

are not useful for NCI because CAS Registry Numbers cannot be assigned to the confidential structures which constitute about half of the NCI database.

- (13) The Screening Laboratories that are currently under contract to NCI are The Southern Research Institute (Birmingham, AL), Battelle Columbus Labs (Columbus, OH), Illinois Institute of Technology Research Institute (Chicago, IL), Mason Research Institute (Worcester, MA), Institut Jules Bordet (Brussels, Belgium), Arizona State University (Tucson, AZ), and the University of California (Los Angeles, CA).
- (14) Heller, S. R.; Milne, G. W. A.; Feldmann, R. J. "A Computer-Based

Chemical Information System". *Science (Washington, D.C.)* 1977, 195, 253-259.

- (15) Milne, G. W. A.; Heller, S. R.; Fein, A. E.; Frees, E.; Marquart, R.; McGill, J. A.; Miller, J. A.; Spiers, D. S. "The NIH-EPA Structure and Nomenclature Search System". *J. Chem. Inf. Comput. Sci.* 1978, 18, 181-186.
- (16) Confidential compounds in the DIS are described as "discreet", and their NSC Numbers are prefixed with a "D". The misuse of this word has long been recognized; it is a practice, however, that is so ingrained that no amount of reference to dictionaries can be expected to change it.

The NCI Drug Information System. 2. DIS Pre-Registry

G. W. A. MILNE* and ALFRED FELDMAN

Information Technology Branch, Division of Cancer Treatment, National Cancer Institute, Bethesda, Maryland 20205

J. A. MILLER, G. P. DALY, and M. J. HAMMEL

Fein-Marquart Associates, Baltimore, Maryland 21212

Received April 21, 1986

The Pre-Registry Module of the Drug Information System (DIS) is a staging area through which all new compounds are passed prior to acquisition and testing. Several methods are available for the entry of structures into the Pre-Registry; all involve built-in data validation. Newly entered structures are examined by computer programs for structural novelty and potential for anticancer activity. For those compounds that proceed to acquisition, the various acquisition steps, such as letter writing and record updating, are performed automatically. When a sample is obtained, the entire Pre-Registry record is updated and moved forward into the permanent DIS chemistry files.

INTRODUCTION

The goal of the Developmental Therapeutics Program at the National Cancer Institute (NCI) is to identify compounds that possess utility in the treatment of human cancer. In order to identify promising compounds, the program maintains liaisons with several thousand research organizations around the world and also supports two literature surveillance efforts, one focused on natural products and the other on synthetic chemicals. All the data associated with these activities are stored in the NCI Drug Information System (DIS), which is described in this series of papers.

When a promising compound comes to the attention of the NCI drug screening program, its structure and other relevant information are entered into the Pre-Registry subsystem of the DIS. Upon further examination, only about 40% of these structures are determined to be of sufficient interest to merit testing, and with only a fraction of these, i.e., some 30% of the original entries, are acquisition efforts successful. The Pre-Registry is therefore a staging file, and compounds in the Pre-Registry are only assigned temporary identification (TID) numbers. No more than 30% of all Pre-Registry entries are subsequently passed through to the permanent files of the DIS where each is assigned a permanent identification number.

As a first step then, the Pre-Registry must ascertain that a compound is new to the program. Compounds that have been tested are not reacquired.¹ If the compound is new, it must further be determined whether there is any reasonable expectation that it may have any activity against cancer. If the compound is not eliminated by either of these tests, a decision is made as to whether or not it should be acquired. If this decision is affirmative, the Pre-Registry sets the acquisition process in motion.

When the acquisition is complete and a sample of the compound has been received, the programs must log in the

sample and acknowledge its receipt. At this point, the Pre-Registry record is moved to the permanent DIS files and the compound is assigned an NSC Number.² The actual sample, meanwhile, is forwarded to a storage facility³ which is responsible for inventory and shipping data, as is described in part 4 of this series. A housekeeping task for the Pre-Registry involves the disposition of all the records for old compounds, whether they be selected and acquired, selected and not acquired, or not selected.

SOURCE OF CHEMICALS

The screening of chemicals for antitumor activity has been continuing at NCI since 1955. In the early years of the program, there were many compounds available and of interest in connection with this screening effort. As a result, selection and acquisition of compounds was not a major task. The screening capacity of the program, however, has always been considerable, and to date, over 400 000 distinct compounds have been tested.

A substantial proportion of the compounds tested by NCI have never been published and therefore are not included in the 7.9 million compounds registered⁴ by Chemical Abstracts Service (CAS). It follows therefore that NCI has examined less than 10% of all published structures, and there are some 7 million published compounds which have not been tested. In spite of this, it is no trivial task for the NCI to find thousands of compounds that are (a) available in half-gram quantities and (b) not closely related chemically to compounds which have already been tested.

The funding and staffing of the NCI program currently allow the screening of about 10 000 compounds per year. It is a matter of experience that of every three to five compounds considered for testing, only one will actually be acquired, and accordingly the program considers at least 30 000 compounds

each year. To examine so many compounds, the NCI maintains a standing relationship with several thousand laboratories around the world, and in any year, some 20 000 structures are proposed by such groups for testing. In addition, a continuing, semiautomated surveillance of the open literature is carried out, and about 6000 candidate synthetic compounds are so identified each year. Natural products are dealt with separately from synthetic compounds. A group of natural-products chemists at NCI maintains liaison with laboratories around the world and, as with synthetic chemicals, also monitors the literature continuously for novel or otherwise interesting structures derived from natural sources. In addition to these major sources, a variety of other organizations submit compounds to NCI in support of the program. These include universities and research foundations as well as numerous government laboratories, several of which are in the intramural laboratories of the NCI itself.

DATA ENTRY

When DIS development began in 1982, it was concluded that the state of the art did not permit chemical structures to be drawn in an acceptable form by a computer from connection tables. A prerequisite⁵ of the DIS was the capability to produce satisfactory chemical structures at output time, and accordingly the approach the DIS takes is to enter chemically and esthetically satisfactory structures which are then stored as sets of vectors and carried forward in the Chemistry record for subsequent output. The vector set is also used as the source of the connection table, which is needed for the searchable files. The entered structural diagram contains all the known stereochemical features, but these are used only for structure display; they are not incorporated into the connection table. A textual field containing the stereochemical descriptors is generated and may, if necessary, be used for searching. Our subsequent experience reveals, however, that stereochemistry is rarely important as a search criterion, suggesting perhaps that searching with three-dimensional structures is an unnecessary complication. Searching with purely two-dimensional structures very occasionally leads to false positives, but these can always be resolved by reference to either the textual stereochemical qualifier or the nomenclature fields.

