USL / DBMS NASA / RECON

WORKING PAPER SERIES

Report Number

DBMS.NASA/RECON-23

The USL/DBMS NASA/RECON Working Paper Series contains a collection of reports representing results of activities being conducted by the Center for Advanced Computer Studies of the University of Southwestern Louisiana pursuant to the specifications of National Aeronautics and Space Administration Contract Number NASW-3846. The work on this contract is being performed jointly by the University of Southwestern Louisiana and Southern University.

For more information, contact:

Wayne D. Dominick

Editor
USL/DBMS NASA/RECON Working Paper Series
Center for Advanced Computer Studies
University of Southwestern Louisiana
P. O. Box 44330
Lafayette, Louisiana 70504
(318) 231-6308

A SURVEY OF CHEMICAL INFORMATION SYSTEMS

Aneesa Bashir Shaikh

USL/DBMS NASA/RECON Working Paper Series

Report Number DBMS.NASA/RECON-23

December 5, 1985

CONTENTS

I-1 : Abstract :

Rapid access to accurate, up-to-date chemical information is essential for the functioning of modern society. Large corporations and governments have spent millions of dollars in developing chemical information systems. This paper surveys some major systems. It also determines the R&D trends that have been occurring. The main focus is not on the systems themselves, but on the novel research ideas and improvements introduced in each. The variety that exists in chemical IS&R systems is tremendous because chemistry is a vast and diverse science. To simplify matters, chemical IS&R systems were studied under the following classes :

patents and bibliographies

pharmacology and toxicology

networks

spectra

crystals

physical properties

Where necessary, explanatory notes have been provided for readers with little background in chemistry.

I-2 : Statement of the Problem :

This paper is a far cry from an exercise in academic futility. To begin with, chemistry is the foundation stone of our modern, materialistic society. Cars, bridges, toys, clothes, medicines, computer chips... the list is endless.

Rapid access to accurate, up-to-date chemical information is rapidly gaining importance. Countries and companies with such facilities have a definite economic and political edge over those that don't (eg: chemical warfare).

The development budgets, the time and development effort put into, and the sheer size (CAS Registry contains 10 million records) of such chemical IS&R systems are awesome. Such systems cannot be ignored.

In my opinion, it is essential for any well-informed data base scientist to be aware of :

(a) what work has been done and the characteristics of currently available systems

(b) ongoing research activities, current development trends, and likely future developments

Discovering the answers to the above 2 questions is the objective of this paper. A range of systems has been examined. Stress has been placed on new ideas introduced in each system, rather than on mundane details.

I-3 : Motivation :

I am currently a graduate student in Computer Science. My class mates and I study little else besides computer science. In graduate programs everywhere, the main focus is on theory (as it should be). However, sometimes it is too easy to get so involved with computers that we students completely lose perspective and instead of seeing computers as part of the whole human endeavour toward progress, we may view our profession as an end in itself.

Computer science is much more than an introverted field. Where does it fit into our society ? How are computers used ? What is their impact on other disciplines ? What are the practical applications of all this wonderful theory ? It is to answer these questions, in at least a minute way, that I have elected to study Chemical IS&R systems and examine the general impact computers have had on chemists and the way they do things.

I-4 : What Is Chemistry About :

What are the concerns of chemists ? What kinds of data do they like to collect and study ? What special things do they do? Without some understanding of the basic chemical issues at stake, it will be impossible to fully understand this project. Therefore, some very simple concepts will now be explained in a very easy way.

Chemists are interested in the physical properties of elements, compounds and mixtures. Melting points, boiling points, specific heats, specific gravity, color, odor, density, electrical and thermal conductivity, electronegativity, van der Waal forces, London forces ... The list of physical constants they use is endless.

Pharmacologists and environmentalists record the properties of chemicals and their effects on people, animals, plants and the ecosystem. "What chemicals are manufactured by which companies at what prices" type of information may seem trivial enough but is vitally important in industry. This information must be continuously updated.

Chemists must know the structure of chemical compounds. Inorganic compounds are quite simple. Organic compounds are another story. Organic molecules with 10,000 atoms are considered small. Structural isomers are chemicals with the same formula but different atomic arrangements. Stereoisomers are chemical structures with identical topologies but differing bond angles and bond lengths. The number of variations is mind boggling. All this information must be handled nonredundantly and precisely.

Industrial chemists need information about reaction pathways to select the best possible way of manufacturing a substance. For inorganic compounds, this is relatively simple. Organic chemicals, however, are highly complex. Each molecule may have hundreds of reaction sites. There are thus countless permutations of reaction pathways and intermediate compounds. All this information must be stored in a form usable by human beings and expert systems.

My freshman chemistry professor studied enzyme reactions in soap bubbles for 20 years. The branch of chemical kinetics concerns itself with the effects of activation energy, concentration, catalysts, temperature, pressure etc. on the rate of chemical reactions. Information about other techniques, such as fractionation and distillation, must also be stored.

Crystallographers are interested in crystals. Ionic solids often arrange themselves in crystalline forms. They can be very pretty. They shatter only along certain planes. Their structures can be deduced using X-rays. The X-rays must be stored and interpreted properly.

Metals also form crystals but these are a different kind of crystal. They do not shatter easily. In defect crystals, impurities cause weaknesses by disrupting the crystalline structure. The metals must be treated to strengthen the structure. For instance, Carbon is added to iron during steel making. This is the kind of information our society depends on. Think of bridges and buildings and spaceships and cars.

Electrochemists concern themselves with chemical changes produced by electric current and with the production of electricity by chemical reactions. Details needed here are numerous and will not be delved into. Electrochemical information is vital in the manufacture of batteries and many metal making processes (sodium, aluminium).

Nuclear chemistry is a highly specialized branch; nuclear reactions_ are different from ordinary chemical reactions. One element is converted into another; the whole atom takes part in these reactions; tremendous amounts of energy are involved. Fission occurs when a large atom splits into smaller ones; fusion when many little nuclei fuse to form a larger one. Nuclear reactions are employed in medicine, agriculture, industry and research. Many complex details have to be recorded.

An area of very exciting research in chemistry is spectroscopy (infrared, ultraviolet, mass). The entire field is based on some fundamental ideas from quantum mechanics. Electrons spin around nuclei in orbitals. Every orbital is associated with an energy level. Electrons can occupy only certain discrete energy levels; they absorb or emit discrete amounts of energy when they move from one level to another. When atoms are bombarded with energy, they absorb only certain wavelengths, get excited, and move to higher energy levels. This causes instability and they then fall back to lower energy levels, releasing energy. The released energy is recorded in atomic spectra of which there are many kinds : infra-red, ultra-violet, mass spec. Also, electrons create their own magnetic fields because they are charged particles in motion. This can be stored in nuclear magnetic resonance spectra.

If there are so many kinds of chemists, it is necessary for them to communicate with each other. A unique identification and indexing system universal to all branches of chemistry is necessary since many compounds have multiple names. There are many classification schemes but most of them are unsuitable for computer use or designed for specific subareas only.

The world's chemical knowledge is increasing at a rate of 20% per year. Thus, the storage of up-to-date chemical patents and bibliographies is very important. Duplication of effort would be avoided and researchers would have a much easier time if good systems were available. A serious language barrier exists; good abstraction techniques are not available; standard procedures are not followed. Yet, such information is vital for pharmaceuticals where drug development is a very, very expensive proposition involving hundreds of people over time spans of 20 years or more.

II-1 : Peculiarities of Chemical Information :

In the field of chemistry today, there is an intensively used worldwide information network. Large national information systems have arisen and are beginning to cooperate at the international level. They exhibit a great capacity to bring together relevant information. Local activities have developed where the need is particularly great or specific. Sometimes, chemical companies (normally competitors) cooperate. All this raises the question of what features are peculiar to chemical information to enable it to achieve a position of such prominence in science and technology. The answer to this question can lead to a better understanding, and better utilization of chemical information (Fugmann, 1985).

Chemistry has at its disposal a language of great uniformity and clarity. This is the structural formula. Through its pictorial character, it conveys a structure concept in its greatest conceivable comprehensibility, clarity and definitiveness. The structural formula also lucidly displays every kind of structural relationship between more or less closely related substances. This alerts the chemist to analogy inferences and at the same time overcomes barriers to international communication. Natural language, even expert terminology, is very vague. (eg : there are at least 10 different meanings of corrosion). The differences in meaning are so great that an indexer must not permit an ambiguous term to enter a search file without clarification and revision. This is a problem.

The development of chemical information has attracted particularly great efforts and occasioned unusually large expenditures on the part of the scientific community. The long-lasting value of chemical information has been the motivation. In chemistry, it is worthwhile to recover the knowledge gained through experimentation, even in times long past. Generally, experiments are so precisely described (in the last century), that they can be used as a foundation for further work. When new theoretical ideas are being developed, this spells a vast saving of experimental work. Consequently, search files of chemical information are particularly large. (eg: The CAS published over 10 million abstracts). Not only are they extraordinārily large but they are also normally searched from beginning to end. This is not always the case in other fields. Sophisticated search techniques are needed.

Chemical information files are probably searched more frequently and intensively than the files of other disciplines. This is due to the intensive research work being conducted. Information that is published in a highly specialized field may be of value to a specialist in another field. Chemists frequently change their area of emphasis. Thus, chemists need to familiarize themselves quickly with their new area and thus require rapid access to relevant literature.

High precision ratios of searches are required. Otherwise, searchers would get fed up at not being able to access relevant information. Basic research in chemistry demands the most complete information possible. Using a different data base or reformulating the query should, in the ideal case, have no effect on the completeness of the answer. This is not always so. It is unique to chemical documentation that loss- and noise- avoiding retrieval has already been achieved in structural searches; expensive indexing and retrieval techniques have to be used.

II-2 : Storage and Retrieval of Spectra :

Identification and structure elucidation of organic compounds is today mostly done with spectroscopic methods. Interpretation of spectra for this type of application is still a largely empirical process and thus relies heavily on the use of previously accumulated reference data. In order to clearly define the requirements that have to be met by reference data compilations, it is worthwhile to analyze the spectra interpretation process in detail (Wolff and Parsons, 1985).

What the analyst does when he identifies an unknown sample by spectroscopic methods can be looked at as a transformation of information. The spectrum of the sample is a very complicated function of its structure. The spectrometer can be looked upon as an analog computer that implements this function. There is more than just one independent variable. The spectrum depends upon molecular structure, the sample preparation technique, the operator and various other parameters, not always obvious. For the following discussion the contribution of these other factors is neglected.

In principle, the interpretation process is quite simple :

apply the_inverse function to the spectrum. In reality, the function is not explicitly known. With quantum chemical calculations, one can, to a certain extent, simulate the function F. It is possible to approximately predict the spectral properties of a sample when all its structural parameters are known. However, the calculations are very involved and accuracy in most cases is not adequate for practical applications. The variations-range for observed chemical-shift-values is of the same order of magnitude as the estimated errors of the calculation.

Inversion of the function is in general not possible, the only exception of practical significance being X-ray diffraction. Therefore, the function is simplified by splitting it up into partial functions and relating partial spectra to partial structures, where the total spectrum is a combination of the various partial spectra.

The partition of the total function into partial functions is done in such a way that inversion is possible. The inverse transformation generally gives many different partial structures for one partial spectrum. There is, however, an additional constraint. The partial structures inferred from the spectral data have to form a consistent set. The spectral data predicted for the inferred substructural elements have be to consistent with the experimental spectrum. For the consistency check, non-invertible functions may be used which are inherently more accurate than the inverse function. In most cases, several different internally consistent sets of partial structures can be identified.

The members of each set are then combined to to form chemically reasonable total structures. Sometimes, excessively large numbers of candidate structures are found. Prediction of spectral data from a given structure can be done with significantly better precision than the reverse. Therefore, the spectral data is generated and compared to the experimental spectrum. If large discrepancies are noted, the respective candidate structure is eliminated. Hopefully, only one tentative structure survives this process which is then thought to be the structure of the compound at hand.

Thus, the steps are :

recording of spectra

correlation (inferring possible structural fragments)

consistency checks (selecting a suitable subset)

structural assembly (combining the partial fragments)

spectrum prediction

spectral comparison

Compilations of reference spectra are necessary prerequisites for research and development. Due to the lack of suitable reference-material investigations in computerized spectra, interpretations have in the past concentrated on the consistency checks and structural assembly where a rather high standard has been reached. However, as long as the other steps, in particular the correlation step, do not perform with greatly improved precision and selectivity, total system performance will not meet the standards required for the solution of real problems.

A library search technique is a shortcut. Here the correlation step and the structural assembly step are by-passed. All structures represented in the spectra library are selected. The spectra prediction step then becomes trivial as the library spectra are available. The only steps actually performed are the spectra registration and comparison.

The storage space requirements will now be estimated. To encode the full structure of a chemical compound with n atoms requires about 3n parameters for describing the geometry, n parameters to identify the atom types and about 2n parameters to describe the existence and type of the bonds. In all, the order of magnitude is 6n parameters for the full description of the structure. However, the full structure of a sample is hardly ever known : only the connectivity is given. To describe the connectivity, about 3n parameters are required (atom type, neighbor atom, bond type). If hydrogen atoms are implied and if we assume 16 bits per parameters, then 1,000 bits are required to describe the connectivity of an average organic compound with about 20 non-hydrogen atoms.

IR spectroscopy will be used as an example for the storage space required for spectral data storage. If the range recorded is from 4,000 to 400 /cm with a resolution of about 1/cm then, roughly 4,000 data points must be stored. If the resolution on the transmission axis is set to 1 part in 1,000, each value requires 10 bits. Thus, an infra-red spectrum requires about 40 bits.

There is additional information that should be recorded. (name of the compound, various technical parameters describing the spectra recording condition, the source of the sample, and its estimated purity). For these, some 400 characters are required. This corresponds to about 3,000 bits. Thus, the full amount of information available in an average infra-red spectrum requires about 45,000 bits. A floppy disk typically holds 2M bits of information : this corresponds to 40 to 50 fully documented IR spectra.

This brings us to the next question : how many spectra are required for a spectral data compilation to be useful in a given context ? This is a very difficult question that cannot be answered precisely. It is, however, possible to give estimated upper and lower levels.

In a library search, the contents of the spectra library critically limits the performance of the system. If no suitable reference compounds form part of the library, no system, however sophisticated, can ever produce a useful answer. This calls for a very comprehensive collection that includes as many different compounds as possible. On the other hand, if many similar compounds are in the library, all positions in the hit list will be occupied by reference compounds of the same general type. Only the first entry will bear some useful information : all following entries will just repeat the same message. In addition, processing and storage costs are directly proportional to the size of the library. Thus, an optimal spectra library will contain a few carefully selected reference compounds for every compound class, but not duplicates and large numbers of closely related compounds (eg : the homologous series). Furthermore, simple model compounds are to be preferred over highly complex molecules.

