures for use of the information facilities were transferred to the index cards from the records generated in Chapter 5. These figures included the number of substructure searches requested, the number of visits made to the sample location service, the number of compounds requested by the visit, and the number of compounds obtained from these requests. From these results it was possible to calculate the remaining use factors - the success rate in sample location and the number of compounds requested per visit. Similarly, the 1972 figures were taken from the results collected in Chapter 5. These were then compared with the 1973 figures and the relative use measures calculated as percentage increases or decreases.

The number of compounds made by each chemist in 1972 and 1973 were also taken from the data recorded in Chapter 5. The 1973 figure was used as found, and the 1972 figure was used to estimate the percentage difference in productivity in 1973 over 1972.

The use of novel chemistry was estimated from the chemists' entries on the compound registration forms. Part of the information given was whether the compound had been made by a standard method or by a novel reaction. All novel reactions were noted and a count made for each chemist.

The activity pattern for each chemist and each test had next to be established. This was done by analysing data contained in existing computer files. Using a standard software package, all the data pertinent to chemists' testing patterns in 1973 was abstracted and written to a magnetic tape file. Printouts of all data relating to

each chemist were obtained. The data was first searched to find all references to individual chemist initials, and then a utility program was used to print all tests with their results against each compound submitted. Each list was manually examined to find the following data for each chemist: tests to which the chemist was most commonly submitting compounds, i.e. his testing programme, the average number of tests per compound, compounds which were made by him and which compounds had been acquired from other sources. From these results it was possible to establish the percentage of compounds tested by him that were made by him, the activity patterns for the various types of compound and the percentage of compounds made and tested in his main testing programme The next stage was to establish the normal activity patterns for each test. Counts of the compounds active in each test and of the compounds tested in each test were made and the ratio active/tested calculated for each test. Each chemist's activity pattern was then compared with average values and the percentage variations recorded.

The retrospective ICRS file was searched for journal articles submitted by ICI Pharmaceuticals Division in 1973. The results were analysed into the number of publications per chemist. The number of assistants for each chemist were estimated from staff lists. Finally, the number of ICRS searches for each chemist was obtained from information scientist's logbooks.

6.3.2 Input to the computer

The data was abstracted from the index cards and transferred onto punched cards. The format chosen was that suitable for input to the ICL Statistical Analysis Package (81). The data input is shown in Appendix V, with each chemist identified by a two-digit number.

6.4 Statistical Analysis

6.4.1 Introduction

The user survey had collected together data relating to thirty variables on a user population of eighty. The first stage of the analysis was to measure the extent of the relationship between one variable and any one other, i.e. the correlation between two factors. This was done by the generation of correlation coefficients between each pair of variables. The second stage of the analysis was an attempt to group the variables measured into more significant larger groupings. To do this, factor analysis techniques were applied.

6.4.2 Statistical analysis parameters

The parameters used are given in Appendix VI and the results of the computer run are given in Appendix VII.

6.4.3 Interpreting the results from the statistical analysis

The correlation matrix is used to measure direct relationships between any two factors amongst a larger set of possibly inter-related factors. To do this, the ICL Statistical Analysis Package uses the Students T function (81). The mathematically derived formula is:

$$t = \sqrt{\frac{(n-2)R^2}{(1-R^2)}}$$

where t = Students t function
R = correlation coefficient
and n = number of degrees of freedom, i.e. the number of observations made.

Looking up Students t tables for the thirty observations used and to achieve a significance of greater than 99%, R, the correlation coefficient must be greater than 0.29. Hence, from the correlation matrix generated by the computer run, a value of greater

than ±0.29 indicated that the two factors involved were related directly with a confidence of at least 99%. Scanning the matrix allowed direct relationships between pairs of measurements to be found. Each parameter is reported with the significant related factors in Figure 48.

Figure 48 - Related Factors in the Correlation Matrix

FACTOR	RELATED FACTORS	DISCUSSION
Section.	None.	
Position.	None.	
Time in section.	Time in department. Activity of acquired compounds.	The longer a chemist is in a section the better he is at selecting compounds not made by him which may be active. He tends to concentrate on known contacts working in his research area.
Time in department.	Time in section. Inversely number of compounds acquired.	Older chemists tend not to randomly acquire large numbers of compounds for testing.
Position in testing programme - lead chemist.	Number of searches. Number of visits. Number of compounds made.	Responsibility for a testing programme appears to increase the chemists use of the facilities.
Number of searches.	Lead chemist. Number of compounds requested. Number of visits. Number of samples obtained. Number of compounds per visit.	Much of the use of the sample location service is a direct consequence of a substructure search request.
	Compounds made by the chemist.	Perhaps intermediates found by a substructure search reduce the synthetic effort required to prepare a compound.
	Increase in searches. Increase in visits. Increase in compounds requested.	Even high users of the service were increasing their use of the services.
Number of visits.	Lead chemist. Number of searches. Number of compounds requested. Number of compounds located. Sample success.	Use of other services.
	Compounds made by the chemist.	More use for the service if more compounds made.
	Increase in visits, sample requested and in compounds obtained.	Satisfied customers were perhaps realising potentials of services.

FACTOR	RELATED FACTORS	DISCUSSION
Requests for compounds.	Lead chemist. Number of searches. Number of compounds obtained. Number of visits. Number of compounds per visit.	Other services.
	Compounds made by chemist.	More use for the service if more compounds are made.
	Increase in number of requests. Increase in samples obtained.	See above.
Samples obtained.	Number of searches. Number of visits. Number of requests. Sample success. Number of compounds per visit.	Other services.
	Compounds made by chemist.	Samples obtained for intermediates reduce preparative effort.
	Increase in number of requests. Increase in samples obtained.	See above.
Sample success.	Number of visits made. Sample obtained.	Used to calculate sample success.
	Inversely novel chemistry.	Non-availability of intermediate leads to novel method of preparation.
Compounds/visit.	Samples requested. Samples obtained. Relative increase in requests/sample. Number of searches.	Used in calculation. Chemist using sub-structure search service requests more compounds per visit.
Compounds made.	Various factors indicating use of the services.	
	Inversely to novel chemistry.	More time to prepare compound - less output.
Novel chemistry.	Inversely to compounds made. Sample success. Compounds for test.	Chemists working in novel areas of chemistry are less productive in drug design areas.

FACTOR	RELATED FACTORS	DISCUSSION
Compounds acquired.	Inversely to time in department.	See above.
	Tests per compound.	Compounds acquired are sent for a larger range of tests.
	Total activity.	Possibly a consequence of a number of tests for which submitted.
	Compounds per test.	Used in calculation.
Tests per compound.	Compounds acquired. Compounds per test.	
Activity, own compounds.	Other activity measures and relative activities.	No relationships to other factors.
Activity, acquired compounds.	Other activity measures.	
	Length of service in section.	After some time on a project, chemists tend to acquire from productive sources.
Activity, all compounds.	Other activity measures.	
	Number of compounds acquired.	Acquired compounds providing the distinguishing factor in activity determination.
Compounds per test.	Inversely to novel chemistry.	
	Compounds acquired. Tests per compound.	See above.
Increase in searches.	Number of searches. High increase in compound location services.	
Increase in visits.	Other relative measures. Number of searches. Number of visits.	
	Inversely to increase in activity.	More compounds being made led to a decrease in overall activity.
Increase in samples requested.	Other relative measures.	
	Inversely to increase in activity.	
Increase in compounds obtained.	As increase in samples requested.	

FACTOR	RELATED FACTORS	DISCUSSION
Increase in compounds made.	Increase in number of searches.	Substructure search could be having direct effect on compounds made.
Increase in activity, own compounds.	Other activities. Other increases in activities.	
	Inversely to increase in sample location services.	Could sample location services be most useful when looking for new leads?
Increase in activity, acquired compounds.	Other activities. Other increases in activities.	
Increase in activity, all compounds.	As for increase in activity, his compounds.	
Publications.	None.	
Assistance.	Lead chemists.	More assistance given to lead chemists.
Literature searching.	None.	

Use of the services appeared to be well correlated with productivity in terms of compounds made, but there seemed to be little or no relationships with the biological activity parameters. This may have been due to the problems involved in measuring biological activity parameters across the 170 biological tests. Good users of one service appeared to be good users of other services whether computerised or manual. The methodology was therefore immaterial to the user. Chemists background seemed not to relate to use of the services. This supports previous studies such as those of Back (17) and others (18). The role of a lead chemist seems the one prominent environmental factor. Responsibility for a test appears to encourage a chemist to make use of the service. A measure of the acceptability of the computerised services may be taken by the way in which good users of the system in 1972 were still good users of the system in 1973.

In general terms, the use of the various services seemed to be closely related to each other, pointing to the possible integration of the various services. Use of the services appeared to be independent of background, making the way forward to 100% usage more difficult. The problems of each non-user would have to be considered separately. There was a very close correlation between the use of the services and a chemists productivity in terms of compounds made, perhaps justifying the existence of the services. No relationships between use of the services and biological activities could be found - this was not unexpected because of the problems of generalising drug activity.

6.4.4 Interpreting the factor analysis

The aim of the factor analysis was to reduce the dimensionality of the variables measured by taking advantage of their inter-relations. Some of these relationships existed because of the dependence of one variable on another. For example, there will be some dependence between the total compound activity and the activity of the compounds made by him since the second is a subset of the first. The factor analysis used the symmetrical correlation matrices, R, produced above. The statistical package used (81) resolves these R matrices into pxk factor matrices where the number of factors k is

very much smaller than the number of variables, p.

The relationship of the correlation matrix to the factor matrix can be expressed as:

$R = VoVo^{1}$ where Vo is the matrix of orthogonal factors.

The computations carried out involve analysing the matrix R for its latent roots and vectors. For any set of variables, only the off-diagonal correlations in R are obtained (see Appendix VII, Tables 5-10).

The final stage in the factor analysis was that of factor rotation and interpretation. Once the set of factor loadings, corresponding to a set of hypothetical variates, had been obtained the final step was to try to interpret them in a way which gave a meaningful summary of the original data.

A factor loading of 0.4 was considered to be significant, and three factors emerged. These are given in Figure 49.

Figure 49 - Significant Results from the Factor Analysis

FACTOR	CONTENTS	POSSIBLE GROUP NAME
1	Activity - his, acquired and total. Relative activity - his, acquired, total.	ACTIVITY
2	Searches, visits, requests, samples, sample success, compounds per visit, compounds made, relative requests, relative samples.	USAGE/COMPOUND PRODUCTIVITY
3	Increase in use of services - searches, visits, requests, samples.	RELATIVE USAGE

All the remaining factors had only two or three significant loadings and no reliable interpretation could be placed on the possible groupings. Hence, the factor analysis had produced three independent factors - activity, usage/compound productivity and relative usage. The relationship "activity" can be seen as being a function of the relationships between the variables used, but the other results supported the conclusions reached from the correlation matrix:

(a) There appeared to be no correlation between activity and information usage or the number of compounds made.

(b) Usage of information services was closely related to the number of compounds made.

(c) Chemist's background appeared to have no significant effect on their use of information services or on the number of compounds made.

The most important result of the factor analysis is the grouping together of the number of compounds made with the use measures for the services provided. This supports the relationships found in the correlation matrices.

Chapter 7 - Conclusions and Suggestions for Future Developments

7.1 Objective of the Study

The study set out to examine in depth the development of a specialised information service within an industrial company, combining computer technology with an understanding of the human needs of the immediate users, the information staff, and the final customers, the chemists. The study was essentially in three parts:

(a) The design of an information system for the storage and retrieval of chemical structure information generated within the company.

(b) The study of the effects of the application of increasing levels of computer technology on the use made of the service.

(c) The search for relationships between the use made of the service and the customer's own background and his success in drug design.

The work was carried out within the Research Department at ICI Pharmaceuticals Division, and hence the study examines a real-life situation. It was felt that the best results would be obtained if the users were unaware that they were being studied. This together with the need to maintain established communication patterns and working relationships put a number of constraints on the conduct of the study.

The author worked within the Data Services Section which was responsible for the collection, storage and exploitation of research data generated during chemical, biological, toxicological and clinical research. She worked as a member of the group providing chemical structure retrieval services and was responsible for the systems development. She worked alongside the information staff and was therefore able to study their behaviour without their knowledge.

Such a close working relationship was not possible with the customers of the service - members of the Chemistry Department. This user group consisted of some 80 research chemists, each having an average of two graduate assistants. The chemists were working in multi-disciplinary project teams and were designing drugs for specialised research targets.

One could expect the study to measure the variations in use of the information service by chemists working in the various therapeutic areas. However, other evaluations of research behaviour had concluded that there were very significant differences in research behaviour between individual scientists, even those working in the same subject area (8-10).

Garvey (66-72) also showed that these variations in behaviour also led to differences in information use. This study therefore sought to identify some associations between the individual behaviour of its user chemists and examined whether differences in use could be associated with social conditions, e.g. background, status, age, or with the individual research area. Garvey (66-72) had found some relationship between information use and the various stages in a research project.

The study of human variables and the development of the computer system had to be carried out in parallel. A company-wide chemical retrieval system was urgently required to reduce obvious duplication of effort and consequently to lower manpower levels. The systems design work was given priority and resources were allocated to the development of the chemical structure handling system. There were constraints placed on this stage of the study. Any system developed had to take into account the available data bases and the facilities already functioning from these sources, the structure handling facilities already developed by ICI, and the divisional computer facilities with their particular orientation towards on-line working.

At the same time as the system was being developed, the author was studying the user characteristics. Reactions of the information staff were directly observed as the system became operational and whenever practical the system was manipulated to aid their job satisfaction.

Modifications to the system to meet the human requirements of the chemist were much less direct. Firstly it was necessary to collect information relating to their information use and to their behaviour. Once again, the environment imposed constraints. It was felt that any method of collecting this information which involved user participation was undesirable. This excluded the most commonly used forms of user study - questionnaires, interviews and diary methods. The decision was made to measure and analyse the user characteristics covering the three areas of concern - the use of the service, the human variables of the user and finally the user's success in drug design. Measurements on the use of the service were easily obtained by asking the information staff to maintain appropriate log books. Human variable data was more difficult since it had to be collected without the users participation. However, by consulting staff lists, etc. it was possible to build up a background for each user based on length of service, length of time on a project, status, administrative responsibilities, etc. Measurement of success in drug design was most difficult since a common measure was required across all therapeutic areas. The criteria chosen was the chemist's success in finding biological activity, both generally and in his own therapeutic area. The success of these measurements, however, depended on the biologist's assessment of activity and whether the biologists concerned had common standards.

It was not possible within the constraints of the study to examine the chemist's use of other information facilities such as the library, technical literature search services and patent search systems. Data collected from such areas might have given added insight into how a chemist conducts drug research.

The study examined a working industrial information service in which users could be individually monitored. Operational measures were sought to characterise user behaviour and these measures were compared with data on information use. The users were unaware the study was being carried out and could not, therefore, influence the outcome.

7.2 Design of the Information System

Before beginning systems design, it was necessary to examine the relevant information requirements of the chemical users. Garvey (66-72) had established the existence of patterns in scientific research and an understanding of the basic information needs in drug research in ICI had to be established. This was done by informal conversations with chemists and gradually a model was developed.

Chemists' needs for company-generated structural information varied with the stage of their research project. For example, when a chemist considered a new approach, he had to first check whether the compounds involved were already known to the company and whether they had already been subjected to biological evaluation. The chemist would usually ask for information on a class of compounds since there were many structural variations which would give similar biological effects. Having ascertained the novelty of his approach and examined any existing relevant data, he would then select one or more possible drug candidates. In making his choice, he would take a number of variables into consideration. These may have been biochemical and based on a knowledge of the biological process, they may have been physico-chemical, e.g. dependent on solubility or polarity or they may have been based on organic chemical constraints, e.g. difficulty of preparation. His next need for structural information would probably be in his search for suitable intermediates and he would now be looking for specific molecules from which to prepare his target molecule. Where possible, he would also be

interested in the availability of the samples and often in the method of preparation. Finally, having made and purified his compound, the chemist needed to register it so that the sample could be appropriately weighed, labelled and sent for biological testing. These points illustrated some of the basic needs for structural information in the drug design process. There were many other individual variations. For example, if a compound was shown to be active by random screening in a biological test, the chemist may have scanned the existing compounds for samples of the same type that could also be submitted for test. This enabled a class of compounds to be evaluated without any need for preparative effort.

From this understanding of the research process, it was possible to define a basic system which catered for the following functions:

(a) Stored information on all compounds known to the company in such a way that structural information could be obtained at the specific or at the class level.

(b) Allowed compounds to be registered into the collection so that information available was always up-to-date and correct.

(c) Allowed individual compounds to be looked up in response to chemists' requests.

(d) Allowed classes of compounds to be isolated from the total data base.

The next stage was to design the computer system so that these requirements could be achieved. Both on-line and batch facilities were made available by the computer department.

There were considerable advantages in making use of the on-line facilities. The data base would always be up-to-date, using VDU terminals immediate answers could be obtained for urgent requests, and any input errors could be corrected effectively. However, there were economic constraints preventing the development of a fully on-line system. For an on-line system, the data base had to be readily available to the computer system whether users were interacting with it or not.

To fit in with other computer users and to achieve a reasonable level of cost-effectiveness, the on-line system was designed to ensure that the data took up as little disc space as possible, the files were frequently used by a number of people to perform a number of tasks and that each task was carried out as efficiently as possible.

