proprietary data from a given date forward, McNeil's management realized that initially this system would be a drain on resources. With the addition of the Theilheimer and Current Literature File databases, the system has already shown positive results through decreased library search times. In addition, the in-house literature file is proving helpful by showing all users those areas of interest to other chemists at McNeil.

Two other areas, not originally envisioned, in which chemists are applying REACCS are retrieval of articles pertinent to a given area of chemistry and "trouble shooting". Previously, once developmental chemists at McNeil found a leading article, they would request author searches over the CAS databases and cited-article searches over the ISI databases to find as much pertinent information on a reaction as possible. The CLF database on REACCS is becoming one of the primary sources of preliminary reaction citations. Chemists use REACCS to trouble shoot by searching the databases for reactions in which substrates with the same functional groups as those being studied react with the same catalysts or reagents. Occasionally the chemistry obtained in a search of this type can be used to help explain the observed chemistry.

At McNeil, REACCS is rapidly becoming an indispensable tool for our chemists. Although data entry into our proprietary databases began only a few years ago, chemists are already saving time. Often searches reveal chemical reactions that had been optimized earlier and can now shed light on current projects; with REACCS, chemists find this information even though these earlier projects may produce molecules with little obvious similarity to those molecules being prepared in the current projects. As these databases continue to grow, their utility is expected to expand also.

ACKNOWLEDGMENT

We acknowledge all chemists at McNeil Pharmaceutical who helped in the development of our current databases. We are especially grateful to Julie Spink, who did much of the original abstracting for our proprietary database, to William Bullock, who helped in the database creation and data entry, and to Barbara Baughman, for her continuing efforts to keep this database current. Our thanks also to Cynthia Dunn, Leonard Herring, and other members of the Scientific Computer Services Department for their help in bringing this work to fruition.

REFERENCES AND NOTES

(1) The filing system was to be implemented on a DEC-10 computer system.
(2) System 1022 is available from Software House, Cambridge, MA.
(3) MACCS-II and REACCS are available from Molecular Design Ltd., San Leandro, CA.
(4) SYNLIB has been marketed by SmithKline/Beckman, Philadelphia, PA.
(5) Barcza, S.; Kelley, L. A.; Lenz, C. D. "Biosynthesis-Metabolic Pathways Database". Presented at the 192nd National Meeting of the American Chemical Society, Anaheim, CA, Sept 1986.
(6) For examples of software written to interface MACCS with other databases see: (a) Legatt, T.; Saltzman, A. "Development of a Chemical Structure Display System and Its Interface with the Associated Biological Database". *Drug Inf. J.* **1986**, *20*, 51. (b) Adamson, G. W.; Bird, J. M.; Palmer, G.; Warr, W. A. "Use of MACCS within ICI". *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 90.
(7) Mills, J. E.; Maryanoff, C. A.; Sorgi, K.; Stanzione, R.; Scott, L.; Herring, L.; Spink, J.; Baughman, B.; Bullock, W. *J. Chem. Inf. Comput. Sci.* (following paper in this issue).

REACCS in the Chemical Development Environment. 2. Structure and Construction of Proprietary Databases†

JOHN E. MILLS,* CYNTHIA A. MARYANOFF, KIRK L. SORGI, ROBIN STANZIONE, LORRAINE SCOTT, LEONARD HERRING, JULIE SPINK,‡ BARBARA BAUGHMAN, and WILLIAM BULLOCK§

McNeil Pharmaceutical, Spring House, Pennsylvania 19477

Received September 11, 1987

The utility of a chemical database management system is dependent both upon the ease of use of the software and upon the information stored in the database. The process used at McNeil Pharmaceutical to select datatypes for proprietary REACCS databases is discussed. Problems associated with the storage of negative as well as positive results, storage of several structures for a given molecule, and storage of incompletely characterized molecules using REACCS are discussed. Useful techniques to circumvent some potential problems are provided.

