Chemical Information from DIALOG Information Services

PETER F. RUSCH

DIALOG Information Services, Inc., Palo Alto, California 94304

Received June 3, 1985

The DIALOG information retrieval software was designed in the 1960s to provide interactive searching of large databases by alphanumeric search terms, in both Boolean and proximity operations. Chemical information databases were adapted to this storage and retrieval system in the 1970s, leading to the first on-line availability of *Chemical Industry Notes*, *CA Index Guide*, *CA Patent Concordance*, and the *CA Subject Index Alert* (CASIA) in combination with the *CA Condensates* file. Complimentary chemical substance databases were created with unique features and enhanced content. The sources, derivation, and uses of these on-line databases and many of the search terms are described.

INTRODUCTION

The DIALOG† Information Retrieval Service had its origins in the early 1960s as a research project of the Lockheed Missiles and Space Company Research Laboratories in Palo Alto, CA. After the creation of several versions of information retrieval software, two government agencies took delivery of the search system and began operating it for their own internal purposes. One of these, the NASA installation, remains active today and is accessed frequently by both agency employees and contractors. The other system, sent to what is now the U.S. Department of Energy, is also still active and is operated on a daily basis from the computer center at Oak Ridge, TN. Each of these government agency versions of the software has evolved in a way somewhat different from that of the DIALOG on-line commercial service.

By the late 1960s, work was well under way to modify the computer program for commercial purposes. By 1970 the commercial version of the DIALOG Information Retrieval software had been created and was made available with one database, ERIC (Educational Resources Information Center), which was made available to five education centers sponsored by the National Institute of Education.

Subsequently, the DIALOG on-line program was modified to allow access to two databases, ERIC and NTIS (National Technical Information Service), accessed by more than two customers, which was a major milestone. By 1972 the DIALOG Information Retrieval Service[1] was a commercially available service operating, primarily through a leased-line network, throughout the U.S. Overseas access was available through direct telephone communications. In 1981, the service was separately incorporated as a wholly owned subsidiary of Lockheed, DIALOG Information Services, Inc.

Over the years, the DIALOG Information Retrieval Service has grown substantially to the point where now more than 50000 customers have DIALOG passwords and there are more than 200 different databases available on-line during the complete hours of the service. Although many of these databases are scientific and technical in nature and, therefore, contain chemical information, the purpose of this paper will be to concentrate on a few specific chemical information databases. Among those of greatest interest in this will be databases licensed from Chemical Abstracts Service (CAS) and transformed for the on-line offering through DIALOG Information Retrieval Services. In many cases, unique algorithms have been developed for processing the incoming data to yield new search terms, and modifications have been made to the DIALOG search system to accommodate chemical information needs.

†DIALOG is a registered trademark of DIALOG Information Services, Inc.

Peter F. Rusch received his B.S. and M.S. degrees in chemistry. He graduated from the University of Texas at Austin in 1971 with a Ph.D. in physical-inorganic chemistry. After postdoctoral research appointments in France and the U.S., he joined Chemical Abstracts Service as an information scientist in 1973. In 1975, he joined DIALOG Information Services, where he is now Director of the Chemical Information Division. He has been Assistant Secretary and is presently Chairman of the American Chemical Society's Division of Chemical Information.

EARLY DATABASES

In the early 1970s, Chemical Abstracts Service began to make available for license its CA Condensates file. This file contained bibliographic information and the keyword phrases from each of the weekly issues of *Chemical Abstracts*. This file was primarily made available from the beginning of the *Ninth Collective Index* period in January 1972. By 1973, at least one on-line offering of the CA Condensates file was available commercially. In 1974, DIALOG brought up its version of the CA Condensates file as File 3. One of the distinguishing features of this implementation of CA Condensates was the fact that the title words, in addition to keywords and other bibliographic information, had been made available for direct search. In comparative tests, the DIALOG implementation of CA Condensates provided considerably greater numbers of items retrieved for any given search statement. Due to storage constraints in the early 1970s, some of the proximity information, although available in the internal

0095-2338/85/1625-0192$01.50/0

files, was removed from the on-line version of the files available to the public. This constraint in the CA Condensates files was removed by 1975. Interestingly, this plan of making all fields available with proximity search where appropriate characterized the development of the chemical information files on the DIALOG system as a unique feature of that system until the beginning of the 1980s.