Experience with conventional data collections has shown that even libraries with only 1,000 reference compounds can definitely be useful. However, the not-so-common compound types would be almost completely missing. If these were included, the size of the library would have to be increased considerably to about 3,000. This is considered the minimum required to document virtually all compound classes of practical and theoretical interest with a few carefully selected examples.

If this number is increased, the various compound classes may be documented more thoroughly and exotic species may be included. However, the increase in performance of a library search system will eventually level off and will even decrease with increasing library size. Some researchers believe that libraries holding substantially more than 30,000 carefully selected spectra are not worth their respective expenses in time, money and storage space.

The number of compounds required for a truly useful spectra library is estimated to be in the order of magnitude of 10,000.

If this number is compared to the number of spectra that can be stored on a diskette, one realizes that storage of the full information of a spectrum is currently not feasible. It is obvious that attempts at data compression should be directed towards the spectral data, as these occupy about 80% of the storage space. Finally, data compression should be performed in such a way that no relevant information is lost.

If the library is used exclusively for library search, the spectral data section can be reduced by storing only those spectral features that are used during spectras comparison. However, every other use of the spectra library is severely restricted and the system loses all its potential for further development. The encoding of the spectral data is always irreversible and necessarily based on the present level of knowledge. Consequently, a data collection encoded in this way is firmly scheduled for early obsolescence.

Data compression is extremely critical to the quality of the data collection. Every piece of information lost during this step is lost for good and can never be recovered except by recording the spectrum again. Only irrelevant species of data should be dropped and all relevant data has to be retained. This implies that one has to specify exactly which data items will be relevant to the solution of future, still undefined problems. It is difficult, if not impossible. Thus, enough information is stored to enable the original curve trace to be reconstructed from the compressed data. If the analyst considered this reconstructed curve trace as virtually equivalent to the original one, it may safely be assumed that no important information has been lost. However, this test contains no indication of the amount of irrelevant data still contained.

Several data compression algorithms have been tried and evaluated as to their suitability on a trail and error basis. In the case of IR spectra, storage space can be significantly reduced if suitable interpretation algorithms are used during reconstruction. At present, 12,500 bits are needed for encoding an average IR spectrum. If the band shape in a spectrum does not bear analytically useful information, data compression is much easier and can be made more efficient. Only the coordinates of the band maxima or their equivalent have to be recorded.

Even if the spectral data is highly compressed, the storage space requirements remain very high. Today's lab and instrument computers cannot economically have the necessary mass storage capacity. Thus, spectral data collections, designed to meet the needs of the future, will have to be implemented on large main-frame computers or on dedicated systems.

Spectroscopic data, as well as supplemental information is, in many cases, unreliable. Users generally consult with a reference data collection when in doubt about the values and significance of spectroscopic parameters. They are in no position to judge the quality of the retrieved data sets. Therefore, credibility becomes the central factor governing user acceptance. Furthermore, in many pattern recognition algorithms, and, to a lesser degree, in cluster-analysis programs, decision parameters are extremely sensitive to a few outliners in the data base which, in most cases, correspond to erroneous entries. Thus, thorough and careful verification of all data included is of utmost importance.

It is necessary to procure a large number of top quality spectra of suitable model compounds. To do this at reasonable cost and with acceptable efficiency, world-wide cooperation of analytical laboratories in universities and industries is needed. Recent trends seem to indicate that international cooperation efforts have finally met with success.

II-3 : Heuristic Clustering :

Heuristic clustering methods for text data can be applied to chemical data bases (Bernin, 1985). Clusters are sets of chemical compounds defined by similarity of both chemical structure and activity. Activity and structure information can be stored together in the same data file in essentially the same format and can be processed identically by the same computer hardware.

Text is intricate and imprecise. In a free-text search, any string in the entire body must be found. If the natural text format is changed to a strict format, it is easier to search, but much of the information is lost. Also, the information available would be incomprehensible and thus, unusable. Therefore, natural language must be retained.

One experiment involving a subset of the Merck Index has shown that 20% of the compounds fall into one of four usage categories : antimicrobial, antiseptic, analgesic, and sedative. A count was made of medical term usage. Another set of common names of all compounds and synonyms (including Geneva and trade names) was made. A shortened thesaurus was constructed.

Obviously similar terms were grouped together. The richness of the language was thus retained; search procedures were simplified. However, imprecision still caused a great deal of ambiguity and redundancy. Inverted indexes were used.

Information on each compound was originally stored as continuous text in a series of variable length fields. Errors occurred during data entry. There were no error-checking routines. If data is stored in separate files, line tags (common names) facilitate correlation. One master file available with an internal system of index numbers retreives all information.

Clustering in non-parametric spaces is much more difficult. Usually, the more properties two items share, the closer they are. This is not so in medicine. Two substances with completely different structures may produce similar effects on people. Thus, the concept of using disimilarity as a clustering guide was experimented with. Numbers from +10 to -10 were used to indicate degree of- disimilarity, measured by counting pairs of exclusive properties present.

Chemical and medical uses were placed in independent sets. Sometimes, certain chemical structures with medical uses were masked by semantic noise. The purpose of clustering to search for meaningful relationships was sometimes betrayed by the deeper structure-action affinities. Intelligent hueristic searching to identify relationships was needed.

The iterative heuristic approach produced interesting results. Hierarchic relations, behavioral and psychological effects were studied using heuristic approaches. Abstract hyperspaces based on text data are typically non-parametric in their properties and open-ended. They are amenable to heuristic clustering not only by well-known algorithmic methods. The two methods used were : similarity of pairs of properties and dissimilarity of pairs of opposing properties. One step was insufficient to perform clustering. Iterative searching involving successive steps, each dependent on the former, required a chemist to be in control all the time.

III-1 : Derwent's Patent System :

The family of Derwent's patent information services forms one of the most complex information systems offered to the public today (Kaback, 1977). There are alerting services in the form of expanded titles, short abstracts and long abstracts. There is retrieval via classified browsing files, various printed indexes, punch card sorting, computer tapes and since 1976, on-line interactive searching. In the on-line files alone, the subject retrieval parameters include Derwent classes, manual code classes, multi-punch codes and title keywords. There are multiple options everywhere.

Derwent's Central Patent Index and World Patent Index cover virtually all of the world's chemically related patent publications as well as a large percentage of nonchemical patents. The volume of information involved is very large. It is offered in a variety of forms and is accessible in a variety of ways. In using any tool, it is essential to know just what that tool's characteristics and capabilities are. This is a highly complex system.

The present character of Derwent products is highly dependent on the company's history. Derwent began in 1951 by recognizing a need for rapidly produced alerting abstracts of British patents. Over the next dozen years or so, similar products were begun for a number of other countries. At this point, alerting was the only focus; no means was provided for retrospective retrieval. The jump into the world of retrieval took place in 1963 with the start of Farmdoc, which covered pharmaceutical patents. In 1965, there was Agdoc on pesticide and fertilizer patents. In 1966, Plasdoc for polymers.

Farmdoc, Agdoc and Plasdoc had two main methods for retrieving subject information : manual codes and multi-punch codes. Manual codes were developed to form classified files of abstracts to be searched by browsing. The manual code systems have been subject to various modifications and subdivisions over the years to adapt them to changing and growing technology.

The multi-punch codes provided a deep indexing system, based on information in the full patent specification, originally for use on punch card sorters and later adapted to computer tape searching. The Farmdoc-Agdoc code is primarily a structural fragmentation code, coupled with terms describing compound properties and uses. The Plasdoc multi-punch code covers all aspects of polymer information. These codes too have been modified as needed over the years.

In 1970, Derwent expanded its documentation services to cover the entire range of chemistry and chemical technology. Farmdoc, Agdoc and Plasdoc became three of twelve sections of the new Central Patents Index. Manual code retrieval was developed for each of the new CPI sections and the Farmdoc-Agdoc multi-punch code was adapted to Chemdoc, the general chemical section of CPI. No comparable retrieval system was developed for the remaining eight sections of the CPI. When the nonchemical WPI services were started in 1974, only international patent classification was used for subject retrievals.

A result of this history is that various sections of CPI and WPI have sharply differing retrieval capabilities. The reason for this disparity is basically economic : the most elaborate retrieval techniques have been developed in areas in which there were sufficient subscribers prepared to pay enough money to finance the system.

The CPI covers the full range of chemistry and chemical technology. The volume of patents covered is very high and the cost for handling this large volume of material is correspondingly high, particularly if there is to be any great sophistication involved in the retrieval system. The only viable way to handle this was to offer CPI in a series of independent sections since few organizations are large enough or have broad enough interests to be able to justify acquiring the whole system.

Patents received by Derwent are checked to see whether they are new to the system or equivalent to known patents. The checking is done primarily on the basis of priority dates and application numbers. Non-convention patents are also checked and Derwent frequently succeeds in identifying non-convention equivalents. The on-line file can be used to retrieve all members of a patent family given any member or the priority application number. Coverage of chemically related patents is, to all intents and purposes, complete.

New patents are classified into one or more broad Derwent classes which are used in allocating each patent to one or more sections of the CPI-WPI system. CPI classification is done by Derwent pèrsonnel; WPI classification is based on international patent classes. A given patent may appear in as many as 4 sections of the CPI. There are times when one wishes for more generous multiple classification. In general, there is a bias in CPI to concentrating on end products rather than on starting materials or processes; this traces back to the fact that it was originally designed for pharmaceuticals and pesticides. There are obvious economies in Derwent for restricted classification but also obvious potential pitfalls for information users.

Completeness is one of the very great virtues of CPI and WPI. A second great virtue is timeliness. Within 5 to 6 weeks of publication, all patents from most key countries are covered in a WPI gazette, one for each section. Listings in the WPI gazettes appear in two sections : one organized by patentee and the other by international patent class. Thus, they can be screened for key companies and for subject matter concepts.

A great deal of information is crammed into a very small space accomplished in part by the use of a highly telegraphic style that can appear confusing. This is no problem for information specialists who will take the trouble to find out what all the data elements mean but for the inexperienced user, this can be a problem.

The basic material from the WPI gazette is used as the heading for a series of abstract publications : relatively short alerting abstracts arranged first by country and then in subject groupings by Derwent class; longer, basic documentation abstracts arranged _by country; and for polymers and other high interest areas, basic abstracts arranged in subject profile headings. They are aimed at different users : those interested in the state of the art, lawyers, chemists and engineers (no duplicates).

The manual codes are available across the entire CPI. They are used to provide files of abstracts to search through and the quality of the basic abstracts is one of their greatest strengths. A collection of clear, detailed summaries of related technology may not be the most modern method of information retrieval but it is still a highly effective one. The human mind can spot concepts that can never be articulated to a computer.

Efforts have been made to subdivide codes that produce large numbers of items but problems still exist. Users are cautioned to check generic manual code classes as well as more specific ones. Manual code cards take time to file and space to store but they are worth it. They are obtainable on microfilm but cards are easier and quicker to search. Multi-punch code retrieval was used only with the advent of the on-line file. The on-line file has greatly enhanced retrieval from Derwent's database. There are many searchable parameters in this file.

Title keywords were not used since experience suggested they would not work well. International patent class was a failure with poor recall. Both manual and multi-punch codes gave excellent recall, the latter with very high relevance. Each retrieved some items not found by the other. Multi-punch code may be very effective for complicated molecules but is relatively inefficient for simple ones. Other retrieval problems stem from the fact that products are not starting materials. Catalyst retrieval is sketchy. Title keywords would be more effective if the misspelling rate was less. New hyphenation rules will help. Users should truncate terms to allow for variant endings and use both American and British spellings. Tests are underway to add keyword indexes. So many changes have occurred. Better documentation is needed. Retrieval presents more problems than alerting.

III-2 : The IFI Comprehensive Data Base :

This is the result of the merger of the IFI Uniterm Index and the Du Pont Index to U.S. Chemical Patents in 1971 (Donovan and Wilhide, 1977). The Uniterm Index was developed in 1955 as a coordinate term index. It was first published in printed form as a coordinate term index. It was first printed as a dual dictionary that permitted simple Boolean logic intersections by comparing inverted index lists for matches. The vocabulary was open-ended and uncontrolled. In addition to the dual dictionary of major terms, a single dictionary of minor terms was provided for index entries with low postings : these were primarily chemicals not in computer format. In 1964, USPO classification for patents was added to the file. Subscriber demand also produced a controlled vocabulary and a review of the minor term vocabulary for new candidates to the major vocabulary.

The Du Pont Index to U.S. Chemical Patents was begun in 1964. A controlled open-ended vocabulary was used for both chemical and non-chemical concepts. A chemical fragmentation scheme was designed to index both specific chemicals and generic chemical structures. A system of roles was developed to specify whether the chemical indexed was present, reacting or a reaction product. An elaborate role scheme was also designed for the indexing of polymers.

The Comprehensive Database was created in 1971 when IFI/Plenum Data Co. acquired the Du Pont Index and merged it with their Uniterm Index. It now contains 400,000 U.S. chemical patents dating back to 1950. The file is updated quarterly. Lag time is 6 months. The growth rate is 20,000 patents per year. The selection of patents is done on an automatic basis for a specified list of USPO subclasses. The Official Gazette is scanned for patents of chemical interest appearing in other subclasses.

The searchable elements of a database are accession number, patent number, assignee, USPO subclass, compounds, chemical fragments and general terms. The accession number indicates the date. Both the patent number and the accession number can be used to retrieve the complete index record. The index elements can be used in any combination for search strategies.

The assignee is that appearing on the face of the patent. This requires the searcher to investigate the name, history, acquisition and merger history of any company to be searched. At present, a major revision of the assignee vocabulary is in progress. Name changes of companies are being collected to a standard format. However patents issued to companies which have been merged or acquired are not posted to the parent company.

The USPO Classifications assigned to the patent are searchable. Both the original reference (OR) and the cross reference (XR), subclasses and main classes are searchable; they represent subject indexing of the patent by the Patent Office. They are used as concepts in the construction of search strategy. Another valuable use of patent class and subclass is to partition the file into subsets. This saves cpu time. The classifications are updated annually to reflect classification revisions in the USPO.

The General Term Vocabulary is controlled and open-ended. It includes general chemical concept terms, trade names, nonchemical terms and search term only (old indexing philosophy). There are 10,000 terms 25% of which are nonchemical with generic class terms, trade names, general polymer terms and the discontinued search term only (STO) indexing terms from IFI. STO was very general, being used for the 1950-1964 time frame when there were far fewer patents.

The Compound List Vocabulary is a register of chemical names that have five or more patents indexed to them. There are 14,000 entries. Substructure searches can be performed separately on these.

The chemical fragment indexing scheme is applied to three types of chemical structures : Markush formula, indexer generated generic structures and specific chemicals which are not on the Compound List. The chemicals in the minor term index in the Uniterm are indexed with the fragmentation scheme. This work was completed in 1975. Four kinds of fragment terms are used : atoms present, functional groups, configuration descriptors and ring descriptors.

A must-possible approach is used. Possible are alternative components of the structure. At search, the possible fragments are asked for and the must fragments which do not apply are negated. The must fragment list contains 151 structure and configuration codes. Irrelevant answers can be eliminated. A link technique is used to prevent scrambling of the fragmentation indexing tō two or more chemical structures from the same patent.