All structures considered for acquisition and testing must first be entered into the DIS Pre-Registry, and this raises the first of several operational questions. If 5 min is allowed for structure entry and 3 min for verification, then an annual load of 50 000 structures will require 3.3 entry operators for input and verification. This is, moreover, a "best-case" calculation, because it assumes that an operator can function for 8 h with little break. In practice we find that people can work well for one hour out of two at these tasks, and so the manpower estimates must be doubled. In practice, the data entry and verification in the DIS, using specially designed programs which are described in this section, require about 4 min per structure; an annual input of 50 000 can be handled by 3-4 entry operators.

Several programs have been written in the DIS for entry of chemical structures. It is necessary to enter a structure as a preliminary to a structural search, and the DIS has a series of programs that allow any user to assemble a chemical structure. These are based upon the structure generation programs that are used in the NIH-EPA Chemical Information System.⁶ The user can employ different commands to build a ring, add side chains to it, designate specific element and bond types, and otherwise modify the growing structure. The major advantage of this approach is that one can use any type of computer terminal. Building structures in this way is, however, relatively slow. For a typical searcher, this is a minor problem and is outweighed by the convenience that a

specialized terminal is not required. Consequently, this method of structure input is used in the DIS for the generation of query structures, as will be seen in the next paper in this series. For those concerned with high-volume structure entry, however, slow throughput represents a far more serious problem than the need for a specialized terminal.

The first structure entry program that was designed at the NCI for database building employs a Hewlett-Packard "graphics" terminal (2623, 2648, or equivalent) and takes advantage of the soft keys on the terminal. A second program uses a Victor-9000 with a touch pen as well as a keyboard for input. A third approach uses keyboard entry on an IBM PC(XT) or PC(AT).

The first of these programs, conceived primarily as a demonstration project, uses the remote time-shared DEC-10 computer to process the input. Each keystroke on the Hewlett-Packard terminal is transmitted to the mainframe where it is interpreted and processed, with the result being returned to the terminal. With such older terminals, there are few alternatives to this approach. The concomitant I/O traffic is not conducive to a high-volume production operation, however, and consequently, this method of structure input was not seriously considered for this purpose.

In the second approach, programs were written that use a grid on the screen of a Victor-9000. The structure is built on the grid with the help of a touch pen and a menu. The menu may contain, in addition to commands, various commonly used structural fragments. This program proves to be very popular with operators because it allows them to "draw" with the touch pen much as they write with a pencil. A further significant advantage of the Victor program is that the user's terminal is, in fact, a microcomputer and is "smart" enough to handle command interpretation without reference to the mainframe. This type of distributed processing not only provides for more satisfactory response at the terminal but saves mainframe time. With this approach, a complete structure can be built locally at a speed which approximates that of drawing on paper. Once the structure is fully assembled at a Victor microcomputer, the vector set and the connection table that correspond to it are passed to the mainframe in a single package.

In the third method,⁷ the program previously used on the distant host was run on an IBM PC(XT); the chemical structures are generated completely on this microcomputer. It also has a "data packaging" feature and in addition employs a powerful set of softkey definitions, allowing one to create with a single keystroke, for example, a specific ring or chain of atoms. This method requires only about 25% of the keystrokes per structure used by other methods. In this approach, furthermore, use of a light pen is deliberately avoided so as to preserve the typist's cadence and so accelerate the entry process. Since it runs as a live job on the IBM PC, it is sufficiently fast that an operator can adopt a typist's rhythm and enter structures usually in less than a minute. The IBM PC(XT) method has not yet been put into production; it is currently used for nonproduction entry of structures into the DIS.

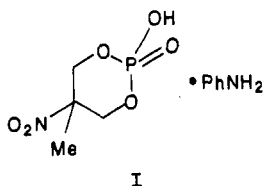
When the structure has been completely entered with either the Victor-9000 or the IBM PC program, it is uploaded to the mainframe, and in response the DIS prompts the operator to enter a molecular formula. The DIS maintains stringent quality control at this data entry stage, comparing the entered molecular formula to that which it calculates from the structure. Any discrepancy is referred back to the operator, and when the two formulas agree, the program goes on to prompt the operator for other data pertaining to the compound. These data include the identity of the putative supplier, any known biological activity, the date of entry, the identity of NCI staff associated with the compound, and so on. The entire

record is then assigned an identifying number (TID) and stored in the Pre-Registry area of the DIS, and the next entry is begun. Completed records are passed to data checking operators who retrieve and review all newly entered Pre-Registry records.⁸ When review is complete and approval is signified, the new record is moved on to the preselection review process. The connection table for the compound is retrieved and submitted to a complex series of operations in which it is broken into numerous chemical structure search keys which are ultimately used in the chemical structure searching programs and which are described in the next paper in this series. Selected search keys are equipped with index pointers and merged into (or concatenated onto) the search files in the Pre-Registry of the DIS. At the end of this operation, the new compound takes its place in the Pre-Registry file and will respond to searches with many of the parameters that are used in the DIS.

COMPOUND EVALUATION

Once a chemical structure has been entered into the DIS Pre-Registry, the first step in its evaluation is to determine whether or not the compound is unique to the system. In order to settle this point, two searches are done, first in the DIS Chemistry files for the full structure and then in the Pre-Registry files. These searches are automatic and in the majority of cases are quite straightforward, providing a "yes" or "no" answer to the question. The system response is also quite simple: if the full structure is found to be already in either file, it is not pursued further.

A complication arises when the structure at hand is a so-called "dot-disconnected" structure. Structures containing noncovalently bound moieties are usually represented in a dot-disconnected form.⁹ Thus aniline (PhNH_2) is not dot-disconnected, but aniline hydrochloride (PhNH_2HCl) is. The problem here is that the two moieties differ greatly in significance. The full structure search could be limited to retrieve only the exact match, aniline hydrochloride, but most would agree that aniline hydrobromide, while not an exact match, is very closely related to aniline hydrochloride, as is aniline itself. Another facet of the problem appears when aniline hydrochloride is contrasted with the aniline salt, NSC 110688 (I). To regard these as an exact match would seem far-



fetched, but they could be viewed as differing from one another no more than do aniline hydrochloride and aniline hydrobromide.