Beginning in late 1975, DIALOG Information Services took an increased interest in licensing of databases from Chemical Abstracts Service. At the second DIALOG customer meeting in 1974, discussion between DIALOG and CAS staff revealed that the printed publication *Chemical Industry Notes* (CIN) was available as an in-house file through one of the large Italian chemical companies. CAS, recognizing the desirability of a computer-readable version of the CIN file, decided to begin keyboarding operations in order to capture the back-file and continue the availability of CIN as a computer-readable file in the future. By late 1975, DIALOG had obtained the rights to use this file by way of a license from a third party. The implementation of the CIN file was carried on through late 1975, and by January of 1976, the CIN file was available on-line for the first time commercially anywhere in the world. The early success of the CIN file was probably somewhat mixed, drawing only about 20 h of usage per month in the entire customer community. The major significance of the availability of CIN on-line from DIALOG was that a great deal of experience was gained with the Standard Distribution Format (SDF) that had been adopted by CAS for all of its computer-readable products. As had been anticipated by CAS and others, the availability of SDF would permit licensees to use certain transformation programs repetitively for different CAS databases. This proved to be the case, and work was begun at DIALOG on a variety of other files that became available throughout 1976.

INDEX AND REGISTRY DATABASES

Among the files available from CAS in the mid-1970s was the CA PATENT CONCORDANCE. Because of the internal processing of documents by CAS, this was the only file that contained equivalent patent numbers for those chemically related patents covered by CAS. The file was licensed and made available from DIALOG in late October 1976. Initially, the file was constructed as essentially an additional index to the CA Condensates file covering the same time period. Something close to 200 000 equivalent patent numbers had been made available in this way, and with the DIALOG "file switch" feature it was possible for a customer to enter the CA PATENT CONCORDANCE, select an equivalent or covered patent number, switch to the CA Condensates file, and produce an output of the CA CONDENSATES record for which the equivalent patent number had been found. This proved to be an efficient and low-cost means of making available this unique set of data. Throughout 1976, however, another collection of CA CONDENSATES data, covering the final 2 years of the *Eighth Collective Index* period (1970–1971), became available. DIALOG made this file available to the customer community in November of 1976. With portions of two collective index periods available in two separate CA CONDENSATES files, the CA PATENT CONCORDANCE had to take on a different formulation. In 1977, the file was reformulated to become its own DIALOG file, permitting direct search of the equivalent patent numbers with an accession number pointer in each record to the appropriate CA CONDENSATES file containing the patent covered by CAS. This form of the CA PATENT CONCORDANCE file became widely used.

By late 1979, it was clear that Chemical Abstracts Service would withdraw this file from license because of arrangements that had been made with the International Patent Documentation Center (INPADOC) of Vienna, Austria. In January 1980, the CA PATENT CONCORDANCE was withdrawn from license and removed from the DIALOG Information Retrieval Service.

Among other files that were developed by Chemical Abstracts Service and were available for license during 1976 was the Chemical Abstracts SUBJECT INDEX ALERT (CASIA). This file was the result of a large amount of planning and processing by Chemical Abstracts Service. What made this file so valuable for consideration as an on-line file was that it contained information organized by CA Abstract Number, the same organization as the CA CONDENSATES file. The CASIA file from CAS contained the volume general subject and chemical substance index entry information produced by CAS for its own printed indexes. Due to internal processing considerations, the data were not available until at least 6 weeks after the appearance of the bibliographic and keyword information in the corresponding CA CONDENSATES file.