The search aids are General Term Vocabulary, Thesaurus, Compound Term List, Fragment Term List, Assignee List, frequency tables, Indexer's and Searcher's Guide and a manual describing programs and i/o. Other aids are USPO Manual of Classification.

A character string search program is used to search vocabularies and texts. It has been very useful in identifying subclasses for identification and search. It is an important tool for teaching USPO classification. It permits the use of And and Or and allows the specification of hierarchies.

The data base permits the recall of indexing for the searcher's view. When the prior cut is available, the complete index for it printed out. This is a great help in building search strategies. Since indexing vocabulary has changed significantly, it is very helpful to look at the indexing of patents in older portions of the file. It is also an educational tool.

There are four parts to the data base management system. The first part is an editor that validates the search input. A search program is the second part. The third part sorts the output according to instructions. The fourth part is the report writer. The editor, searcher and report writer are frequently operated independently as well as in conjunction with a full search retrieval request.

The retrieval system is based on weighted term searching. The index terms are grouped to logically express the various parts of the search inquiry. The terms in each group are given individual positive values. Each group is assigned a threshold value, minimum acceptable weight, MAW. The sum of the values of the weight index terms must meet or exceed the MAW in order for the patent to be retrieved.

There is an implied And between groups. The weighted search term technique has proved to be most valuable. It allows the searcher to control the relative importance of search terms to the strategy. In the output, the total weight for each patent retrieved is given. This enables evaluation of the hit in relation to the search strategy. A unique feature is the answer frequency table available as an option for USPO subclasses, assignees and total answer weights.

Boolean logic can be simulated using the weighted term technique. An OR exists between each term within a group that has a weight value greater than or equal to MAW. The And can be forced between two sets of terms within a group by assigning values less than MAW to these terms but which together add up to a value greater than or equal to MAW. A list of synonyms can be identified by the MAW for the first term and 0 weight for all subsequent terms. The AND relationship between groups can be altered by use of the search chaining technique. Within a group of related requests, a chain can be established. There are two types of chained outputs : in the first, all searches identified on the chain eliminate duplicate answers from higher up the chain. In the second chain techniques, all answers found in the chain of searches are printed in the output of the first search in the chain. With this technique, it is possible to establish an Or relationship between groups.

If the MAW is set at zero and the term weights are positive, the group does not enter into the logic of the search However, any answer hits which contain indexing given in the zero group will show these term weights in the total weight which accumulates for a hit. This technique is used frequently when the patents of a company are reviewed. Known technology is signalled in a zero groups so that review time can be concentrated on unknown aspects of the company's technology. It is also used to review the content of a specific subclass. Furthermorē, it permits the searcher to use additional indexing to the retrieved patent which is not essential to the logic of the search but may help evaluate the answer hit.

An average of three strategies is prepared for each search topic. A separate strategy is usually prepared for the old IFI Uniterm Indexing. Both broad and narrow strategies are designed for searches. The chaining technique is used for multiple strategies so that redundancy is removed. Searching the file in-house permits unlimited use of alternate strategies. The development of alternate strategies is encouraged. The search results are screened by the searcher. The total answer weight and title are the first screen. The CA abstracts are reviewed. All patents since 1965 are on microfilm, copies of which are given to the client. The printout is given to the client for his evaluation and use. Copies of the search strategies is maintained for two years.

The search output can be sorted by accession number, patent number, assignees, patent subclass or total weight. Alternate sorts of the same answer list can be done. The output can be recorded on tape (save tapes). These are constructed in areas of technology which is searched repeatedly. Cpu time save by searching mini-databases is considerable. R&D of small information systems just for R&D work is being done. Effective searchers must consider all indexing options : the old Uniterm indexing, USPO classification and the current Du Pont indexing. The primary use of the IFI data bases is in performing prior searches for the research group at idea stage and when the invention record is prepared. The state of the art data base is updated semi-annually. Experiments are being carried out for use of the data base for marketing. An IBM 370/150 running the operating system VS1 with RJE/RJO is being used. Batches of 3 to 8 searches each are used. Large batches are economical to run.

Many problems are encountered. Frequent changes in vocabulary and indexing make the system hard to learn. The fragmentation scheme has a high level of false retrieval if common structures are used. The lag time between reality and the indexing vocabulary is high but this is a problem in other data bases also.

Among its nice features are the chemical structure searching which is provided. The search of USPO subclasses is allowed. Polymer chemists have their own subdivisions. Much time is saved. It is cost effective. The search work is confidential. The information specialist is familiar with the user needs. Searches can be restructured and updated. It is fast and comprehensive.

III-3 : PULSAR :

When a scientist requires information, he will often draw first upon personal resoures. These resources will include all the specific contributions of that individual as well as any relevant literature information which he can recall. It is at this latter stage that major retrieval problems occur. A researcher acquires a rapidly increasing accumulation of data as his career proceeds. The specific manner in which these data are stored will affect their subsequent availability.

A variety of techniques have been traditionally employed for this task; most usually, some variant of the well-known card file system. In this case, information is recorded on an index card, the files grows and the cards are resorted as new categories are created. The system begins to become inefficient when the cards number in the thousands. At this point, the categories usually have become too general and substantial cross-indexing is needed. There are numerous inconveniences associated with all card filing systems.

A number of general and extensive chemical information systems are available to facilitate literature searching. Even with these systems readily accessible, there is strong justification for maintaining a personalized system restricted to the interests of an individual and based on the individual's own choice of keywords. Due to the availability of microcomputers wtih diskettes, Purdue University Literature Search And Retrieval system (PULSAR) has been developed (Smith et. al, 1981).

PULSAR is implemented on a Radio Shack TRS-80-II with 64K bytes of memory, a printer, and from one to four 500K byte disk drives for a total storage capacity of 2M bytes. The program is written in BASIC and occupies 64K. The various options available are : add articles, search for keywords, display routines (single/multiple articles, keywords, journal, disk directory, free space, system status), data checking routines, editing routines and disk file management routines.

Every record has the following information : entry number, journal name, volume, year, page, type of article, keywords. Logical boolean operators are available. Upto eight keywords can be specified in one search expression.

The program uses five files to store the data. Each of the files is divided into records of equal length. The records may be read/written in any order. It is random access. The keywords are stored in a binary search tree format for efficient searches, additions and deletions. One record is designated the head of a tree. The left pointer of each keyword contains either the record number of a keyword alphabetically preceding it or 0. The right pointer of each keyword contains either the record number of a keyword alphabetically succeeding it or 0.

The links pointer contains the number of a record in the link file where the article numbers associated with the keyword are stored. The link file contains records which each hold 10 different article numbers. For more than 10 articles, additional link records can be linked together via the link continue pointer. This allows the system to be as dynamic as possible. Internal limitation currently require a maximum of 1000 references for each keyword.

The article file is used to store information pertaining to each article; each record of the file contains all information about a single article. This includes pointers containing the record numbers in the keyword file of the keywords which refer to the article. This is primarily for display purposes. The volume, year, page, type and a pointer contining the number of a record in the journal file are stored as well. Rather than storing a 40-character journal name, a code is used. Disk space is conserved.

The free record file contains the record numbers which are available when a keyword is deleted. When a new keyword is added, it will be stored in one of these free spaces. If no space is available, it will be added to the end of the file. PULSAR contans many subroutines to perform all these file manipulations.

The synthetic organic chemist faces a special problem in dealing with scientific literature. The very nature of the discipline demands exceptionally broad coverage of the literature. During the course of a total syntheis, he may require information about virtually any facet of organic chemistry.

PULSAR offers considerable assistance in the information handling ability for the scientist. There are two compelling reasons for adopting such a system. Firstly, it is a personalized system. Thus, it is perfectly matched to the thought processes of the individual. As a keyword system, it is by its very nature programmed to respond in the exact language of the individual user. It is this specific ability that makes it more responsive than systems that are more general in nature (CA, Lockheed). The purpose of PULSAR is not to replace those systems which currently provide extensive access to the literature but rather to provide an organizational framework upon which to develop an individual's own perspective. The second reason for adopting PULSAR is that in the process of deciding which keywords to assign to any given article, the reader is forced to carefully specify how each particular article fits into the current scientific discipline. Thus, the literature is more thoroughly read and understood.

When an article is read, up to eight keywords are chosen. Keyword order is designated so that a crude narrative can be inferred from examination of the sequence. The most critical determinant in optimizing the system efficiency lies in the careful selection and maintenance of the keyword data base. In its current configuration, PULSAR can stored about 20,000 articles and 3,000 keywords. After 3,000 keywords are added, the keyword file expands much more slowly. New keywords are added interactively. Editing routines help users to identify misspellings, redundancies, unnecessary synonyms, and underused or overused keywords.

A specific problem in using a keyword-based system for literature retrieval in organic chemistry is associated with the variability of organic nomenclature. There is much personal latitude in the use of hyphens and spaces in designating organic functional groups. Since the computer would perceive each variant as a distinct keyword, a routine was written to remove all blanks and hyphens. Although the compacted form is never seen, the net effect is that the program prevets accidental redundancies from occurring. In conclusion, PULSAR provides a highly versatile tool for the organization and retrieval of literature and should be applicable in many areas outside organic chemistry.

IV-1 : Chemfile :

The BIOSIS Chemfile is an information system developed to assist in the assignment of CAS Registry Numbers to biological compounds (Graham, 1977). Only 23,000 or 10% of all BIOSIS articles contain unique compounds but the work helped in understanding the requirements for indexing other BIOSIS products. CAS Registry Numbers are index terms in the printed version of Abstracts on Health Effects of Environmental Pollutants. Tapes are sent to TOXLINE.

The number of different substances in HEEP is relatively small and the frequency of occurrence is relatively high : this is why CHEMFILE was developed. Every record includes the BIOSIS accession number, the CAS REGISTRY Number, molecular formula and synonyms such as trade, generic and systematic names.

The serially assigned BIOSIS accession numbers tie all the information together. It is used rather than the CAS Registry Numbers so that substances and publication items can be processed further while the Registry Number is being located. Also, since not all toxic chemicals are registerable, information has to be retained and effort should not be wasted in searching for non-existant Registry Numbers.

Originally, the data was stored on magnetic tape. Each record contained the accession number, a code for data type (10 for Registry Number, 20 for molecular formula), and chemical information. The tape record was periodically listed to give a hard copy of all substances in alphabetical order. Numerically ordered lists of all Registry Numbers were checked manually to see no Registry Number was entered more than once. Chemfile was updated by batch.

Various validation checks such as accession number, Registry Number and proper format were performed at input. Problems were : errors were detected only when batch runs were made and the whole job had to be aborted, corrections made and the job rerun. Since listings were lengthy, infrequent runs were made and it was hard to know what the computer contained between runs as there were no interactive capabilities. The process for error correction was cumbersome. Incorrect records had to be keyed in exactly as they had appeared so they could be located and then the correction could occur; all this was batch. run batch.

A new system became available in 1976 on an IBM 370/145. The file now includes data of record creation, date of last change and employee number of the operator. Each data type may have up to 99 modifications or subfields, (7010-01 .. 7010-99) Further flexibility has been incorporated to permit 99 submodifications of each synonym (7010-01-01 .. 7010-01-99)

Use of the CHEMFILE maintenance system requires that the operator know the accession number, data type and modification and submodification numbers which will serve as record identifiers for the new information or to retrieve a record already in the system.

The interactive user interface is great; like QBE, the selection screen is especially formatted to help the user formulate the query and retrieve a record. The record screen has three subfields. The uppermost field is where information about the data type, date of creation, date of last modification and id of the operator is displayed. Some items may be null if the record is just being entered.

The second field is the data field; it is used to display or enter one data item at a time; it contains registry number or substance name. The third field is a menu of functions; it deals mainly with paging within and between records. At any one time, the entire screen displays information pertaining only to one record. Every time a record is entered, invalid data type, modification and submodification numbers are automatically rejected. BIOSIS Accession Numbers and CAS Registry Numbers are also checked.

The following printed reports are available; the data base maintenance record is printed nightly; it is a log of all transactions. Listings with information ordered by Registry Numbers or by alphabetical listing or by Accession Number are available.

Access to the data base was limited. It had been designed for verification only. It was unavailable to searchers or indexers. It had no search capabilities nor could it display all information about a given substance at once. This situation was partially rectified by the purchase of a STAIRS package for on-line retrieval. Modifications are not permitted. Users can use Chemfile to search. A given substance is retrieved through constructing the known name and displaying hits. Simple boolean operation searches based on names and fragments are allowed. Text searching is available. Searches may be printed.

IV-2 : PAGODE :

Chemical information must be completed by biological information for pharmaceutical data bases. Creating a numerical comparative biological data base is not easy. Chemical information usually exists in a standard form ready for computer processing. Biological information is quite another cup of tea.

The first step is to store in-house chemical structures for computer handling. This in-house documentation problem requires a coding scheme more accurate than the huge international chemical systems of documentation because it serves a much narrower purpose. To get a precise description of any sequence of atoms in a molecules, sophisticated coding schemes must be used.

Dubois developed the DARC system and this is used in the general organization of PAGODE and the coding of compounds (Berdago et. al, 1978). The whole project was divided into several steps, the first of which was to create the data base and test it in batch processing. The DARC system met these initial requirements.

Before choosing the DARC system as a foundation, other systems were studied. Factors examined were : a one-to-one correspondence between structure and code; the possibility of automatically constructing search-keys or search-screens from the code and vice versa; the description of molecular topologies. DARC was found to be the best in these respects.

DARC is based on a topological code : a chemical formula is considered to be a graph with three main parts; DEX is the existence descriptor; DLI is the bond descriptor; DNA is the nature of the atom descriptor. The starting point of the description is the focus and its choice is determined by formal rules. The description covers the molecule by successive layers of two bonded-atoms each. When the first environment is described, the terminal atoms are new origins for the next description and the process continues until the molecule is completely described. The topological screens are called FRELs and they are the search keys.

In the first phase of implementing the system, only these topological screens were considered. Other chemical screens such as side-chain screens will be introduced later. The structure searching generalization is based on inverted files. The TOPOCODEUR program was used to carry out the coding and generation of topological screens. It generates the connectivity matrix and writes results on tape. PAGODE consists of only 30 assembly programs (6,000 lines) implemented on IBM 360/40 running DOS. Data is stored in three files :

The DD (dynamic data file) contains all the data involving a (the DARC encoding formula, molecular formula and biological information). There is one record per compound retrievable by 5 digit CM number. By means of hash-coding, the approximate address can be located. For this reason, the DD is divided into two subfiles : a basic subfile where all prime addresses are stored and an overflow subfile for duplicate addresses and synonyms. Each record includes a pointer to the next synonym.

The screen file contains all the screens present in all the structures. There are many FRELs for each molecular formula. The DARC code of isomers varies. The screen part is divided into 55 memory regions, one per DEX. Each region of variable size contains all the screens within this DEX. With each screen, one can see its address in the INV file. Blocking is used.