The DIS offers no solution to this quandary; instead it has adopted a pragmatic empirical procedure. The first fragment entered is considered to be the "main" fragment, and a full structure search is always done for this fragment, irrespective of its identity. If 50 or more hits are obtained from this search of the main Chemistry file, no other fragment searches are done and the Pre-Registry is not even checked for the compound. Fragments beyond the first one are not searched if they (a) contain no carbon or (b) are single-atom fragments (a hydrochloride has a single-atom fragment—Cl—because hydrogen atoms are implicit). When all the fragment searches have been done, the results lists are intersected together. All the compounds that are in all of the results lists will be found by the intersection and a check by molecular formula is done to find any exact matches of the starting structure. Such compounds are reported as "identical" and the remaining re-

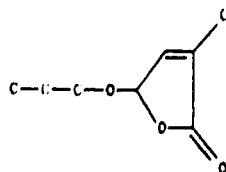
trievals are reported as "similar".

This proves to be a workable procedure. If aniline hydrochloride is entered as an example, one identical compound (aniline hydrochloride, NSC 7910) is returned and 22 related compounds—free aniline (NSC 242949) and 21 other salts of aniline—are returned as "similar". This information is carried forward and is weighed in the final decision concerning selection of the compound.

A second evaluative step that is handled within the Pre-Registry leads to an estimate of the octanol/water partition coefficient for the compound. The logarithm of the partition coefficient of any compound between water and *n*-octyl alcohol is often regarded as an indicator of its availability in biological systems.¹⁰ Several attempts have been made to develop computer programs that, given the structure of a compound, can estimate the octanol/water log *P*. The DIS has used the procedure developed by Hopfinger¹¹ and more recently has begun to use the procedure of Hansch and Leo.¹² The latter procedure is generally more useful because it can handle a wider variety of structural features. Both programs work by analyzing the compound's connection table and breaking it down into a complete set of smaller fragments. Each of these contributes in known measure to the solubility of the compound in water or *n*-octyl alcohol. These contributions, along with a series of combination terms and other structure-based correction factors, are used to compute the partition coefficient and hence its logarithm. The estimated value of the log *P* is saved, and any factors limiting the confidence of the estimate are also stored for future use.

The last evaluative step in the Pre-Registry is used to develop data pertaining to the expected activity of the compound. This is done by means of the methods developed by Hodes¹³ and draws upon the testing data that have been developed during the 30 years of the NCI program. A Hodes "training set" is derived by taking a group of compounds together with their measured activity in a tumor system. In this case, the model used data derived from the murine P388 leukemia screening, in which hundreds of thousands of compounds have been tested;¹⁴ the resulting model predicts, of course, only for this cancer. The structure of every compound in the set is decomposed to a set of atom-centered fragments or "keys", and each key, together with the P388 performance of the compound containing it, is stored. This file is sorted so as to find all the occurrences of each key, and the average activity associated with each key is recorded. The probability of occurrence of each key in the entire database is also recorded. It is then a fairly simple matter to build a table in which there is, for each key, a probability of P388 activity as well as an indication as to how frequently the key occurs across the whole database. As a new structure proceeding through the Pre-Registry arrives at the Hodes model stage, it too is broken into keys, and for every key, the activity and structural novelty scores are ascertained. Summation of these key scores in either category gives respectively an estimate of the projected activity and the overall structural novelty of the compound. Such "Hodes scores" may be seen in the Pre-Selection Report which is shown in Figure 1.

These estimates are combined with the log *P* estimate and the information concerning similar and identical structures in the existing files. To this are added other data concerning, for example, the source of the compound. Finally, using criteria established by NCI staff, the computer program makes a determination as to whether or not the compound should be acquired and tested. The sum of all this information for each compound, as well as the computer's recommendation, is printed in a report, an example of which is shown in Figure 1. This Pre-Selection Report is provided by the DIS on a weekly basis to NCI medicinal chemists for their review.



9219681

```

TID: 9219681-A
HODA: I
HFLG: OFF
HODN: 92
SONH: NO
OSWT:
HODM: 1
HODT: A
HDTS: 0
HOD1: I
HD1S: 12.29
HOD2: I
HD2S: 5.53
HOD3: H
HD3S: 4.46
HODL: J
HDLS: 24.21
HDAS: 32.3
NMAT:
MATC: 85278-Z\SIMILAR\SELECTED
      230337-Y\SIMILAR\SELECTED
      230340-C\SIMILAR\SELECTED
      230341-F\SIMILAR\SELECTED
      249797-A\SIMILAR\SELECTED
      271634-N\SIMILAR\SELECTED
      284621-F\SIMILAR\SELECTED
      286312-R\SIMILAR\SELECTED
      289946-R\SIMILAR\SELECTED
      329968-T\SIMILAR\SELECTED
      329969-U\SIMILAR\SELECTED
      350598-S\SIMILAR\SELECTED
      402957-G\SIMILAR\SELECTED
      600195-T\SIMILAR\SELECTED
      600199-X\SIMILAR\SELECTED
      600201-Z\SIMILAR\SELECTED
      601078-F\SIMILAR\SELECTED

SSPL: U.S. DEPT. AGRICULTURE
SLBN: 1951-A
CRIC: Y011-3;Y01M-5;Y02I-4
LOGP: 0.94
LPFL: CALCULATION COMPLETE
DTSL:
DTEN: 03-MAY-84
OTCK: P. LIWANAG
DTCK: 4
CONF: OPEN
ACAT: DFDV
EMRS: NO
COMI: 905
MOLF: C18 H12 O3
STXT:
DSOR: 14-MAY-84
NSOR: 280
SREC: SELECT
MTXT:
UHAZ:
HFST:
QRSN:

```

Figure 1. DIS Pre-Selection Report.

Since the weekly Pre-Selection Report contains 400–500 structures, each with a quantity of supporting data, the review process is necessarily fairly rapid. The computer's selection recommendation is provided as a concurrence/nonconcurrence option. If the medicinal chemist reviewer accepts the recommendation, no action is necessary; the computer's recommendation is adopted. Failure to accept the recommendation is signified by nonconcurrence. This causes the computer to ignore its recommendation and proceed in the other direction.

ACQUISITION OF COMPOUNDS

Once a structure has been entered into the Pre-Registry and has been designated for acquisition, as described in the previous section, a formal effort to acquire a sample is begun. This is done by the Acquisition section of the DIS Pre-Registry.

A central function of the Acquisition section of the DIS is the ordering of compounds. An order for a sample of a compound may be an initial order or a refill request. The Order module of the DIS can accept and handle either of these different types of orders. For initial acquisitions, the Order module identifies compounds by means of the Temporary Identifier (TID) used in the Pre-Registry. When this number is presented to the program, it retrieves the appropriate records for the compound from the Pre-Registry, ascertains the identity of the supplier, and then constructs a letter to be sent by the NCI to the supplier to request a sample of the compound.