Some early test results in 1975 indicated that it might be possible, with the proper understanding of the database content and format, to offer the CASIA file on line. To assist with this process, CAS staff met with several of the on-line services, including DIALOG, to discuss the possibilities. Among the major problems that were considered at this time were the number of data fields available on the CASIA file, the size of the file itself and the on-line storage that would be required to offer it commercially, and the fact that the data appeared asynchronously with the bibliographic component of the CAS database. DIALOG staff began working in early 1976 to overcome the three major problems.

The first problem to be overcome was the asynchronous appearance of the data on the CASIA file with respect to the CA CONDENSATES file. The title, author, other bibliographic and keyword information would be made available, grouped by CA Abstract Number in the CA CONDENSATES file. From 6 to 18 weeks later, the volume index entries would be made available on the CASIA file for the same CA Abstract Number. At that time approximately 80% of the items had their volume index entries available on the CASIA file within 10 weeks of appearance of the CA CONDENSATES record. The DIALOG Information Retrieval System was particularly well adapted to permit the asynchronous appearance of data. The CA CONDENSATES file and CASIA file could be independently delivered and converted to their respective search terms. This would be done in a way so that accession numbers would be uniquely assigned for each abstract number, thereby permitting the search terms from either file to retrieve the same original reference. In the earliest versions of the DIALOG search software delivered to NASA, provision was made for the appearance of two distinct display files, each keyed by the same accession number. Consequently, the CA CONDENSATES data could be immediately made available on its display file while the CASIA information could be independently loaded to another display file as it became available.

During the mid-1970s, IBM-compatible disk storage became more widely available at prices that were decreasing with time. In spite of this benefit, the amount of storage that would be potentially required for the additional CASIA information was quite large. In order to evaluate the amount of storage and prepare for it, a thorough analysis was given to the content of the data fields on the CASIA file. CASIA offered the clear advantage of having numerous unique fields that would influence both the recall and precision of searching for items in the combined CAS database. Prior to this time, the National Library of Medicine (NLM) had instituted a chemical name dictionary[2] (CHEMLINE) as an adjunct file to the TOXLINE database. In this implementation, the literature

file (TOXLINE) had the CAS Registry Number as the only controlled-vocabulary reference to chemical substances. A similar solution[3] could be applied to the implementation of the CASIA file in conjunction with CA CONDENSATES. To do this, it was necessary to remove from the CASIA file the chemical-substance identification fields dealing with nomenclature and molecular formulas, leaving only the CAS Registry Number in a chemical-substance index entry with the corresponding modifying information, if present. By adopting such a solution, both the size of the CASIA file was reduced and the CHEMNAME file was created as the largest single chemical name dictionary of its kind.

The general-subject information of the CASIA file was reduced by taking only single terms from the index modification field attached to either the general-subject or chemical-substance index headings. In this way, an independent unit record from CASIA was constructed. This record contained complete, qualified, CA general-subject index headings, CAS Registry Numbers, and the unique modifying words from the index entries. The first implementation of this CASIA file came about as DIALOG File 30 in April 1976, containing items from the CASIA file corresponding to CA CONDENSATES records in File 3. Under this implementation, a searcher could search in either the File 3 or File 30 and, by using the "file-switch" mechanism of DIALOG, continue the search, by selecting terms or by combining sets in either of the files. The file-switch feature of DIALOG had been available for a number of years and permits a searcher to change from one file to another without deleting sets.

As the index entries from CA chemical-substance information were reduced by retaining only the CAS Registry Number and unique words from the modifying phrase, it became clear that some significant information was being lost. To compensate the customer community for this loss, a development was undertaken to restore some of the information in an abbreviated form through the use of letters as suffixes to CAS Registry Numbers.

Many of the chemical-substance index entries appeared in printed form and in computer-readable form at that time with no modifying phrase whatsoever. It was the intent of CAS in the production of such an index entry to reduce space required in the print product and to infer from such an entry the preparation of the chemical substance being referenced. Alternatively, preparation of the chemical substance could have been indicated through use of the qualifier data element in which the word "Preparation" would appear for any one of a predetermined group of CAS Registry Numbers, or the term "preparation", its abbreviation, or an appropriate synonym, such as "production", might be part of the modifying phrase. By means of a syntactical analysis of eac of the modifying phrases for a chemical-substance index entry, it was possible to add the suffix "P" to the CAS Registry Number for those entries clearly indicating the preparation of the chemical substance represented by that CAS Registry Number. This algorithm proved highly successful in its ability to discriminate among chemical-substance index entries concerning preparation and those that were concerned with other properties or uses. Even though all of the modifying phrases had been reduced to a set of single terms, this important context, preparation of chemical substances, was restored to the search terms with good recall and high precision while reducing the storage required to contain the entire CASIA record.