The chemical structure handling system was designed within these constraints. On-line techniques were developed where appropriate and additional facilities were designed to run in batch mode. Within the on-line system, the information on each compound was structured to accommodate existing data with minimum storage requirements and to give fast responses to any part of the data base. Costs were spread by operating the system on a company-wide basis and making the system available on three different sites, each site working on a different research problem. The data base allowed on-line access to the structural information by a variety of keys. These included four types of reference number, molecular formula for use by chemists and WLN for use by trained information staff.

Initially all the on-line routines were designed for use by information staff with or without the user chemist being present. This enabled the input structural information and its display as WLN and permitted the information staff to translate to and from the structural diagram manually.

When the system became operational, it was found that the response times for the interactive routines (specific compound location and compound registration) were slower than had been anticipated. This was not satisfactory since users had to wait several seconds from

asking the questions to obtaining an answer and they found the system time-consuming and frustrating. Some system modification was carried out so that questions could be batched and users could be typing in the next question whilst awaiting a previous reply. However, this did not solve the basic problem - a faster response was required for each query. This could possibly have been achieved by more sophisticated manipulation of the random access keys. Further work would be needed to examine both molecular formula and WLN keys to identify selective characteristics which would enable shorter and more precise indexes to be set up. This type of investigation could also lead to a reduction in the total file size and hence lower the data storage costs.

The substructure search system developed used both batch and on-line facilities. The user's request was first translated into a complex association of molecular parts using one or more levels of the multi-level search system.

The first level, the fragment search, required that the system analysed each compound in the files for each question. Any compounds containing the required fragment associations were then subjected to the next search level, the WLN/molecular formula string search. Both fragment and string search parameters were complex and on-line validation allowed the information staff to correct errors quickly. The execution of the searches themselves could be carried out in on-line mode for company-generated compounds, but the process took minutes rather than seconds and could not be said to be interactive. The more detailed search technique - atom-by-atom search was carried out in batch mode and although the search was time-consuming, the parameters were simple and on-line validation was not vital.

The system would work effectively providing it was used by information staff with a good knowledge of the system, and the combination of on-line and batch facilities ensured the searches were carried out as quickly and efficiently as possible.

The chemical information system developed is still in day-to-day operational use, both for compound registration, specific compound location and for substructure search. However, the manual molecular formula indexes are still being maintained for the chemists own use. This use is small and many chemists have been persuaded to use the services provided. Initial discussions with chemists to remove the manual indexes have shown one major deficiency in the computer system - the lack of browsability. Although some chemists are only too willing to pass a request on to information staff, others have a need to carry out the search themselves. If the system was to be used by the chemists themselves, a number of developments would be necessary. The first stumbling block is the system's reliance on WLN - a language chemists generally do not know and have shown little enthusiasm to learn. The first stage in the development must be therefore to make the WLN transparent to the user. Questions could be asked using molecular formula and answers could be in the form of the two-dimensional structure diagram. Thus the chemist could simulate his use of the manual indexes. Some incentive would be necessary for him to change to the new facility - a terminal in his own work area perhaps?

Although on-line facilities have enabled the development of systems to meet some of the chemists' needs, batch processing has not been completely superceded. One of the selling points of the ICI system was the chemist's ability to search a variety of data bases at the structure and substructure level. In his research, the chemist commonly needed access to a compound or a class of compounds. His requirement was often the quickest way of obtaining a sample of that compound. He first had to determine whether the compound was available within the company, next whether he could buy it from a suitable starting material from commercial suppliers, or finally whether its use or preparation had been reported in the literature. The chemical search system

developed at ICI allowed the chemist to satisfy these needs.

In summary, the chemical information system developed and now in operational use combined on-line and batch techniques. On-line routines were used where a fast response was required, e.g. location of specific compounds or where data required immediate validation, e.g. compound registration or search parameter formulation. Batch routines were used where large amounts of computer processing were required, e.g. structure generation from WLN or to extend the services provided, e.g. to less commonly-used data bases.

Existing information staff were to operate the new facilities and so it was important that they felt involved in the project. The specific compound enquiry procedures, the registration procedures and all the substructure search procedures have all been substantially modified in response to the use by information staff. It was not possible to merely introduce facilities which were better from the systems economics point of view, they had to bring substantial work advantages to the information staff. Reactions of the individuals were very different, most responses reflecting previous experience. Constant and individual training was necessary to show each person the advantages of the system in day-to-day use.

One major problem was the ease with which the on-line system could be abused. The introduction of on-line facilities needed careful monitoring, particularly if users had no knowledge of the consequences of their actions inside the computer. Certain measures had to be taken to reduce the freedom available through the system, since excessive computer time was being wasted on the least important questions. In addition, the information scientists tended to regard the terminal as a prestige piece of equipment and often tried to find ways to "beat" the system.

Working with the user information staff, it was possible to develop a system which allowed the information staff to participate at the design stages and to specify their own individual requirements. This need for systems design to cover both technical and human aspects has been a concern in the computer world for the last few years. Mumford (84) working at the National Computing Centre tried to develop a model to enable computer systems to be designed in both human and technical terms. The main problem in its implementation was that few computer technologists had received any training in the behavioural sciences and this made it difficult for them to identify social needs and incorporate mechanisms to cater for these in their design procedures. This might also be true of many information scientists, particularly in industry. Here much emphasis is placed on understanding the technology of the subject, e.g. chemistry, biology rather than understanding the scientists and the way in which they work. Garvey (66-72) has pioneered some of the work in this area, but it may be a long time before most industrial information units move from their present technical standpoint.

7.3 Study of the Use Made of the Facilities

Chemists' use of the various services was found by analysing the logbooks maintained by information staff over the period January 1972 to December 1973. A number of measurements were made for each of the three main categories of service provided - substructure searching for classes of compound, registration of new compounds and location of samples for specific compounds - and the results are summarised in Figure 50.

These figures show the overall increases in the use made of the services. However, more information on the trends can be seen by comparing the data obtained for the first three months of 1972 with the last three months of 1973. These figures are given in Figure 51.

Figure 50 - Summary of Chemist's Use of the Information Facilities in 1972 and 1973

SERVICE	1972	1973
A. Substructure Search		
1. Number of users	52	58
2. Number of questions asked	306	385
3. Average number of questions per user	5.9	6.8
B. Compound Registration		
1. Number of users	72	80
2. Number of compounds registered	3617	5076
3. Average number of compounds registered per user	50	63
C. Sample Location		
1. Number of visits to the service	1238	1325
2. Number of samples requested	3658	4078
3. Number of samples requested per visit	2.9	3.1
4. Number of samples obtained from requests	1746	1952
5. Average number of samples obtained per visit	1.4	1.5

Figure 51 - Comparison of Chemist's Use of the Information Facilities Early in 1972 and Late in 1973

SERVICE	JAN-MAR 72	OCT-DEC 73	DIFFERENCE
A. Substructure Search			
Number of questions	103	147	+43%
B. Compound Registration			
Number of compounds	963	1389	+44%
C. Sample Location			
Number of visits	369	363	-2%
Number of compounds requested	1071	1311	+22%
Number of samples obtained	478	607	+27%

There was therefore an increase of over 40% in the use made of the substructure search service at the end of the two-year period over the use made at the beginning. More chemists were using the service and each chemist was asking more questions. In addition, each question was usually asked against two or more data bases. For example, if a chemist was interested in synthetic intermediates, he would ask for information on both the company data and the commercially available compound data.

This increased use of the substructure search service seemed to indicate users were accepting and making use of the service. This was gratifying since attention had been given both to the technical suitability of the facilities and to the human preferences of the scientists. For example, output from a substructure search was in the form of index cards, showing all the available structural and non-structural data. This was important since it allowed the chemist to sort and select the data into associations meaningful to him in his current research problem.

Much attention was also given to creating the right environment for chemists to discuss their substructure search problems. This interface area had been stressed by Lancaster (49) in his appraisal of the Medlars system. In the ICI environment, chemists were encouraged to bring

their questions to the information staff, to discuss the information required and to formulate a realistic question. Lancaster (49) had found users asked questions which were often too narrow and gave them insufficient answers, or alternatively asked questions which were too broad. This was also the author's experience, and chemists were found to have definite anticipations of the information requested. This was often influenced by previous answers and by manual search techniques. Individual chemists reacted differently to the approaches made and gradually the information staff built up patterns of working with the individuals.

The increase in use of the substructure search service was a gradual one and hence supported the results of Rubinstein and Schultz (50) that researchers gradually accepted a service once its usefulness had been proved to them.

The increases in the numbers of compounds being registered showed similar trends to those for use of the substructure search service, i.e. each chemist produced an average of 25% more compounds in 1973 over 1972 and there was a gradual growth from early 1972 until late 1973. It was probable that there was some relationship between these two activities and that the increase in compound registration could be attributed to the move towards substructure searching rather than the availability of more sophisticated registration facilities.

Similar assumptions can be made about the association of the increase in sample location service usage and the increase in numbers of substructure searches. Here the increases were found not in the number of times a chemist visited the service, but in the number of compounds requested. Instead of asking for a specific compound, chemists were requesting examples from a broader class. Attempts to directly relate the two services were limited by the information available. However, it was found that at least 20% of the sample requests came directly from substructure searches.

Inspite of the overall increase in use of the substructure services, some 30% of chemists never used the service. Some of these chemists were not involved in preparative effort, but there was still a large number who might benefit from some use of the substructure facilities. It was hoped that some guidance could be obtained from the results of the use vs chemist background survey.

7.4 Relationships Between Use Made of the Service and Characteristics of Individual Scientists

Studies of the chemist population had to be carried out indirectly since it was felt that direct interaction, e.g. by questionnaire would endanger the delicate communication channels being set up between information scientists and chemists. Interrogation of chemists to find out why they used the service and what value they gained from the facilities would perhaps turn them away from using the service.

It was necessary therefore to define a number of performance ratings so that these could be directly related to usage of the services. This was difficult because of the variations in compound preparation and in biological activity measurements. Since there was no well-defined way to assess each chemist's performance, several factors were recorded which, together, might give some indication of performance. These eight measures included the novel reactions, the numbers of compounds acquired from other sources, the extent of biological testing carried out and the activity of the compounds submitted for test.

The use made of the information services was also studied in relation to the chemist's background. Four indications of background were measured - status, length of company service, length of time in project and the role against biological test. To supplement these, several other factors were also measured. These included the number of publications, the number of graduate assistants and the use of other computerised structure services.

In total there were 80 chemists and there were a total of 30 measurements for each chemist. The data on use, performance and background was collected from manual records, transferred to the computer and subjected to an automatic statistical analysis (81). Firstly, correlations were sought between the individual measurements. Secondly, factor analysis techniques were used to try to relate these measurements to more meaningful grouping, e.g. those relating to performance.

Correlation matrices were used to measure the relationship of one factor with another and a significance of greater than 99% was used to find relevent direct relationships between pairs of measurements. The results for the use vs performance part of the survey are summarised in Figure 52.

Figure 52 - Summary of Relationships Between Factors Showing Chemists Use and Performance

FACTOR	RELATED FACTOR	INTERPRETATION
1. No. of substructure searches.	A. Use of compound location services.	A. Much of the use of the sample location service is a direct consequence of a substructure search request.
	B. Compounds made by the chemists.	B. Availability of intermediates reduce the synthetic effort required to prepare a compound.
2. Use of sample location services.	A. No. of sub-structure searches.	See 1.A.
	B. Compounds made by the chemist.	See 1.B.
3. Success in sample location, i.e. compounds obtained/ compounds requested.	A. Inversely to novel chemistry.	A. Non-availability of intermediates leads to novel methods of preparation.
4. No. compounds requested per visit to sample location service.	A. No. of sub-structure searches.	A. Use of substructure searches leads to broader questions and more compounds being requested.
5. No. of compounds made.	A. Various factors indicating use of services.	A. Chemists using the services widely are making more compounds.
	B. Inversely to novel chemistry.	B. More time to prepare compounds - less output.
6. Involvement in novel chemistry.	A. Inversely to compounds made.	A. Chemists working in novel areas of chemistry appear to be less productive in some aspects of drug design.

FACTOR	RELATED FACTOR	INTERPRETATION
	B. Inversely to success in sample location.	B. See 3.A.
7. Compounds acquired.	A. No. of tests on each compound.	A. Compounds acquired are sent for a larger range of tests.
	B. Total nos. of compounds given as biologically active.	B. The larger the range of tests the greater the chance of finding activity.
8. Biological activity on own, acquired and all compounds.	A. Other activity measures.	A. Internally consistent, but not related to any use measurements.

Firstly, the use factors relating to the substructure search service, the compound location service and the compound registration service were consistent amongst themselves. This supports the earlier observation that the rise in use of all the services stemmed from the increases in use of the substructure search services.

These use factors were also related to the synthetic productivity of the user as shown by the very significant relationship between the numbers of compounds made and the number of substructure searches requested, and the ensuing requests for chemical samples. It seems reasonable to assume that the availability of intermediates helps chemists by reducing the effort required to prepare a compound. In addition, the number of compounds prepared by a chemist was inversely related to his use of novel chemistry. Chemists appeared to use novel methods of preparation only when intermediates were not available, their prime objective being to obtain a given molecule for biological evaluation. Use of the information facilities to obtain intermediates for synthetic work seemed to be desirable.

Another suggested use of the information facilities was to obtain compounds for test which were already available within the company. This would enable ideas to be evaluated without expensive preparative effort being spent. However, there was no significant correlation between the number of substructure searches requested and the numbers of compounds acquired for test.

Attempts to relate the use of the various services with a chemist's success in obtaining biologically active compounds were not conclusive, although the activity measures chosen appeared to be internally consistent. In addition, these measures did not relate to the factors chosen to characterise chemist's behaviour, e.g. age, status. The lack of correlations would appear to be associated with the choice of parameters. A broad classification of activity was used across a wide variety of biological tests. Some tests denoted activity whenever a biological response was found, however slight the response. At the other extreme, no compounds had ever been found active in some tests. Hence, the chance of a chemist finding activity varied with the tests being carried out in addition to his ability to prepare active compounds. An attempt was made to overcome this by comparing a chemist's performance against the overall performance for that test. The success of this, however, depended on the variety of compounds being submitted and the availability of that test to all chemists. This was not always the case: often a biological test was limited to compounds from one or two chemists. For the use vs activity study to have been more fruitful, a detailed analysis would first be required to quantify the various activity

coefficients used in all the biological tests. This was not possible within the constraints of this study.

The value of the information services appears to be evident only in the increase in the numbers of compounds prepared by chemists using the services. The need to prepare and test more and more compounds is becoming one of the prime targets of a drug research group. Techniques in biology have developed making the screening of large numbers of compounds efficient and relatively cheap. It therefore makes economic sense to test as many compounds as widely as possible.

Around 30% of chemists, however, never used the substructure search services and hopes were that information gleaned in the use vs background study would help find any reasons why. The results of this part of the study are summarised in Figure 53.

Figure 53 - Summary of Correlations Between Use of Information Services and Background Factors

FACTOR	RELATED FACTOR	INTERPRETATION
1. No. of substructure searches.	A. Responsibility for biological test.	A. Chemists in charge of the provision of compounds for test were under pressure to provide more compounds.
2. Use of sample location service.	A. Responsibility for biological test.	A. See 1.A.
3. No. of compounds made.	A. None.	A. Background factors did not influence nos. of compounds made.
4. Therapeutic area (section).	A. None.	A. Use of services did not relate to the type of biology being evaluated.
5. Time spent on project.	A. Time spent in company.	A. People who had been in company for some time, often stayed on same project.
6. Time spent in company.	A. Time spent on project. B. Inversely to no. of compounds acquired.	A. See 5.A. B. Older chemists tended not to acquire compounds for test, but concentrated on preparation.
7. Responsibility for biological test.	A. Use made of the services. B. No. of compounds made.	A. See 1.A. B. Responsibility for finding compounds for test appeared to affect positively the no. of compounds made.

FACTOR	RELATED FACTOR	INTERPRETATION
8. Status.	A. None.	A. Chemists status did not affect his use of the services
9. Assistance.	A. Responsibility for biological test.	A. More assistance given to chemist in charge of biological targets.
10. Computerised structure services on literature services.	A. None.	A. Good users of compound services were not necessarily good users of literature-based services.

There seemed to be little correlation between the chemist's use of the information services and the measures chosen to depict the chemist's background. The only positive correlation was between the chemist's use of the service and his responsibility for a biological testing programme. A chemist in this position had to ensure the biologist was supplied with the right compounds at the right time.

Use of the information services appears to be independent of age, status and background, and the wide range of use pointed to the individual way chemists perform their research function. This result supports the conclusions of other studies, such as that of Crane (8). He also found that scientists identify individually with a research problem and react each in their own way.

The last part of the analysis was an attempt to relate the individual measurements into larger, more meaningful groups. This was done by applying factor analysis, a multivariate analysis which attempts to explain the common variance of observations on a set of variables as the sum of variances of a smaller ordered set of uncorrelated factors. The significant factors which emerged by considering a factor loading of 0.4 are given in Figure 54.

Figure 54 - Summary of Significant Groupings from the Factor Analysis

FACTOR	CONTENTS	POSSIBLE GROUP NAME
1	Activity - his compounds, acquired compounds and total compounds.	Activity.
2	Substructure searches, sample location service parameters, no. of compounds made, increase in use of sample location service.	Usage/Compound Productivity.
3	Increase in use of information services.	Relative use.

All the remaining factors had only two or three significant loadings and no reliable interpretation could be placed on the possible groupings. Hence the factor analysis produced only three significant groupings - one centred around the measures used to define biological activity, one centred around use of the information services and the number of compounds prepared, and the third centred around the increase in use of the information services.