The Pre-Registry is the only source of initial orders of compounds. Additionally, a substantial number of reorders come, as will be seen, from the Screening area, initiated when positive preliminary test results have been obtained on a compound. Since the initial supply (600 mg) of a compound is usually adequate only for preliminary testing, further testing will require an additional supply of the compound. In such cases a reorder, or refill request, may be initiated. Such an


order, triggered by the combination of a request for testing and an inadequate inventory, is placed manually by staff of the Acquisitions contractor. The reorder will, by definition, involve a registered compound, which will be identified by an NSC Number. The Order module accepts a "re-order" command, queries the operator for the necessary information, and initiates the order. The operator may designate the earlier supplier or a different supplier and may also suggest the amount of material to be requested. This transaction is then reviewed by a supervisor, and if approved,¹⁵ it is "placed". The order then enters the letter-generation phase where it will be mingled with all other initial orders and reorders. The program allows appropriately authorized staff to place orders directly, and it also permits the deferral of initiated orders as an alternative to approval or cancellation.

Compounds slated for acquisition or reorder may be confidential or not, and they may be synthetics or natural products. Furthermore, suppliers may or may not use their own identifying codes for compounds. To accommodate these variations, nine standard request letters have been compiled, from which the DIS will use the one appropriate to the case at hand. If more than one compound is dealt with in a letter, the changes to the text are such that a new "plural" letter is needed. There are therefore 18 letters to draw from. Initial acquisition letters are written in English unless they are directed to France, Germany, Japan, or Latin America and Spain, in which case they are written in the recipient's language. Something approaching one-quarter of all letters are written in a language other than English. The number of standard letters has thus swollen rapidly to 90, but these are all "canned", and it is quite trivial for the program to retrieve and use the correct one for any purpose.

In addition to being written in a language of choice, the body of the text may contain a number of variables. Prime among these is the number of compounds mentioned in the letter and the date of any prior correspondence that may be cited. Both of these variables are passed to the letter-writing program and incorporated into the appropriate letter. If only one compound is involved, a "singular" letter is produced and no number is printed.

The letter generator thus goes through a series of steps in preparing a file for printing. First, the number of compounds mentioned in the letter must be identified and the code for the letter type determined. The body of the letter is then assembled. Variable numbers within the letter are substituted, the letterhead is added, and, finally, the attachments are assembled. Many letters, particularly initial requests, carry attachments, each page of which provides information, including the structure, pertaining to one of the chemicals requested in the covering letter. The attachments also carry the name and address of the recipient, but they are printed in English and on blank stock rather than letterhead.

An assembled print file containing perhaps 100 letters is directed to a Hewlett-Packard 2380 laser printer, which is not restricted to defined character sets. This printer is provided solely with blank stock, and for each letter it must first print the NIH letterhead, which is a graphic. The printer must then revert to the English alphabet in a pica font in order to print the dateline and recipient address, the latter positioned to be readable through a window envelope. Next, the necessary alphabet must be determined and used for the body of the text, which includes the variable digits mentioned above. The signature block is printed in English pica. The printer then skips to the next page. If the letter has no attachments, the next page is a new letter and the process is repeated. If, on the other hand, the first letter required attachments, then these are printed without letterhead. These typically will contain a structure, and to draw this the printer must be provided with



DEPARTMENT OF HEALTH & HUMAN SERVICES
Public Health Service
National Institutes of Health
Bethesda, Maryland 20205

22 August, 1985.

Dr. Frederick J. Hampton,
Department of Chemistry,
Academic University,
Rivertown, KS, 66106

Dear Fred,

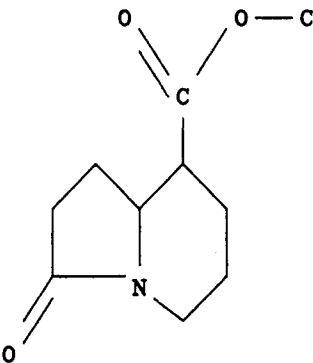
The Preselection Committee has completed its review of the compounds offered in your letter of 2 August, 1985. We would be interested in testing the compounds indicated on the enclosed sheet.

Initial sample size has a significant effect on the type of testing which can be performed. A sample of 600 mg is most adequate for initial assays to determine if a compound is active. A sample of at least 250 milligrams is required if your supply is limited. In general, the larger the sample size, the more meaningful the screening results will be.

Thank you for your interest and participation in the Cancer Chemotherapy Program.

Sincerely,

V. L. Narayanan, Ph.D., Chief,
Drug Synthesis & Chemistry Branch
National Cancer Institute
Landow Building, Room 5C-18B
Bethesda, Maryland 20205



92206221

Molecular Formula: C₁₀ H₁₅ N O₃

Figure 2. Acquisition letter.

the vector set generated for that compound during its initial preregistration. Any other data on the attachment are printed in English. A typical letter with a one-page attachment is shown in Figure 2.

With this program and the laser printer, over 200 letters are printed each week to order or reorder samples of about 450 compounds. Each letter and its attachment sheets are contiguous and can be mailed as they emerge from the printer.

RECEIPT OF SAMPLES

When an initial acquisition is completed, i.e., when the solicited sample is received by NCI, a formal receipt process is initiated. The material is identified to the Pre-Registry in terms of its Temporary Identifier, and the time and date of its receipt are recorded. The sample weight is generally not known at this point, but if it is, it can be entered.¹⁶ Any other pertinent data are also entered here, and the record is then passed to a senior reviewer. This person confirms that the data are all correct and releases the compound into the registration process.¹⁷ This begins with a check of the DIS Chemistry database to see if the identical structure has been registered since the structure was examined by the Pre-Registry. This is an unlikely event; should it occur, the compound is returned to the supplier. If the compound is still "new", it will be

registered immediately and assigned the next available NSC Number. At this point, a label carrying a barcoded version of the NSC Number is printed automatically and is affixed to the sample container.¹⁸ Almost the entire Pre-Registry record for the compound is written into the Chemistry database, and at the same time, an entry corresponding to that sample is written into the Inventory database. Finally, the sample is added to a "Courier List", which accompanies each package of samples sent to the Storage facility, and an entry in the Shipping database is made noting the transfer of material from the Supplier to the Acquisitions contractor. The entire Pre-Registry record is then expunged, as is the Order record. With reorders, the receipt process is simpler because the Chemistry record has already been established and is unaffected by a refill. The proper changes are made to the Inventory and Shipping databases, and the receipt is complete.

ORDER MODULE

Once the selection process is complete, the first acquisition of a compound is handled largely automatically by the DIS, as has been described. A reorder, on the other hand, may raise a number of uncertainties. Months, or even years, may have elapsed since the initial order, and the first supplier may no longer be able to supply more compound. In some cases,

OPTION? REORDER

NSC Number? 373853-T

Reorder Amount? 1.5 GM

Supplier Code? K26C-2

Reorder Reason (terminate with a blank line)?