Following the general principles of the availability of search terms in a DIALOG file, the general-subject index headings were made available for search as the complete general-subject phrase, the phrase with any qualifiers that might be attached, or the individual words from the entire qualified general-subject heading. In a similar manner, the CAS Registry Numbers were made available for search with and without the letter suffix "P". The individual words from the modifying phrases were collectively searchable as part of the CASIA record. Originally, these modifying words did not have any proximity in order to reduce the selection of records based upon a potentially spurious proximity that might otherwise be introduced.

Approximately 3% of the chemical-substance index entries in the *CA Volume Chemical Substance Index* actually contained no CAS Registry Number. This fact is not evident from a casual scan of the printed index but was quite clear in the presentation of the computer-readable file. As the *Ninth Collective Index* closed in 1976, a decision was made by CAS to restore some of this information to the CASIA record. Consequently, beginning in 1977, it was possible to retain for the on-line CASIA file the CAS Registry Numbers for nonspecific derivatives. These CAS Registry Numbers were then added to the file with the letter suffix "D" to indicate a nonspecific derivative.

CHEMICAL SUBSTANCE DATABASES

Although storage for the CASIA file was greatly reduced by compaction of the record and removal of some chemical-substance information, it was necessary to provide an alternate means for searching that chemical-substance information. The first such commercially available file containing that chemical information was the CHEMNAME file, made available from DIALOG during April 1976, simultaneously with the CASIA file. During the development of the conversion and file-loading process for CASIA, a parallel process was developed for the CHEMNAME file. As an incoming CASIA tape was being processed and converted to provide general-subject and chemical-substance information, each of the CAS Registry Numbers in that tape was checked against a master file. If the CAS Registry Number had not been previously encountered, then the chemical-substance information for that CAS Registry Number, including the *CA Index* (or type I) name, molecular formula, and CAS Registry Number were written to another in-process file. The CHEMNAME file first appeared with more than 129 000 chemical substances corresponding to unique occurrences of the Registry Numbers in the corresponding CASIA file. Since CASIA covered the entire 80 sections of *Chemical Abstracts*, the substances contained in the CHEMNAME file were the full range from elements to alloys to polymers to organic chemicals. The first version of the CHEMNAME file contained thousands of each of these chemical-substance types. These files were introduced at the Centennial ACS Meeting in New York City in April 1976. The process of checking each incoming CAS Registry Number against a master file was continued throughout 1976.

The development of the CHEMNAME file from CASIA required that consideration be given to how an on-line distribution service would handle the wide variety of chemical substances in the CASIA file.

Using CASIA as the source of a chemical-substance file meant that there were really only three possible types of data that could be searched and displayed; systematic names, CAS Registry Numbers, and molecular formula information. For each of these, special algorithms were created to take greatest advantage of both the information content and the DIALOG text-searching software. There had been a large amount of research[4] at CAS and NLM on the searching of systematic *CA Index* names. This was during the *Ninth Collective Index* period, for which a new systematic nomenclature had been introduced by CAS for use in its volume chemical substance indexes. A major characteristic of the nomenclature was that it consistently used the same character strings to represent the same structural features. As a result, it was possible to use

text-searching methods to retrieve chemical substances with common structural features. Although eschewed by many chemists as being unnatural compared to the uncontrolled vocabulary of previous years, the ninth collective nomenclature was necessary and sufficient for text-search methods. In order to exploit this as much as possible, the rules for producing search terms from character strings were modified in the DIALOG system. With regard to systematic nomenclature, two characters of punctuation are significant in their original context: the comma and the period. Commas are used in systematic nomenclature to separate locants in a string. It became evident that these locants were most meaningful in their original grouping rather than being made available for search individually. Like all systematic nomenclatures, the CAS *Ninth Collective Index* names favored the lower numbered locants. Consequently, almost any name requiring locants had the numeral 1. Thus, 1 by itself became a search term with very little value; however, in combination with those other locants to which it was attached by commas, it was much more valuable.