Each screen of the screen file has an entry in the INV file. The inverted list of all compounds that go with this screen is recorded here. The list is of variable size. To avoid unnecessary storage the list is divided into physical records of fixed length. When additional storage is required, pointers are provided.

Chemical structure retrieval is done using DARC screens. The user can have a combination of screens and boolean operators. Parentheses can be as complex as you like. A DARC screen may be open or closed. Descriptors of a closed screen are clearly defined. Indetermination leads to a set of several screens (ie: variations).

The PAGODE query subsystem links different operations as follows. Decoding and syntax control of input screens is the first step. The question screen is retrieved from the SCREEN file. The second step is that in the case of open screen retrieval, PAGODE generates the maximum number of possible graphs and selects the regions which might contain suitable screens. The search is carried out with these screens or equivalently, expressions. In the third step, expressions are translated to inverted Polish notation. In the fourth step, query processing occurs. While a pointer P scans the expression, specified PAGODE modules read the corresponding lists in the INV file at locations and store them in a stack scanned by pointer Q. Each time a Boolean operator is encountered, the requested operation is performed on the two earlier operations and the result replaces the two lists. It offers real advantages such as speed in processing, omission of parentheses and possible intermediate results if requested. Then the result is printed.

Utility programs are available to simplify querying, compound entry and. To make operations as streamlined as possible, PAGODE offers utility programs to do hash coding, screen retrieval, closed screen generation from open screens, read/write of any record, logical boolean expression translation to polish notation, etc.

When the system was evaluated, it was found that query time depends on the number of screens that have to be searched for, the question structure and the operators (And and Not are slower than Or because m.n comparisons are needed versus simply appending). Also, if the screen occurs frequently, then it has long lists of INV files to be processed. Trying to evaluate question execution time was difficult because of all these various factors.

Considerable noise was noticed. This is because there are no screens in some carbon chains and some of the FRELs existing are badly arranged. If screens overlap the structure properly, the noise rate is low. FREL encoding requires much attention.

Otherwise, FRELs are generally satisfactory although noise is the biggest disadvantage. New screen types could be introduced for certain unfavorable situations. Noise is not considered a major drawback for an inquiry system as long as it is not excessive. This is because sometimes unexpected results can provide new insights into the problem.

IV-3 : CBF :

CBF (Computer handling of Biological and chemical Facts) has been successfully used by various research centers of C. H. Boehringer Sohn, and Ingelheim for 16 years (Becker et. al, 1981). It was conceived as a data base for storing biological screening results from drug research. CBF provides printed information about chemical compounds, and/or screens results continuously or on demand. It contains 170,000 structures and 260,000 test results.

Every record has a condensed connectivity table to store all the structural information. A series of screens is machine generated from each table. Only unambiguously defined structures are stored in a retrievable form. Equivocal substances are stored by name or print format. They cannot be retrieved by structure search but are included in the printout.

A microprocessor-controlled semi-graphic device was developed for formula input optimization. It has an effective price/performance ratio. To simplify matters, abbreviations are used. Connectivity tables of the abbreviations are stored in a special file and incorporated into the general connectivity table. For a minimization of time consuming searches, five screens are used. Basic screens store the number of atoms, heteroatoms, rings, and bonds. In empirical screens, only the type and number of atoms are stored. In modification screens, additional non-structural information about the chemical is stored. The bits screen stores information about fragments of the structure. The ring screen contains detailed information about the rings in the structure. All this stored chemical information can be printed on a graphics printer or displayed on the terminal.

Scientists have to fill out a special form to enter a compound into the data base. This form contains four sections. The test section contains the modification number, date, experimental method and precautions taken. The observation section contains the data that was collected. In the results is stored the conclusions. The scientist can record his/her bright ideas and insights in the commentary section. The values of the observation and results are stored as descriptor codes. Codes are also available for application modes, side effects and units. Using codes saves storage.

Terms may be combined. Retrieval depends on boolean logic. Chemical structures or substructures can be searched for. In one search, topological searches of up to 26 different substructures in one stored structure can be performed. Positive or negative search logic can be used. (ie : X MUST be present; Y must NOT be present). Partial identity can be stipulated or forbidden for any substructure. Substructure searches and partial identity operations can also be combined without restrictions with boolean logic. In biological searches, individual information elements can be invoked by means of keywords. Stipulations above or below particular levels may be made. Definite side effects can be requested or excluded. The substance code is hierarchically constructed with activity profiles feasible.

In printout processing, the entire biological and chemical information available to the editor post-processor program is printed in highly flexible, user specifiable format.

Conclusions are that this system is easy to use to store and retrieve information; it is modular and flexible; now it is on-line not only off-line.

IV-4 : HEEDA :

The Office of Toxic Substances (OTS) of the EPA is charged with the making of regulatory decisions under the Toxic Substances Control Acts (TCSA) concerning 43,000 commercial chemicals listed in the inventory. Chemicals must be ranked for regulatory concern or selected for testing. Thus, OTS developed the Health and Environmental Data Analysis System (HEEDA) (Lefkovitz et. al, 1981). HEEDA focuses on structure-activity prediction in toxicology and contains validated or reviewed toxicological data that can be correlated structurally with the chemicals. Quantitative Structure-Activity Relationships (QSARs) focus attention on chemicals of concern.

In the initial phase, data was acquired and organized; appropriate sources of valid data were identified; data was classified both biologically and chemically. The second phase involved the development of data accessing mechanisms; substructure searches were an essential design feature; OTS simply connected HEEDA to CSIN; all search (including substructure) capabilities are derived from the Chemical Structure and Nomenclature Search System (CSNS). The final phase was developing data correlation methodologies, mathematical modeling techniques and report generation programs.

A primary goal was to form a collection of standardized, reviewed data that could be subjected to various statistical tests to correlated biological end effects with structure. Correlation of biological effects with each other has to be done. Potential hazards of new substances is estimated by comparing the unknown chemicals with known compounds along a variety of axes. The statistical models used must be validated. Various chemical information and structural descriptors are needed. Key considerations are : completeness, statistical validity and conformity to current laboratory practices. Code is needed to indicate the quality and completeness of the data, and also to compute additional chemical information.

The initial concern was predicting carcinogenecity. A reliable carcinogenesis data base was needed : one that was balanced with respect to positive and negative end effects, different structural classes, and animal vs. human data. The International Agency for Research on Cancer Monographs, and the Bioassays of the National Cancer Institute did not meet these standards, but were used because little else was available. This data was augmented with anti-neo-plasticity and mutagenicity data from GENETOX (another data base).

The data base design is unique. It combines completely different but meaningful observational categories of information of regulatory concern in an open-ended manner in one system. All information related to a specific agent is uniquely identified by its CAS RN. Every record has pointers to two file systems : the Auxiliary Chemical Data File, and the Experimental Data File. ACDF has chemical data needed for identification (names), modeling (structure and substructures), and cross-referencing to related chemicals (metabolic or reaction products).

EDF contains experimentally derived data. It is organized as a strict hierarchy. At the top of the hierarchy is the unique chemical identification code, the CAS RN. At the second level is a series of observational classes. Four classes are currently supported : end effects, environmental measurements, biological measurements, chemical/physical properties. However, more classes can be added as needed. The third level divides each observational class into subclasses. The fourth level stores single references (papers) or documented test summaries. The fifth level contains the remaining experimental and other descriptive details in a series of data items called qualifiers. Qualifiers are organized along the lines of a scientific paper with citation, method, materials, detailed results and discussion. At any level, any number of repeated terms can be entered.

This organization enables searches of the following kinds to be performed : all test results of chemical X; effects of chemical X; carcinogenic results and qualifier details of chemical X; carcinogenic results of species Y using chemical X.

HEEDA is designed to accept data with varying degrees of detail and differing format at the qualifier level depending on the scientific needs of the source. Qualifier definitions are open ended. Therefore, the above data classification was necessary to prevent chaos. It is shallow but familiar and simple. Deep classification hierarchies have the disadvantage that information becomes highly differentiated and cross-referencing difficult. It is easy to tie relevant data together at the lower levels of the hierarchy. The user can readily identify analogous types of data across all test systems even if specific details vary widely. Bibliographies are not stored in the citation. For inquiry, the user can specify any combination of the top three facets (or levels). Open ended features are : an unlimited number of observational classes and subclasses; an unlimited number of test references; any number and type of qualifier details.

Classification of chemicals by substructures and the comparison of biological effects within and between classes is important. A prior classification scheme was developed to express some scientific observations. This scheme has 69 classes (codes). The codes are entered into the "Names" record. Any number of class names can be assigned per compound and separately searched or retrieved. Ad hoc classifications is where the computer automatically assigns substructural fragments to a chemical from a connection table. This is supported in HEEDA.

Data definition is an important part of any data base management system. In HEEDA, there are three steps in data definition. First, the data element and its source are identified, described and assigned a faceted code. Then, the attributes are given. Finally, information is entered into the DD table. Where the code is already present, only new information is added. The data base administrator is responsible for preventing redundancy and assigning common field names to the attributes-so the system is coherent. Many fields are of variable length.

HEEDA was designed to accept data from many sources in many formats, interactive or batch. Conversion routines ensure that internal representation is uniform. Care was taken to ensure efficient retrieval by the correlation tools (for modeling and prediction). Modeling and prediction are truly powerful. Specially formatted quantitative values can be forwarded to a model. Models are constructed on the basis of mathematical or statistical procedures that operated on continuous or discrete data. Physical-chemical and structural properties are independent variables; end effects are dependent variables. Reports provide a range of correlation techniques based primarily on visual effects. Any data field can be selected for display or for forwarding to the report generator. Data can be sorted, columnated, array into matrices and graphed.

The file maintenance module performs the format conversion, controls modification (access being tightly controlled), and incorporates chemical descriptors automatically generated from special purpose programs. Compounds are uniquely identified by their CAS RN. In fact, random access is possible only through the CAS RN. It may be sequentially accessed by predefined chemical nāmes, substructures (CNCS from CSIN) and all kinds of combinations using boolean operators. Individual fields or entire tests can be modified (ie: granularity can be selected). Programs generate structural or chemical descriptors from connection tables. Outputs from these programs serve as input to the modeling programs. Data is passed through the file maintenance module for reformatting.

IV-5 : NAPRALERT :

NAPRALERT is a data base that contains a textual-numeric collection of records regarding the chemistry and pharmacology of natural products and their appropriate taxonomical data (Loub et. al, 1985). It is presently accessible only off-line through the Program for Collaboration Research in the Pharmaceuticals Sciences (PCRPS) at the University of Illinois, Chicago. However, efforts are being made to get it on-line.

The NAPRALERT file was specially designed to be of value in drug development and contains information relevant to the research efforts of natural products chemists. It covers the chemistry and biological activity of extracts and/or secondary constituents isolated from, and identified in plants, marine organisms, microbes, and to some extent, animal data. Chemical information regarding vertebrate animals (lipids, enzymes, proteins) is not included. However, certain reptilian toxins (snakes) are included. One important design feature of this data base is the recording of information capable of predicting or rank ordering organisms as to their probability of having specific biological properties if properly investigated.

Considerable development effort went into the NAPRALERT data base. More than 150 journals of natural products, and various comprehensive abstracting services were scanned by PhDs for relevant articles. Other articles were encountered nonsystematically from books, reviews, and personal communications. Data was then entered on special forms.

There are four major types of data : citation, taxonomy, chemical information, pharmacological data. All this is stored in a vertical hierarchy. A horizontal hierarchy exists to enter one-to-many records at each level. Records are associated with each other through a common citation number assigned to demographic data, and an occurrence number for each record created under the single parent record. All this data is maintained with Indexed Sequential Access Method file management for rapid retrieval. Currently, more than 43,000 citations have been entered with more than 105,000 organism names, 195,000 pharmacological results and 190,000 compounds. The numbers do not represent unique names since a given organisms may have been investigated many times and the same chemical compound exist in many sources.

Information in the "Demographic" record is similar to that found in most bibliographies but also contains data unique to

NAPRALERT. (One example is the address of the senior author or the publishing company. This greatly facilitates correspondence with the author for additional information or reference material. Updating this information is not practical, so it is of use in the current awareness field only).

The "Organism" record contains a full taxonomical description of the organism structure. It excludes the organism class, common names and geographic origin. The part studied, the condition of the material and the amount of material examined is all carefully recorded. This can have major considerations on research results. Logistical problems are simplifed.

When chemical constituents are isolated or evaluated, data is recorded in a "Compound" record. Trivial names are used whenever possible since many products are complex molecules needing highly evolved systematic nomenclature. IUPAC names are a last resort. A code is then assigned to each compound. A binary numerical code identifies the major chemical class of the natural code. Thus, it is possible to retrieve/display entire classes of natural products with a single sort. There are 80 major types. Following the chemical class code is a 3-digit code which defines the carbon skeleton or the major substructure. It can be used to sort compounds by unique carbon bases but their graphic presentation requires the use of a non-computerized base-code dictionary. Finally, up to 12 different binary alpha-numeric designations can be used for functional groups present. A mechanism has been incorporated to reveal negative data. Last, but not least, the yield is stored.

When biological activity is present, positive or negative effects reported are recorded. A "Pharmacology" record is prepared. A variety of alpha-numeric work-type codes are used. Besides reducing typing errors and saving storage, retrieval is faster. Biological studies are differentiated into one of 16 major pharmacological categories, each with specific biological effects. Additional data items recorded are : type of extract used, mode of administration, test animal, sex, dosage, qualitative or quantitative results and the pathological system or substrate evaluated. A "Textual" record permits an unlimited number of modifying remarks and observations. Other fields subjectively evaluate conditions of the study and assign point values for prediction. Much of the scientific information computerized is repetitive. Textual retrieval of codes for report generation is available. A permanent file of all relevant articles and abstracts is maintained.

The most frequently requested information is concerned with ethno-medical (folklore) information, biological activities for extracts, and listing of chemical constituents. Also needed, albeit less frequently, are biological activities of natural products, identification of taxonomic sources, yields and geographic distributions. An important design feature is the ability to rank order organisms as to probability of having specific useful biological activities. The predictive analyses program was developed by WHO. By initially assigning numerical values of interest and negative values to undesirable effects, unbiased simulation can be performed. During the initial phases of drug development, the entire data base can be scanned for useful information. NAPRALERT is used by the National Cancer Institute, and various herbal, pharmaceutical and cosmetic industries. In the future, handbooks will be automatically generated by it.

IV-6 : MAÇCS :

Every organization that makes decisions on the basis of data obtained from biologically active substances is confronted with the problem of organizing and retrieving a multiplicity of data elements on a large number of compounds. Sandoz Pharmaceuticals (New Jersey) and all Swiss branches of ICI chose the Molecular Access System (MACCS), a data management system based on molecular structure, to manage their large chemical and biological data bases. MACCS was developed by Molecular Design Limited (Adamson, et. al, 1985).