INTRODUCTION

The storage and retrieval of chemical reaction information are areas in which dramatic advances have been achieved over the past decade. Although works such as Houben-Weyl's *Methoden der Organische Chemie*, Theilheimer's *Synthetic Methods*, and Beilstein's *Organische Chemie* have been used with limited success for the storage of reaction data, all of these methods suffer from their lack of generality; each is designed for recovery of reaction data through a single specific search strategy. Through its subject index, *Chemical Abstracts* provides somewhat limited search capabilities for reactions and for information on the preparation of a molecule whose structure is known. Advances in computer technology have made the formulation of search queries via graphics input commonplace.[1-5] These advances have simplified the process to such an extent that it is not uncommon for "bench chemists" to learn how to formulate queries using software such as REACCS after a training period of 1-4 h.

IDENTIFICATION OF DATABASES

The need for an efficient means to store and retrieve chemical reaction data has been presented elsewhere.[4,6] This

† Presented in part at the MDL Software User's Group Meeting, San Francisco, CA, April 2, 1987, and at Reaction Indexing Seminar, Saddle Brook, NJ, June 1986.

‡ Present address: Ciba Corning Diagnostics, Medfield, MA 02052.
§ Present address: Department of Chemistry, Emory University, Atlanta, GA 30322.

0095-2338/88/1628-0155$01.50/0

paper will present the considerations made in developing proprietary databases for use at McNeil Pharmaceutical. The need for chemical reaction database management systems can be easily established. However, the ease of use and flexibility of the database management software are only two of the factors to consider when one is constructing a proprietary database system. Two other factors that contribute significantly to the usefulness of these systems are (1) the exact types of information maintained in proprietary databases and (2) the detail that a chemist may readily obtain through a search of the databases. Database definition is one of the major obstacles the designers of any proprietary database must surmount before creating the database. Assuming they desire an active user group, when defining a database, the designers must, of necessity, progress through several steps before educing the final database form and content.

In discussions, representatives from both the Chemical Development and Chemical Research departments established several different anticipated uses. They also determined that no single database could conveniently meet everyone's desires. After carefully defining needs, the representatives identified several tentative databases to meet the requirements for each proposed use. Then they more clearly defined the purpose of each database, and they prepared lists of all reaction variables needed to fulfill those purposes. In additional discussions the representatives ranked each datatype as essential, helpful, or desirable but nonessential for the stated purpose of each database. In this way, they gradually culled from the large number of datatypes originally proposed for each database those deemed most useful to meet the databases' definition.

Table I. Summary of Datatypes

name	datatype	name	datatype
	Reaction		
date	date	solvent	
notebook	integer	name	text
report	text	amount	real
chemist	text	product	
department	text	amount	real
project	integer	yield	real
literature reference	text	isolated	text
reactant		conditions	
source	text	size	real
lot no.	text	atmosphere	text
amount	real	pressure	real
catalyst		time	real
name	text	temp	real
source	text	comments	text
lot no.	text	hazard	text
amount	real	summary	text
	Molecule		
CPD no.	integer	isomer	
stock no.	integer	specific rotation	real
name	text	solvent	text
hazard/comments	text	temperature	real
molecular weight	real	concentration	real
recrystallization		wavelength	real
solvent	text	stereochem	text
volume	real		
notebook	integer		
melt pt	real		
boil pt			
temp	real		
pressure	real		

PROPRIETARY DATABASE CONSTRUCTION

The need to store and retrieve data generated at McNeil was the original justification for purchasing REACCS. It is therefore not surprising that the first database created was designed specifically for this purpose. Chemists desired a database that would provide negative as well as positive results. They also wanted enough information readily available to allow them to decide whether reactions that had been performed should be reinvestigated due to advances in the understanding of the reactions' critical parameters. Chemists required scale and reaction times to determine volume efficiencies and manpower needs. They desired some indication of the amount of work done on a given reaction. Reference to the laboratory notebook was necessary so chemists could obtain more extensive evaluation of a reaction. They required a method to retrieve data on specific projects. The representatives concluded that, although desirable, inclusion of complete experimental details was not cost effective. Reference to original literature was highly desirable. Table I gives a list of datatypes deemed most useful for McNeil's Chemical Development proprietary database. With this list of datatypes, the group drew relationships linking related pieces of information. Since "department" pertains only to "chemist", the group linked these two pieces of data together. Other related datatypes are apparent from the table; for example, "amount", "yield", and "isolated" are linked through "product".