By the same token, von Bayer nomenclature for polycyclic ring systems required that the length of each of the ring systems from a bridge-head atom be cited in brackets by arabic numerals separated by periods. Thus, retaining the period in its original context between these numbers was of great significance; otherwise, the numbers so produced would conflict with locants. Normally, in a bibliographic database, all punctuation characters would be used to delimit search terms. This rule was modified in the case of systematic nomenclature to permit commas and periods to remain in place. All other punctuation including blanks, hyphens, parentheses, and brackets was used to create individual search terms with appropriate proximity. Search terms so produced were quite useful but were quite frequently themselves composed of several terms representing different structural moities.

To provide for a finer detail of search in systematic nomenclature, it was decided to apply a systematic segmentation process to all of the *CA Index Names*. This process was applied in such a way so that a search term such as BENZENAMINE would consistently be segmented into the chemically significant segments BENZEN and AMINE. The retention of the proximity of these segments with the noncommuting DIALOG operator (W) permitted the segments to be searched in such a way as to reconstruct their original context.

All nomenclature search terms were labeled as to their systematic name field origin such as heading patent, substituent, name modification, or stereochemical descriptor by prefixes. To reduce confusion when these precise prefixes were not required for searching, all nomenclature terms were also made available for search in the basic index without the need for any field qualification.

Molecular formula information offered a rich source for searching if it was considered both in its entirety and for its significant constituent parts. Chemical Abstracts Service has, for many years, depended upon and used a dot-disconnectd form of a molecular formula. Each portion of a dot-disconnected molecular formula is, in general, a valid molecular formula itself. Thus, the task at hand for the CHEMNAME file was to enrich these rather long, preferred molecular formulas so as to provide more search terms.

Each of the molecular formulas was made directly searchable in its original form minus any leading nonalphabetic or nonnumeric characters. From a dot-disconnected molecular formula, each portion of the "dot disconnect" was also made available for search. This provided, for example, the ability to group salts of an acid by merely selecting the molecular formula for the parent acid. Given the potential ambiguity in that search term, methods were employed to permit the selection of a term as a complete original term or as a partial term. A further decomposition of molecular formulas into their individual element counts proved to be useful. An examination of the distribution of elements in molecular formulas within the CAS database guided this process. Even though approximately 96% of the chemical substances indexed by CAS contain carbon, it is easily illustrated that no more than 6–7% of the substances contained any specific number of carbon atoms. This led to development of the Element Count search terms, which could then be searched as individual terms or in ranges in addition to the other information in the file. Each element from a complete molecular formula was entered into the database as an Element Count search term. Summations of these element counts across dot disconnects were also made available for search on the presumption that they provided useful information that was otherwise not available. Some confusion results from the element counts derived from the molecular formula for polymers, but this far outweighs the lack of such information.