MACCS is currently considered to be one of the best systems available for the storage, searching and retrieval of both molecular structure information and associated data. Chemical and biological information is stored together to accomodate modern drug research needs. Storage is open-ended. The system is graphic, interactive and user friendly. It provides storage, searchability and prompt retrieval of molecular information, administrative and biological data.

The chemical-biological information retrieval system is organized as a matrix of compounds vs. data fields. The data fields contain information about each compound. The list of compounds and data fields is sequential and extendable. The data base lacks hierarchy. Three formats are available for each data field : numerical data types and comments (searchable separately); formatted data types (up to 20 characters long) with comments; totally unstructured text data type.

Subsets of these fields may be sent to the report generator. Data types may be searched for a string of characters. Each line is examined sequentially using And or Or. Column searches for value ranges can be performed for active compounds. The column searching capability permits some limited hierarchy in the data base.

Zoned data types may be used to represent disease groups with each zone being two characters long. Each zone is designated for a qualitatively different entry. Blank spaces function as delimiters. The zoned data types create a new substructure within the relational data base. Uniform structuring of data types provides efficient search capabilities. The entries in each line are internally related (eg : they are all obtained at the same time or by the same measurement). Subzones further divide zones. Subzones represent disease subgoals. They are separated by hyphen delimiters. These new data types were implemented in MACCS's text format because there are no format restrictions in the text format. This allows the entire data base to be search in one fell swoop. Subzones are further divided into individual records of test data. Each record has a 3-character long identifier.

At the highest level of the hierarchy, all data belonging to all data types within a disease group can be retrieved or searched by specifying only the first 2 characters or the abbreviation and the wild card character. At the second highest level, all data belonging to a disease subgoal is retrieved or searched by specifying the first 5 characters (including delimiter) of data type name followed by the wild card character. Any test data can be individually retrieved by entering the 7 identification characters (2 delimiters, 5 code characters and the wild card character). In the case of range searches, a hit list is produced. If column numbers are specified, only column search queries may be performed using simple boolean logic.

The format provides neat, easily read tables when printed. It stores and delivers much information. Hard copy reports of the data base can be obtained. Molecular Access Report Generator (MARGEN), also from Molecular Design Limited, is used for this. Each page contains the data types specified. There is header type information with captions. The reports are themselves hierarchical. 24,000 compounds are stored in the data base. They have 200 hierarchically ordered data fields. All in all, there are 40,000 compound/data field combinations, many with multiple line entries.

MACCS has not method of entering text data : a means outside MACCS had to be formed. The first attempt to control the data entry function was a routine called Command Procedural Language (CPL), running under the Prime O/S. CPL allows data to be entered directly into MACCS using a standard system editor. It prompts for the MACCS internal or external registry number and the name of the file. Job steps are used. There are two mores : the input mode and the edit mode. The user can enter data in columns (like QBE). When the file is complete, it may be printed and/or placed in the data base. Only one compound may be entered at a time. Thus, for multiple compounds, the user must make separate entries. Since there is no automatic error checking, the user must be careful.

Another data entry program was written in BASIC, also on the Prime computer. It contains tables for all the biological data type names and numbers. The user is prompted for data. Automatic error checking is done. Help files even contain the phone number of the author and explanations of the most common problems encountered. The output is a data file properly formatted for entry into the MACCS data base.

V-1 : Introduction :

One area of modern chemistry which is receiving considerable attention is the identification of chemical substances from laboratory measurements. NIH/EPA Chemical Information System has been undergoing development for a decade (Milne and Heller, 1980). It is the result of cooperation between many groups in the U.S. and elsewhere. It permits fingerprint recognition in a variety of efficient and inexpensive ways and is used very heavily in this manner by scientists all over the world.

It consists of a collection of chemical data bases together with a battery of computer programs for interactive searching through these disk stored data bases. In addition, CIS has a data referral capabilities as well as a data analysis software system. It has four main areas : numerical, spectral, and toxicological data bases; data analysis software; structure and nomenclature search system; data base referral.

The numerical data bases that are part of the CIS include files of mass spectra, carbon-13 nuclear magnetic resonance, X-ray diffraction data for crystals and powders, mammalian acute toxicity data, hazardous chemical data, water pollutant data, and aquatic toxicity data. There are bibliographic data bases associated with the X-ray crystallography and nuclear magnetic resonance spectroscopy areas. The analytical programs include a family of statistical analysis and mathematical modeling algorithms, programs for the second order analysis of NMR spectra, chemical modeling and energy minimization of conformational structures. Programs that design chemical synthesis are being tested and may, if viable, become part of the CIS in the future.

The center or hub of the CIS is the Structure and Nomenclature Search System SANSS which allows the user to search through data bases of structures such as those associated with collections of mass spectra for occurrences of a specific structure or substructure.

The entire CIS structure can be viewed as a network of independent numerical data bases, linked together through the SANSS hub using the Chemical Abstracts Registry Number (CAS RN) as the unique universal chemical identifier for each compound. Telenet is the telecommunications network used by the CIS.

Over the past few years, a general protocol for the updating of CIS components, or addition of new components has been established. In the first phase of this protocol, a data base is acquired from one of a variety of sources. Some of the CIS data bases have been developed specifically for CIS (eg: the mass spectral data base). Others (eg: the Cambridge Crystal File) are leased for use in the CIS. Yet others (eg: the X-ray powder diffraction file) are operated within the CIS by their owners, in this case the Joint Committee on Powder Diffraction Standards. Sometimes, the information comes from other Government Agencies which retain responsibility for the file, its contents and its maintenance (eg: the NIOSH Registry of Toxic Effects of Chemical Substances. RTECS).

If the data base is to be made searchable, some reformatting, sorting and inversion of files is usually required; this is carried out on an IBM 370/168 which is well-suited to processing large files of data. Once inverted lists have been prepared, they are transferred to the CIS PDP-10 computer which is primarily a time-sharing computer, and programs for generating searchable files and for searching through them. Analytical, data-base-independent programs of CIS are usually written entirely on the PDP-10.

Out of this work, there finally emerges a pilot version of each CIS component. Access to this pilot version on the CIS

PDP-10 is provided to a small number of people. These users are permitted free use of the test component and in return, they test it thoroughly for errors and deficiencies. When testing is complete, the entire component is made available through the operational CIS networked PDP-10. At this point the component is available to the general scientific community, including Government Agencies and is used on the standard CIS fee-for-service basis. In this phase, the system is maintained by a Government contractor. However, efforts are made to generate sufficient interest and use in the system so that other organizations would take over the responsibility of maintenance.

Use of a networked PDP-10 is straightforward and is favored because the alternative philosophy of exporting programs and data bases to locally operated PDP-10 or DEC-20 series is less workable and contains a number of deficiencies that are overcome by a network. Most important among these is the fact that use of a networked machine means that the data bases need be stored only once at the center of the network. Further, a single copy of a data base is easy to maintain, whereas updating a data base that resides on many computers is virtually impossible. Finally, communications between systems personnel and users is very simple in a network environment, as is monitoring of system performance.

CIS is a government supported chemical information project. It is available to all users on a publically available computer system in the private sector. This is in accordance with standard U.S. government regulations. In the middle of 1980, there were over 400 organizations in 17 countries (including Easter Europe), representing over 900 users who accessed the system on a daily basis. CIS is available 24 hours a day, 7 days a week. Each month, about 30,000 searches and related transactions are carried out.

With the current high level of interest in chemical ionization mass spectrometry, there is need for a reliable file of gas phase proton affinities. No data base of this sort has been previously assembled. The task of gathering and evaluating all published gas phase proton affinities was completed by Rosenstock and co-worders at NBS. This file, which has about 400 empirically evaluated gas phase proton affinities drawn from the open literature, can be searched on the basis of compound type or the proton affinity value.

Among other developments in progress is the Congen/Genoa program designed under the Sumex/DENDRAL project. This will be merged into CIS. This program generates structures corresponding to a specific empirical formula. It will be extremely useful in a strategy for structure searching. A reduced set of structures could be produced for consideration by Congen/Genoa. Each structure could then be used as an input in the substructure search system. The various compounds whose registry numbers are so retrieved could be considered to be answers.

The structures proposed from these programs could then be confirmed from the spectral data bases. One can even speculate further to the day when synthetic pathways to any likely but unavailable candidates could be designed by the computer system. It would be easy to add the very practical touch of checking that any starting materials for such syntheses are commercially available at appropriately low costs.

In a different approach, the power of pattern recognition techniques could be assessed within some of the very large files contained in the CIS. This is a very useful exercise because there is little reported work of this sort on large files. Work was begun to explore the value of such methods in handling the problem of identification of true unknowns, such as water pollutants. Programs designed to test mass spectra for the presence of elements or groups (eg : halogens, aromatic rings), have been written and their utility as pre-filters on mass spectral data prior to searching is being tested.

CIS is a powerful networking tool designed to speed access to a wide variety of chemical and non-chemical data bases, coordinate searches, reduce errors, generate properly formatted reports, and reduce search time. Three different search modes are available. In addition, there are a number of utilities to support searches and permit users to customize procedures. Users may share information within a defined user group. The three search options available are : Script mode, and two varieties of Direct mode.

Scripts are merely guided searches. The system leads the user through a series of menus. In contrast to Direct mode, script searches are mode automated. Menus are carefully designed. The user has to learn very little about languages or mechanisms. The name of a script gives a clear idea of its purpose. Scripts are flexible. They search query-list files when looking for terms. Query lists tell remote data bases precisely what terms should be present in the information sought. There are a number of standard query lists designed by professional searchers. Users may use and even modify these query-lists through utilities.

In Script mode, the system permits the user to select the data base, provides query-lists of acceptable terms, helps to formulate the query, automatically dials the system, searches the data base and returns a nicely formatted result. In Direct mode, the system links the user directly to the remote system. More user interaction and knowledge are necessary. At the highest level of user involvement, the system connects the user who then has free rein with the search. However, the user is never without system support. At any time, the user may invoke helpful utilities and procedures without losing his/her work or invoking penalties. This is called Enhanced Direct mode.

In Enhanced Direct mode, the user's query-lists can be used, reformatting can be performed, and editing of data done. Utilities can be used in both script and direct modes. File operations enable the users to list file names, size, last date of modification, contents, and other information. File transfers can be performed. CIS can access over 100 data bases !

V-2 : CAS Chemical Registry System :

This is a computer based system that uniquely identifies chemical substances on the basis of their molecular structure. Stereochemical representation has been an integral part of the CAS Chemical Registry System since its inception as Registry I in 1964 (Blackwood, Elliot and Stobauch, 1977). Stereoisomers are considered to be different individual substances and each is recognized as unique. The unique identification of stereoisomers permits the storage and retrieval of information collected from scientific literature about a specific isomer. In addition, it permits some degree of generic searching by the computer for structures that have a specified stereochemistry. The representations of stereoisomers have evolved in scope and complexity from relatively simple beginnings to sophisticated and highly systematized treatment in the present Registry System.

In Registry I, to accomodate descriptive terms for stereochemistry, a separate segment was added to the atomic bond description or the connection table in the computer structure record. Initially, these descriptive terms consisted of conventional prefixes commonly used in chemistry, names of substances that imply their stereochemistry, and * to denote that stereochemistry was shown in the structural diagram in the original article but was not described in the text.

In Registry I, machine registration was based on matching the unique connection table and stereochemical description against the Registry Master File. Registration was completed by the machine if there was an exact match or if the unique table was different from any other. When there was an exact match on the unique table but not on the stereochemical descriptor, registration was completed only after review by a chemist. The disadvantages were : uncontrolled vocabulary, total lack of machine editing and validation, and machine registration of stereoisomers only after manual intervention.

Building a standardized, controlled descriptor vocabulary began soon after installation of Registry I. For each class of steroids, alkaloids, terpenes and carbohydrates, word roots were established which implied specific stereochemistry at several locations in a basic unit. In addition to the sets of descriptors for these four classes, common stereochemical terms were included in the control file as descriptors. For many products, it was necessary to use prefixes with the terms to indicate the presence of additional or modified stereochemistry. Since the prefix terms were not a controlled vocabulary, their use required the complete descriptor to be considered potentially ambiguous and subject to review.

A machine editing program became possible and necessary. In Registry II, an important feature was that once the descriptor had been machine validated, complete machine registration of structure representation containing valid descriptors could be done without a manual review.

However, more extensive, more exactly defined stereochemical descriptors were needed. Many descriptors were quite complex and consisted of more than one type of information. In most cases, they had to be added to the file with a code indicating editing was to be bypassed. Registration could not be completed without manual intervention by a chemist. The explicit identification of each type of information was thought to be a means for allowing more complete machine editing and providing better control of the terminology. Registry III is a system that edits and validates all stereochemical descriptors and greatly expands automated registration of stereoisomers.

In addition to 7 specific descriptor types, an NS is used when no specific stereoisomer was reported and an * when it could not be expressed by descriptor conventions. An identifier is associated with each type to allow different edits to be applied and to see whether combinations are acceptable.

During input processing, unit-by-unit edits are performed on the individual descriptors. These edits verify that the format used is acceptable, check whether the terms are valid for the indicated descriptor types, check the locants to see if they fall in the accepted range and check whether combinations are valid for complex descriptors. A BNF is used to allow precise definitions. Validation does not necessarily prove that the descriptor is the correct one for that substance. However, valid input is passed onto the processing step. Descriptors containing errors are listed for the chemist to review and the corrected descriptors are sent back through the input edits.

During registrations, Unique Chemical Registry Records (UCRRs) are generated and matched against the Registry Master File. In cases where the topology is matched, the stereochemistry must be matched. Both those UCRRs which have topologies new to the file and those which match both on topology and stereochemistry are completely processed by the machine. Those UCRRs which match on topology but not on stereochemistry go to the In-Context Editing program. The input descriptor is compared to all the descriptors already on file for substances with the same topology. The in-context edits determine descriptor type compatability as well as descriptor content compatability.

The CAS Registry in 1977 contained records for over 3.6 million substances of which 750,000 were stereoisomers.

V-3 : SANSS :

All the compounds in the files of the CIS have been assigned a registry number. This is used to retrieve from the CAS Master Registry (7 million entries), all synonyms that the CAS has identified for the compound. Further, the registry number can be used to locate in the CAS files, the connection table for the compound's structure. This is a two-dimensional record of all the atoms in the molecule, which atom each is bonded to, and the nature of the bonds. The connection table is the basis of the substructure search component of the SANSS (Milne et. al, 1978).

The purpose of the SANSS is to permit a search for a user-defined structure or substructure through the data bases of the CIS. If a substructure is in a CIS data base, then armed with its CAS Registry number, the user can access that file, locate the compound and retrieve the relevant data available.

There- are a number of ways to search the CIS Unified Data Base. The main ones are : Name/Fragment Name search (NPROBE); Nucleus/Ring Search (RPROBE); Fragment Search (FPROBE); Structure Code Search (SPROBE); Molecular weight- molecular formula-partial formula search; total atom-by-atom, bond-by-bond search (SUBSS);

total or full structure search (IDENT).