: For testing in tumor panel.

:

Tumor System? 3LL39

Overdue Date? 1-MAR-86

Reorder Requested By? Jones

Comment?

Specification Completed

(ORDR) Order Number: 22174-K

(CMPD) Compound Identifier: 373853-T

(AMTO) Order Amount: 1.5

(UNIO) Units for Order Amount: GM

(SPLR) Name Code of Supplier: K26C-2

(ORSN) Reason for Order: For testing in tumor panel.

(TUMR) Tumor System: 3LL39

(DOVR) Overdue Date: 01-MAR-86

(DOVE) Elapsed Days Until Order is Overdue: 54

(RQTR) Person Who Requested Reorder: Jones

IS ALL INFORMATION CORRECT (Y/N) (Y)? Y

The order is to be placed (C) or to be left pending (N)? N

OPTION?

Figure 3. Reorder session.

alternative suppliers are known; in other cases, alternative compounds, whose supply is adequate, may be available. For these reasons, the reorder operation is handled by means of an interactive program which allows the staff to change any parameters that were associated with earlier orders of the compound.

An interactive use of the reorder facility is shown in Figure 3. In this case, the operator has initiated an order for 1.5 g of NSC 373853. The supplier that provided the first sample will be contacted again, with a letter which refers to one compound only and which will be written in German. When initiation of the order is complete, it must be approved by a supervisor,¹⁵ and then the letter and any attachments will flow into the DIS print stream. The order, in the meantime, is appended to the Order database of the DIS and will, in all its particulars, become searchable at the next update of this database, which takes place during the next weekend.

SEARCHABLE DATABASES

The basic function of the DIS Pre-Registry is to assist in the selection and acquisition of chemicals and to document that acquisition effort. In fulfilling these objectives, two databases are generated. These are the Pre-Registry database itself and the Order database. In addition, the Namecodes database is largely built as a result of acquisition efforts. All three of these databases are maintained along with the other larger DIS databases. They are updated on a regular schedule and are fully searchable in their own right.

a. Pre-Registry Database. The Pre-Registry database is searched automatically for every new structure entered, as has been described above. This database may also be searched interactively whenever necessary. Such interactive searching is accomplished in much the same way as searching in the main

OPTION? DTSL/SEP-85

422 hits were found and stored in file 1 associated with the PREREG database

OPTION? PRGSUBS 1

Doing sub-structure search
Type E to Exit

| | | |
|------------|-----------------|---|
| File entry | 20, Hits so far | 0 |
| File entry | 40, Hits so far | 0 |

:

| | | |
|------------|------------------|----|
| File entry | 300, Hits so far | 0 |
| File entry | 320, Hits so far | 10 |
| File entry | 340, Hits so far | 11 |
| File entry | 360, Hits so far | 11 |
| File entry | 380, Hits so far | 11 |
| File entry | 400, Hits so far | 11 |
| File entry | 420, Hits so far | 11 |

File = 2 Successful sub structures = 11

Figure 4. Search for pyrrolonaphthalenes in the Pre-Registry database.

Chemistry file, a subject which is discussed in detail in the next paper in this series. The search options in the Pre-Registry, however, are somewhat limited. Full structure and substructure searching is possible, and searches for molecular weight and molecular formula are supported. The Pre-Registry file is not screen-searchable for compounds having specific rings or atom-centered fragments because the necessary keys are generated for the first time when the structures are moved to the Chemistry file. There is, however, a wealth of other information in the Pre-Registry database, and much of this is extremely useful for searching. The database carries all the acquisition data, such as dates when various transactions occurred, the identity of the potential supplier, and the scoring of that structure in the Hodes model and the log *P* estimate. This information can be very useful when one is trying to discover NCI's experience with a particular structural type, as is shown in the following example.

In September 1985, samples of 422 compounds were solicited by NCI. These can be retrieved by means of the DTSL field, in which is stored the date of the "submittal letter":

OPTION? DTSL/SEP-85

422 hits were found and stored in file 1 associated with the PREREG database

When a search is made through this group of 422 for all compounds containing a pyrrolonaphthalene substructure, a total of 11 compounds are discovered, as shown in Figure 4. These compounds, it transpires, were all requested in a single letter, whose "submittal letter batch number" (SLBN) was 1729, as can be seen from the Pre-Registry entry for the first compound, TID 926619, which is shown in Figure 5. Batch number 1729 can be retrieved with its SLBN and has 36 compounds.

OPTION? SLBN/1729

36 hits were found and stored in file 3 associated with the PREREG database

All of these were requested in a letter dated 9 September, 1985, addressed to supplier Q42U:

OPTION? FORMAT SSPL DTSL

OPTION? T3//30

File: 3 Entry: 30

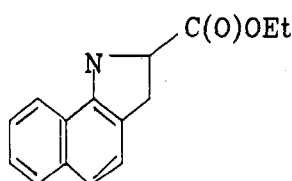
TID Number 926607-P

(SSPL) First Supplier Code: Q42U-4

(DTSL) 9-Sep-85

Certain of the crucial data items, it will be noted, are accompanied with alphanumeric check characters. Thus the TID Number 926607 is expressed as 926607-P. This alphabetic character allows the DIS to guard against mistyped identification numbers and is used only when data are being written into the databases. When searching, check characters are not used.

TID Number 926619-C
 (TID) Temporary Identifier: 926619-C
 (SLEN) Submittal Letter Batch Number: 1729-L
 (SSPL) First Supplier Code: Q42U-4
 (CONF) Confidentiality Status Flag: NON-DISCREET
 (DTSL) Date of Submittal Letter: 09-SEP-85
 (SUBM) Submission Category Code:
 D Donated Synthetic
 U University
 D Domestic
 E Starks Encouraged
 (COMI) Supplier Designation
 (MOLF) Molecular Formula: C15 H13 N O2



(STER) Structure Error Flag: NO ERRORS
 (MP) Melting Point: 170
 (DTEN) Entry Date: 17-SEP-85
 (OTCK) Checking Operator: 1131.2627
 (DTCK) No. of Days from Entry to Check: 7
 (LOGP) Log P Coefficient: 1.62
 (SREC) Selection Recommendation: REJECT
 (USOV) User Override of Selection Recommendation: YES
 (PROC) Pre-Registry Processing Flag: SELECTED

Figure 5. Pre-Registry record for a pyrrolonaphthalene.

Thus, in a matter of minutes, a user can search the entire Pre-Registry area for a particular structural type and, having found a "family" of compounds, can learn when and from whom they were requested.