Molecular formulas contain information about the kind and number of elements within a chemical substance. By inference, the data can be interpreted in terms of the periodic classification of the elements. Grouping of elements in accordance with the periodic classification provided additional search terms with two distinct properties. First, the elements could be aggregated with the meaning of the groupings of the periodic classification, and second, the search terms could be so constructed as to indicate both inclusion and exclusion of these groups of elements simultaneously. To exploit the information properties of the periodic classification, two types of search terms were created specifically for the CHEMNAME file, based upon molecular formulas. The first of these was the Group Number term, which simply used the periodic group number rearranged to be represented by a letter, A or B, followed by an arabic numeral, 1–8. The source of this information was a simple Deming periodic classification[5] well-known to most chemists. In this way, a chemical substance containing sodium would also be assigned the search term GN=A1, and a chemical substance containing Niobium obtained the search term GN=B5. For each element recognized in a molecular formula, a group number term was assigned. These terms permitted the collection of substances with elements from the same groups, but they did not indicate the presence or absence of other groups within the same chemical substance. To extend this idea and apply fully the periodic concept, another set of terms, known as Periodic Index terms, was created. These terms were composed of precoordinate group numbers, such that both inclusion and exclusion of groups were simultaneously specified. For example, sodium chloride was assigned two group number terms plus the Periodic Index search term PI=A17 that simultaneously expresses the presence of one or more elements from groups A1 and A7 and no other "A-group" elements. A more complex molecular formula such as $C_{28}H_{34}Cl_4CuN_6O_2Zn$ is assigned the partial Periodic Index Terms PI=A567 and PI=B12 plus the complete form PI=A567B12.

Both the Group Numbers and Periodic Index terms in their original formulation have survived into the most recent versions of the CHEMNAME file providing added search terms that convey chemical information carried by a molecular formula as interpreted through the periodic classification of the elements.

The CHEMNAME file illustrated that it was possible to conduct an on-line search by using a combination of both nomenclature and molecular formula information simultaneously. The probabilistic occurrence of search terms was such that even crude Boolean searches permitted a reasonably rapid

reduction of the file to groups of compounds of interest. What was missing in all of this was the ability to search on more familiar, less systematic, chemical-substance names. For example, it was not possible to search the CHEMNAME file in this formulation for commonly known synonyms. To provide such a capability, DIALOG obtained license for the first computer-readable version of the *CA Index Guide*.

The *CA Index Guide* is that publication of CAS that contains information about indexing terminology, both general subject and chemical substance. It is in the *Index Guide* that one gets clarification of the use of a general-subject heading or alternate general-subject headings. Likewise, it is a source of commonly used chemical-substance synonyms. Using the computer-readable version of the *CA Index Guide*, which contained the CAS Registry Number, the systematic CA index name for the ninth collective period, and the synonyms, it was possible to enrich the CHEMNAME file with synonyms. To do this in a way less systematic than the *CA Index Guide* provided, an algorithm was developed to "uninvert" the synonym names. Curiously, the *Index Guide* would contain synonym names such as "BENZENE,AMINO-". This seemed to be a rather uncommon expression for the synonym AMINOBENZENE, and thus, an algorithm to uninvert the name to provide a search term AMINOBENZENE was created. Although thousands of synonyms were added to the CHEMNAME file in this way, many synonyms never found their way into the file because they were neither in the *CA Index Guide* nor were they used as "systematic" nomenclature in the CASIA file.

The CHEMNAME file went through several updates with CASIA and the *CA Index Guide* as sources with the enrichment of search terms until the end of 1976 when CAS indicated that it would provide the Registry Nomenclature File (RNF) of more complete chemical substance information for all CAS Registry Numbers appearing in the CASIA file. Some advantages of the RNF were that it contained synonym names, nomenclature segmentation points, and ring system information where appropriate for a given chemical substance. DIALOG Information Services began working on conversion of the RNF as the source of a reloaded CHEMNAME file. By mid-1977 it was clear that some division of this very large file would be desirable. At that time, four possible divisions were proposed, but only the division into more and less- frequently cited chemical substances was acceptable to both the DIALOG Chemical Information staff and the customer community. This turned out to be an extremely popular division of the file, leading to a reloaded CHEMNAME file of just over one-million chemical substances that had been cited two or more times in the CASIA file from 1967 to the present time. Subsequent updates have increased the CHEMNAME file to well over 1.5 million chemical-substance records.

This division of the chemical substances and redefinition of the CHEMNAME file provided for relatively low-cost chemical-substance searching and the opportunity to create several other chemical-substance files containing records for CAS Registry Numbers cited one time and grouped by the collective index period. Each of these files benefitted from the algorithms that had been developed in the original CHEMNAME file derived from CASIA and the *CA Index Guide*.