The representatives added certain datatypes to those given in Table I to facilitate database administration, to simplify searches for specific reaction conditions, and to simplify data display. They included the "audit" datatype for use by the system administrator. Once a reaction or molecule and its associated data have been entered into the database, the auditor reviews the data for accuracy and then approves them by entering his/her initials into the database. If additional data variations on a reaction are added, the audit datatype is deleted until the new data have been checked. Thus, this datatype provides a simple, effective method to ensure that more than one individual has seen all data contained in the database.

Occasionally, chemists isolate isomers of unknown relative or absolute stereochemistry. To allow chemists to store information on these compounds as efficiently as possible, the group added a datatype to indicate that the stereochemistry of the compound is unknown. Thus, if this datatype is filled without a chiral label on a molecule, it indicates that the relative stereochemistry of one or more stereogenic centers is unknown. If the datatype is filled for a molecule containing a chiral label, it indicates that two epimers have been resolved (or at least enriched) but that the absolute stereochemistry is unknown. This datatype allows for simplified updates to molecules that have been entered into the database, while enabling entry of physical data on molecules whose structures have not been established unambiguously.

Other datatypes are used to establish whether a reaction has been performed under optimal conditions and, in the case of catalysts, to serve as an indication of potential problems caused by variations in catalyst lots. Some of these datatypes, such as "reflux" and "technical memo", require only a positive or negative response.

Since REACCS is a hierarchical database management system, the representatives added extra datatypes in order to facilitate storage of related pieces of data. For example, they added "variation" to link all data relevant to one reaction. REACCS defines a reaction only in terms of the reactants and products; when all the reactants and products are identical but different catalysts, solvents, and/or conditions are used, the data are stored in the database as variations of the same reaction.

Figure 1 shows relationships between the final datatypes. Although a chemist may find a piece of data by entering the complete tree name into REACCS, when one uses care in construction of the individual datatype names, only the last name in the entire tree is necessary to find that piece of data. For example, if "comments" is only used in the complete datatype, "rxn:variation:comments", then the same piece of in-