The interpretation of these results supports the observations made in other parts of the study:

(a) Activity measures, although consistent amongst themselves, did not relate to measures indicating use, the number of compounds made or the chemists background.

(b) A chemist's use of one of the services was closely related to his use of the other services. Increases in use of the substructure search service was thought to have caused corresponding increases in use of the other services.

(c) Use of the information services was closely related to the number of compounds made by any chemist.

(d) Background factors such as age, status, research area, appeared to have no relationship to the use made of the information services or to a chemist's success in drug design.

7.5 Overall Considerations

This sociotechnical study of information systems design was carried out in an industrial environment where users could be readily identified. The closed community enabled users to be closely studied over a two-year period, and the effects of increasing degrees of computer technology quantified.

The system was designed to improve facilities offered to chemists for specific compound location, compound registration and for substructure search. The study showed the changes in the substructure search area to be most significant and chemists making more use of the substructure search facilities also made more use of the other services. The impact of the introduction of the substructure search service and its continuing development and enhancement can be seen by studying Figure 55. This shows the total use of the substructure search service over a five-year period.

Figure 55 - Number of Substructure Searches Carried Out in the Period 1971-1975

YEAR	NO. OF QUESTIONS
1971	400
1972	480
1973	979
1974	1563
1975	2789

The period studied, 1972 and 1973, was the beginning of a growth period which has led to increases of nearly 700% in the total numbers of searches being carried out. Pharmaceuticals Division chemists asked over 680 questions in 1975 as compared with 300 in 1972 and 385 in 1973, with a very small rise in the number of chemists. This means that each chemist is now asking at least twice as many questions as he was in 1972. However, use of the chemical search facilities still varies widely, some chemists making little or no use and others making extensive use.

The impact of the information services can only be measured in terms of the increase in the numbers of compounds prepared by chemists who made extensive use of the services. No significant relationships could be found between the information facilities and a chemist's success in preparing biologically-active compounds. In addition, age, status or background of chemists did not influence their acceptance of computerised information facilities, or affect the numbers of compounds prepared by him.

It is possible that better correlations between information use and chemist's background could have been obtained by using different

methods of data collection. For example, more detailed information may have been obtained by asking chemists to fill in questionnaires. However, there are disadvantages to such an approach including interference with the operational information scientist/chemist interface and a false picture being obtained by the chemists knowing they were being studied. The use of operational measures enabled the chemists and information scientists to be studied without their knowledge and a realistic study to be made of an established information service. The value of the service must be seen in the chemists' willingness to make more and more use of the facilities offered.

The findings of this study confirmed the great variations in chemists' information seeking behaviour found in other studies (66-72). Each chemist had unique information needs and used available information sources to different degrees. This presents challenges to the designer of on-line information systems for use directly by chemists themselves.

In batch systems or in on-line systems using intermediaries, the information scientist interface can manipulate the system to meet individual user needs. This is not possible in on-line systems used directly by the chemist. However, before such a system could be designed for use by ICI chemists, more research is required to understand their scientific methods and their information needs.

Appendix I - Distribution of Molecular Formulae in ICI Compound File

POPULATION	NUMBER OF FORMULAE	CUMULATIVE NO. OF FORMULAE	PERCENTAGE OF FORMULAE	CUMULATIVE % OF FORMULAE
1	26827	26827	60.82	60.82
2	6676	33503	15.14	75.96
3	3104	36607	7.04	83.00
4	1813	38420	4.11	87.11
5	1232	39652	2.79	89.90
6	765	40417	1.73	91.63
7	589	41006	1.34	92.97
8	439	41445	0.99	93.96
9	405	41850	0.92	94.88
10	313	42163	0.71	95.59
11	258	42421	0.58	96.17
12	202	42623	0.46	96.63
13	186	42809	0.42	97.05
14	125	42934	0.28	97.33
15	121	43055	0.27	97.60
16	105	43160	0.24	97.84
17	106	43266	0.24	98.08
18	79	43345	0.18	98.26
19	78	43423	0.18	98.44
20	66	43489	0.15	98.59
21	62	43551	0.14	98.73
22	56	43607	0.13	98.86
23	58	43665	0.13	98.99
24	36	43701	0.08	99.07
25	39	43740	0.09	99.16
26	26	43766	0.06	99.22
27	27	43793	0.06	99.28
28	29	43822	0.07	99.35
29	25	43847	0.06	99.41
30	24	43871	0.05	99.46
31	25	43896		
32	18	43914		
33	14	43928	0.18	99.64
34	14	43942		
35	7	43949		
36	11	43960		
37	11	43971		
38	8	43979	0.11	99.75
39	8	43987		
40	10	43997		
41	10	44007		
42	9	44016		
43	9	44025	0.10	99.85
44	8	44033		
45	7	44040		
46	6	44046		
47	0	44046		
48	5	44051	0.03	99.88
49	3	44054		
50	0	44054		
51	6	44060		
52	4	44064		
53	4	44068		
54	5	44073		
55	1	44074	0.07	99.95
56	3	44077		
57	4	44081		
58	2	44083		
59	2	44085		
60	1	44086		

POPULATION	NUMBER OF FORMULAE	CUMULATIVE NO. OF FORMULAE	PERCENTAGE OF FORMULAE	CUMULATIVE % OF FORMULAE
61	2	44088		
62	0	44088		
63	0	44088		
64	2	44090		
65	1	44091	0.03	99.98
66	3	44094		
67	1	44095		
68	2	44097		
69	1	44098		
70	2	44100		
71	0	44100		
72	1	44101		
73	0	44101		
74	0	44101		
75	0	44101	0.01	99.99
76	1	44102		
77	0	44102		
78	0	44102		
79	1	44103		
80	0	44103		
81	0	44103		
82	1	44104		
83	0	44104		
84	0	44104		
85	0	44104		
86	1	44105		
87	0	44105		
88	0	44105	0.01	100.00
89	0	44105		
90	0	44105		
91	0	44105		
92	0	44105		
93	1	44106		

Appendix I (Cont.)

Appendix II - Statistics of Occurrence for Automatically Generated Fragment Code - ICI Company File

DESCRIPTION	COUNT	%
UNUSUAL HETERO	7221	4.60
POSITIVE CHARGE	6646	4.24
NITRO - NO2	14663	9.35
DIOXO - NOT NO2	13601	8.67
TERM.O(NOT C=O)	5615	3.58
1 Y BRANCH	27686	17.65
4 BRANCH CARB	12419	7.92
H-FREE N	25322	16.15
>3 BRANCH N	1313	0.84
1 S (NOT C=S)	19429	12.39
>1 S (NOT C=S)	5569	3.55
1 -C=S	6283	4.01
>1 -C=S	879	0.56
1 D BOND NOT NU	11967	7.63
>1 D BOND NT NU	2563	1.63
TRIPLE BOND	868	0.55
1 CH METHYL	28696	18.30
>1 CH METHYL	39216	25.01
CH ETHYL	33184	21.16
ALKYL CH 3-9 C	12181	7.77
ALKYL CH > 9 C	2205	1.41
CH HALOGEN	10568	6.74
1 CH CHLORINE	3051	1.95
>1 CH CHLORINE	3055	1.95
CH BROMINE	1513	0.96
CH FLUORINE	3511	2.24
CH IODINE	140	0.09
1 CH -NH-	15645	9.98
>1 CH -NH-	6061	3.86
1 CH -NH2	8753	5.58
>1 CH -NH2	1210	0.77
1 CH -N=	8284	5.28
>1 CH -N=	2432	1.55
CH UNUSUAL CARB	5206	3.32
1 CH O (EX VO)	4664	2.97
>1 CH O(EX VO)	3484	2.22
1 CH OH(EX VQ)	13454	8.58
>1CH OH(EX VQ)	3759	2.40
1 CH V(EX VOVQ)	17011	10.85
>1CH V(EX VOVQ)	3718	2.37
1 CH -C=O.OH	6248	3.98
>1 CH -C=O.OH	709	0.45
1 CH -C=O.O-	10027	6.39
>1 CH -C=O.O-	2969	1.89
1 SUBST METHYL	38795	24.74
>1 SUBST METHYL	23298	14.86
SUBST ETHYL	9464	6.03
SUBST ALK 3-9C	5345	3.41
SUBST ALK >9C	719	0.46
SUBST HALOGEN	36920	23.54
1 SUBST CHLOR	17842	11.38
>1 SUBST CHLOR	11800	7.52
SUBST BROMINE	5289	3.37
SUBST FLUORINE	3300	2.10
SUBST IODINE	1115	0.71
1 SUBST -NH-	28521	18.19
>1 SUBST -NH-	7541	4.81
1 SUBST -NH2	10381	6.62
>1 SUBST -NH2	2546	1.62
1 SUBST -N=	5217	3.33
> 1 SUBST -N=	3588	2.29

Appendix II (Cont.)

DESCRIPTION	COUNT	%
SBST UNUSUAL C	2805	1.79
1 SB O (EX VO)	19383	12.36
>1 SB O (EX VO)	7347	4.68
1 SB OH (EX VQ)	15984	10.19
>1 SB OH(EX VQ)	5143	3.28
1 SB V(EX VQVO)	15554	9.92
>1SB V(EX VQVO)	2052	1.31
1 SB -C=O.OH	5410	3.45
>1 SB -C=O.OH	944	0.60
1 SB -C=O.O-	8414	5.37
>1 SB -C=O.O-	1944	1.24
1 O IN 1 RING	13018	8.30
>1 O IN 1 RING	2089	1.33
1 O IN >1 RING	1782	1.14
>1O IN >1 RING	225	0.14
1 N IN 1 RING	32812	20.92
>1 N IN 1 RING	29152	18.59
1 N IN >1 RING	8876	5.66
>1 N IN >1 RING	4998	3.19
1 S IN 1 RING	10938	6.97
>1 S IN 1 RING	546	0.35
1 S IN >1 RING	1216	0.78
>1 S IN >1 RING	32	0.02
1 =O IN 1 RING	17842	11.38
>1 =O IN 1 RING	8366	5.33
1 =O IN >1 RING	1613	1.03
>1 =O IN >1 RNG	580	0.37
1 Y IN 1 RING	6725	4.29
>1 Y IN 1 RING	1044	0.67
1 Y IN >1 RING	651	0.42
>1 Y IN >1 RING	32	0.02
1 X IN 1 RING	657	0.42
>1 X IN 1 RING	125	0.08
1 X IN >1 RING	126	0.08
>1 X IN >1 RING	6	0.00
AROMATIC 6	113163	72.16
CARBOCYCLIC 5	4768	3.04
CARBOCYCLIC 6	11347	7.24
CARB NOT 5 OR 6	1553	0.99
HETEROCYCLIC 5	38430	24.51
HETEROCYCLIC 6	49105	31.31
HET NOT 5 OR 6	3511	2.24
1 HETERO 1 RNG	31475	20.07
2 HETERO 1 RNG	34376	21.92
>2 HETERO 1 RNG	9177	5.85
1 HETRO >1 RNG	9063	5.78
2 HETRO >1 RNG	5665	3.61
>2 HET >1 RNG	793	0.51
1 SNGLE HET RNG	42065	26.82
>1 SNGL HET RNG	8097	5.16
1 SNGLE CRB RNG	6855	4.37
>1 SNGL CRB RNG	788	0.50
1 C/C 1 RG SM	8442	5.38
>1 C/C 1 RG SM	4570	2.91
1 C/C >1 RG SM	706	0.45
>1 C/C >1 RG SM	188	0.12
1 C/H 1 RG SM	21486	13.70
>1 C/H 1 RG SM	5141	3.28
1 C/H >1 RG SM	1600	1.02
>1 C/H >1 RG SM	150	0.10
1 H/H 1 RG SM	7776	4.96

Appendix II (Cont.)

DESCRIPTION	COUNT	%
>1 H/H 1 RG SM	472	0.30
1 H/H >1 RG SM	274	0.17
>1 H/H >1 RG SM	6	0.00
SPIRO	1887	1.20
TRUE BRIDGE	1611	1.03
1 MULTICYCLIC	1493	0.95
>1 MULTICYCLIC	885	0.56
BI-LINKAGE	25361	16.17
CHELATE	514	0.33
METALLOCENE	199	0.13
INORGANIC	387	0.25
1 RG SYSTEM	75313	48.02
2 RG SYSTEMS	17717	11.30
>2 RG SYSTEMS	2087	1.33
1 BENZENE RING	57027	36.36
2 BENZENE RGS	25889	16.51
>2 BENZENE RGS	5727	3.65
1 CARBOCYCLIC	26117	16.65
2 CARBOCYCLICS	13279	8.47
>2 CARBOCYCLICS	7058	4.50
1 HETEROCYCLIC	57936	36.94
2 HETEROCYCLIC	19409	12.38
>2 HETEROCYCLIC	3293	2.10
POLYPEPTIDE	780	0.50
POLYMER	499	0.32
> 1 Y BRANCH	10609	6.76
M REFERENCES	110050	70.17
SC REFERENCES	52187	33.28
R REFERENCES	90890	57.96
LIVE RECORDS	156824	57.96

Appendix III - Fragment Code Automatically Generated From the Notation - 122 Fragment Version

A. Miscellaneous

(1) Atoms other than C, H, O, N, S or halogens.
(2) Positive charge (indicates quarternary salt present).
(3) Branching terminal nitro-group.
(4) Dioxogroup (not NO_2).
(5) Terminal oxygen (not carbonyl).

B. For All Non-Cyclic Parts of the Molecule

(6) Trisubstituted carbon.
(7) Nitrogen bonded to more than three atoms.
(8) Trisubstituted nitrogen.
(9) Tetrasubstituted carbon.
(10) One sulphur atom.
(11) More than one sulphur atom.
(12) One double bond.
(13) More than one double bond.
(14) Triple bond.

C. Chain Fragments

(15) One methyl/methylene group.
(16) More than one methyl/methylene group.
(17) Ethyl/ethylene group.
(18) Alkyl chain with 3-9 carbons.
(19) Alkyl chain with ten or more carbons.
(20) Generic halogen.
(21) One chlorine.
(22) More than one chlorine.
(23) Bromine.
(24) Fluorine.
(25) Iodine.
(26) One -NH- group.
(27) More than one -NH- group.
(28) One primary amine.
(29) More than one primary amine.
(30) One linking N (no hydrogens) i.e. -N=.
(31) More than one linking N, i.e. -N=.
(32) Cyanide group.
(33) One linking -O- atom.
(34) More than one linking -O- atom.
(35) One -OH group.
(36) More than one -OH group.

(37) One carbonyl group.

(38) More than one carbonyl group.

D. Ring Substituents

(39) One methyl/methylene group.

(40) More than one methyl/methylene group.

(41) Ethyl/ethylene group.

(42) Alkyl chain with 3-9 carbons.

(43) Alkyl chain with ten or more carbons.

(44) Generic halogen.

(45) One chlorine.

(46) More than one chlorine.

(47) Bromine.

(48) Fluorine.

(49) Iodine.

(50) One -NH- group.

(51) More than one -NH- group.

(52) One primary amine.

(53) More than one primary amine.

(54) One linking N (no hydrogens) i.e. -N=.

(55) More than one linking N, i.e. -N=.

(56) Cyanide group.

(57) One linking -O- atom.

(58) More than one linking -O- atom.

(59) One -OH group.

(60) More than one -OH group.

(61) One carbonyl group.

(62) More than one carbonyl group.

E. Ring Heteroatoms

(63) Single occurrence of oxygen.

(64) Multiple occurrence of oxygen.

(65) Single occurrence of oxygen in more than one ring.

(66) Multiple occurrence of oxygen in more than one ring.

(67) Single occurrence of nitrogen.

(68) Multiple occurrence of nitrogen.

(69) Single occurrence of nitrogen in more than one ring.

(70) Multiple occurrence of nitrogen in more than one ring.

(71) Single occurrence of sulphur.

(72) Multiple occurrence of sulphur.

(73) Single occurrence of sulphur in more than one ring.

(74) Multiple occurrence of sulphur in more than one ring.

(75) Single occurrence of carbonyl.

(76) Multiple occurrence of carbonyl.

(77) Single occurrence of carbonyl in more than one ring.

(78) Multiple occurrence of carbonyl in more than one ring.

(79) Single occurrence of any other heteroatom.

(80) Multiple occurrence of any other heteroatom.

(81) Single occurrence of any other heteroatom in more than one ring.

(82) Multiple occurrence of any other heteroatom in more than one ring.

F. Ring Types

(83) Aromatic 6-membered ring.

(84) Carbocyclic 5-membered ring.

(85) Carbocyclic 6-membered ring.

(86) Heterocyclic 5-membered ring.

(87) Heterocyclic 6-membered ring.

(88) Carbocyclic rings other than 5- and 6-membered.

(89) Heterocyclic rings other than 5- and 6-membered.

G. Ring Fusions

(90) One carbo/carbo fusion.

(91) More than one carbo/carbo fusion.

(92) One carbo/carbo fusion in more than one ring system.

(93) More than one carbo/carbo fusion in more than one ring system.

(94) One carbo/hetero fusion.

(95) More than one carbo/hetero fusion.

(96) One carbo/hetero fusion in more than one ring system.

(97) More than one carbo/hetero fusion in more than one ring system.

(98) One hetero/hetero fusion.

(99) More than one hetero/hetero fusion.

(100) One hetero/hetero fusion in more than one ring system.

(101) More than one hetero/hetero fusion in more than one ring system.

H. Ring Linkages

(102) Spiro ring indicator.

(103) True bridge indicator.