TRAINING AND PRACTICE FILES

As the offering of chemical-substance and chemical-literature information available from DIALOG Information Services grew, the customer community also grew and sought additional opportunities for training and practice. For some time, DIALOG and other on-line distributors had requested the right to use all or significant portions of the CAS databases for such educational purposes. By 1978, Chemical Abstracts Service had defined an educational package consisting of two issues, approximately 15 000, of literature items from the CASIA file and a corresponding chemical substance file. In November 1978, these databases were made available on-line as the CA CONDENSATES/CASIA and CHEMNAME ONTAP files, respectively.

COMBINATION OF DATABASES

In early 1977, CA CONDENSATES for the *Tenth Collective Index* period was made available from DIALOG as File 4. Rather than open another, separate CASIA file, the decision was made to relax this requirement and merge CA CONDENSATES and CASIA into a single search file. The CHEMNAME file had become a successful adjunct, permitting the reduction of storage by retaining only CAS Registry Numbers as chemical-substance identifiers in the CASIA record.

Experience with a separate CASIA file in the previous year had indicated that there was only one problem to be overcome in providing a unified file of search terms from both CA CONDENSATES and CASIA.

The CA CONDENSATES file contains a data element that lists the document type, one of which is Patent. The CASIA database for the same original item contains a different data element listing one of three possible document types, none of which is Patent. Thus, the unified file required that CA CONDENSATES be converted first and that a master file of patent accession numbers be retained for use in the conversion of the CASIA file. For display purposes, the data element of CA CONDENSATES was chosen since it discriminated among six different types. The significance of the patent data element indication is that in the merged CA CONDENSATES/CASIA file the traditional limit to patent or nonpatent document type was retained for all search terms.

The success of the merger of CA CONDENSATES and CASIA into a single on-line search file was well received by both the customer community and by the database producer. By 1978, Chemical Abstracts Service had completed their plans to adjust internal processing so as to make available a unified database under the name of CA SEARCH. The unified database required that the bibliographic, keyword, and indexing information be made available simultaneously for a document, rather than permitting the time lag until the appearance of the index entries. As a consequence of this processing, it was possible that certain index entries for a document would not be ready at the time the file was to be packaged. Consequently, CAS adopted the policy of packaging all of the data that was available for an item at that time and instituted a semiannual addendum feature for the CA SEARCH file. This is still the practice and policy of CAS, and approximately 6000–8000 items within a given volume of information appear on the addendum tape for one of several possible reasons. Late-arriving index entries can be added to those already produced for a given document, and the entire group of data will be reissued on the addendum file. Alternatively, data may be corrected or deleted if found to be in error. It has always been the practice of DIALOG Information Services to add these addendum data as soon as they are available. These data completely replace all of the previous data available for a given item in the CA SEARCH database. In keeping with DIALOG philosophy of making information available to customers and also giving them the opportunity to disregard it at their own request, these data are frequently found in the Selective Dissemination of Information (SDI) profiles at peculiar times. Since the data arrive approximately 3 months after the close of a CA volume, the SDI profiles run at that time can contain information from 6 to 9 months "old"

although it is truly new since it has only recently arrived on the addendum file.

Since 1976, the various developments in chemical information at DIALOG have led to a series of robust files available on-line with frequent updating and unique value-added features. Perhaps one of the more significant of these features developed in the last several years has been that of ring system information. CAS elected to remove this information from the most recent versions of the Registry Nomenclature File. Experience with the DIALOG Information Services' CHEMNAME and other chemical-substance files clearly indicated that this information provided a cost-effective, inexpensive way to group chemical substances on the basis of structural features. As a result, complete new algorithms were developed by chemical information staff at DIALOG to generate the ring information in accordance with the rules and definitions published in the CAS *Parent Compound Handbook*. These data have been the subject of several presentations and will be more fully described in a subsequent publication.