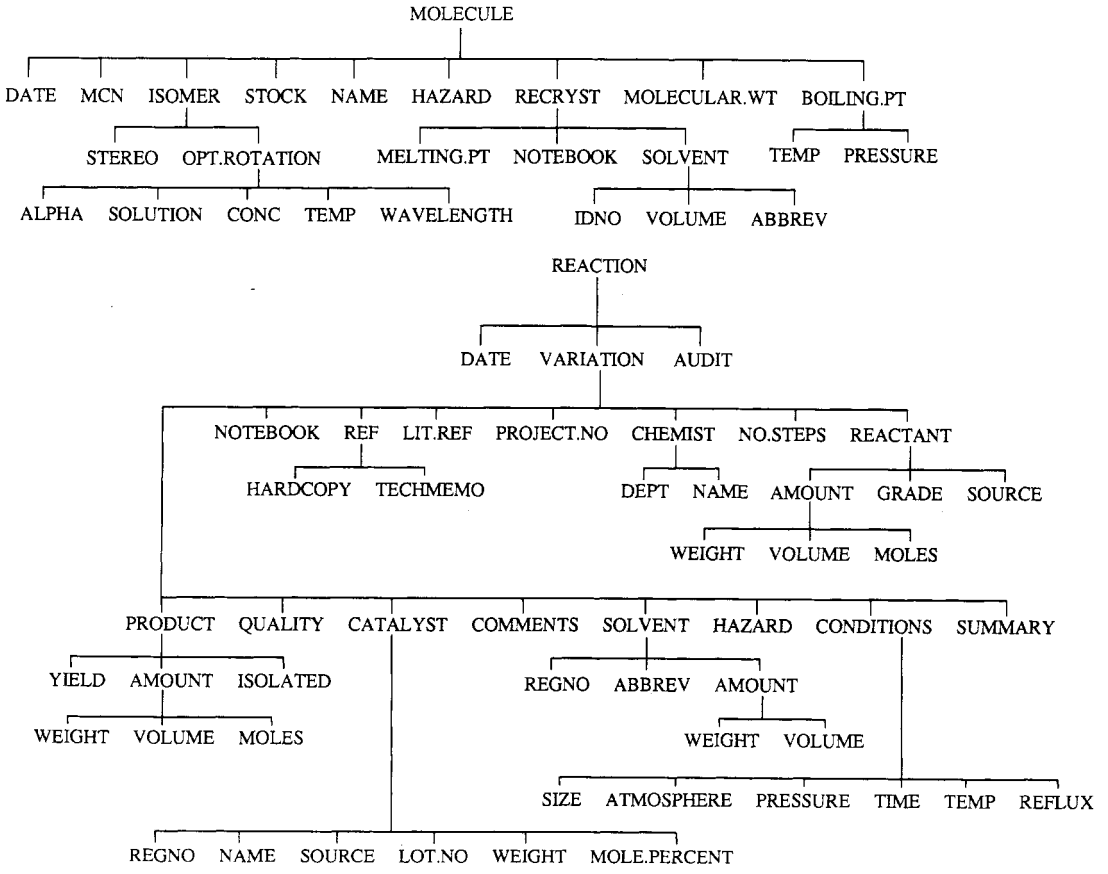

Figure 1. Complete hierarchical structure for datatypes in McNeil Pharmaceutical's proprietary database.

Table II. Multiple Datatypes

	Reaction	
(1) variation	(2) chemist	(3) literature ref
(4) solvent	(5) catalyst	(6) conditions
(7) hazards	(8) comments	
	Molecule	
(1) name	(2) hazard	(3) recrystallization
(4) solvent (recryst)	(5) boiling pt	(6) opt rotation

formation can be obtained simply by entering "comments" as the datatype during execution of REACCS.

Once the representatives placed all datatypes in a hierarchy, they evaluated each to determine whether to store single or multiple pieces of data in that datatype. Certain datatypes may contain multiple pieces of data and yet be identified as single. An example of such a datatype is "yield", when the yield is expressed as a range of values. Through the use of a multiple "reference" datatype and a single yield datatype allowing for a range of values, users can enter several reactions run under the same reaction conditions but giving different yields as a single variation rather than as multiple variations of the same reaction. Table II provides all multiple datatypes.

Once the group established the hierarchy shown in Figure 1, the information was translated into a database specification file. During translation, all datatypes used to store binary information (i.e., datatypes that required a yes or no response) received special treatment. The group designed prompts to request a single letter for use in data display. Entry of the letter represents a positive response. For example, "reflux" stores data indicating whether a given reaction is run in a refluxing solvent. Entry of the letter D constitutes a positive response. Data output is formated to print "REFLUXE" before the actual data. The net result is that "REFLUXED" is printed whenever a user enters data into this datatype; nothing is printed in the absence of data. Using a "Y" or "N" response with the data to be printed after REFLUXED would have resulted in the display of unnecessary data when a reaction was not heated at reflux.

After the scientists created the new database using the specification file, they wrote boxfiles, which control the content and format of the data displayed by REACCS. These boxfiles allow individual users to display data in a personalized fashion. The data-display flexibility is most obvious when the program displays the data associated with one reaction using different boxfiles on a graphics terminal. This capability enables the program to display only those datatypes that are most useful to a given user, resulting in more rapid screening of hit lists. If desired, chemists may view databases without the use of these files. In the absence of these files, only the reactants and products are displayed.