(104) One multi-cyclic point.

(105) More than one multi-cyclic point.

(106) Ring of rings.

I. Total Ring Features

(107) One ring system.

(108) Two ring systems.

(109) More than two ring systems.

(110) One benzene ring.

(111) Two benzene rings.

(112) More than two benzene rings.

(113) One carbocyclic ring.

(114) Two carbocyclic rings.

(115) More than two carbocyclic rings.

(116) One heterocyclic ring.

(117) Two heterocyclic rings.

(118) More than two heterocyclics.

J. Non-Chemical Fragments

(119) A collection.

(120) M collection.

(121) SC collection.

(122) R collection.

Appendix IV - Specification For 152 Fragmentation Code Automatically Generated From The WLN

Fragments and Their Meaning

The fragments are divided into groups and the method of identification from the WLN is discussed.

A. All Parts of the Molecule

(1) Atoms other than C, H, O, N, S or halogens.
Character sequence -aa- or the character B (not ∇B) or the character P (not ∇P) found anywhere in the molecule.

(2) Positive charge.
Character sequence ∇&Q∇ indicating quarternary salt present. Only pertains to ICI generated files.

B. All Non-Cyclic Parts of the Molecule

Character sequences must be outside the ring signs (i.e. T...J and L...J).

(3) Branching terminal nitro-group - NO2.
The character sequence NW (or WN at the start of the notation).

(4) Dioxygengroup (excluding NO2).
The character sequence W but not NW or WN. Any substituent W found within ring signs is also included here.

(5) Terminal oxygen (not carbonyl).
The character sequence O& or O∇, or the letter O starting the notation, e.g. N-oxide, sulphoxide.

(6) One 3-branch carbon atom.
The character Y (but not ∇Y) occurring once only.
(Note: More than one 3-branch carbon is fragment 148.)

(7) 4-branch carbon atom.
The character X (but not ∇X).

(8) 3-branch nitrogen atom.
The character N, but not ∇N or NW or WN or NU or UN. This definition also includes unusual conditions of nitrogen, e.g. in cyanide, isocyanide, etc.

(9) Nitrogen atom with four or more co-ordinates.
The character K but not ∇K.

(10) One sulphur atom.
The single occurrence of S, but not ∇S or US or SU.

(11) More than one sulphur atom.
The multiple occurrence of S, but not ∇S or US or SU.

(12) One -C=S group.
The single occurrence of the groups US∇ or US& (or SU at the start of the notation only).

(13) More than one -C=S group.
The multiple occurrence of the groups US∇ or US& (or SU at the start of the notation only).

(14) One double bond, excluding -C=S, -N= or -C=O.
The single occurrence of the letter U, but not in any of the following groups: ∇U, UU, US, SU, NU, UN, MU, UM.

(15) More than one double bond, excluding -C=S, -N= or -C=O.
The multiple occurrence of the letter U, but not in any of the following groups: ∇U, UU, US, SU, UN, NU, MU or UM.

(16) Triple bond.
The occurrence of the symbol combination aUUa.

C. Chain and Substituent Fragments

The type of fragment in this class is subdivided into two subsets:

(a) Chain fragments (17-44). Here the character sequences must not be immediately attached to a ring system (i.e. preceded by a locant or prior to entry into a ring system).

(b) Substituent fragments (45-72). Here the fragments must be directly attached to a ring of some kind, and may be found after a locant in the notation or attached to a trailing ring system.

CHAIN	SUBSTITUENT	
(17)	(45)	One methyl/methylene group. Single occurrence of the number 1 not followed or preceded by a numeral.
(18)	(46)	More than one methyl/methylene group. Multiple occurrence of the number 1 not followed or preceded by a numeral.
(19)	(47)	Ethyl/ethylene group. Occurrence of the number 2 not followed by or preceded by a numeral.
(20)	(48)	Alkyl chain (CH2)n or CH3(CH2)n-1 where n = 3-9. Occurrence of a number in the range 3-9, but not followed by or preceded by a numeral.
(21)	(49)	Alkyl chain (CH2)n or CH3(CH2)n-1 where n = 10 or mor Occurrence of the sequence nn or nnn.
(22)	(50)	Generic halogen. Occurrence of any of the characters E, F, G, H, I.
(23)	(51)	One chlorine. Single occurrence of the character G.
(24)	(52)	More than one chlorine. Multiple occurrence of the character G.
(25)	(53)	Bromine. Occurrence of one or more E symbols.
(26)	(54)	Fluorine. Occurrence of one or more F symbols.
(27)	(55)	Iodine. Occurrence of one or more I symbols.
(28)	(56)	One-NH- group. Single occurrence of the symbol M, but not UM or MU at the start of the notation.
(29)	(57)	More than one -NH- group. Multiple occurrence of the symbol M, but not UM (or MU) at the start of the notation.
(30)	(58)	One -NH2 group. Single occurrence of the symbol Z.
(31)	(59)	More than one -NH2 group. Multiple occurrence of the symbol Z.
(32)	(60)	One -N= or N= group. Single occurrence of the symbol sequence UN or NU or UM (or MU at the start of the notation).

CHAIN	SUBSTITUENT	
(33)	(61)	More than one -N= or N= group. Multiple occurrence of the symbol sequence UN or NU or UM (or MU at the start of the notation).
(34)	(62)	Unusual carbon atom. One or more occurrences of the symbol C. Usually found in triple bonds, such as cyanides, isocyanides, etc.
(35)	(63)	One -O- group. Single occurrence of the symbol O, but not in the sequence OV or VO, e.g. ethers.
(36)	(64)	More than one -O- group. More than one occurrence of the symbol O, but not in the sequence VO or OV.
(37)	(65)	One -OH group. Single occurrence of the symbol Q, but not in the sequence VQ (or QV at the start of the notation).
(38)	(66)	More than one -OH group. Multiple occurrence of the symbol Q, but not in the sequence VQ (or QV at the start of the notation).
(39)	(67)	One -C=O group. Single occurrence of the symbol V, but not in the sequence VQ or VO or OV (or QV at the start of the notation).
(40)	(68)	More than one -C=O group. Multiple occurrence of the symbol V, but not in the sequence VQ or VO or OV (or QV at the start of the notation).
(41)	(69)	One -C(=O)-OH (acid) group. Single occurrence of the symbol combination VQ (or QV at the start of the notation).
(42)	(70)	More than one -C(=O)-OH (acid) group. Multiple occurrence of the symbol combination VQ (or QV at the start of the notation).
(43)	(71)	One -C(=O)-O (ester) group. Single occurrence of the symbol combination VO or OV.
(44)	(72)	More than one -C(=O)-O (ester) group. Multiple occurrence of the symbol combination VO or OV.

D. Ring Heteroatoms

Each ring system in the molecule is analysed; each ring is isolated and assigned a heteroatomic description. This description lists the heteroatoms present in the ring. The following fragments are set according to the analysis of that ring description.

(73) Single occurrence of oxygen.
A ring description contains only one oxygen (O).

(74) Multiple occurrence of oxygen.
A ring description contains more than one oxygen.

(75) Single occurrence of oxygen in more than one ring.
More than one ring description each containing only one oxygen.

(76) Multiple occurrence of oxygen in more than one ring.
More than one ring description containing more than one oxygen.

(77) Single occurrence of nitrogen.
A ring description contains one nitrogen (N, M, K).

(78) Multiple occurrence of nitrogen.
A ring description contains more than one nitrogen.

(79) Single occurrence of nitrogen in more than one ring.
More than one ring description containing one nitrogen.

(80) Multiple occurrence of nitrogen in more than one ring.
More than one ring description containing more than one nitrogen.

(81) Single occurrence of sulphur.
A ring description contains only one sulphur atom (S).

(82) Multiple occurrence of sulphur.
A ring description contains more than one sulphur.

(83) Single occurrence of sulphur in more than one ring.
More than one ring description contains one sulphur.

(84) Multiple occurrence of sulphur in more than one ring.
More than one ring description containing more than one sulphur.

(85) Single occurrence of carbonyl.
A ring description contains one carbonyl (V).

(86) Multiple occurrence of carbonyl.
A ring description contains more than one carbonyl.

(87) Single occurrence of carbonyl in more than one ring.
More than one ring description contains one carbonyl.

(88) Multiple occurrence of carbonyl in more than one ring.
More than one ring description containing more than one carbonyl.

(89) Single occurrence of exocyclic double bond.
A ring description contains one exodouble bond (Y).

(90) Multiple occurrence of exocyclic double bond.
A ring description contains more than one exodouble bond (Y).

(91) Single occurrence of exocyclic double bond in more than one ring.
More than one ring description contains one exodouble bond.

(92) Multiple occurrence of exocyclic double bond in more than one ring.
More than one ring description contains more than one exodouble bond.

(93) Single occurrence of any other heteroatom.
Occurrence of any letter other than H, K, M, N, O, S, T, V, U, X or Y.

(94) Multiple occurrence of any other heteroatom.
Occurrence of any letter other than above more than once in the same ring description.

(95) Single occurrence of any other heteroatom in more than one ring.
More than one ring description contains a letter other than those given above.

(96) Multiple occurrence of any other heteroatom in more than one ring.
More than one ring description contains more than one letter other than those given above.

E. Ring Types

On analysis of the WLN ring record, a ring type description is set up which gives information on the size of each ring and the saturation/unsaturation value of that ring. The ring descriptor gives the atom types in each ring and this is used to determine whether hetero/carbo.

(97) Aromatic Carbocyclic 6-membered ring.
The presence of at least one 6-membered ring, fully aromatic and no heteroatoms present in the ring description

(98) Carbocyclic 5-membered ring.
The presence of at least one 5-membered ring saturated or partially saturated (specifically excluding aromatic case) and no heteroatoms present in the ring description.

(99) Carbocyclic 6-membered ring.
The presence of at least one 6-membered ring, saturated or partially saturated, and no heteroatoms present in the ring description.

(100) Carbocyclic rings other than 5- and 6-membered.
The presence of at least one ring (not 5- or 6-membered), saturated or partially saturated and no heteroatoms in the ring description.

(101) Heterocyclic 5-membered ring.
The presence of at least one 5-membered ring, saturated or unsaturated, and at least one heteroatom in the ring description.

(102) Heterocyclic 6-membered ring.
The presence of at least one 6-membered ring, saturated or unsaturated, and at least one heteroatom in the ring description.

(103) Heterocyclic rings other than 5- and 6-membered.
The presence of at least one ring (not 5- or 6-membered), saturated or unsaturated, and at least one heteroatom in the ring description.

F. Heteroatom Count

Count of total number of heteroatoms of any type occurring in one ring.

(104) One heteroatom in one ring.
Total of one heteroatom in one ring.

(105) Two heteroatoms in one ring.
Total of two heteroatoms in one ring.

(106) More than two heteroatoms in one ring.
Total of three or more heteroatoms in one ring.

(107) One heteroatom in more than one ring.
Total of one heteroatom in more than one ring.

(108) Two heteroatoms in more than one ring.
Total of two heteroatoms in more than one ring.

(109) More than two heteroatoms in more than one ring.
Total of three or more heteroatoms in more than one ring.

G. Ring Fusions

A set of ring descriptions is set up for each ring system in the order in which they occur. These are compared to find the fusion types.

(110) One single heterocyclic ring.
A heterocyclic ring unfused to any other ring.

(111) More than one single heterocyclic ring.
More than one heterocyclic ring unfused to any other ring.

(112) One single carbocyclic ring.
A carbocyclic ring unfused to any other ring.

(113) More than one single carbocyclic ring.
More than one carbocyclic ring unfused to any other ring.

(114) One carbo/carbo fusion.
A carbo ring (saturated or unsaturated) fused to a second carbo ring (saturated or unsaturated).

(115) More than one carbo/carbo fusion.
More than one carbo ring attached to another carbo ring within the same ring system.

(116) One carbo/carbo fusion in more than one ring system.
One carbo ring attached to a second carbo ring occurring in more than one ring system.

(117) More than one carbo/carbo fusion in more than one ring system.
More than one carbo/carbo fusion occurring in more than one ring system.

(118) One carbo/hetero fusion.
A carbo ring (saturated or unsaturated) fused to a hetero ring.

(119) More than one carbo/hetero fusion.
More than one carbo/hetero fusion occurring in the same ring system.

(120) One carbo/hetero fusion in more than one ring system.
One carbo/hetero fusion in more than one ring system.

(121) More than one carbo/hetero fusion in more than one ring system.
More than one carbo/hetero fusion occurring in more than one ring system.

(122) One hetero/hetero fusion.
Two hetero rings fused to each other.

(123) More than one hetero/hetero fusion.
More than one hetero/hetero fusion occurring in the same ring system.

(124) One hetero/hetero fusion in more than one ring system.
One hetero/hetero fusion in more than one ring system.

(125) More than one hetero/hetero fusion in more than one ring system.
More than one hetero/hetero fusion occurring in more than one ring system.

H. Ring Linkages

(126) Spiro ring indicator.
Sequence "ba-&b" or "b&-&b".

(127) True bridge indicator.
Within ring signs any of the character sequence bab, ba&&&, ba&&T, ba&T, ba&b, ba-b, baT, baJ, ba&-T, Ba&-&x, but not preceded by bn within the ring.

(128) One multi-cyclic point.
Within any ring signs sequence bna where n=1.

(129) More than one multi-cyclic point.
Within any ring signs sequence bn where n > 1, or sequence bnn.

(130) Bilinkage.
Two ring systems (including benzene) are linked together.

I. Unusual Conditions

(131) Chelate.
Ring beginning with character D.

(132) Metallocene.
Ring containing character O, not within hyphens.

(133) Inorganics.
Notation begins with a space.

J. Total Ring Features

Used to indicate the presence of ring features in the molecule.

(134) One ring system.
Occurrence of one ring system (not benzene).

(135) Two ring systems.
Occurrence of two ring systems (not benzene).

(136) More than two ring systems.
Occurrence of more than two ring systems (not benzene).

(137) One benzene ring.
Occurrence of one phenyl group.

(138) Two benzene rings.
Occurrence of two phenyl groups.

(139) More than two benzene rings.
Occurrence of more than two phenyl groups.

(140) One carbocyclic ring.
Occurrence of one individual, fused or aromatic ring (excluding non-fused benzenes) in total molecule.

(141) Two carbocyclic rings.
Occurrence of two carbocyclic or aromatic rings (excluding non-fused benzenes) in total molecule.

(142) More than two carbocyclic rings.
Occurrence of more than two carbocyclic or aromatic rings (excluding non-fused benzenes) in total molecule.

(143) One heterocyclic ring.
Occurrence of one individual heterocyclic ring in total molecule.

(144) Two heterocyclic rings.
Occurrence of two heterocyclic rings in total molecule.

(145) More than two heterocyclics.
Occurrence of more than two heterocyclic rings in total molecule.

K. Special Compound Types

(146) Polypeptide.
Notation begins with #.

(147) Polymer.
Notation begins with /.

L. Extensions

(148) More than one 3-branch carbon atom.
The character Y (but not ∇Y) occurring more than once.

Note: See fragment 6.

M. Fragments Relevant to Company Data Base

(149) M reference.
Set if the CR number has a M reference on the main record or any suffix record.

(150) SC reference.
Set if CR number has a SC reference on the main record or any suffix record.

(151) R reference.
Set if the CR number has an R reference on the main record or on any suffix record.

(152) Live records.
Set for all CR numbers which are present on the file.

Fragments 149-152 may be used for different purposes depending on the data bases.