SUMMARY

Chemical information at DIALOG Information Services has developed through licensing of databases from Chemical Abstracts Service and providing the first on-line implementations of several of these. *Chemical Industry Notes* and *Patent Concordance* were first made available in 1976. Also, the CA CONDENSATES file for several collective index periods and the CASIA indexing file were first made available on-line in 1976 by DIALOG Information Services. In that same year, the CHEMNAME chemical-substance file was created from CASIA and the *CA Index Guide*. As a result of that work, the *Index Guide* was also used to enrich the combined CA CONDENSATES/CASIA file first offered on-line in 1977 by DIALOG Information Services.

REFERENCES AND NOTES

(1) Summit, R. K. "Lockheed Experience in Processing Large Databases for its Commercial Information Retrieval Service". *J. Chem. Inf. Comput. Sci.* **1975**, *15*, 40-42.
(2) Schultheisz, R. J.; Walker, D. F.; Kannan, K. L. "Design and Implementation of an On-Line Chemical Dictionary (CHEMLINE)". *J. Chem. Inf. Comput. Sci.* **1978**, *29*, 173-179.
(3) Callahan, M. V.; Rusch, P. F. "Online Implementation of the CA SEARCH File and the CAS Registry Nomenclature File". *Online Rev.* **1981**, *5*, 377-393.
(4) Fisanick, W.; Mitchell, L. D.; Scott, J. A.; Vander Stouw, G. G. "Substructure Searching of Computer-Readable Chemical Abstracts Service Ninth Collective Index Chemical Nomenclature Files". *J. Chem. Inf. Comput. Sci.* **1975**, *15*, 73-84.
(5) Fernelius, W. C.; Powell, W. H. "Confusion in the Periodic Table of the Elements". *J. Chem. Educ.* **1982**, *59*, 504-508.

Evolution of Industrial Chemical Information Systems

CARLOS M. BOWMAN* and PAULA B. MOSES
The Dow Chemical Company, Midland, Michigan 48674

Received January 21, 1985

A modern industrial chemical information system is an integrated system that provides for the storage, retrieval, and manipulation of internal and external information to meet the needs of the organization. It incorporates some of the most modern tools and methods available in an effort to be cost effective. The various components of an industrial chemical information system are enumerated, and their evolution is described and anticipated.

INTRODUCTION

Unlike many other areas of chemistry, most of the initial work in chemical information was carried out in an industrial environment rather than at universities and research institutes. Chemical companies pioneered the development of indexing and retrieval techniques. Many of the early advancements in manipulation and handling of chemical structure information were also made in the information groups or laboratories of industrial firms. The introduction of computers into chemical information was also an area where the industrial information chemist led the way.

A detailed description of the many advances made in the chemical industry is beyond the scope of this paper. Instead, we will describe a typical information system as it is found in the industrial environment. In so doing, we will point out those areas that are unique to industry and will then recount the developments that have led to the present structure and function of an industrial chemical information system.

MISSION OF AN INDUSTRIAL CHEMICAL INFORMATION SYSTEM

The mission of an industrial chemical information system is not very different from any other information system. It is charged with providing a means of preserving information that may be of value in the future, maintaining a retrieval process that assures timely and accurate recovery of the stored information, and providing access to tools for manipulating that information. However, there are some very important distinctions between an industrial information system and a public or government system.

One of the main differences is that an industrial information system must include in its collection the technical information generated internally in the firm. Much of this information is proprietary in nature and must be protected from untimely or unauthorized disclosure. Another difference is that time is, in some instances, a much more important factor in the storage, indexing, and retrieval of information in the industrial environment. Failure to obtain the needed information in a timely manner could result in a substantial economic penalty to the organization.

An overriding constraint on the operation of an industrial information system is that it be cost effective. The determination of what is cost effective is a difficult one, but one that the manager of such a system must be ready to defend and expound frequently and regularly. This pressure to maintain cost effectiveness has made industrial information managers seek out services from many sources as they make every effort to avoid unnecessary duplication of effort in either operating or development efforts.

EXTERNAL INFORMATION

The greatest portion of the chemical information used by an industrial chemical firm comes from outside the organization. It typically comes from journals, books, government

0095-2338/85/1625-0197$01.50/0