SPECIAL DATA STORAGE PROBLEMS

The storage of both negative and positive results raised some problems that do not arise in many other applications. In experiments in which no reaction is observed or in those where some of the starting material is recovered, representing the same molecule as both reactant and product could lead to a large number of "false hits" when a substructure search for a product is performed. The representatives overcame this problem by defining special atom types: Rsm (for recovered starting material) and Nr (for no reaction). Registering these special atoms as molecules allows for their use in reactions. Where more than one reactant may be recovered, McNeil's REACCS system uses text datatypes to define which reactants are recovered. In addition to these special atoms, Mpr (for multiple products) was also added to the periodic table. Use of Mpr as a product indicates that a reaction was run under the given conditions and that several unidentified products were formed with loss of starting materials.

An unanticipated problem was the focus of several discussions on the structure of proprietary databases. REACCS does not currently have the capability of storing and displaying

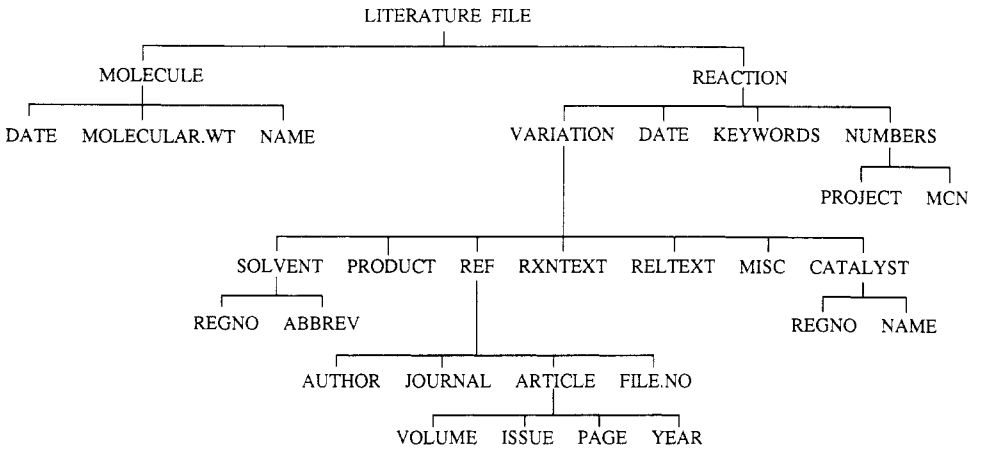

Figure 2. Complete hierarchical structure for datatypes in McNeil Pharmaceutical's in-house literature database.

intermediates in a single reaction. For example, treatment of an amide with phosphorous oxychloride, followed by an amine to yield an amidine, is known to occur through formation of an intermediate Vilsmeier complex. Consequently, to present the complete reaction mechanistically, a series of three reactions is typically employed in commercial REACCS databases: The first is the reaction of the amide with the phosphorous oxychloride to yield the Vilsmeier complex; the second is that of the Vilsmeier complex with the amine to yield the amidine; and the third is that of the overall reaction from amide to amidine. The group decided to present only the overall transformation in our proprietary databases and to label the reaction as a multiple-step, one-pot reaction. The benefits of this decision are a more efficient use of system resources and fewer hits in substructure searches for products. A disadvantage is that a user cannot search the chemistry of specific reactive species without isolating and characterizing the intermediate. Once such species have been isolated and characterized, their preparation and subsequent reactions can be added to the database.