Note: In the text the following abbreviations are used:

a = alphabetic

b = space

n = numeric

Appendix V - Input Data for Statistical Analysis

INPUT DATA FOR STATISTICAL ANALYSIS

CHEMIST	DATA
CHEM10	1 2 2 3 1 1 4 6 1 17 1.5 37 52 87 6.5 39 18 36 97 -50 -76 -83 -94 250 13 -8 10 0 1 0
CHEM11	1 3 2 3 0 7 15 27 12 44 1.8 11 9 75 1.9 100 50 87.5 100 16 -47 -53 -83 125 74 28 60 1 2 0
CHEM12	1 2 2 3 0 1 7 30 16 53 4.3 38 3 43 4.6 54 41 49 96 -50 -51 -25 -20 131 32 15 25 0 1 0
CHEM13	1 3 3 3 1 0 9 18 10 56 2.3 122 1 53 9.7 41 77 66 91 -100 -58 -65 -28 93 17 34 40 0 2 0
CHEM14	1 3 3 2 0 0 0 0 0 0 0 53 3 89 1.1 66 83 68 100 0 -100 -100 -100 377 24 40 25 0 0 1
CHEM15	1 1 3 3 0 0 3 14 2 14 4.7 58 0 70 5.1 56 39 51 69 0 -81 -83 -75 100 13 -8 7 0 0 0
CHEM16	1 4 3 2 0 4 20 123 94 76 6.2 55 5 57 6.2 63 78 69 94 25 45 324 344 130 20 34 26 0 2 0
CHEM17	1 3 1 1 0 8 6 95 37 39 16 44 0 86 1.9 83 2 48 96 0 0 0 0 100 57 -81 1 0 1 0
CHEM20	2 3 3 3 0 0 0 4 3 75 1.0 71 8 92 5.8 17 57 27 64 -100 -60 -88 -83 230 -5 32 1 0 2 0
CHEM21	2 3 3 3 1 1 9 13 11 85 1.6 12 8 24 1.3 73 59 62 100 0 -53 -80 -47 150 54 40 44 0 2 0
CHEM22	2 3 2 3 1 6 14 27 18 67 1.9 119 3 50 3.1 40 70 56 76 -14 -55 -57 -30 165 29 41 34 0 2 0
CHEM23	2 3 3 2 0 0 11 15 10 67 1.4 96 4 86 7.2 18 50 23 73 -100 -61 -72 -71 70 0 32 0 0 1 0
CHEM24	2 4 2 2 0 8 5 9 7 78 1.8 26 8 71 7.6 32 0 32 100 700 67 40 250 24 14 0 14 00 0 0
CHEM25	2 3 3 3 1 5 14 27 18 67 1.9 54 3 30 2.8 22 39 28 51 -16 -85 -67 -30 97 8 21 10 0 2 0
CHEM26	2 1 3 3 0 1 40 65 56 86 1.7 59 2 71 3.5 38 60 45 70 -40 -24 -48 -37 166 20 42 27 0 1 0
CHEM27	2 3 3 3 1 0 11 16 11 69 1.5 108 0 78 2.1 51 50 51 83 -100 -54 -56 -54 211 33 32 33 0 2 0
CHEM30	3 3 1 1 0 24 10 30 14 50 3 36 0 100 4.8 43 0 43 100 0 0 0 0 100 39 0 39 0 1 0
CHEM31	3 3 2 3 0 7 8 20 7 35 5 30 13 50 8.4 7 7 7 84 600 100 100 100 333 3 3 3 1 2 0
CHEM32	3 3 1 1 0 9 10 23 14 61 2.3 76 0 83 3.9 26 0 22 100 0 0 0 0 100 -4 0 -4 0 1 0
CHEM33	3 3 3 3 0 5 19 49 20 41 2.6 155 8 91 11.7 11 20 11 100 25 -13 0 18 477 6 15 7 0 1 0
CHEM34	3 4 2 2 0 15 25 73 51 68 3.1 42 10 86 9.3 14 17 14 100 0 0 0 0 100 5 3 5 1 0 0
CHEM35	3 3 1 2 1 22 31 198 74 37 6.4 140 4 73 7.1 20 16 19 100 175 92 91 21 205 17 11 16 0 2 1
CHEM36	3 1 3 3 1 2 25 49 24 53 1.9 106 4 74 5.9 14 4 12 100 -33 144 206 225 126 11 3 9 3 2 0
CHEM37	3 3 3 2 1 23 18 103 22 21 5.7 216 0.5 58 6.5 31 46 39 100 -17 -65 -31 -74 140 17 34 26 0 1 0
CHEM38	3 3 3 3 1 10 41 239 140 59 5.8 165 5 67 5.4 14 33 24 47 30 -58 -24 -14 43 -17 3 -5 1 1 0
CHEM39	3 3 3 2 0 1 2 7 5 71 4.5 120 1 98 7.4 27 57 30 93 -93 -98 -97 -98 47 -1 27 0 0 1 0
CHEM40	3 2 3 3 0 1 12 14 11 69 1.3 95 1 90 10.4 69 55 51 100 100 71 75 492 153 29 34 22 0 1 0
CEHM41	4 3 2 3 0 3 5 15 8 53 3.0 38 0 71 4.3 11 60 18 93 200 153 47 43 121 -19 30 -12 0 2 0

Note

Variables in order of occurrence: section, status, time in section, time in department, position in testing programme, number of searches, number of visits, requests for compounds, samples obtained, sample success, number of compounds/visit, compounds made, novel chemistry, compounds acquired, tests per compounds, activity for own compounds, activity for acquired compounds, activity for all compounds, compounds per test, increases in searches, increases in visits, increases in samples requested, increases in samples obtained, increases in compounds made, increase in activity for own compounds, increase in activity for acquired compounds, publications, assistance and literature searching.

Appendix V (Cont.)

INPUT DATA FOR STATISTICAL ANALYSIS

CHEMIST DATA

CHEM42 4 3 2 2 0 3 32 70 39 55 2.2 160 0 96 10.4 40 60 40 100 -50 -20 -48 -35 112 18 39 13 0 2 0

CHEM43 4 3 2 2 0 8 11 21 11 48 2.1 77 0 89 6.8 15 0 14 89 700 267 567 939 219 -15 0 -16 0 2 0

CHEM44 4 4 3 3 0 4 16 24 17 65 1.6 75 0 77 4.6 43 25 39 89 -56 -43 -66 -71 442 23 3 17 0 0 0

CHEM45 4 1 3 3 0 1 12 18 7 39 1.5 49 16 53 12 43 92 61 100 -56 -56 -61 -73 240 22 60 40 0 2 3

CHEM46 4 2 3 3 0 0 2 3 2 67 1.5 77 5 87 7.4 73 92 74 97 -100 -83 -93 -91 260 43 51 44 0 2 0

CHEM47 4 4 3 3 0 1 4 5 1 20 1.3 49 0 74 3.8 78 90 79 100 -53 -53 -88 -95 240 57 59 58 0 0 0

CHEM50 5 3 3 2 1 0 11 13 11 85 1.2 95 5 86 4 4 25 7 97 -100 -52 -87 -68 150 3 24 5 0 2 0

CHEM51 5 4 3 3 1 4 9 29 18 62 3.2 136 1 91 7.7 48 30 47 89 -60 -47 -37 -6 142 13 -5 11 0 1 0

CHEM52 5 3 2 2 1 5 7 18 9 50 2.4 118 4 75 8 0 3 1 88 0 -13 -5 125 100 -1 2 0 2 2 1

CHEM53 5 3 3 3 1 5 25 55 29 53 2.2 131 0 94 1.9 58 33 58 83 67 -26 -38 -45 91 -12 -37 -13 0 2 0

CHEM54 5 4 3 3 1 24 42 105 34 51 3.9 73 0 93 3.2 92 44 80 83 -48 -45 -32 -5 55 12 -26 10 0 3 1

CHEM55 5 4 3 3 0 4 23 43 25 52 2.1 65 20 83 6.4 44 46 49 96 300 293 700 2510 220 -26 -24 -26 0 1 0

CHEM56 5 3 2 2 1 0 4 9 2 25 2 44 25 100 7.2 2.4 0 2.4 100 0 0 0 0 100 1 0 1 0 1 0

CHEM57 5 3 1 1 0 2 0 0 0 0 0 73 3 96 3.2 4 0 4 100 0 0 0 0 0 100 3 0 3 0 1 0

CHEM58 5 1 3 3 0 0 3 50 31 62 14.7 0 0 0 0 0 0 0 0 0 -100 -25 614 933 100 0 0 0 0 1 0

CHEM59 5 3 1 1 0 0 0 0 0 0 0 4 100 100 4.0 0 0 0 100 0 0 0 0 100 0 0 0 0 1 0

CHEM60 6 3 2 2 0 0 0 0 0 0 0 2 0 100 4.5 50 0 50 100 0 0 0 0 100 29 0 29 0 1 0

CHEM61 6 1 3 3 0 1 0 0 0 0 0 22 4 69 5.4 36 40 38 100 100 -100 -100 -100 1300 15 19 17 0 2 0

CHEM62 6 3 3 3 0 0 0 0 0 0 0 0 7 14 50 2.7 0 33 17 100 0 0 0 0 200 0 12 -3 0 1 0

CHEM63 6 4 3 3 0 0 0 0 0 0 0 0 4 25 100 19.0 25 0 25 100 0 0 0 0 100 4 0 4 0 2 0

CHEM64 6 4 3 3 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 100 0 0 0 0 0 0

CHEM65 6 4 3 3 0 0 0 0 0 0 0 1 100 0 0 0 0 0 0 0 0 0 0 100 0 0 0 0 0 0

CHEM66 6 2 3 3 0 0 0 0 0 0 0 4 100 0 0 0 0 0 0 0 0 0 0 100 0 0 0 0 3 0

CHEM67 6 3 3 3 1 4 9 11 7 64 1.9 89 12 56 7.9 13 12 13 100 -20 -18 -69 -56 146 12 11 12 0 1 0

CHEM68 6 2 3 3 1 4 1 1 0 0 1 40 0 40 4.3 0 0 0 100 400 -80 -80 100 150 0 0 0 0 2 0

CHEM69 6 3 3 3 0 5 13 21 12 52 1.4 46 1.5 38 4.6 15 15 17 100 0 -24 -30 -20 56 15 18 17 0 2 1

CHEM70 7 3 3 3 1 0 3 13 7 54 1.4 15 0 100 2.25 0 0 0 100 -100 -27 -69 -72 400 0 0 0 0 1 0

CHEM71 7 3 3 2 1 5 5 10 2 20 1.7 78 2.6 89 1.7 47 71 58 97 100 -65 -57 -79 915 20 24 20 0 2 0

CHEM72 7 3 3 2 1 0 3 3 7 67 1.5 19 5 52 1.9 60 80 70 94 0 -67 -57 -13 400 12 32 22 0 2 0

CHEM73 7 4 3 3 1 5 10 21 6 29 2.1 91 1 72 1.4 45 52 47 100 25 -48 -73

Appendix V (Cont.)

INPUT DATA FOR STATISTICAL ANALYSIS

CHEMIST	DATA
	-84 96 -3 4 -1 0 2 3
CHEM74	7 3 2 2 0 2 1 7 2 29 7 51 2 100 5.6 68 0 68 86 0 0 0 0 100 20 0 20 0 2 0
CHEM75	7 3 3 2 0 4 2 2 1 50 1 61 0 91 1.7 51 0 49 90 100 0 0 0 300 3 0 0 0 2 0
CHEM76	7 1 3 3 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 100 0 0 0 0 0 0
CHEM80	8 4 3 3 0 0 21 27 20 74 1.3 53 9 72 7.0 33 25 30 96 0 -28 -45 -33 100 20 15 20 1 0 0
CHEM81	8 3 1 1 0 0 3 11 6 54 3.7 12 0 100 2.5 50 0 50 100 0 0 0 0 100 40 0 40 0 1 0
CHEM82	8 4 3 3 0 1 6 13 8 44 2 18 16 56 6.5 43 33 40 100 0 0 63 500 100 17 8 15 1 1 0
CHEM83	8 3 2 2 1 8 15 165 62 38 11 107 0 39 4.2 35 30 32 92 164 -25 259 816 100 9 3 5 0 2 0
CHEM84	8 3 3 3 1 7 19 70 34 49 9.5 65 0 67 5.9 18 44 25 75 250 -10 125 88 250 8 34 15 0 2 0
CHEM85	8 3 3 3 1 34 40 232 98 42 5.9 98 1 59 3.0 7.3 5 6.3 99 750 330 1050 1533 100 2 0 1 1 2 0
CHEM86	8 3 2 2 1 15 13 47 14 30 3.6 73 4 78 6.2 18 91 18 93 36 -10 176 133 67 -7 -4 -6 0 1 0
CHEM87	8 3 3 3 1 19 27 48 16 33 1.8 82 0 34 4.0 5 5 5 91 1900 800 960 700 1200 -5 -5 -5 0 1 0
CHEM88	8 1 3 3 0 0 8 14 9 64 1.8 37 19 84 11 0 0 0 100 -100 -50 -50 13 133 -10 -10 -10 2 1 0
CHEM89	8 3 3 3 1 3 2 2 1 50 1 33 3 65 5.7 38 25 33 95 50 -75 -82 -50 100 13 0 8 1 0 0
CHEM90	8 3 3 3 1 4 10 22 15 68 2.2 73 1 62 3.8 36 29 33 100 300 25 100 650 480 0 -6 -2 0 2 0
CHEM91	10 3 3 3 0 0 5 6 2 33 1.2 57 7 84 4.5 92 71 90 96 -100 -62 -87 -95 918 61 40 58 0 2 0
CHEM92	10 3 3 2 0 0 12 14 6 43 2.3 30 3 96 7.3 73 0 69 76 0 -4 0 -25 71 14 0 38 1 2 0
CHEM93	10 4 3 3 0 0 3 3 1 33 1.0 52 0 95 8.2 62 57 63 100 0 -67 -80 -90 117 30 26 29 0 0 0
CHEM94	10 3 1 2 1 2 8 10 3 30 1.3 80 5 91 5.4 74 53 74 84 100 -60 -64 -70 35 45 17 42 2 2 0
CHEM95	10 3 3 3 0 4 6 37 5 14 6.2 84 3.6 77 4.8 91 91 91 94 100 -14 23 -80 103 80 80 80 0 2 1
CHEM96	10 3 3 3 1 3 15 39 14 41 2.4 37 13 67 4.2 97 87 93 88 200 -40 -34 -36 200 65 50 62 0 2 0
CHEM97	10 3 3 2 0 0 6 8 3 39 1.3 70 3 85 5.2 56 71 61 80 0 -70 -78 -68 215 52 41 50 1 2 0
CHEM98	10 4 3 3 0 0 3 6 1 17 2.0 7 0 100 1.6 33 0 33 100 0 0 0 0 100 2 0 2 1 0 0

Appendix VI - Parameters for Statistical Analysis

```
CHEMSVLP
OBSE,,CHEMAM
COL,CHEMAM
   SECTN POSTN SERVSCSERVDPLEADCHSEARCHVISITSREQSTSSMPLESS
   MSCSSCMPVSTCPDSMD
   NOVELCCPDSAQTSTCPDACTHISACTACQACTTOTCPDTSTRELSEARELV
   ISRELREQRELSMPRELCMP
   RELAHSRELAAQRELATTPUBLICASSISTLITSEA
MATR,,CHEMAM
Data Cards
END OF DATA
POBS,33,CHEMAM
PMNS,45,CHEMAM
PMNS,45,CHEMAM,,S
NORM,CHEMAM
PNRM,26,CHEMAM
CROS,CHEMAM
PCRP,66,CHEMAM
COVA,CHEMAM
CORR,CHEMAM
PCOR,16,CHEMAM
PCOR,16,CHEM,,X
FACT,CORR,48,CHEMAM
   NUM,12
   PRI
   ITE,C
   CON,20
   VAR,0.0
   RES,0.000001
   MAX,80
   PSCR,44
GET OFF
****
```

1. Correlation Matrix for Section, Position, Time in Section, Time in Department, Position in Testing Programme, Number of Substructure Searches Requested, Number of Visits to Sample Location Service and Number of Compounds Reqeusted from Sample Location Service

19/32/15 05/08/74 ICL 1900 STATISTICAL ANALYSIS X052/33

CORRELATION MATRIX CHENAM

	SECTN	POSTN	SERVSC	SERVDP	LEADCH	SEARCH	VISITS	REQSTS
SECTN	1.000000	0.152683	0.141780	0.071261	0.058589	-0.095804	-0.140186	-0.116538
POSTN	0.152683	1.000000	-0.079827	-0.136552	0.039448	0.167873	0.039650	0.065699
SERVSC	0.141780	-0.079827	1.000000	0.703643	0.092202	-0.181012	0.080198	-0.063739
SERVDP	0.071261	-0.136552	0.703643	1.000000	0.169231	-0.126803	0.135605	-0.024242
LEADCH	0.058589	0.039448	0.092202	0.169231	1.000000	0.356527	0.342642	0.338530
SEARCH	-0.095804	0.167873	-0.181012	-0.126803	0.356527	1.000000	0.657221	0.792704
VISITS	-0.140186	0.039650	0.080198	0.135605	0.342642	0.657221	1.000000	0.792869
REQSTS	-0.116538	0.065699	-0.063739	-0.024242	0.338530	0.792704	0.792869	1.000000
SMPLES	-0.196322	0.085885	-0.004172	0.009448	0.287155	0.698252	0.819741	0.948174
SMSCSS	-0.269297	0.057542	0.116218	0.043966	0.153384	0.486304	0.417335	0.492504
CMPVST	-0.126156	-0.098680	-0.215962	-0.182791	0.051083	0.323179	0.195038	0.338737
CPDSMD	-0.243614	0.062704	0.092961	-0.049452	0.414170	0.402084	0.513332	0.473635
NOVELC	0.002835	-0.004245	-0.072229	-0.004497	-0.130851	-0.182218	-0.209028	-0.170846
CPDSAQ	0.076371	0.205337	-0.251408	-0.367325	-0.040880	-0.046408	0.050058	-0.111426
TSTCPD	-0.059729	-0.020010	0.020957	0.042104	-0.091852	-0.015936	0.070403	-0.010802
ACTHIS	0.042763	0.149740	0.023624	-0.017401	-0.072245	-0.095973	0.009673	-0.019031
ACTACC	-0.107241	0.015117	0.353390	0.243917	0.020690	-0.151704	0.104209	0.007102
ACTTOT	0.053372	0.136795	0.102542	0.033878	-0.067464	-0.100728	0.027717	-0.033063
CPDTST	-0.024696	0.187760	-0.214322	-0.218661	0.130377	0.107833	0.100450	0.081978
RELSEA	0.155489	0.088622	-0.028832	0.058176	0.104149	0.403315	0.219494	0.181004
RELVIS	0.101192	0.078329	-0.052941	0.035232	0.060298	0.347687	0.316494	0.202496
RELREQ	0.106522	0.007611	0.010799	0.039266	0.068445	0.420460	0.353034	0.413387
RELSMP	0.046262	0.090207	0.080978	0.075499	-0.005048	0.246779	0.241215	0.315067
RELCMP	0.054126	-0.179111	0.205675	0.180077	-0.015569	0.020709	0.001602	-0.092275
RELAMS	0.044433	0.014300	-0.102858	-0.061404	-0.112068	-0.129505	-0.105388	-0.097795
REL/AQ	-0.057012	-0.095451	0.265090	0.162356	-0.105033	-0.229972	-0.044441	-0.141997
RETATT	0.069976	-0.010203	0.021376	0.014103	-0.077628	-0.146511	-0.065381	-0.125715
PUBLIC	0.189317	-0.118345	-0.056615	0.059829	0.111111	0.067846	0.180011	0.093678
ASSIST	0.061193	-0.155564	0.047377	0.043051	0.302634	0.105345	0.192582	0.180667
LITSEA	-0.049421	-0.158049	0.032025	0.044538	-0.023156	0.085902	0.113838	0.113042

Note: Significant correlations are underlined.