While structure searching is very important and cannot be replaced by other methods (such as fragment searching, linear notations or name searching), the ability to search for a chemical by name or partial name is very useful. In particular many drugs and pesticides have simple and short trivial names. Name searches are likely to be the best method because such compounds are often complex cyclic structures, difficult to draw. The program (SSHOW) prints out the files containing relevant data (name, molecular formula, structural diagram, correct Chemical Abstracts Index name, and synonyms).

As the first step in a substructure search, the user must define the substructure of interest to the computer. This is done with a family of structure generation commands They can create a ring of a given size, a chain of a given length, a fused ring system and so on. Branches, bonds and atoms can be added and the nature of bonds and atoms can be specified. In the absence of a definition, an atom is presumed to be carbon. As the query structure is developed, the computer stores the growing connection table. If the user wishes to view the current structure at any point, the display command (D) uses the current connection table to generate a structure diagram.

When the appropriate query structure has been defined, a number of search options can be used to find occurrences in the data base. The two most useful search options are : fragment probe and ring probe. The fragment probe will search through the assembled connection tables of the data base for all occurrences of a particular atom-centered fragment (ie: a specific atom together with all its neighbors and bonds).

The ring probe search is a search for all structures in the database containing the same ring or ring system as the query structure. A ring that is considered to be an answer to such a query must be the same size as that in the query structure. It must also contain at least as many non-carbon heteroatoms as the query structure, but the nature of the heteroatoms can be required by the user to be the same or different to that in the query structure. This type of bonding is not considered in RPROBE. Thus, with a query structure of furan, the only exact answer is furan, but the user may permit the retrieval of other answers, including tetra-hydro-furan and thiophene.

In addition to these structural searches, there are a number of special property searches that often prove to be very useful as a means of reducing a large list of answers resulting from structure searches. The special properties include : a specific molecular weight, a range of molecular weights, and a given number of rings of a given size. Searches may also be conducted for the molecular formula corresponding to the query structure, or for other user defined molecular formulas. This may be specified completely or partially. The number of atoms of any element may be entered exactly, or as a permissible range.

If one's purpose is to determine only the presence or absence in a data base of a specific structure, this can be accomplished with the search operation IDENT. IDENT includes the query structure connection table and searches through a file of hash encoded connection tables for an exact match. The search which is very fast with respect to both cpu and elapsed time. It was designed specifically for those users who, to comply with the Toxic Substances Control Act, have to determine the presence or absence of specific compounds in EPA files.

Finally, a link between the CIS structure search system and the vast scientific literature has been established using the Lockheed Dialog system. The link consists of using an intelligent terminal with local memory to store the CAS RN (found in a CIS SANSS search) for later transmission to the Lockheed system (after the user logs off CIS and logs into Dialog) without having to redial the phone (since Lockheed and CIS both use Telenet). This semi-automatic interfacing of different computers holds considerable promise for the future of linking of computers throughout the world.

V-4 : MSSS :

Developed as a joint effort between NIH, EPA, NBS and the Mass Spectrometry Data Centre (MSDC) in England, the current MSSS data base contains about 34,000 mass spectra (Heller, 1985). Computer techniques have been employed to assign every spectrum a quality index. Where duplicate spectra appear in the archive file, the best spectrum is selected for use in the working file.

Searches through the MSSS data base can be carried out in a number of ways. With the mass spectrum of an unknown in hand, the search can be carried out interactively. Extra identifying information can be used. If there are still a large number of answers after entering the peaks, the search can be further reduced to a manageable number of spectra by entering further peaks. In addition, the data base can be examined for all occurrences of a specific molecular weight or a partial or complete molecular formula. Combinations of these properties can be used in searches. Thus, all compounds containing x atoms of type y with a base peak at a particular value m/z, can be identified.

In contrast to interactive searches, which are of little appeal to those with large numbers of searches to carry out, there are available two batch type searches. These accept the complete spectrum of the unknown and examine all spectra in the file sequentially to find the best fits. These are the KB forward search and PBM reverse search algorithms. Spectra can be entered from a teletype. In a more powerful approach, a user's data system can be connected to the network for this purpose and the unknown spectra down-loaded into the computer network for searching.

Once an identification has been made, and the name and CAS RN of the data base compound have been reported to the user, the data base spectrum can be listed or plotted. This facilitates direct comparison of the unknown and standard spectra.

V-5 : CSEARCH :

CSEARCH is a program for the identification of organic compounds and fully automated assignment of C-13 NMR spectra (Kakhauser and Robien, 1985). The number of published reference data exceed many thousand spectra per year. Therefore, computerized data bases are needed. Each reference data set has : entry number; compound name; comment; experimental condition; structure; atom type, connectivity and bond type matrices; solvent, molecular formula; literature; chemical shifts; multiplicities; assignment of resonance lines. Exact duplicate spectra have been removed from the file. The data base (currently 10,000 entries) will benefit considerably from recent international agreements to the effect that all major compilations of CNMR data be pooled.

From the data, several subfiles containing special information can be created allowing efficient execution of different search strategies. Each search function is called by a three letter name. A program called C-13 ADD adds, modifies, deletes, records, and generates lists sorted by name, bibliography or molecular formula. CSEARCH is designed for interactive use but batch is available. In interactive search, a user enters a shift with an acceptable deviation. The algorithm reports the number of file spectra fitting this criterion. The names of the compounds whose spectra have been retrieved can be listed. Alternatively, the list can be reduced by the entry of a second chemical shift. A search for spectra of compounds having a specific molecular formula can also be carried out but there is no capability for searching on molecular weight, a parameter of little relevance in CNMR spectroscopy.

To institute a batch search, a user enters all chemical shifts from the unknown and starts the search. The entire unknown spectrum is compared to every entry in the file. The best fits are noted and reported to the user. The program searches for the absence of peaks in a given region, as well as the presence of peaks. It thus has the capability of finding those compounds which are structurally similar to the material that gave the unknown spectrum.

Commands available are :

ISA :: finds all molecules with a specific molecular formula

MOF :: defines the upper and lower limits
for molecular formulae
This is calculated from a starting point in a
homologous series and an increment.

LIN :: finds spectra with uncommon lines

SUB :: finds spectra with given shift values
and spectral similarities

GRO :: finds groups. It is time consuming
because it involves matrix manipulation
and searches the whole spectrum.

SPH :: performs spectrum interpretation
using all the functional groups

QUI :: produces spectra from formulae

PAR :: performs substructure searches

RIN :: finds ring series

NAME:: finds name fragments.
String searches are accelerated
by screening files encoded in bit patterns.

PLO :: is the graphics package

LIT :: finds all stored bibliographies

ZIT :: finds all cross references

ACC :: records accounting information about operating time, etc.

When a search is complete, the user is provided with the CAS RN of compounds whose spectra fit the input data. The names of the compounds in question are also given. If more information is required, the complete entry for a given CAS RN can be retrieved (structural formula; name; molecular formula; registry number;

experimental data; the spectrum; single-frequency, off-resonance decoupled multiplicities; relative line intensities and assignments).

An interface between CNMR SSS and SANSS allows a user to define a substructure and then examine the chemical shifts associated with particular carbon atoms of interest. The shift data are neatly plotted out to the user with appropriate standard deviations. This should be quite helpful in structure elucidation problems.

Automatic assignment of C-13 NMR spectra can be done. It is a very tedious task. Data is compared to literature and experiments. Simple algorithms then use the structure and resonance positions of atoms to generate a connectivity table. The connectivity table is converted into code. Chemical shift values are encoded from the substructure file. The matrix is mathematically processes so that every row and column has exactly 1 element. Values of 0,1 indicate presence/absence of that single element. Theoretically, this is valid because shifts occur in discrete multiples only.

After extensive testing, it was found that the algorithm works even for complex compounds. Furthermore, the number of errors decreases strongly with a growing reference data set. Erroneous assignment was caused by insufficient representation of the query structure and missing information. A further source of errors were user-defined limits for extension.

V-6 : GINA :

Many proton NMR spectra can be satisfactorily analyzed by hand. Such first-order analysis is a quite satisfactory way of assigning chemical shifts and coupling constants to the various nuclei involved. In certain cases, however, second-order effects become important. As a result, more or fewer spectral lines than are indicated by first-order considerations will result. A way to analyze such spectra is to estimate the various coupling constants and chemical shifts and then, using any of a variety of standard computer programs, calculate the theoretical spectrum corresponding to these values. The calculated spectrum can be compared to the observed spectrum and a new estimate of the data can be made. In this way, by a series of successive approximations, the correct coupling constants and chemical shifts can be determined.

The CIS component GINA (Graphical Interactive Nmr Analysis) (Heller, -1985) permits these operations interactively in real-time. The program was designed for use with a vector-cathode-ray-tube terminal upon which each new theoretical spectrum could be displayed for comparison with the observed spectrum.

V-7 : NMRLIT :

One of the most recent components to be added to the system is NMRLIT (Heller, 1985). It searches a data base of over 34,000 citations drawn from the Preston NMR Abstracts and Index. Currently, abstracts published from 1964 are covered and quarterly updates are planned. Searching the NMRLIT system is identical to the procedures used for the Mass Spectrometry Bulletin system. In NMRLIT, author, subject, journal, and nucleus searches are available. Up to three authors are retained from the original article, along with journal name, volume and page reference. The subject coding is done at the time of abstract reporting by the NMR Abstracts Board of Editors. It covers some 150 categories ranging from NMR parameters, and experimental techniques to characterizations of the substances studied. Each NMRLIT citation gives the Preston Abstracts number for easy access. Perhaps the most useful aspect of the system is the nucleus search.

V-8 : CRYST :

CRYST is a series of search programs working against the Cambridge Crystal File (15,000 compounds), and over 27,000 bibliographic entries dealing with published crystallographic data, mainly for organic compounds (Heller, 1985). The entry for each compound contains the compound name, molecular weight, CAS RN, space group of crystallization, parameters of the unit cell of the crystal as well as full atomic coordinate data. The file may be searched on the basis of any of these parameters. Since all the compounds have been registered by the CAS, structural and substructural searches are available.

Once an entry of interest in the Cambridge X-ray file has been located by one of the search programs, its crystal sequence number can be used to retrieve the appropriate literature reference, structure or atomic coordinate data. The complete literature references information is the basis of a system for searching by author(s), title, journal, date, etc. Free-text searching is available. Boolean logic operators are used to formulate queries.

Once a paper of interest has been identified, all the crystallographic information in that paper can be examined. The crystal sequence serial number associated with the paper is used to retrieve information. Alternatively, the CAS RN of any particular compound can be used to retrieve any data of interest from other files of the CIS.

V-9 : XTAL :

NBS has collected a file of data pertaining to 45,000 crystalline materials (including those in the Cambridge file) (Heller, 1985). The data in the XTAL file includes : cell parameters; number of molecules per unit cell; measured and calculated densities of the crystal; determinative ratios. Every compound is identifiable by its name, molecular formula and CAS RN. In addition, the file can be structurally searched by the CIS SANSS. Searches through this data base for crystals with specific space groups or densities have been developed. It is possible to locate crystals with reduced cells of given dimension. It is hoped that this will prove to be a very rapid method of identifying compounds from readily measured crystal properties.

V-10 : PDSM :

PDSM (Powder Diffraction SysteM) contains 33,000 powder diffraction patterns (Heller, 1985). This work is a direct descendent of Hanawalt's pioneering work. A problem that arises in connection with this particular component stems from the fact that powders are frequently mixtures of different crystalline phases : the patterns obtained experimentally are often combinations of one or more file entries. To solve this difficulty, a reverse searching program examines the experimental data to see if each entry from the file is contained in it. A subtraction routine to help in identifying mixtures has also been implemented.

V-11 : CAISF :

The volume subject indexes to Chemical Abstracts refer to compounds by means of highly systematic nomenclature. The computer readable data base is called Chemical Abstracts Integrated Subject File (CAISF) (Heller, 1985). It is divided into two parts : the General Subject File and the Chemical Substance File. Within each file, arrangement is alphabetical upon parent headings. CAISF differs from most computer readable information files in that different pieces of information referring to a document are not found together, but are scattered throughout an alphabetic index.

CAISF permits substructure searches, using the systematic nomenclature as the structure file. It is linked to searches of general concepts in the remainder of the file. CAISF contains the CAS RN for the compounds indexed so one can obtain the connection tables from the CAS Registry System. From such a subset, a file of bit masks based on structural fragments can be derived.

The scope existed for comparatively studying the effectiveness of a structure text search (using CAISF nomenclature files) vs. a parallel structure text search (using a fragment bit mask file). Lynch and his co-workers developed a sophisticated technique for deriving a fragment code from a connection table file. Fragments are selected for inclusion in the code on the basis of their observed frequency of occurrence. Furthermore, Lynch believed that fragments centered on a bond (rather than an atom) would give superior retrieval performance.

Programs were written to take the incoming CAISF and divide it into two files : the nomenclature file (each entry contained name and CAS RN) and text file (rest of information). Both files were inverted to give search and dictionary files. A permuted dictionary was installed for the nomenclature file. (ie : entries for "sulfuric" would be "sulfuric", "ulfuric", "lfuric", down to "ic"). Leading truncation could be employed. The dictionary was very bulky but it was essential to find embedded terms. Only the document address was provided in the inverted files with no indication of the precise word position. The file was shortened because duplicates were thus removed.

It was found that nomenclature searches gave better precision and the same recall power, but structure searches were better able to handle large files on a small computer in little time. Precision and performance were acceptable in both cases.

Structure-profile writing is not easy but the process can be automated. Nomenclature profiles are also hard to write but since a deep understanding of chemistry and CAS nomenclature is needed, it cannot be computerized. However, the non-availability on a routine commercial basis of the connection table data base tips the scales towards nomenclature searches.

V-12 : RTECS search system :

The National Institute for Occupational Safety and Health (NIOSH) is legally required to prepare a list containing known toxic effects of all chemicals. The Registry of Toxic Effects of Chemical Substances (RTECS) was created (Heller, 1985). It is updated annually. In 1980, RTECS contained 45,000 different chemicals with 70,000 toxicity measurements associated with each of these chemicals. The RTECS data base can be searched in a number of ways : NIOSH number, CAS RN, type of animal tested, route of dosage, LD50, LCLO, and so forth. The NIOSH RTECS file is linked to CIS so that structure-activity relationships may be examined using SANSS.

V-13 : AQUATOX :

Owing to the importance of fish in human nutrition, the danger posed to fish by chemicals was recognized as a major activity of EPA and other U.S. government groups. A data bank of aquatic toxicity was developed by EPA and ASTM (Heller, 1985). This data bank was available on CIS in the middle part of 1981. It has information on chemicals found in fish, reported toxicities, literature citations, common and scientific names of species studied, temperature, ph and hardness the water in the study, salinity of the water and a comments section for other desired information related to the study.