Search capability of this sort fulfills a number of program objectives. It allows NCI staff to make decisions in the light shed by prior experience. Eleven compounds in this series were tested. In six cases, the decision to acquire and test the compounds was made over the recommendations of the Pre-Registry, which urged rejection of the structures. One of these selections is shown in Figure 5. The overriding decision was made because this ring system had not been tested systematically, although compounds containing it had been acquired as early as 1970 and as recently as 1984. This supplier had synthesized a number of compounds in the series, and it was decided to examine the whole group. Meeting a second program objective, the DIS makes a reliable audit trail available to NCI management. When a compound is found to possess antitumor activity, such audit trails become vitally important in establishing the ownership of a compound.

A Pre-Registry record is shown in Figure 5. This summarizes the data that have been collected on the compound

in question, including items such as the confidentiality of the compound (CONF), the supplier's designation for the compound (COMI), and the various applicable submission category codes (SUBM). The SUBM of E (Starks encouraged) represents an acquisition which followed primarily from the efforts of the acquisitions contractor (Starks) and is used by NCI to aid in monitoring the contractor's performance.

b. Order Database. The Order database of the DIS is used to place and track all orders for chemicals. It contains almost purely management data such as dates when specific events occur, the identity of the supplier, the reason for the order, and so on. As are all the DIS databases, the Order database is interactively searchable. A simple search of the request data field (DTRQ), for example, reveals that 709 orders were placed during September 1985.

OPTION? DTRQ/SEP-85

709 hits were found and stored in file 2 associated with the ORDER database

Of these, 616 represented initial acquisitions (OTYP/1) and the remaining 93 were reorders (OTYP/3).

OPTION? #1 AND OTYP/1

616 hits were found and stored in file 3 associated with the ORDER database

OPTION? #1 AND OTYP/3

93 hits were found and stored in file 4 associated with the ORDER database

A simple Boolean test confirms that every one of the 709 distinct compounds processed during September represented either an initial acquisition or a reorder.

OPTION? #2 AND (#3 OR #4)

709 hits were found and stored in file 5 associated with the ORDER database

Typical examples of each of these order types are shown in Figure 6. All of the fields shown in these examples can be displayed upon request, and most of them may be searched in the standard DIS manner by using the field/value format that was discussed in the previous paper.

The order that resulted from the pyrrolonaphthalene Acquisition effort can be retrieved from the Order database by using any of a number of parameters. Here it should be noted, however, that as a piece of data moves from one database to another in the DIS its field name changes. This is necessary so that the DIS can discern from the field name which database is to be searched. Thus while the supplier's identity is stored in the SSPL field of the Pre-Registry, it will be in the SPLR field of the Order database. Likewise, the TID number that is assigned a compound in the Pre-Registry is carried forward to the Order database, where it is stored in the CMPD field. Any of these pieces of data will retrieve the order in question:

OPTION? CMPD/T926607

One hit found and stored in file 6 associated with the ORDER database

OPTION? FORMAT ORDER

OPTION? T 6

File: 6 Entry: 1
 Order Number 19449-W
 (ORDR) Order Number: 19449-W
 (CMPD) Compound Identifier: T926607-P
 (OSLB) Submittal Letter Batch Number: 1729-L
 (DTRQ) Date of Original Order Request: 13-OCT-85
 (OTYP) Type of Order: INITIAL ACQUISITION
 (OSTA) Status of Order: PLACED
 (DOST) Date of Order Status: 13-OCT-85

| Order Number 18507-X | |
|----------------------|---|
| (ORDR) | Order Number: 18507-X |
| (CMPD) | Compound Identifier: T924612-V |
| (OSLB) | Submittal Letter Batch Number: 1622-T |
| (DTRQ) | Date of Original Order Request: 15-SEP-85 |
| (OTYP) | Type of Order: INITIAL ACQUISITION |
| (OSTA) | Status of Order: PLACED |
| (DOST) | Date of Order Status: 15-SEP-85 |
| (OAMT) | Amount Ordered: 800.0 MG |
| (DOVR) | Overdue Date: 14-DEC-85 |
| (DOVE) | Elapsed Days Until Order is Overdue: 90 |
| (DOUT) | Date Order Letter Sent: 15-SEP-85 |

| Order Number 17975-U | |
|----------------------|--|
| (ORDR) | Order Number: 17975-U |
| (CMPD) | Compound Identifier: 376738-H |
| (DTRQ) | Date of Original Order Requi: 04-SEP-85 |
| (OTYP) | Type of Order: REORDER BY MANUAL TRIGGER |
| (OSTA) | Status of Order: RECEIVED |
| (DOST) | Date of Order Status: 13-NOV-85 |
| (OAMT) | Amount Ordered: 50.0 MG |
| (SPLR) | Name Code of Supplier: 267A-2 |
| (ORGN) | Reason for Order: TO CONTINUE IN PS |
| (TUMR) | Tumor System: 3PS31 |
| (DOVR) | Overdue Date: 03-DEC-85 |
| (DOVE) | Elapsed Days Until Order is Overdue: 90 |
| (RQTR) | Person Who Requested Reorder: MRI |
| (OCMT) | Order Comment: TOXIC, RETEST |
| (AQMT) | Acquisition Method: FREE |
| (DOUT) | Date Order Letter Sent: 04-SEP-85 |
| (DINA) | Date Order Was Inactivated: 13-NOV-85 |
| (DTRC) | Date Order Was Received: 12-NOV-85 |
| (FAMT) | Amount Received: 57.6 MG ESTIMATED |
| (SAMT) | Amount Shipped: 57.6 MG |

Figure 6. Records for an initial acquisition and a reorder.

(OAMT) Amount Ordered: 600. MG
 (SPLR) Name Code of Supplier: Q42U-4
 (DOVR) Overdue Date: 11-JAN-86
 (DOVE) Elapsed Days Until Order is Overdue: 90
 (DOUT) Date Order Letter Sent: 13-OCT-85

c. Namecodes Database. The Namecodes database is a file of names, addresses, and affiliations of over 10 000 individuals with whom NCI has collaborated during the 30 years of the program. Each entry is identified by means of a "namecode" which is used for all reference to that entry. Other fields in the Namecodes database contain information as to the role played by the individual (supplier, screener, NCI staff, etc.) and a geographic code that permits ready access to the type of geographical statistics often requested of NCI, for example, the number of compounds received from domestic, as opposed to foreign, sources. Finally, by means of a cross-referencing system, Namecodes keeps track of changes that occur with the passage of time. Thus current and previous contact persons within a given organization are tracked—the former for attribution, the latter for future correspondence. Copies of all screening data are automatically sent to the supplier of a compound, or his designee, and this is facilitated by a current Namecodes database. In the pyrrolonaphthalene acquisition which has been illustrated in this discussion, the submission letter was addressed to supplier Q42U: reference to the Namecodes database reveals Q42U to be a Dr. Hosmane of the University of Maryland in Catonsville, MD.