2. Correlation Matrix for Samples Obtained from Location Service, Success in Obtaining Samples from Location Service, Number of Samples Requested per Visit, Number of Compounds Made, % of Compounds made by Novel Chemistry, Number of Compounds Acquired for Test, Number of Tests per Compound and Activity of Own Compounds

19/32/14 05/06/74 ICL 1900 STATISTICAL ANALYSIS XDS2/33

CORRELATION MATRIX CHEMAM

	SMPLES	SMSCSS	CMPVST	CPDSMD	NOVELC	CPDSAQ	TSTCPD	ACTHIS
SECTN	-0.196322	-0.269297	-0.128156	-0.245614	0.002835	0.074371	-0.059729	0.042763
POSTN	0.083985	0.057542	-0.098660	0.062704	-0.004245	0.205357	-0.020010	0.149740
SERVSC	-0.004172	0.118218	-0.215962	0.092981	-0.072229	-0.281486	0.020957	0.023624
SERVOP	0.009448	0.043966	-0.182791	-0.049452	-0.004497	-0.367325	0.042104	-0.017061
LEADCH	0.267153	0.135584	0.051083	0.414170	-0.130851	-0.046880	-0.091832	-0.072243
SEARCH	0.698232	0.080304	0.323179	0.402084	-0.182218	-0.046408	-0.015956	-0.059973
VISITS	0.819741	0.417333	0.195038	0.513352	-0.209028	0.050058	0.070403	0.009873
REQSTS	0.726174	0.192564	0.538757	0.473655	-0.170846	-0.111428	-0.010602	-0.019831
SMPLES	1.000000	0.324104	0.466840	0.409277	-0.162598	-0.106509	-0.004394	-0.025370
SMSCSS	0.324104	1.000000	0.170533	0.310115	-0.357874	0.196108	0.077271	0.099817
CMPVST	0.466840	0.170533	1.000000	0.169800	-0.231272	-0.215355	-0.121816	0.110798
CPDSMD	0.409277	0.310115	0.169800	1.000000	-0.299746	0.216214	0.224946	0.013341
NOVELC	-0.162598	-0.357874	-0.231272	-0.299746	1.000000	-0.200948	-0.061127	-0.265812
CPDSAQ	-0.106509	0.196108	-0.215355	0.216214	-0.200948	1.000000	0.396681	0.278400
TSTCPD	-0.004394	0.077271	-0.121816	0.224946	-0.061127	0.396681	1.000000	-0.074684
ACTHIS	-0.025370	0.099817	0.110798	0.013341	-0.265812	0.278400	-0.074684	1.000000
ACTACQ	0.042701	0.219521	-0.084978	0.271302	-0.210830	0.061102	0.007557	0.604012
ACTTOT	-0.021555	0.113439	0.010586	0.061869	-0.277749	0.291905	-0.040743	0.970100
CPDTST	0.050074	0.195693	-0.090308	0.264657	-0.328604	0.072717	0.431487	0.289357
RELSEA	0.106391	-0.103915	0.028779	-0.055870	-0.084575	-0.156285	0.005391	-0.190831
RELVIS	0.150269	-0.001919	0.046406	-0.045306	-0.015743	-0.100220	-0.006135	-0.208250
RELREQ	0.359863	0.043673	0.341421	-0.007160	-0.056556	-0.192558	-0.068508	-0.232883
RELSMP	0.315953	0.137334	0.249705	-0.004844	-0.024745	-0.165550	-0.002802	-0.140400
RELCMP	-0.117592	-0.115328	-0.150255	-0.033170	-0.074681	-0.037436	0.061037	-0.041984
RELAMS	-0.130784	0.009073	0.080627	-0.112869	-0.140011	0.097025	-0.071199	0.709939
RELAAQ	-0.115534	0.103667	-0.211496	0.160445	-0.129821	0.000608	0.105473	0.322018
RETATT	-0.132614	0.059453	-0.076157	-0.028011	-0.156306	0.120959	0.021279	0.721622
PUBLIC	0.111740	0.096953	-0.030472	0.010818	-0.034815	0.098750	0.144977	-0.090162
ASSIST	0.157091	0.135051	0.076569	0.210304	-0.008054	0.029603	0.052456	0.182041
LITSEA	0.063758	-0.075507	0.017640	0.090222	-0.020093	-0.037250	0.155079	0.081742

Note: Significant correlations are underlined.

3. Correlation Matrix for Activity for Acquired Compounds, Activity on all Compounds, Compounds per Test, Increase in Number of Substructure Searches, Increases in Visits to Sample Location Service, Increase in Samples Requested, Increase in Compounds Obtained and Increase in Number of Compounds Made

19/32/17 05/08/74 ICL 1900 STATISTICAL ANALYSIS XDS2/33

CORRELATION MATRIX CHEMAM

	ACTACQ	ACTTOT	CPDTST	RELSEA	RELVIS	RELREQ	RELSMP	RELCMP
SECTN	-0.107241	0.053372	-0.024696	0.155469	0.101192	0.106522	0.046262	0.054126
POSTN	0.015117	0.138795	0.187760	0.088622	0.078529	0.007611	0.090207	-0.179111
SERVSC	0.353340	0.102842	-0.214322	-0.028852	-0.052941	0.010799	0.080978	0.205875
SERVDP	0.243917	0.033878	-0.218681	0.058176	0.035232	0.039266	0.075499	0.180077
LEADCM	0.020690	-0.067444	0.130377	0.164149	0.060298	0.068445	-0.005048	-0.015569
SEARCM	-0.151704	-0.106728	0.147853	0.403315	0.347687	0.420480	0.240779	0.020709
VISITS	0.104269	0.027717	0.160450	0.219494	0.316494	0.353056	0.291213	0.001802
REQSTS	0.007102	-0.035083	0.081978	0.181004	0.202496	0.413567	0.315069	-0.092273
SMPLES	0.042701	-0.021555	0.050074	0.106391	0.150289	0.359883	0.315933	-0.117342
SMSCSS	0.219521	0.113439	0.195693	-0.063915	-0.001919	0.043673	0.137334	-0.115528
CMPVST	-0.084978	0.010506	-0.090368	0.028779	0.046406	0.341421	0.249005	-0.130255
CPDSMD	0.271302	0.061809	0.264637	-0.055870	-0.045308	-0.007180	-0.004644	-0.055170
NOVELC	-0.210830	-0.277749	-0.328264	-0.084575	-0.015745	-0.056556	-0.024745	-0.074681
CPDSAQ	0.001102	0.291905	0.072717	-0.156265	-0.100220	-0.192558	-0.123530	-0.037456
TSTCPD	0.007557	-0.040745	0.431437	0.005391	-0.006135	-0.008506	-0.002802	0.001657
ACTMIS	0.264012	0.970100	0.289357	-0.190831	-0.268256	-0.252863	-0.146400	-0.041984
ACTACQ	1.000000	0.713866	0.191388	-0.213594	-0.285854	-0.247496	-0.119733	0.076268
ACTTOT	0.713866	1.000000	0.286941	-0.212282	-0.294171	-0.276516	-0.161559	-0.020403
CPDTST	0.191388	0.286941	1.000000	0.091531	0.045564	-0.073365	-0.007455	0.134257
RELSEA	-0.213594	-0.212282	0.091531	1.000000	0.655845	0.686543	0.460197	0.477594
RELVIS	-0.285854	-0.294171	0.045564	0.655845	1.000000	0.805108	0.606604	0.379771
RELREQ	-0.247496	-0.276516	-0.073365	0.686543	0.805108	1.000000	0.661734	0.192393
RELSMP	-0.119733	-0.161559	-0.007455	0.460197	0.606604	0.661734	1.000000	0.082419
RELCMP	0.076268	-0.020403	0.134257	0.477594	0.379771	0.192393	0.082419	1.000000
RELAMS	0.448315	0.719263	0.204155	-0.207564	-0.289967	-0.319053	-0.329532	-0.030039
RELAAQ	0.776318	0.450999	0.113454	-0.196464	-0.263125	-0.275687	-0.273203	0.075151
RETATT	0.379770	0.758062	0.179640	-0.236574	-0.323812	-0.349972	-0.330657	-0.026048
PUBLIC	-0.184408	-0.116550	0.109636	-0.008057	0.078628	0.058008	0.032381	-0.147018
ASSIST	0.277614	0.223225	0.084251	0.043669	0.022957	0.063379	0.047404	0.045967
LITSEA	0.222776	0.145651	0.100964	-0.075468	-0.066013	-0.083580	-0.092818	-0.006016

Note: Significant correlations are underlined.

4. Correlation Matrix for Increase in Activity for Own Compounds, Increase in Activity for Acquired Compounds, Increase in Activity for all Compounds, Number of Publications, Number of Assistants and Use of Computerised Compound-Oriented Literature Services

19/32/18 05/08/74 ICL 1900 STATISTICAL ANALYSIS XDS2/33

CORRELATION MATRIX CHEMAM

	RELAHS	RELAAQ	RETATT	PUBLIC	ASSIST	LITSEA
SECTN	0.044433	-0.057012	0.069976	0.189317	0.061193	-0.049421
POSTN	0.014300	-0.095431	-0.010203	-0.118345	-0.153564	-0.156029
SERVSC	-0.102858	0.265090	0.021376	-0.056615	0.047377	0.052025
SERVDP	-0.061404	0.102536	0.014103	0.059829	0.043051	0.044330
LEADCH	-0.112068	-0.105053	-0.077828	0.111111	0.302434	-0.023156
SEARCH	-0.129505	-0.229972	-0.146511	0.067846	0.195345	0.085302
VISITS	-0.105360	-0.044441	-0.063561	0.180011	0.192382	0.113838
REGSTS	-0.097795	-0.141997	-0.125715	0.093676	0.186669	0.113042
SMPLES	-0.130784	-0.115536	-0.132014	0.111740	0.157091	0.063758
SMSCSS	0.009073	0.163667	0.059453	0.098353	0.133651	-0.075567
CMPVST	0.086627	-0.211496	-0.076157	-0.030472	0.076589	0.017040
CPDSRD	-0.112869	0.160445	-0.020311	0.010418	0.010364	0.090222
ADVELC	-0.146011	-0.129861	-0.156306	-0.034815	-0.008054	-0.020093
CPDSAQ	0.097625	0.000068	0.126959	0.098750	0.029663	-0.057250
TSTCPD	-0.071199	0.105473	0.021279	0.164977	0.052456	0.155079
ACTHIS	0.769959	0.522018	0.721622	-0.090182	0.186441	0.081742
ACTACQ	0.446515	0.778318	0.579770	-0.184408	0.227614	0.222776
ACTTOT	0.719203	0.650944	0.758062	-0.118550	0.223225	0.145651
CPDTST	0.204155	0.113454	0.179640	0.109656	0.084231	0.100964
RELSEA	-0.207564	-0.190484	-0.230594	-0.098037	0.043869	-0.073408
RELVIS	-0.289967	-0.263163	-0.329412	0.078628	0.022957	-0.066013
RELREQ	-0.319033	-0.275867	-0.349972	0.058506	0.063379	-0.083580
RELSMP	-0.325552	-0.275205	-0.350637	0.032361	0.047404	-0.092618
RELCMP	-0.036039	0.073151	-0.020048	-0.147016	0.045967	-0.006016
RELAHS	1.000000	0.493570	0.910955	0.002067	0.070496	0.089756
RELAAQ	0.493570	1.000000	0.695767	-0.131411	0.161304	0.240319
RETATT	0.910955	0.695767	1.000000	-0.010461	0.147962	0.176910
PUBLIC	0.002067	-0.131411	-0.010461	1.000000	0.020728	-0.012964
ASSIST	0.070496	0.161304	0.147962	0.020728	1.000000	0.181913
LITSEA	0.089956	0.240319	0.176910	-0.012964	0.181913	1.000000

PCOR,16,CHEMAM,,X

Note: Significant correlations are underlined.

5. First Stage Factor Analysis Giving Factors 1, 2 and 3 for First 21 Variables

```
                    19/57/18    05/08/74     ICL  1900 STATISTICAL ANALYSIS     ADSC/33
                    FACTOR ANALYSIS OF  CORR  OF  CHEMAN

MAXIMUM NO OF ITERATIONS REACHED   80
NUMBER OF ITERATIONS   80
WEIGHTED SUM OF SQUARES OF RESIDUALS  20,9406913

MATRIX OF FACTOR LOADINGS
```

VARIABLE	COMMUNALITY	SPECIFIC VARIANCE	FACTOR 1	FACTOR 2	FACTOR 3
EIGENVALUES OF R MATRIX			1188,17733313	387,12579724	139,10009723
VARIANCE			0,07451394	0,02197884	0,01503220
SECTN	0,24947964	0,75052036	0,04939466	0,07318879	0,21584104
POSTN	0,16023448	0,83976552	0,10217466 -	0,11059451	0,06429313
SERVSC	0,67765924	0,32234076	0,07204597	0,01794239	0,12100041
SERVOP	0,80626324	0,19373676	0,01951053 -	0,00404988	0,08176710
LEADCN	0,39824976	0,60175024 -	0,09016022 -	0,28758064 -	0,15810934
SEARCH	0,78193529	0,21806471 -	0,17101069 -	0,78195195 -	0,13570470
VISITS	0,87615186	0,12384814 -	0,04613513 -	0,78871357 -	0,10429770
REQSTS	0,99634527	0,00365473 -	0,10916715 -	0,94891958 -	0,27230380
SMPLES	0,98377736	0,01622264 -	0,10102158 -	0,90887422 -	0,28800236
SMSCSS	0,70899552	0,29100448	0,09286598 -	0,22853100 -	0,08223133
CMPVST	0,91742178	0,08257822 -	0,01894425 -	0,55428592 -	0,03000749
CPDSMD	0,71521705	0,28478295	0,01928004 -	0,41004281 -	0,27634542
NOVELC	0,34329560	0,65670040 -	0,24457814	0,22784102 -	0,05924102
CPDSAQ	0,82457848	0,17542152	0,26737185	0,06081341	0,04430490
TSTCPD	0,47732128	0,52267872 -	0,03237299	0,03750143 -	0,07933547
ACTNIS	0,99151341	0,00848659	0,95784620 -	0,11087899	0,10937398
ACTACQ	0,96875510	0,03124490	0,70384193 -	0,06759894 -	0,01871012
ACTTOT	0,99876488	0,00123592	0,97469935 -	0,10500369	0,10343213
CPDTST	0,75871154	0,24128844	0,27271182 -	0,11138280	0,00340378
RELBEA	0,93397502	0,06602498 -	0,27174232 -	0,31222275	0,35263350
RELVIS	0,87497730	0,12502284 -	0,30525187 -	0,35100699	0,01659479

6. First Stage Factor Analysis Giving Factors 1, 2 and 3 for Variables 22-30

19/57/19 05/08/74 ICL 1900 STATISTICAL ANALYSIS R082/33
FACTOR ANALYSIS ON CORR OF CHEMAP

VARIABLE	COMMUNALITY	SPECIFIC VARIANCE	FACTOR 1	FACTOR 2	FACTOR 3
RELMED	0.99439075	0.00560925	- 0.37304341	- 0.88094225	0.04302300
RELSHP	0.84696072	0.15303328	- 0.27533326	- 0.49738793	0.03191224
RELCPP	0.37643520	0.62356480	- 0.03837722	0.02350765	0.28270512
RELANS	0.99339126	0.00660880	0.82549047	0.05030407	- 0.17569125
RELAAQ	0.92965772	0.07034228	0.52929263	0.12460342	- 0.17421017
RETATT	0.99527057	0.00472943	0.85733042	0.07500837	- 0.18273994
PUBLIC	0.19429662	0.80570338	- 0.09888472	- 0.06092028	- 0.09336100
ASSIST	0.15844879	0.84155121	0.19088252	- 0.22135334	0.02960773
LITSEA	0.15299981	0.84700019	0.14830369	- 0.08959735	- 0.12952635