V-14 : WDROP :

The Distribution Register of Organic Pollutants in Water contains : the chemical found, sampling site, data, reporting laboratory, analytical method used and date of record entry. WDROP currently contains over 20,000 citations (Heller, 1985). WDROP is published by the EPA. It is searchable under SANSS and special software that was developed for it. The system is part of CIS. It is possible to answer questions such as : In how many locations is a given chemical or a class of chemicals found ? Are there patterns of distribution of chemicals found in water that indicate problems with plant effluents ? Answers to these and similar questions, coupled with the toxicity data from RTECS and other CIS sources should provide valuable technical information to enable governments to regulate and control pollutants more effectively.

V-15 : OHMTADS :

OHMTADS contains detailed information on over 1,000 chemical substances (Heller, 1985). The information, numerical data as well as interpretative comments, has been assembled from scientific literature. It underwent a major update in 1981. It is expected that over 50 new chemicals per year will be added to the system. The OHMTADS system emphasizes the effects chemical substances can have when spilled. Much more information is provided (synonyms; major procedures of preparation; common means of transportation; flammability; methods of analysis; chemical, physical and biological properties; other parameters). There are a total of 126 fields of information. OHMTADS was first used in a June, 1971 fire in an agriculture chemicals warehouse at Farmville, North Carolina. Since then, it has been used on a regular basis in emergencies by over 100 groups and spill response teams throughout the world.

V-16 : MLAB :

MLAB is a program set which can assimilate a file of experimental data (such as a titration curve) and perform on it any of a wide variety of mathematical operations (Heller, 1985). Included amongst these are : differential and integral calculus; statistical analysis (mean and standard deviation, curve and distribution fitting, linear and non-linear regression analysis). Output data can be presented in any form by a PDP-10 resident graphics program. Data can be displayed in the form of two- or three-dimensional plots. These can be viewed and modified on a crt terminal prior to pen and ink plotting.

V-17 : Chemlab :

A problem of long standing has been to estimate the relationship between the conformation of a molecule in a crystal vs. in solution (where barriers to rotation are greatly reduced). A sophisticated program set for conformational analysis of molecules in solution by empirical and quantum mechanical methods has been developed for this purpose. Chemlab can run batch or interactive (Heller, 1985). As input data, it requires the structure of the compound and this can be provided as a set of coordinate data from X-ray measurements. Alternatively, it can be entered interactively in the form of a connection table, or the program can simply be provided as a CAS RN and if the corresponding connection table is in the files of the CIS, it will use that.

VI-1 : OCETH :

OCETH is a collection of spectra files and associated processing programs developed at the University of Zurich (Wolff and Parsons, 1983). It is organized as a data bank (ie: the user programs do not access the data directly but through a central processing program). The strict separation between the user and the data base helps to secure the integrity of the data; it effectively isolates the user from the internal structure of the data base.

The internal structure of the system is designed to give almost unlimited possibilities for combined processing of all data items without regard to the data type or spectroscopic method. This is realized by having a rigorously standardized format common to all spectroscopic methods presently applied and suitable for future expansion. For practical reasons, data related to different spectroscopic methods is kept on separate files. Updating is much easier with separate files. Also, it provides the scientist with specialized data compilations for his own pet analysis method.

The set of files encompass the full spectroscopic and supplemental information are referred to as Library Files. Their internal structure is optimized for easy maintenance and unlimited future development, not faster access. Each spectrum documented corresponds to one data record in the library file. Each record consists of three segments.

The first segment is of fixed length and holds all information about the chemical (identification number; chemical name; CAS RN; structure; size; empirical formula; nominal molecular mass). Furthermore, it holds the key for interpreting the second segment. The format of the first segment is exactly the same for all library files.

The second segment includes data about the spectrum registration and the sample. It is again of fixed length. However, some entries have different meanings in different library files. The key to their interpretation is part of the first segment. The second segment holds the key for interpreting the third segment.

The third segment is of variable length and contains the spectrum in compressed form. The length of the third spectrum is specified in the second segment, along with the mode of compression used.

The file is preceeded by a header record. This identifies the file and gives its length, source, history and other necessary/convenient data. This cascading data structure (where each part holds the information necessary to correctly interpret and process the following part), makes it possible to write a unified set of programs to handle data from all files. To retrieve any item from any field, the user accesses the central processing program which automatically take cares of the various codes and compression schemes used. The central processing program has access to a library of subprograms designed for various applications. These include routines to output full or partial data sets in standardized formats on various peripherals, programs to generate images of the library in formats suitable for data exchange and for generating index files.

For most standard applications, the library files are not accessed directly. Rather, specialized index files containing appropriate subsets of the full data are used. For example, the permuted WLN code, empirical formula and nominal mass are directly available by means of index files. Experience has shown that many chemists prefer to deal with a hard copy rather than doing an on-line computer search at a terminal. Consequently, the index files are supplied in printed form whenever economy and technology permit. Truncated spectral catalogues are also provided whenever this seems appropriate. This admittedly conservative approach is justified by the fact that chemists are not comfortable with computers. Also, a large proportion of all queries involve a search for a single, fully specified entry, where a manual, telephone-directory type search is adequate.

More complex search problems are done by computer. The most common of these are : library searches and substructure searches. Library search procedures use special search files where selected spectral attributes are included as a binary code. The spectral attributes are selected so as to emphasize structural similarities rather than individual differences between reference spectra. The system accepts a spectrum as input, converts the spectrum into binary source code, and compares it to all reference spectra in the file. A self-optimizing search strategy is used. It produces a ranked list of 10 reference compounds believed to be structurally most similar to the sample. In addition, the spectra may be plotted. Substructure search programs are still in the planning stages. However, programs to convert WLN codes into connectivity tables and pictorial diagrams suitable for output on a line printer have been acquired and integrated into the system.

The spectroscopic data bank provides the base for various research projects. For example, one project aims at a spectrascopy-oriented classification scheme for organic compounds. The classification scheme universally used today dates from the beginning of the century. It is primarily based on the reactivity of the compounds and on functional groups; this makes it optimal for syntheses, reactions, and reaction mechanisms. However, the modern analytical chemist sees his sample from a spectroscopic point of view. In a large compilation of spectra from compounds covering a wide range of compound classes, natural clusters of compounds with common spectral properties are sought. These clusters correspond to spectroscopic compound classes. Initial studies have shown that applying cluster analysis methods to mass spectra represent compounds exhibiting structural similarities not easily described by the classical functional-group vocabulary.

VI-2 : ASTM :

The ASTM file contains 100,000 infra-red spectra (Rumble and Lide, 1985). A text oriented search allows information supplemental to peak positions and intensities to be included in the search (compound name; fragments; functionality; empirical and molecular formulae; strength of the entry; melting point; boiling point; serial number; abundance of chemical classification data). Statistical and pattern recognition techniques are used for a specific subset of spectra. Data was obtained from the American Society for Testing and Materials.

ASTM is slower than many other data bases, but its capabilities are unique. Spectral compression, file inversion and hash ooding are used. Data is encoded in a binary format. A considerable savings of storage space is achieved and an execssive number of matches is avoided. Algorithms are available for retrieving spectra of mixtures. It is available to the Triangle Universities Computation Center.

Because the format is non-standard, a fairly complex PL/1 program is needed to make it compatible with existing mini-computer search systems. The available data is written in 2 different files : the spectral file and the name file. Spectral searches utilize hardware for speed. The following information is converted into 1 character string (profile or code) for each spectrum : (serial number; partial molecular formula; compound name; strong absorptions; peak positions; chemical classification data).

WARPSET compiles the input data (search profiles) into a set designed for sequential and canonical searching of the data base. WARP-8 is a multiple profile search that makes all matches, solves boolean expression and outputs spectral hits. SORPRINT sorts the output according to the profile numbers and prints them. Truncation capabilities (analogous to substructure searches) and wiggle options (permit standard deviations to be specified) are available.

VII-1 : CRYSRC :

CRYSRC is a generalized chemical information system applied to a structural data file (Villareal et. al, 1975) (Bergehoff et. al, 1985). The interactive retrieval system is based upon modular design of input, query and output routines. CRYSRC was implemented and tested on the Cambridge Crystal Data Centre files. Use of 3-d display facilities to create models of the retrieved data are available.

CRYSRC employs the inverted list method of file organization and requires a directory. The directory consists of ordered tables containing pointers to locations of data in the file. Search of the directory is analogous to search of a library card catalog. File generation generally represents building the data file and/or creating a directory to access this file. The pointers in the directory are ordered by keys. It is possible to reduce qualifiers into smaller subsets that involve secondary as well as primary qualifiers.

The number and types of keys are specified during initialization. Screens are permitted to accompany prime key qualifiers. it give a substructure search capability. Screens can be represented as a series of switches that may be one or off depending on the particular structural characteristics that are being represented. The keys permit retrieval on sequence accession numbers, compound names, molecular formulas, author names, journal entries and substructure fragments. Sequence accession numbers are considered important for retrieving previously accessed entries without having to repeat the search and selection process. Accession numbers are assigned to each entry during file generation.

Developing qualifiers for compound names required study. 400 6-character words and word stems are used as qualifiers in the compound-name key-list. Qualifiers for molecular formula retrievals are generated from elements appearing in formulae for various entries. The system is initialized to take into consideration the natural order of occurrence of chemical elements for this particular key and can thus respond more quickly to queries specifying the most frequently occurring elements. Chemical elements are the primary qualifiers and the number of atoms per element are the secondary qualifiers.

The surnames of authors (coded) are primary qualifiers. The first names (initials) are secondary qualifiers. There is a 3-letter numeric code for journals. This reduces the number of entries. The final type of key are predefined substructurew, each associated with a label and a screen. Inverted lists are used as qualifiers, and point to entries containing particular fragments. Retrievals on individual substructure fragments or combinations thereof, use only a short time (under 1 minute) but tremendous storage space.

The exact nature of the structural characteristics to be represented by screens was determined by analyzing sample files from original crystallographic chemical files. Direct retrievals on a number of prechosen fragments is permitted. This may lead to the the retrieval of irrelevant fragments but is better than an exhaustive search of the entire data base. The CRYSRC system permits accession number, keyword and screen retrievals, as well as browsing. 4 search methods are available.

In the first and simplest approach, the user specifies the group number (partition number) and the entry number of the required compound. He must have previous knowledge of the coding scheme employed.

The second method is browsing. This can be entered at any starting location. It lets the user just browse through the entries. Once initiated, users can interact with the system to control the display of information, skipping of entries or saving of entry information. Exiting is done at the user's discretion.

The third approach depends heavily on keys defined during system initiation, and combinations of the keys with primary and secondary qualifiers. All queries must be expressed in disjunctive normal form.

The last method is screen searches. This depends on predefined screens. The query screen is input in the form of a connectivity table. Screen searches are very dependent on the nature of the screen-building process, and the quality and range of characteristics represented by screens in the qualifier lists.

These four methods provide several alternatives for retrieval. Very narrow or very broad queries can be specified. The user has the option of processing query results through a flexible set of output-processing routines. The output processing mechanism adds an additional dimension to the system. Submodules of the output processing modules can be added and/or modified to meet changing data or hardware requirements without affecting the rest of the system. 3-d graphics display is available. Interactive graphics make complex structural models available to the user. They help overcome a limitation of textual data to describe special relationships in a holistic form. The user is free to ask general geometric question about the molecular model in 3-d. CRYSRC is easily portable.

VIII-1 : DETHERM :

Acquisition of the physical values of chemical compounds is still one of the main problems for engineers and chemists in designing chemical equipment and plants. It is not certain that an exhaustive search of the literature will produce all of the characteristics required for the desired detailed design. This is particularly true for mixtures. Therefore, in many cases, it is necessary to fit physical characteristics from basic data to the process conditions by suitable computations, especially when it is not possible to carry out measurements oneself.

It is desirable to have a physical properties data bank for similar and recurring problems which will provide the desired data, tailored as far as possible to the application. The data bank, which is to serve chemical engineering and plant construction, must make available, along with stored literature data, data computed for any desired process conditions or design details. It is obvious that this can be accomplished only by using large computers if, in addition to storage, rapid processing of the requests is also desired. DECHEMA has developed such a computer system. It is called DETHERM (Heller et. al, 1976).

Quantifying substance-specific characteristics of chemicals is important in predicting their behaviour. This quantification is done in accordance to the document DIN 1313. Some of the physical properties recorded are : the pressure, volume and temperature behaviour; caloric properties; phase equilibria; transport properties (viscosity, thermal conductivity, electrical conductivity); boundary-surface properties; acoustic properties; optical properties; safety engineering characteristic data; molecular properties.

Sometimes, properties are used to determine other properties. For instance, the velocity of sound can be used to calculate caloric properties. Acquiring, analyzing, storing and retrieving data is very expensive. Calculating data from stored data is a way of minimizing costs. For economic reasons, only 3,000 of the commercially most important compounds are included; unusual compounds of purely scientific interest are totally excluded.

Physical property data is available for only a small fraction of the exceptionally large number of chemical compounds (ie : only for the most commonly used ones). Even for these compounds, data is available over only a small range. Only a very few substances, like water, have been thoroughly examined. The amount of information available is deplorably insufficient. Sources of physical property data are periodicals, secondary literature and tertiary literature. There are also some monographs. Acquisition of data from these sources is expensive and tedious. Often, the data may not be as precise and reliable as necessary.

Determining the quality of the data is too complex for the user; it has to be performed by the information specialist. The expert evaluates the information in the following way :

1. what properties are available ?
2. acquire and read the source literature
3. examine and criticize the collection method
4. decide whether or not the data is sufficiently reliable

Examining the data and deciding if it meets standards is not easy. Tremendous expertise is required. Statistical programs are available to perform regression and correlation. Programs are available to check whether the data fits a pattern (eg : homologous series), or if it is consistent with other data. Evaluation is extremely complex and tedious.

Computer programs must be available to provide calculated data for mixtures. (All possible mixture permutations cannot possibly be stored, yet their properties are needed). It is necessary for the data to be easily accessible by these programs. When physical properties are unavailable, programs must be able to estimate them. DECHEMA therefore consists of three parts : the stored data, computing programs, analysis programs.

The physical property data retrieval system (SDR) stores and retrieves the data. It includes data files for bibliographic data, remarks, components (name, empirical formula, CAS RN), and data (property identifier, value, unit, error, error unit). Descriptors (TT) are used only when the theoretical information is to be encoded or computation methods characterized. Chemical compounds can be retrieved by name, empirical formula and CAS RN. Additional identifiers of the component field indicate what type of compunds (organic, inorganic) and what system (pure, mixture) and whether or not there is a reaction mixture. Remarks are necessary when the equations must be explained or abstracts stored.

The data field is the most important part. All fields are stored identically. Any conversion is carried out by special programs. Data is always stored in a record containing :

property (TAG); value (VAL); unit (UNT); error (ERR); error units (ENT). Searches can be carried out on the property identifier, TAG, with any kind of boolean logic. The result of a search is an identification of the document and a printout of the values obtained. It is possible to search for substances or mixtures with particular properties. This can be extended to estimating physical properties, if these are unavailable.