Registration of a single compound as multiple isomers became an apparent problem during database construction. This problem is especially annoying for compounds containing an amidine or guanidine group. REACCS allows registration of two geometric isomers for these compounds. In addition, amidines may be registered in two tautomeric forms, and many guanidines may be registered as three tautomers. Thus, four structures can be registered for an unsymmetrical amidine and up to six structures may be registered for a trisubstituted guanidine. The program may allow registration of additional structures if more than one of these functional groups are present in a single compound or if the alkyl substituents contain stereogenic centers. Since McNeil's proprietary databases are continuously being updated, it is highly possible that different chemists may draw the same molecule as different conformations, since inversion at nitrogen is typically facile. To minimize this problem, a chemist registers all isomers of these classes of compounds when the preferred conformation is initially registered. The first name assigned to other structures is "USE[Regno]" (where [Regno] is the internal registration number of the preferred conformation). During reaction registration, the molecule name is used as the molecular identifier. Thus, if a user enters a structure other than the preferred structure in a reaction, the program automatically displays the message to replace the molecule.

DATA ENTRY

Data entry at McNeil Pharmaceutical is a four-step operation. The first step is the graphics entry of molecules into the REACCS database. The second step is entry of reactions, utilizing the previously entered molecular structures. The third step is entry of reaction data into the database. To ensure that information in the database is accurate and consistent, one chemist with the authority to set the audit datatype checks reacting centers and data. This check constitutes the fourth and final step in data entry.

The rate of entry for a single reaction depends highly upon the database and the number of new molecules to be entered for each reaction. A general rule used at McNeil is that entry of all structures and data for a reaction that has only one variation should take about 35 min. Both chemical structures and the textual data associated with molecules are being entered by a single individual. Auditing of the complete reaction is performed by a second individual working part-time on the system. This has resulted in more consistent formating than may have been achieved with multiple data entry personnel.

Over the past two years, the proprietary database has experienced an annual growth rate, in terms of the total number of blocks used for storage, of about 30%. Although the number of new reactions has slowed slightly, the database continues to grow at the same rate due to the addition of numerous variations to existing reactions. About 7500 total blocks of disk storage are currently utilized for the storage of two proprietary databases. The databases are backed up incrementally daily, and a complete backup is made weekly. Both backups are made to tape.

Once a database contains about 150 reactions, the ratio of the number of molecules to the number of reactions contained in a database is typically 2. This ratio does not appear to change appreciably as the database continues to grow. Consequently, one should not expect to save data entry time by addition of fewer molecular structures as a proprietary database grows.

As the REACCS technology has developed, McNeil Pharmaceutical has received annual updates to this program. In most cases, installation of an update has required only minor changes to files associated with access to the databases and display of the data in addition to installation of the revised REACCS program. Major updates to the REACCS program have required modifications to the database structures in order to utilize some of the provided search techniques. Major REACCS updates have included software tools that were necessary to automatically generate the database modifications. In the most recent REACCS revision, less than 1 h of CPU time was required to add atom to atom search capabilities to McNeil's proprietary databases.

OTHER DATABASES

In addition to constructing a proprietary database containing reactions performed at McNeil, the company established a literature database. The purpose of this file is twofold: First, it allows chemists to organize their files in ways not previously

possible; second, it helps make McNeil chemists more aware of areas of interest to their colleagues. The structure of this database is similar to that of the proprietary database; however, the group has deleted many datatypes to simplify data input. Figure 2 gives a complete hierarchy of datatypes used for McNeil's in-house literature database. Use of the literature database for storage of reactions forces each chemist to evaluate their reading and formulate concrete opinions on why specific articles are of interest to them.

Another database the group is constructing contains information on the overall conversion of commercially available starting materials to compounds in development. Unlike the first database, this database contains no experimental details on individual reactions. Its purpose is to make information on overall yields, equipment, and manpower requirements readily available. In addition, commercial suppliers and costs of starting materials will be stored in this database. It is anticipated that this database will simplify scheduling in the chemical pilot plant and make procurement of starting materials more efficient.

Metabolism can be viewed as a chemical reaction occurring under a given set of experimental conditions where important parameters are not temperature, pressure, etc. but rather animal species. Furthermore, metabolites can be classified by their distribution. By constructing a metabolite database, chemists at McNeil may be able to predict the metabolism of new drug entities and the distribution of the metabolites in test animals and man.