7. First Stage Factor Analysis Giving Factors 4-8 for First 25 Variables

19/57/20 03/08/74 ICL 1900 STATISTICAL ANALYSIS X052/33
FACTOR ANALYSIS OF CORR OF CMEMAP

VARIABLE	FACTOR 4	FACTOR 5	FACTOR 6	FACTOR 7	FACTOR 8
EIGENVALUES OF R MATRIX	109.85667602	85.33170600	14.76175143	11.03706944	8.17716001
VARIANCE	0.00284833	0.01070726	0.00446937	0.00133320	0.01010770
SECTN	0.11090953	0.05296220	0.15724097	0.11189640	0.18503043
POSTN -	0.15469130	0.10557019	0.22315521	0.02298757 -	0.08979872
SERVSC -	0.13153297 -	0.46389565 -	0.17734219 -	0.19466102	0.39956044
SERVOP -	0.03912665 -	0.37451748 -	0.11460773 -	0.26328802	0.32479141
LEADCH -	0.05971062 -	0.06825930	0.22367089 -	0.14344185	0.04879738
SEARCH	0.06243107	0.00874024	0.33107500 -	0.02833875 -	0.01667943
VISITS -	0.45036904 -	0.17948800	0.21741562	0.01863268	0.00888005
REQSTS -	0.01939995	0.01237668	0.01527082 -	0.01210797 -	0.00109202
SMPLES -	0.08817074 -	0.05732253 -	0.02442668	0.05423166	0.04844128
SMSCSS -	0.08332643 -	0.19653573 -	0.09656984	0.05725044 -	0.22730274
CMPYST	0.08769901	0.38127445 -	0.47630431 -	0.07150994 -	0.14549343
CPOSMB -	0.22273527 -	0.27575283	0.10310644 -	0.05413805 -	0.30800270
MOVELC	0.06353455	0.02664354 -	0.06722279	0.04343812	0.22802703
CPOSAC -	0.25241029	0.11102402	0.38350074	0.38372241 -	0.57833193
TSTCPB -	0.01907013 -	0.10762509	0.25110015	0.13647410 -	0.32114816
ACTMIS -	0.09114908	0.19340168 -	0.01879959 -	0.04838790 -	0.00570527
ACTACC -	0.13398595 -	0.55737739 -	0.18870507 -	0.27720695 -	0.13917310
ACTTOT -	0.13237191 -	0.00907840	0.04680764	0.01140396	0.00667805
CPETST -	0.07718597	0.04773132	0.42379106 -	0.00795334 -	0.61000544
RELSEA	0.28497130 -	0.02062057	0.48534253 -	0.34032801	0.06001009
RELVIS	0.29577769 -	0.01839551	0.33778730 -	0.15565577 -	0.00588017
RELREQ	0.29633562 -	0.02073070 -	0.05334465	0.03971072 -	0.01099709
RELSKP	0.09010114 -	0.05061013 -	0.19197067	0.11065017 -	0.06531495
RELCMP	0.12089435 -	0.16192050	0.29673582 -	0.36023865	0.02117009
RELANS	0.46775964	0.20065190 -	0.06315488 -	0.09795831 -	0.02921911

8. First Stage Factor Analysis Giving Factors 4-8 for Variables 26-30

15/57/22 05/08/74 TCL 1900 STATISTICAL ANALYSIS XDS2/33
FACTOR ANALYSIS ON CURR CP CHEMAP

VARIABLE	FACTOR 4	FACTOR 5	FACTOR 6	FACTOR 7	FACTOR 8
RELAAR	0,29246086	- 0,67749225	- 0,11014642	- 0,11413895	- 0,16675749
RETATT	0,43595105	- 0,13235084	0,03712907	0,08396514	0,02536781
PUBLIC	0,13578165	0,08314253	0,10701720	0,17054025	0,00607489
ASSIST	- 0,05337580	- 0,10404026	0,04678489	0,02211458	- 0,01107784
LITSEA	0,04184499	- 0,21089635	0,02456632	0,00484513	- 0,02082780

9. First Stage Factor Analysis Giving Factors 9-12 for First 25 Variables

19/57/22 05/08/74 ITL 1900 STATISTICAL ANALYSIS X0S4/33
FACTOR ANALYSIS OF CURR OF CHEMA

VARIABLE	FACTOR 9	FACTOR 10	FACTOR 11	FACTOR 12
EIGENVALUES OF B MATRIX	6,55087486	4,27617121	3,57457810	1,92698119
VARIANCE	0,04155549	0,00205017	0,00200640	0,00111128
SECTN	0,14992532	0,25950126	0,08387630	0,11403191
POSTN -	0,07620779	0,05502361 -	0,17122807 -	0,07421778
SERVSC -	0,20605290	0,14173274	0,31985185	0,08295708
SERVOP -	0,25336819	0,18856074	0,40042552	0,16574719
LEADCH	0,05889580	0,10081271	0,33041965 -	0,27554550
SEARCH	0,13909295	0,02687134 -	0,00240230 -	0,12091070
VISITS -	0,31478884 -	0,05237689	0,11471651 -	0,10354186
REGSTS	0,03461430	0,03986469	0,00502809 -	0,00127370
SMPLES -	0,15132697 -	0,11548837 -	0,06815236	0,04130767
SMSCSS -	0,43220019 -	0,49067580	0,27293705 -	0,16971513
CMPVST	0,28869346 -	0,27995928	0,16947094	0,07186031
CPDSMD	0,07899099	0,05773632	0,32469911 -	0,20474335
NOVELC	0,00470733	0,22296447 -	0,33833231 -	0,01788041
CPDSAG -	0,15621695	0,13117600	0,07149580	0,09009957
TSTCPD -	0,09762885	0,05400478	0,28296874	0,39649824
ACTMIS -	0,03518485	0,00464485	0,01321618 -	0,01407380
ACTACG -	0,02423329	0,03560333 -	0,05206099 -	0,01755556
ACTTOT	0,00810769 -	0,00169039 -	0,00210164	0,00200071
CPDTST -	0,17329143	0,10074116	0,13320892	0,22444466
RELSEA	0,09537909 -	0,14866219 -	0,07100880	0,03870418
RELVIS -	0,06833057 -	0,06626545 -	0,02181013 -	0,05209017
RELREG -	0,00657885	0,01556993 -	0,00059239 -	0,00334200
RELSHP -	0,17514090	0,07523040 -	0,05114300	0,01570094
RELCMP	0,01461125 -	0,02826829	0,05225063	0,17850463
RELANS -	0,03535542	0,02523973 -	0,01513042	0,00112135

10. First Stage Factor Analysis Giving Factors 9-12 for Variables 26-30

19/57/24 09/08/78 ICL 1900 STATISTICAL ANALYSIS XDS2/35
FACTOR ANALYSIS OF CORR OF CHEMAP

VARIABLE	FACTOR 9	FACTOR 10	FACTOR 11	FACTOR 12
RELAAG	0,06813716	- 0,02758146	- 0,05374463	0,03705964
RETATT	0,02750317	- 0,01563605	0,01482732	- 0,00188818
PUBLIC	- 0,25161449	0,07229668	0,18857424	0,03147428
ASSIST	0,12038691	- 0,00035241	0,12934242	- 0,09154049
LITSEA	0,20214660	0,09368839	0,02746893	0,10163687

SET OFF

BIBLIOGRAPHY

1. LEAKE, C D - Primary Journals: Questionable Progress and Present Problems. J. Chem. Doc. V10 (1), p.27-29, 1970.

2. GUSHEE, D E - Reading Behaviour of Chemists. J.Chem.Doc. V.8 (4), p.191-194, 1968.

3. DE SOLLA PRICE, D J - Science Since Babylon. Yale University Press, New Haven, Conn., 1961.

4. DE SOLLA PRICE, D J - Little Science, Big Science. Columbia University Press, New York, 1963.

5. HERSCHMAN, A - The Primary Journal: Past, Present and Future. J.Chem.Doc. V.10 (1), p.37-42, 1970.

6. MERTON, R K - Priorities in Scientific Discovery: A Chapter in the Sociology of Science. American Sociological Review, V.22 (6), p.635-659, 1957.

7. MERTON, R K and LEWIS, R - The Competitive Pressures: 1. Race for Priority. Impact of Science on Society, V.21 (2), p.151 -161, 1971.

8. CRANE, D - Information Needs and Uses. Annual Review of Information Science and Technology, V.6 1971 ed. Cuadra, CA.

9. PERRUCCI, R and ROTHMAN, R A - Obsolescence of Knowledge and the Professional Carreer. The Engineers and the Social System Wiley, New York, 1969, ed. Perrucci, R. and Gerstl, J E.

10. ALLEN, T J - Communication Networks in R & D Laboratories. R & D Management , V.1 (1), p14-21, 1970.

11. GERSTBERGER, P - The Presentation and Transfer of Technology in a Research and Development Organisation. Ph.D. Thesis, Massachusetts Institute of Technology, 1971.

12. SMITH, C G - Scientific Performance and the Composition of the Research Team. Admin.Sc.Quarterly, V.16 (4), p.486-495, 1971.

13. VOIGHT, M - Scientists Approach to Information. ARCL Monograph No. 24, ALA, Chapter 5. 1961.

14. MENZEL, H - Formal and Informal Satisfaction of the Information Requirements of Chemists. Columbia University, New York, 1970.

15. ROSENBERG, V - Factors Affecting the Preferences of Industrial Personnel for Information Gathering Methods. Inform.Storage and Retrieval, V.3 (3), p.119-127, 1967.

16. ALLEN, T J and GERSTBERGER, P G - Criteria used by Research and Development Engineers in the Selection of an Information Source. J.Applied Psychology, V.52 (4), p.272-279, 1968.

17. BACK, K W - The Behaviour of Scientists - Communication and Creativity. Sociological Enquiry, V.32 (1), p.82-87, 1962.

18. Survey of Information Needs of Physicists and Chemists. J.Doc. V21 (2), p.83-112, 1965.

19. ROWLETT, R J , TATE, F A and WOOD, J L - Relationships Between Primary Publications and Secondary Information Services. J.Chem.Doc. V.10 (1), p.32-37, 1970.

20. WIGHTMAN, J P - Chemical Titles as an Aid to Current Chemical Literature. J.Chem.Doc. V.1 (3), p.16-17, 1961.

21. WILLIAMS, M E and STEWART, A K - ASIDIC Survey of Information Centre Services. ITT Research Institute, Chicago, 1972.

22. KENT, A K - The Chemical Society Research Unit in Information Dissemination and Retrieval. Svensk Kemisk Tidskrift, V.80, p.39-45, 1968.

23. HAYGARTH-JACKSON, A R - Utilisation of Mechanised Services. Information Storage and Retrieval, V.6, p.53-71, 1970.

24. SLATER, M, OSBORN, A M and PRESORAS, A - Data and the Chemist. Aslib Occassional Publication No. 10, Aslib, London, 1972.

25. PHILIPS, R - Survey of Research Workers Use of Data. J. of Industrial Research, V.9 (1), p.30-35, 1967.

26. OSTI - Data Activities in Britain. Office of Scientific and Technical Information, London, 1969.

27. CODATA - International Compendium of Numerical Data Projects CODATA, Springer-Verlag, 1969.

28. NSRD - Report of the National Standard Reference Data System NBS, Washington, 1970.

29. EDMONDS, B - An Examination of Data and Information Sources in the Thermodynamics of Dense Fluid Mixtures. Final report, Postgraduate School of Librarianship and Information Science, University of Sheffield, 1974.

30. KENNARD, O, WATSON, D G and TOWN, W G - Cambridge Crystallographi data Centre: I Bibliographic File. J.Chem.Doc. V.12 (1), p.14-19, 1972.

31. ALLEN, F H, KENNARD, O, MOTHERWELL, W D S, TOWN, W G and WATSON, D G - Cambridge Crystallographic Data Centre: II. Structural Data File. J.Chem.Doc. V.13(3),p.119-123, 1973.

32. EAKIN, D R, FAULKNER, D A, HYDE, E and WARD S A - Report on Forum on Internal Data Banks. AIOPI Conference, Nottingham, 1975.

33. DONALDSON, N, POWELL, W H, ROWLETT, R H, WHITE, R W and YORKA, K H - Chemical Abstracts Index Names for Chemical Substances in the Ninth Collective Period (1972-1976). J.Chem. Doc. V.14 (1), p.3-15, 1974.

34. LYNCH, M F, HARRISON, J M, TOWN, W G and ASH, J E - Computer Handling of Structure Information. MacDonald/American Elsevier, 1971.

35. Survey of Chemical Notation Systems. National Academy of Sciences, National Research Council Pub. 1150, Washington, D.C., 1964.

36. Survey of European Non-Conventional Chemical Notation Systems National Academy of Sciences, National Research Council Pub. 1278, Washington, D.C., 1965.

37. SMITH, E G - The Wiswesser Line-Formula Chemical Notation. McGraw-Hill, New York, 1968.

38. HYDE, E, MATTHEWS, F W, THOMSON, L H and WISWESSER, W J - Conversion of Wiswesser Notation to a Connectivity Notation for Organic Compounds. J.Chem.Doc. V.7 (4), p.200-204, 1967.

39. THOMSON, L H, HYDE, E and MATTHEWS F W - Organic Search and Display Using a Connectivity Matrix Derived From Wiswesser Line Notation. J.Chem.Doc.V.7 (4), p.204-209, 1967.

40. HYDE, E and THOMSON, L H - Structure Display. J.Chem.Doc. V.8 (3), P.138-146, 1968.

41. FARRELL, C D, CHAUVENET, A R and KONIVER, D A - Computer Generation of Wiswesser Line Notations. J.Chem.Doc. V.11 (1). p.52-59, 1971.

42. BOWMAN, C M, LANDEE, F A, LEE, N W, RESLOCK, M H and SMITH B P - A Chemically-Oriented Information Storage and Retrieval System III. Searching a Wiswesser Line Notation. J.Chem.Doc. V.10 (1), p.50-54, 1970.

43. BOND, V B, BOWMAN, C M, LEE, N W, PETERSON, D R and RESLOCK, M H - Interactive Searching of a Structure and Biological Activity File. J.Chem.Doc. V.11 (3), p.168-170, 1971.

44. GARFIELD, E, REVESZ, G S, GRANITO, C E, DORR, H A, CALDERON, M M and WARNER, A - Index Chemicus Registry System. Pragmatic Approach to Substructure Chemical Retrieval. J.Chem.Doc. V.10 (1), p.54-58, 1970.

45. LIN, N and GARVEY, W D - Information Needs and Uses. Annual Review of Information Science and Technology, V.7. Chapter 1, ed. Cuadra, C.A., 1972.

46. LIPETZ, B - Information Needs and Uses. Annual Review of Information Science and Technology, V.5, Chapter 1, ed. Cuadra, C.A., 1970.

47. WOOD, D N - User Studies - A Review of the Literature from 1966 - 1970. Aslib Proceedings , V.23 (1), p.11-23, 1971.

48. BARNES, R C M - Information User Studies: 2. Comparison of Some Recent Surveys. J.Doc. V.21 (3), p.169-176, 1965.

49. LANCASTER, F W - MEDLARS: Report on the Evaluation of its Operating Efficiency. American Documentation, V.20 (2), p.119-142, 1969.

50. RUBENSTEIN, R I and SCHULTZ, L - Evaluation of Usage of a Custom Biological Literature Search Service: Three-Year Study Proceedings of American Society for Information Science, V.5, p.317-322, 1968.

51. DUBINSKAYA, S A - Investigation of Information Service Needs of Chemical Specialists. Nauchno-Techni.Inform., V.2 (4), p.3-6, 1967.

52. SLATER, M - Type of Use and User in Industrial Libraries. J.Doc. V.19 (1), p.12-18, 1963.

53. O'CONNOR, J - Some Questions Concerning Information Needs. American Documentation, V.19, p.200-210, 1968.

54. CLEVERDON, C W - Evaluation Tests of Information Retrieval Systems. J.Doc. V.26 (1), p.55-67, 1970.

55. BRISNER, O - A Model for Evaluating an Information Retrieval System in the Case of Result Preparation and User Convenience. Tidsk.Dok, V.30, p.10-12, 1972.

56. FARRADANE, J - The Evaluation of Information Retrieval Systems. J.Doc. V.30 (2) , p.195-209, 1974.

57. FELDMANN, R J , HELLER, S R and SHAPIRO, K P - An Application of Interactive Computing - A Chemical Information System. J.Chem.Doc. V.12 (1), p.41-47, 1972.

58. FELDMANN, R J and Koniver, D A - Interactive Searching of Chemical Files and Structure Diagram Generation from Wiswesser Line Notation. J.Chem.Doc. V.11 (3), p.154-159, 1971.

59. FELDMANN, R J and HELLER, S R - An Application of Interactive Computing: The Nested Retrieval of Chemical Structures. J.Chem.Doc. V.12 (1), p.48-54, 1972.

60. HALL, J L, NEGUS, A E and NANCY, C J - On-Line Information Retrieval: A Method of Query Formulation Using a Video Terminal. Program, V.6, p.175-186, 1972.

61. NEGUS, A E and HALL, J L - Towards an Effective On-Line Reference Retrieval System. Information Storage and Retrieval V.7, p.249-270, 1971.

62. FISHENDEN, R M - Information Use Studies, Past Results and Future Needs. J.Doc. V.21, p.163-168, 1965.

63. MITTMAN, B and DOMINICK, W D - Developing Monitor Techniques for an On-Line Information Retrieval System. Information Storage and Retrieval, V.9, p.291-307, 1973.

64. STANDERA, O R - Cost and Effectiveness in the Evolution of an Information System: A Case Study. J.Am.Soc.for Inf.Sci V.25, p.203-207, 1974.

65. MEISTER, D and SULLIVAN, D J - Evaluation of User Reactions to a Prototype On-line Information Retrieval System, NASA Report, CR-918.

66. GARVEY, W D, LIN, N, NELSON, C E and TOMITA, K - Research Studies in Patterns of Scientific Communication: 1. General Description of Research Program. Information Storage and Retrieval, V.8, p.111-122, 1972.

67. GARVEY, W D, LIN, N, NELSON, C E and TOMITA, K - Research Studies in Patterns of Scientific Communication: 2. The Role of the National Meeting in Scientific and Technical Communication. Information Storage and Retrieval, V.8, p.159-169, 1972.

68. GARVEY, W D, LIN, N and TOMITA, K - Research Studies in Patterns of Scientific Communication: 3. Information Exchange Processes Associated with the Production of Journal Articles. Information Storage and Retrieval, V.8, p.207-221, 1972.

69. GARVEY, W D, LIN, N and TOMITA, K - Research Studies in Patterns of Scientific Communication: 4. The Continuity of Dissemination of Information by Productive Scientists. Information Storage and Retrieval, V.8, p.265-276, 1972.