SDC consists of programs for computing pure substance and mixture properties. It uses the pure-substance data base and calculates pure substance and mixture properties for any desired conditions of state. The computing programs retrieve the necessary data, and with user-input parameters, compute the required results. Several computation methods are available (UNIQUAC, UNIFAC, NRTL, Lee-Kesler). In calculations involving mixture properties, the programs first check whether or not the mixture is in phase equilibrium.

The third element of DETHERM, the analysis (SDA), consists of statistical analysis programs that evaluate literature and experimental data for special properties, correlate the data with any models and create the entry records. SDA is only an aid, not a substitute, for the human expert. Systematic deviations in the data are recognized. The appropriate models are consulted to see how reliable the data is (ie: how unexpected are the values ?). Obviously incorrect data is thus eliminated. The result of a complete analysis is the attainment of constant pools and thermodynamically consistent values. SDA develops its full effectiveness in combination with the retrieval system by subjecting the sets of data to direct analysis.

DETHERM is available through the DECHEMA information system. DECHEMA is accessible on-line through EURONET DIANE. DECHEMA performs all inquiry processing.

IX : Methodology :

Unfortunately, due to the nature of the problem, the methodology used was highly subjective, diffuse, and imprecise. Various chemical IS&R systems were examined. Important and new features and design ideas were stressed. From these items, current trends were extrapolated, and possible future developments predicted.

Some other researcher may have deemed different features important, and drawn different conclusions about current trends and likely future developments.

The methodology is definitly the major weakness of this study. Since methodology is the foundation stone of any study, the validity of this entire paper is questionable.

In all fairness, I will add that the methodology was forced by the nature of the very problem investigated. I could think of no alternative way of achieving my goals.

X : Status :

This project has been successfully completed. Its purpose was to provide an introductory tutorial to chemical IS&R systems, and to discuss major development trends. This was done.

This survey is not comprehensive, nor was it ever intended to be. Only some major chemical IS&R systems that contributed new research ideas, and novel design features were examined.

A major problem was the language barrier. A lot of papers are written in foreign languages, like German, French, Polish and Russian. As I do not know these languages, and cannot read these papers, I could have easily overlooked a major development. This is quite a serious shortcoming of this survey.

XI-1 : Summary Characteristics :

Derwent's Patent System (III-1) :

1- abstracts of chemical patents

2- retrieval by inverted indexes of patent classes, codes, and title keywords

3- 12 sections to serve specialized chemical needs

4- different search and retrieval capabilities in each

5- different types of reports for various types of users (engineers, lawywers, scientists)

6- very large database

7- spelling problems

8- emphasis on alerting services and collecting abstracts suitable for browsing

9- system was not designed for retrieval

10- data compression used

11- fixed length records

The IFI Comprehensive D.B. (III-2) :

1- massive database

2- merger of Du Pont and IFI Uniterm Indices

3- inverted indexes with retrieval by patent class, compound and structural components

4- controlled, open-ended vocabulary

5- retrieval by accession number

6- ISAM

7- free-text searching

8- complex weighting system to increase search efficiency

9- fixed-length records

PULSAR (III-3) :

1- microcomputer system

2- data stored in a large binary tree

3- personalized keywords

4- inverted index on keywords, and all components of a bibliography (date, author, journal)

5- 20,000 articles

6- standardized spelling routines provided

7- fixed with overflow records

Chemfile (IV-1) :

1- sequential storage of records

2- within records, hierarchy

3- input error-checking done

4- user interface grahical

5- designed for verification of data

6- not designed for retrieval

PAGODE (IV-2) :

1- hash coding

2- topological code used

3- central feature are the topological screens

4- screens correspond to structures and substructures

5- searching and retrieval done via screens

6- inverted indices of screens

7- fixed length with overflow records

8- only 6000 lines of assembler code

9- noise

CBF (IV-3) :

1- ISAM

2- within records, hierarchical

3- based on connectivity tables

4- not based on screens

5- search arguments may be entered as screens

6- information specialists perform searches

7- output may be given in screen form

8- codes used for compression

9- codes correspond to structures

10- inverted indices of codes

HEEDA (IV-4) :

1- node of CSIN

2- shallow but strict hierarchy

3- searches done by CSNS

4- focus on structure-activity relationships, and toxicity rankings

5- variable length records with delimiters

6- open-ended qualifier lists

7- codes for compression

8- pretty printing

9- modelling, prediction, and statistical packages

NAPRALERT (IV-5) :

1- hierarchical

2- inverted indices

3- controlled, open-ended vocabulary

4- codes for compression

5- fields for storing numerical indicators of data quality

6- quality indicators internal to system

7- used by prediction and analysis programs

MACCS (IV-6) :

1- relational data base

2- individual data fields support hierarchy

3- extendable, open-ended

4- fixed-length records

5- coding for compression

6- graphical and non-graphical user interfaces

Networks (V) :

1- collection of data bases
   holding a variety of information
2- linked together and accessed via a network
3- shared library programs for common needs
   (searching, statistics)
4- uniformity of data
5- standard interfaces between data bases and programs
6- data referral
7- massive investment by governments and corporations
8- cooperation between competitors
9- gigantic databases
10- tremendous help facilities for end users
11- classification scheme unambiguous (CAS RN)
12- economic feasability
13- relevance of data vs. low cost

OCETH (VI-1) :

1- data partitioning

2- cascading data structures

3- easy maintenance and unlimited future expansion

4- rigorously standard formats

5- fixed and variable length fields

6- codes for compression

7- central search/retrieval programs

8- users denied direct access

ASTM (VI-2) :

1- hash coding

2- data partitioning

3- non-standard format

4- inverted indices for non-spectral data

5- tremendous spectral search capabilities

6- complex pattern recognition algorithms

7- hardware searches

8- complex compression algorithms

CRYSRC (VII-1) :

1- inverted list organization

2- directory is central feature

3- controlled, open-ended vocabulary

4- screens for inputting search arguments

5- browsing

DETHERM (VIII-1) :

1- relational

2- insufficient data

3- data quality highly variable

4- analysis programs evaluate data quality

5- computing programs deduce unrecorded information

6- data fields may be null if value unknown

XI-2 : Current Trends :

-1- The range of data, capabilities provided by chemical IS&R systems, and their organizations (hierarchical, sequential, relational) is so great that inter-system comparisons are difficult to make.

-2- User friendliness and ease of use is a guiding design principle of all systems. Input routines are frequently graphical, and always user-friendly. Output routines stress tabulation and pretty printing.

-3- The CAS RN system of compound identification is used in almost all systems. It is an internationally accepted standard. It is suitable for computer use.

-4- Chemical IS&R systems in current use are constantly updated to incorporate technological advances. For instance, when time-sharing became economically viable, all major batch systems were modified within 4 years.

-5- Chemical information systems based on networks are proving to be economically viable and extremely popular.

I think that networks will be the main source of chemical information in the future.

-6- Networks are promoting standardization of data format and interfaces between data bases and programs. International agreements about standards have been reached.

-7- Topological screens and screen-based searches were a major breakthrough. Currently, they are gaining in popularity. They are often incorporated into systems where data is organized by some other means, to provide input/output facilities.

-8- The provision of search capabilities in hardware vastly reduces search time. The greatly reduced cost of hardware makes it economically viable. However, it is easy to implement only if considered at design stage. Thus, I think it will be a feature of future systems, but will not be added into current systems.

-9- Microcomputer-based chemical IS&R systems are currently an area of R&D. Memory quotas are the main drawback to their widespread use. In the future, technology breakthroughs may occur that cause microcomputer-based chemical IS&R systems to be viable propositions.

-10- A lot of research is going into reliable and efficient data compression algorithms. This is particularly true for spectral files.

-11- Pattern recognition algorithms are yet another major R&D area. Work is being done primarily on spectral files, and structural searches.

-12- A lot of work is being done on clustering algorithms as an alternative way of building data bases (ie: instead of ISAM, relational or hierarchical).

-13- The great variety of data and the tremendous range in their quality is causing severe problems. The estimation of data quality is a vast and complex subject. Little progress has been made in discovering standard quality metrics. I think that much more work will be done in this area in the future.

XII : Bibliography :

Adamson, G. W., J. M. Bird, G. Palmer, and W. A. War, "Use of MACCS within ICI," Journal of Chemical Information and Computer Sciences, Vol. 25, No. 2, 1985, pp. 90-91.

Becker, J., D. Jung, W. Kalbfleisch, and G. Ohnacker, "CBF - Computer Handling of Chemical and Biological Facts," Journal of Chemical Information and Computer Sciences, Vol. 21, No. 2, 1981, pp. 111-116.

Berdago, S., J. Boitard, J. P. Gervois, A. M. Segretain, and O. Pietrement, "PAGODE : The Computer-Based Chemical Information System of CLIN-MIDY Research Center," Journal of Chemical Information and Computer Sciences, Vol. 18, No. 3, 1978, pp.181-186.

Bergehoff, G., R. Hundt, R. Sievers, and I. D. Brown, "Use of MACCS within ICI," Journal of Chemical Information and Computer Sciences, Vol. 25, No. 2, 1985, pp. 90-91.

Bernin, C. L. "Development of Indexing and Indices," Journal of Chemical Information and Computer Sciences, Vol. 25, No. 3, 1985, pp. 164-169.

Blackwood, J. E., P. M. Elliot, R. E. Stobaugh, and C. E. Watson, "The Chemical Abstracts Registry System III Stereochemistry," Journal of Chemical Information and Computer Sciences, Vol. 17, No. 1, 1977, pp. 3-6.

Donovan, K. M. and B. B. Wilhide, "A User's Experience with Searching the IFI Comprehensive Data Base to U.S. Chemical Patents," Journal of Chemical Information and Computer Sciences, Vol. 17, No. 3, 1977, pp. 139-142.

Fugmann, R. "Peculiarities of Chemical Information From a Theoretical Viewpoint," Journal of Chemical Information and Computer Sciences, Vol. 25, No. 3, 1985, pp. 174-179.

Graham, W. "CHEMFILE : An In-House Information System for the Chemical Indexing of Abstracts on Health Effects of Environmental Pollutants (HEEP)," Journal of Chemical Information and Computer Sciences, Vol. 17, No. 4, 1977, pp. 200-201.

Heller, S. R. "The Chemical Information System and Spectral Data Bases," Journal of Chemical Information and Computer Sciences, Vol. 25, No. 3, 1985, pp. 224-230.

Heller, S. R., G. W. A. Milne, and R. J. Feldman, "Quality Control of Chemical Databases," Journal of Chemical Information and Computer Sciences, Vol. 16, No. 4, pp. 232-233, 1976.

Kaback, S. M., "A User's Experience with the Derwent Patent Files," Journal of Chemical Information and Computer Sciences, Vol. 17, No. 3, 1977, pp. 143-147.

Kakhauser, H. and W. Robien, "CSEARCH : A Computer Program for Identification of Organic Compounds and Fully Automated Assignment of Carbon-13 Nuclear Magnetic Resonance Spectra," Journal of Chemical Information and Computer Sciences, Vol. 25, No. 2, 1985, pp. 103-107.

Lefkowitz, D., A. Rispin, C. Kulp, and H. Hill, "EPA Health and Environmental Effects Data Analysis System," Journal of Chemical Information and Computer Sciences, Vol. 21, No. 1, 1981, pp. 18-28.

Loub, W. D., N. R. Farnsworth, D. D. Soejarto, and M. M. Quinn, "NAPRALERT : Computer Handling of Natural Product Research Data," Journal of Chemical Information and Computer Sciences, Vol. 25, No. 2, 1985, pp. 99-102.

Milne, G. W. A., S. R. Heller, A. E. Fein, E. F. Frees, R. G. Marquart, J. A. McGill, J. A. Miller, and D. S. Spiers, "The NIH-EPA Structure and Nomenclature Search System," Journal of Chemical Information and Computer

Sciences, Vol. 18, No. 4, 1978, pp. 181-186.

Milne, G. W. A. and S. R. Heller,
"NIH/EPA Chemical Information System," _Journal of Chemical Information and Computer Sciences_, Vol. 20, No. 4, 1980, pp. 204-211.

Rumble, J. R. and D. R. Lide,
"Chemical and Spectral Data Bases : A Look into the Future," _Journal of Chemical Information and Computer Sciences_, Vol. 25, No. 3, 1985, pp. 76-83.

Smith, S. F., W. L. Jorgensen, and P. L. Fuchs,
"PULSAR : A Personalized Computer Based System for Keyword Search and Retrieval of Chemical Information," _Journal of Chemical Information and Computer Sciences_, Vol. 21, No. 4, 1981, pp. 209-212.

Villareal, J., E. F. Meyer, R. W. Elliot, and
C. Morimoto, "CRYSRC : A Generalized Chemical Information

System Applied to a Structural Data File," Journal of Chemical Information and Computer Sciences, Vol. 15, No. 4, 1975, pp. 63-72.

4.24

1. Report No. IN-82

2. Government Accession No. 183569 ~~[illegible]~~

3. Recipient's Catalog No.

4. Title and Subtitle
USL/NGT-19-010-900: A SURVEY OF CHEMICAL INFORMATION SYSTEMS

160 p.

5. Report Date
December 5, 1985 DATE OVERRIDE

6. Performing Organization Code

7. Author(s)
ANEESA BASHIR SHAIKH

8. Performing Organization Report No.

9. Performing Organization Name and Address
University of Southwestern Louisiana
The Center for Advanced Computer Studies
P.O. Box 44330
Lafayette, LA 70504-4330

10. Work Unit No.

11. Contract or Grant No.
NGT-19-010-900

12. Sponsoring Agency Name and Address

13. Type of Report and Period Covered
FINAL; 07/01/85 - 12/31/87

14. Sponsoring Agency Code

15. Supplementary Notes

16. Abstract

This Working Paper Series entry represents a survey of the features, functions, and characteristics provided by a fairly wide variety of chemical information storage and retrieval systems currently in operation. The types of systems (together with an identification of the specific systems) addressed within this survey are as follows: Patents and Bibliographies (Derwent's Patent System; IFI Comprehensive Database; PULSAR); Pharmacology and Toxicology (Chemfile; PAGODE; CBF; HEEDA; NAPRALERT; MAACS); Networks - The Chemical Information System (CAS Chemical Registry System; SANSS; MSSS; CSEARCH; GINA; NMRLIT; CRYST; XTAL; PDSM; CAISF; RTECS Search System; AQUATOX; WDROP; OHMTADS; MLAB; Chemlab); Spectra (OCETH; ASTM); Crystals (CRYSRC); and Physical Properties (DETHERM). Summary characteristics and current trends in chemical information systems development are also examined within the report.

This report represents one of the 72 attachment reports to the University of Southwestern Louisiana's Final Report on NASA Grant NGT-19-010-900. Accordingly, appropriate care should be taken in using this report out of the context of the full Final Report.

17. Key Words (Suggested by Author(s))
Chemical IS&R Systems, Information Storage and Retrieval Systems

18. Distribution Statement

19. Security Classif. (of this report)
Unclassified

20. Security Classif. (of this page)
Unclassified

21. No. of Pages
160

22. Price*

*For sale by the National Technical Information Service, Springfield, Virginia 22161