OPTION? FORMAT NAMECODES
 OPTION? T (NCOD/Q42U)
 Name and Address Q42U-4

(NCOD) Name Code: Q42U-4
 (ADDR) Name and Address:
 Dr. R. S. Hosmane
 Department of Chemistry
 University of Maryland
 Baltimore County
 Catonsville, Maryland 21228
 (NAME) Name: HOSMANE,RS
 (ORGN) Organization: Univ Maryland-BC
 (CLSC) Classification Code: 5;12
 (GEOC) Geographic Code: E1119
 (NCON) Normal Confidentiality Flag: 0
 (DTNA) Date Name Code Assigned: 29-NOV-84
 (SALU) Salutation: Dr. Hosmane

FLOW OF DATA

The end result of all these transactions is that both the Chemistry and the Pre-Registry databases grow at a rate of about 10 000 new records per year. Some 30 000 new records enter the Pre-Registry annually, but half of these are dropped when a sample is received. The fact that NCI attempted to acquire a compound and that to date the attempt was unsuccessful constitutes valuable information, and it is just those compounds which will remain in the Pre-Registry database. Since policy dictates that the same chemical may not be accepted from different suppliers, then solicitation of a second supplier must not be undertaken until the result of the earlier solicitation is known. This system achieves that end. If a new structure under consideration has already been acquired, or if an acquisition has been attempted, the DIS will detect this, and the Acquisition staff, so alerted, can resolve the problem.

Table I

| category | MCAT | no. of compd | % of total |
|----------------------------|------|--------------|------------|
| synthetic (donated) | D | 7900 | 88.56 |
| synthetic (contract) | S | 197 | 2.21 |
| synthetic (purchase) | P | 42 | 0.47 |
| synthetic (grant) | G | 81 | 0.91 |
| natural product (donated) | N | 367 | 4.11 |
| natural product (contract) | C | 40 | 0.44 |
| natural product (purchase) | K | 8 | 0.09 |
| prep lab (contract) | L | 234 | 2.62 |
| miscellaneous | M | 49 | 0.55 |
| totals | | 8918 | 99.96 |

No decision has been made as to how long such "unfinalized" Pre-Registry records are to be kept. The file grows at an annual rate of only 10000 structures, however, and accordingly it will be allowed to grow for several years before its size becomes a problem.

ACQUISITION STATISTICS

The DIS retains a considerable amount of information concerning the type of compounds acquired as well as their source. These data can be retrieved in the usual way and are very useful in analyzing the performance of the acquisition effort.

Every compound acquired is assigned to one of nine material categories (MCAT). These are broad classifications such as "Natural Product, donated" or "Synthetic, grant supported", and they serve to place each compound in one or another of these administrative categories. The entire group of compounds acquired in a given year can be distributed into the appropriate category with a simple command such as

OPTION? DACQ/1985 AND MCAT/D

7900 hits were found and stored in file 1 associated with the INVENTORY database

With search statements of this sort, the 8920 compounds acquired during the first 11 months of 1985 can be accounted for according to Table I. It is clear from these data that by far the largest category of chemicals is synthetic compounds which are donated to NCI for testing. Natural products acquired in the same way constitute the second largest group. Occasionally, support in the form of either a grant or a contract was used to underwrite the cost of providing compounds to NCI, and in 234 cases during 1985, the NCI contracted with a laboratory to synthesize compounds. These were, in most cases, compounds that had been found to have activity and for which more extensive testing was scheduled. The importance of these categories does not change greatly from year to year and points up, for example, the importance of voluntary donations vis-à-vis contractually supported derivation of compounds for testing.

The source of acquired chemicals is also categorized, the data being stored in a field called SCAT. This field is used to record the origin of chemicals, as industry, university, research laboratories, NCI, government (excluding NCI), independent groups, or miscellaneous. If the 1985 acquisitions are broken down by these categories with search statements such as

OPTION? DACQ/1985 AND SCAT/N

257 hits were found and stored in file 2 associated with the INVENTORY database

the results plotted in Figure 7 may be derived. From these, it can be seen that industry is the major supplier of compounds with 71.58% of the total. Universities are also major suppliers, and there are significant numbers of compounds acquired from

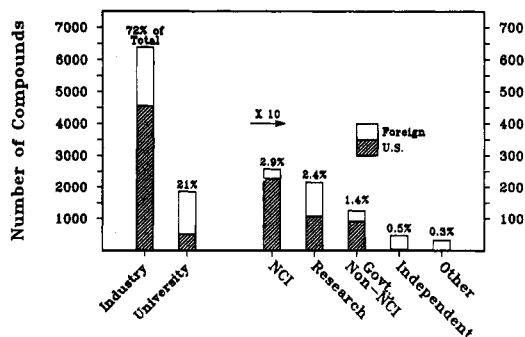


Figure 7. Source of compounds acquired during 1985.

research foundations and from NCI laboratories.

A geographical code (GCAT) is also maintained. This indicates only that an acquisition is from the United States or not and was used to provide the breakdown in Figure 7.

Finally, an "acquisition category" (ACAT) is carried. This represents a reason for the acquisition, and the 106 current such categories include entries such as "protocol analogs of vincristine", "hormonal agent", or "other biological activity (non-cancer)". These data provide a useful view of the circumstances that surround acquisitions. This field can be used as shown below to demonstrate that in 1985 NCI acquired 222 plant products, excluding the 2 plant products actually prepared by NCI itself, and 53 animal products.