SUMMARY

The combination of software and data is what determines the usefulness of any database management system for a specific purpose. Ideally, the database management software should be designed to be "user friendly" and flexible. Unfortunately, these two needs are frequently only partially compatible. In order to make a system user friendly, designers of the system must often sacrifice software flexibility. Alternatively, software written for general use frequently requires specific codes for generating file structures and for searching those files. In order to utilize the database management system, one must first learn how to generate the file structure.

REACCS is no exception to the generalization given above. The flexibility of this database management program requires that some individual or individuals must learn the syntax necessary to generate the desired proprietary database structure. However, the effort to learn this syntax is not nearly as great as that required to define the particular database needs and identify the datatypes that will most effectively fulfill those needs. Without a well-planned database, users will be frustrated by their inability to find information that they are seeking. The need for a considerable expenditure of effort to define the purpose of any database prior to its creation cannot be overemphasized. Designers of the system must spend additional time to clearly identify those datatypes necessary to achieve that purpose. Finally, it is necessary to arrange those datatypes in a hierarchy and to add additional datatypes to organize the database.

At McNeil Pharmaceutical, the combination of well-designed databases with the flexible search strategies allowed by REACCS makes this system highly desirable for use by our chemists.

REFERENCES

(1) Heller, S. R. "Reaction Indexing". *Ind. Chem.* **1987**(Feb), 68.
(2) von Kiedrowski, G.; Eifert, A. "Handling Chemical Structures with a Personal Computer". *Intelligent Instrum. Comput. Appl. Lab.* **1986** (March/April), 110.
(3) Seiter, C. H. "Your PC May Solve Your Chem Lab Problems". *Res. Dev.* **1987**(March), 94.
(4) French, S. E. "Our Reaction Access System". *CHEMTECH* **1987** (Feb), 106.
(5) Hrib, N. J. "Recent Developments in Computer-Assisted Organic Synthesis". *Annu. Rep. Med. Chem.* **1986**, *21*, 303-311.
(6) Mills, J. E.; Maryanoff, C. A.; Sorgi, K.; Scott, L.; Stanzione, R. *J. Chem. Inf. Comput. Sci.* (preceding paper in this issue).

Using Analytical Data To Build Expert Systems

W. A. SCHLIEPER* and T. L. ISENHOUR
Department of Chemistry and Biochemistry, Utah State University, Logan, Utah 84321-0300

J. C. MARSHALL
Department of Chemistry, Saint Olaf College, Northfield, Minnesota 55057

Received September 11, 1987

Generating expert systems can be a long, arduous process. Human experts may be able to use some of their hard-won knowledge to help develop expert systems, providing the domain of the system is modest. As the domain expands, the problem of rule formalization may even baffle a human expert. Inductive methods, implemented by computers, can help generate production rules suitable for expert systems. The ID3 algorithm has been applied to two sets of chemical data, resulting in decision trees useful to classify the data. The resulting decision trees were then transformed directly to a set of production rules for an expert system. This procedure allows the attributes of the data set to be represented directly as objects that are descriptive of the actual data and does not require data transformations. This algorithm is also capable of distinguishing when attributes and associated values do not sufficiently span a given domain.

INTRODUCTION

The classification of complex data objects is frequently of central importance in analyzing chemical data and efficiently retrieving information from data files. Numerical pattern recognition techniques have been widely used for this task.[1-4] These numerical techniques frequently seek to cluster the data by transforming the inherent data attributes to new and possibly composite attributes. This has the conceptual disadvantage that the new attributes may be very difficult to relate to the original data from which they were derived.

Object oriented programming strategies do not require data transformations as all the attributes of the data, both quantitative and qualitative, may be represented as objects that are manipulated directly. This strategy casts the problem into a

0095-2338/88/1628-0159$01.50/0