70. GARVEY, W D and TOMITA, K - Continuity of Productivity by Scientists in the Years 1968-1971. Science Studies, V.2, p.379-382, 1972.

71. GARVEY, W D, LIN, N, NELSON, C E and TOMITA, K - Description of a Machine-Readable Data Bank on Communication Behaviour of Scientists and Technologists. Selected documents in Psychology, V.2, p.3-18, 1972.

72. GARVEY, W D, TOMITA, K and WOOLF, P - The Dynamic Scientific Information User. Information Storage and Retrieval, V.10, p.115-131, 1974.

73. CAMPEY, L H, HYDE, E and JACKSON, A R H - Interconversion of Chemical Structure Systems. Chemistry in Britain, V.6, p.427-430, 1970.

74. SHAW, S R - An Investigation of Some Methods of Improving the Performance of Molecular Formulae in Indexing. University of Sheffield, M.Sc. Thesis, 1973.

75. ASH, J E, HYDE, E and LAMBOURNE, D R - Appraisal of Methods of Representing Organic Molecules: A case for Interconversion. Paper presented at the American Chemical Society Meeting, 1972.

76. ICRS - Radical II Search Package. ISI, Philadelphia.

77. Burroughs Machines - FORTE Manual.

78. CROWE, J E, LEGGATE, P, ROSSITER, B N and ROWLAND, J F B - Searching of Wiswesser Line Notations by Means of Character Matching Serial Search. J.Chem.Doc. V.13 (2), p.85-92, 1973.

79. ICL - FIND II Multiple Enquiry System. ICL User Manual 4187.

80. HYDE, E, LAMBOURNE, D R and MCARDLE, L A - Use of a Multi-Level Substructure Search System: Survey of User Queries. Paper presented at the ACS meeting, April 1972.

81. ICL - Statistical Analysis Package. ICL User Manual 4152.

82. SCHNEIDER, J H, GECHMAN, M and FURTH, D E - Survey of Commercially-available Computer-Readable Bibliographic Data Bases. ASIS Special Interest Group for Selective Dissemination Information, Washington, 1973.

83. HANSCH, C - A Computerised Approach to Quantitative Structure Activity Relationships. Advances in Chemistry Series 114, ACS, 1972.

84. MUMFORD, E - Systems Design for People. National Computing Centre, 1971.

85. WIPKE, W T, HELLER, S R, FELDMANN, R J and HYDE, E - Computer Representation and Manipulation of Chemical Information John Wiley and Sons Inc., New York, 1974.

86. HYAMS, M - Chemical Patents Information. Chemistry in Britain, V.6 (10), p.416-420, 1970.

GLOSSARY

This glossary covers computer, chemical and information terms used in this thesis. In some cases, the term is not explained anywhere in the text. In others, there are many references to a term detailed only on its first mention in the thesis.

A file: The chemical search system developed using a number of standard files. The A file contains WLN, suffix and molecular formula information, and is used to produce the structural diagram and to print the chemical details on final presentation.

A number: The ICI reference number is used for compounds included in the A collection. The latter contains structures for compounds mentioned in commercial catalogues, Organics Division works processes and Organics Division technical reports.

AC number: The ICI reference number used for compounds included in their files of commercially available organic compounds.

Acceptability: The measure of a users willingness to make use of an information facility.

Actual key: The means by which an individual piece of data can be found on a computer file. The address of that data is stored against a search term on a separate computer file.

Atom-by-atom search: Atoms of a molecule are examined in turn to test whether a particular combination of atoms is present in the total molecule. The procedure is repeated until a match is found or until all possible paths in the molecule have been tested.

B3500: Burroughs medium-sized computer mainly used in banking and accounting.

B file: The chemical search system developed uses a number of standard files. The B file contains any supplementary data such as biological results which may require to be presented with the chemical and structural information.

Batch processing: A technique by which data is coded and collected into groups for processing during a single computer run. This is the usual method of computer operation.

Bit: Information is stored in computers in binary code using devices which can exist in two states, usually represented as Ø or 1. The most commonly used device is a spot on a magnetic surface, which can be magnetised or non-magnetised. A bit (binary digit) is one such unit of information, and is the smallest unit that can be processed by the computer.

Bit screen technique: A bit screen is a certain number of bits set aside to record information about a particular item. In its simplest implementation, each bit relates specifically to one characteristic, being set to 1 if the item possesses the attribute, and left at Ø if it does not. The computer can search bit screens very quickly and this technique is often used as a preliminary 'screen' to a more detailed search.

Bit and string search: The system combines together the bit screen technique and the text string search into one program.

Boolean logic: A logic (named after George Boole) based on the operators AND, OR and NOT.

Byte: A sequence of adjacent binary digits operated on as a unit. On the B3500 computer it is eight bits and contains one character of information.

Canonical: Conforming to a previously defined order. For any particular representation (e.g. the notation for a chemical compound) only one alternative will normally conform and a canonical notation is therefore normally the unique form of that notation.

Chained file: A file of records in which, irrespective of the physical position of the records, each record is linked by a pointer to the next in the series. A record can form part of more than one series, in which case it would have an appropriate number of pointers.

Chemical data base: The ICI computer files set-up for compound data generated by Pharmaceuticals Division, Organics Division and Plant Protection Division. The system allows on-line access to the data by compound type or by reference number.

Coarse table: The first index table used when accessing an index sequential file. Each record in this index table contains the value of the first record in the next index table (the fine table).

COBOL: Common Business Oriented Language. It is a programming language in which the source program is written using statements in English of a standard but readable form.

Commercial availability: Chemical compounds purchased from companies whose main purpose is to sell compounds are termed "commercially available". These companies usually provide catalogues in which individual compounds may be found.

Company compound centre: ICI has attempted to collect together in one place all compound information, irrespective of which Division was the originator or producer of the compound.

Computer program: A set of instructions composed to solve a given problem by computer.

Connection table: A method of representing a chemical structure. Each atom in the structure is listed together with all its attached atoms and bonds.

Correlation coefficient: A number between -1 and +1 which indicates the degree of relationship between two sets of variables. A common technique used in the statistical analysis of data.

Correlation matrix: An n-dimensional display of correlation coefficients between a number of variables.

Cost-effectiveness: The cost necessary to achieve a required performance level.

CR number: The ICI company registry number allocated by the Company Compound Centre.

CROSSBOW: Computerised Retrieval of Structures Based on Wiswesser. A set of software which allows computer manipulation of structures, developed by ICI.

Current awareness: A current awareness service is one which provides the user with information about the latest developments in his field of interest as soon as possible after publication.

Data base: A computer-based collection of data. In computer terminology it differs from a conventional file in that it is not designed to satisfy one specific limited application.

Direct access: The process of obtaining data from or placing data into storage where the time required for such access is nearly independent of the location of the data in the store. Usually refers to access to a disc, drum or magnetic card store. Direct access is synonymous with random access.

Disc: Type of direct access store, a continuously rotating circular plate, magnetically coated on both sides and mounted with other discs. Data is stored as magnetically polarised spots representing bits in concentric tracks. There are two basic types. Head-per-track disc is fixed to a disc drive and has one read/write head per track per surface. Disk packs are inter-changeable (i.e. different packs can be mounted on different drive units at different times), and contain only one read/write head per surface.

Disc access: The reading or writing of a piece of data on a direct access disc store. For head-per-track disc, the read/write heads must revolve to the correct location so that the action can be carried out. In addition, disk packs also require that the read/write heads move across the disc to the correct track.

Factor analysis: A statistical technique aimed at finding the relationships between other factors measured. For example, can height and weight be correlated as sex.

False drop: An irrelevant answer retrieved as a response to an enquiry.

Field: A specified area of a computer record used for a particular type of data, e.g. the reference number.

File: A collection of records.

File access: The reading or writing of a piece of data on a file.

File handling: Computer instructions covering the reading/writing of data to files.

File organisation: Records are grouped together in a particular way so that certain types of file access can be achieved. File organisation methods include serial, sequential, random or inverted.

File structure: See file organisation.

FIND II: A text searching and general file manipulation facility provided as part of the ICL software.

Fine table: Second stage of access in an index sequential access. There are a number of fine tables, each giving entries on the file and the addresses where the data can be found.

Flag: An indicator used to signal some condition, such as a deleted record.

Flowchart: This shows diagrammatically the logical relationship between successive steps in a program or system.

Followed-by logic: A search technique where it is possible to find records which have one item of data following another piece of data. For example, a search for A followed by B will produce hits for records containing AB or ACB, but not BA or BCA.

Format: The arrangement or layout of data.

FORTE: A Burroughs software package designed to set up complex file structures.

Fragment code: A method of describing a chemical structure whereby the structure is broken down into groups of atoms and bonds. Each group of atoms is assigned a code number. There are two basic types of fragmentation. Fixed fragmentation relies on the analysis of a molecule to determine the presence or absence of any of a given set of pre-defined fragments. Open-ended fragmentation allows the exact fragments specified to be determined at the analysis stage. Fragment codes of both types are widely used in chemical information systems.

Generic searching: Searching for a general class of compounds, such as alcohols or salicylamides, by name or by their common partial structure.

Graphics terminal: A device linked to a computer which can represent two or three-dimensional concepts. Lines can be drawn at varying angles, for varying lengths and at varying densities, hence, all types of images can be produced.

Hardware: Physical equipment, such as processors and line printers, which make up a computer, contrast with software.

Hit: An item which is retrieved by a current awareness or retrospective search.

ICI number: See CR number.

ICL 1902A: One of International Computers Limited (ICL) small-range computers, largely used for batch systems.

ICRS data base: Cumulation of the ICRS monthly current awareness files. The latter contains references to novel compounds being mentioned in the chemical literature.

Index file: A file set-up from a data file, and containing keys by which individual records on the main file may be accessed. See Coarse Tables and Fine Tables.

Information channel: One method in which a piece of information may be obtained, e.g. journals, library, information retrieval service.

Inverted file: A method of organising a file such that the key to the data is in alphanumeric order and all records containing the data are indicated alongside the key. A subject index at the back of a book is an example of an inverted file.

Key: A field or one of a group of fields used to identify uniquely a record and allow access to the record using that field.

KWIC index: Key Word in Context. A permuted, or rotated, index sorted alphabetically with one entry generated for each keyword in the field. The index is formatted with the keywords vertically aligned and the context of the keyword on either side of each keyword. A keyword may simply be a letter, e.g. in a KWIC index of the Wiswesser Line Notation.

Logs: These record all transactions at the various stages throughout an on-line system. The logs are used to re-create the on-line system up to the part of a failure if such occurs.

M number: Reference number given to all compounds biologically tested at Pharmaceuticals Division.

Machine-readable data: Data in a form which can be read into a computer. Acceptable forms of data include punched cards or tape, magnetic tape or disc, and input for OCR (optical character recognition) and MICR (magnetic character recognition) systems.

Magnetic tape: Plastic tape coated with ferrous oxide and used as a type of sequential access store. Data may be written, read, erased and over-written by read/write heads, and takes the form of frames on rows of magnetic spots across or along the tape.

Main file: Part of the Chemical Data Base. The file contains molecular formulae, reference numbers, and the first 24 characters of the Wiswesser Line Notation.

Miss: An answer to a current awareness or retrospective search which is correct, but has not been found by the system. It usually occurs because of a discrepancy in coding the data or the search question.

Molecular formula: A description of the structure of a compound giving each atom present and the number of occurrences. For example, C10H13N.

Multi-level compound description: A hierarchical description of a compound, where more and more details of the exact structure is being given at each stage.

Multi-level search system: A system which provides the facility to search the structure by several techniques. The techniques are hierarchical, more search detail being included at each stage.

Multiplier and contraction rules: A device within the Wiswesser Line Notation rules to reduce the size of the final notation. It has never been part of the ICI system, but until recently used by most other WLN users.

Node: If a chemical structure is considered as a graph, the atoms of the structure represent the nodes or points, of the graph and the bonds represent the lines, or links.

Nomenclature: Names given to compounds. Usually they can be converted to the complete structure. Used for verbal communication between chemists. For example, paranitrotoluene.

Novelty checking: Establishing whether a compound is already known to a collection. If not known, a new reference number can be allocated.

On-line system: A system in which data is sent to the central computer from the point of origin and output data is transmitted to where it is used. The intermediate stages necessary in batch systems (e.g. punching, transfer to computer location, etc.) are largely avoided.

Open-ended fragment code: See fragment code.

Overflow file: Generally a file used to accommodate records which cannot be located in assigned areas of a direct access file, and which must be stored in another file. When considered as part of the Chemical Data Base, it refers to the file containing Wiswesser Line Notation information over 24 characters and suffix data.

Package: A series of programs which performs an application.

"Pages": If more data requires to be transmitted than can be displayed on a terminal, the data will be separated into pages. Each page contains the maximum information. The first page will be transmitted to the user, and he will have some facility to ask for more, i.e. turn the page.

Parallel running: A method of testing new procedures by operating both the new system and the one it is designed to replace, together for a period of time so the results can be compared.

Pointer: Device by which logically connected records can be found even if they are not physically connected.

Precision: The number of pertinent items retrieved by a search compared with the total number of items retrieved.

Primary publications: Includes scientific periodicals, monographs, reviews, conference proceedings, original reports and patents. (Contrast with Secondary publications.)

Profile: A set of indexing terms and/or character strings which express the scope of a query. The profile is matched against a computer file of entries and 'hits' are selected. The terms in the profile are often arranged in a logical relationship. A profile usually reflects the specialised interests of a user and is used extensively in the selective dissemination of information.

R number: The reference number given to compounds obtained by Plant Protection Division for crop protection testing.

Random access: See direct access.

Recall: The number of pertinent items retrieved by a search compared with the total number of pertinent items in the data base.

Record: One or more items regarded as a unit and the smallest unit to be accessed by a computer program. Records in a file can all be the same length (fixed length) or differ in length (variable length).

Registration: The process of inserting a new item into a file of items in which each is stored once only, e.g. registration of chemical compounds.

Regression analysis: A statistical technique in which a function is fitted to a set of observations.

Re-organisation: The process by which a file is re-structured. This usually occurs when a large number of records have been added to the end of a file and logically connected to various other records in a file. Re-organisation is necessary when the chains of pointers are long and finding a specific record requires a great deal of disc head movement. Searching a file becomes more economic with frequent re-organisation.

Response time: The time the systemtakes to produce a result after the relevant stimulus has been applied.

Retrospective search: Involves a search of the master file or all the files in a system in an attempt to find all the information relevant to an enquiry. For example, to find all compounds containing a given partial structure. Contrast with Current Awareness which involves only a search of the latest information in the system.

Richter: An ordering of the atoms in a molecular formula to give good retrieval in a manual molecular formula index.

SC number: Reference number assigned to compounds stored in Organics Division compound store.

Screening: A general search of a file to select a sub-file of items with the required features for detailed searching.

Secondary publication: Any publication derived from the primary literature and giving a guide to the content of that literature. Usually gives subject-oriented access.

Selective dissemination of information, SDI: A technique for searching information sources and selecting the specific items peculiar to a users interest. The term SDI is commonly used to mean a current awareness service, but this is not implied in its strict definition.

Sequential processing: Processing records in a data file according to some predetermined sequence of keys.

Serial processing: Processing records in a data file in the order in which they occur in a given storage device.

Software: In its most general form, software is a term used in contrast to hardware to refer to all programs which can be used on a particular computer system. More specifically, the term is applied to all those programs which in some way can assist all users of a particular type of computer to make best use of their machine as distinct from the specific programs written to solve the problems of any particular user.

Standard deviation: A statistical term which indicates the extent to which a set of observations are spread about the mean.

String search: The search for a specified string of symbols in a notation or in text. String search usually includes AND, OR and NOT logic, and sometimes more sophisticated options such as followed-by logic. Each character is examined in turn until a match is found or the end of the text or notation has been reached.

Structure card: The output representation from the CROSSBOW system. It is an 8" by 5" index card containing 2D structural, chemical information such as molecular formulae and Wiswesser Line Notation, and finally any related information such as biological results.

Structure display: Automatic generation of the structural diagram from the computer record of the compound.

Structure file: The chemical search system developed uses a number of standard files. The structure file contains the 2D structure display and is generated automatically from the A file.

Substructure searching: Searching of a file of chemical compounds to test the presence of a given partial structure.

Suffix: Used to describe details of a compound in addition to the main structure. For example, salt information or stereochemical details.

Synonyms: In this context, refers to records on file with the same 24 characters at the start of the notation. The overflow file has to be examined to find the difference between two synonyms.

Synthetic pathway generation: The ability to automatically generate the way in which a given compound could be synthesised. A logical extension of the ability of computers to manipulate structures.

Tag: A code used to identify individual searches in a system where more than one can be handled at once.

Terminal: A remote station which is used to transmit and receive data from the central computer.

Transaction: Process from input of query to output of results on an on-line system.

Turnaround time: Time taken for a query to be answered.

Update: To apply new records to a data file in order to amend, add or delete records and thus ensure that the file reflects the latest situation.

Usability: Ease of which an information facility may be used.

Variable length key: See key.

Variable length record: See record.

Visual Display Unit (VDU): An illuminated screen device connected to a computer and used for displaying information stored or generated by the computer. Frequently used in conjunction with a keyboard as an input device.

Wiswesser Line Notation (WLN): A unique and unambiguous code for chemical compounds, frequently used in computerised chemical information systems.

1900: Generic name for ICL 1900 computers, of which ICL 1902A is an example.