OPTION? DACQ/1985 AND ACAT/N

222 hits were found and stored in file 1 associated with the INVENTORY database

OPTION? DACQ/1985 AND ACAT/M

2 hits were found and stored in file 2 associated with the INVENTORY database

OPTION? DACQ/1985 AND ACAT/G

53 hits were found and stored in file 3 associated with the INVENTORY database

These and all other Pre-Registry data are carried forward into the Chemistry database when a compound is registered. In this way, a permanent record of all of this information is maintained and can be searched interactively as is described in the next paper in this series.¹⁹

REFERENCES AND NOTES

- (1) This is the current NCI policy and was established in 1980. Prior to that date, samples from different suppliers of a single compound had been accepted. This, however, led to confounding of patent rights, and accordingly the practice has been abandoned.
- (2) The acronym "NSC" stands for National Service Center, a short form of Cancer Chemotherapy National Service Center, the early name for the program. The "NSC Number" is used by NCI as a Registry Number. Other Registry Numbers, such as the CAS Registry Number, are not useful for NCI because CAS Registry Numbers cannot be assigned to the confidential structures which constitute about half of the NCI database.
- (3) The compounds tested are all presumed to possess acute or chronic toxicity. Consequently, the storage facility must be specialized in that it must be able to tolerate such hazards while preserving a safe environment for staff.
- (4) On 23 June, 1986, the CAS Registry contained 7908 232 entries (STN International online "News").
- (5) Report of the NCI Ad-Hoc Extramural Committee on Information Management, Jacobus, D. P., Chairman, Feb 1979, p 12.
- (6) Heller, S. R.; Milne, G. W. A.; Feldmann, R. J. "A Computer-Based Chemical Information System". *Science (Washington, D.C.)* **1977**, *195*, 253-259. Milne, G. W. A.; Heller, S. R. "NIH/EPA Chemical Information System". *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 204-211.
- (7) Feldman, A. P. U.S. Patent 4476 462, 9 Oct, 1984. Other patents pending.
- (8) Such supervision by senior staff is critical because an error in, for example, a chemical structure, at this stage will be perpetuated through the system and may never be detected. It is for this reason that all entered structures are checked with some care as they are being entered.
- (9) The "dot-disconnect" notation is used in Chemical Abstracts procedures to describe materials which contain more than one moiety not covalently bound together. Salts are the most obvious example of dot-disconnected

- structures. Other examples include simple mixtures, alloys, and polymers formed from a mixture of monomers.
- (10) Hansch, C.; Fujita, T. " ρ - σ - π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure". *J. Am. Chem. Soc.* **1964**, *86*, 1616-1626.
 - (11) Potenzzone, R. Ph.D. Thesis, Case Western Reserve University, Cleveland, OH, Feb 1979. See also: Potenzzone, R.; Hopfinger, A. J. "Structural Correlates of Carcinogenesis and Mutagenesis. A Guide To Testing Priorities". *Proceedings of the 2nd FDA Office of Science Summer Symposium*, 28 August, 1978; U.S. Government Printing Office: Washington, DC, 1978.
 - (12) Hansch, C.; Leo, A. *Substitution Constants for Correlation Analysis in Chemistry and Biology*; Wiley: New York, 1979.
 - (13) Hodes, L. J. "Computer-Aided Selection of Compounds for Antitumor Screening: Validation of a Statistical-Heuristic Method". *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 128-132. Hodes, L. J. "Selection of Molecular Fragment Features for Structure-Activity Studies in Antitumor Screening". *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 132-136. See also: Hodes, L. J.; Hazard, G. F.; Geran, R. I.; Richman, S. "A Statistical-Heuristic Method for Automated Selection of Drugs for Screening". *J. Med. Chem.* **1977**, *20*, 469-475. Hodes, L. J. "Computer-Aided Selection of Novel Antitumor Drugs for Animal Screening". *ACS Symp. Ser.* **1979**, *112*, 583-603.
 - (14) Murine lymphocytic leukemia, strain P388, is the preliminary screen against which all compounds have been tested since 1976. The number of compounds for which there are P388 data in the DIS is now in excess of 200 000.
 - (15) The importance of this check is not so much for data validation (see ref 8) but to guard against the possibility that by misrouting correspondence some confidential information may be transmitted to persons outside the government who are not authorized to see it.
 - (16) The Acquisitions contractor is not equipped to handle hazardous substances and is therefore instructed not to weigh samples that are received. When the sample weight is provided by the supplier, it can be entered and flagged as "estimated".
 - (17) The registration is carried out online, in real time. The registration data will not, however, be searchable until the next subsequent inversion of the Chemistry database, which is a weekly event.
 - (18) From this point on, all movement of this sample is tracked by means of the barcoded label, which must be scanned as it enters or leaves the storage facility. This, coupled with electronic weighing, as is described in a subsequent paper, is crucial to the maintenance of an accurate inventory.
 - (19) Milne, G. W. A.; Feldman, A.; Miller, J. A.; Daly, G. P. "The NCI Drug Information System. 3. The DIS Chemistry Module". *J. Chem. Inf. Comput. Sci.* **1986**, following paper in this issue.

The NCI Drug Information System. 3. The DIS Chemistry Module

G. W. A. MILNE* and ALFRED FELDMAN

Information Technology Branch, Division of Cancer Treatment, National Cancer Institute, Bethesda, Maryland 20205

J. A. MILLER and G. P. DALY

Fein-Marquart Associates, Baltimore, Maryland 21212

Received April 21, 1986

The Chemistry Module of the Drug Information System (DIS) handles a database of 400 000 structures. New or modified records are created in this database on a daily basis and are merged into the file promptly. The Chemistry database is searchable in a wide variety of ways and provides novel methods for both input and output of chemical structures.

INTRODUCTION

In terms of data content, the Chemistry and Biology databases are regarded as the most important in the National Cancer Institute's Drug Information System (DIS). From the point of view of capability, however, the programs that manipulate these databases differ widely. While searching of the Biology database, which is described in part 5 of this series, is straightforward, the search of the Chemistry database is very complex. Corresponding differences are encountered in file management, in input and output processing, and so on. These differences are reflected in the sheer amount of code required by the Chemistry Module,¹ which makes it by far the largest of the DIS and which has required more developmental effort than any other. The large amounts of software reflect the special needs that pervade every aspect of chemical data processing. Here one finds continual mingling of graphics and alphanumeric material, requirements for exhaustive, accurate, and fast searching, complex searching algorithms, many different modes of input and output of data, a need for frequent updating, and a major data security concern. All of these problems come together in the Chemistry area.

For chemical input and output, special hardware is required, operated by specific, device-dependent software, and this software must be interfaced with the rest of the system. Chemical structures further require specialized software for validation, updating, indexing, searching, and so on. Much of this processing is severely computationally intensive, warranting the introduction of techniques capable of alleviating the system's load, but which add to the complexity of the

system. Graphical chemical data undergo only the normal processes of editing, updating, and searching. These functions, seemingly similar to the processes undergone by conventional data, i.e., text and numbers, in actuality are so incompatible that it was deemed best not to let the chemical requirements burden the TDRS—the resident database management system of the DIS. It was decided to maintain a separate database, with separate update, maintenance, and search software for chemical structures, and to interface only the outputs of these two systems. Having to deal not only with the complexities of chemical structure processing but also having to maintain and interface two separate file management systems contributes substantially to the complexity of the Chemistry module.

Systems of the magnitude of the DIS are not created from scratch, but built on experience gained with earlier systems. Thus the DIS benefited considerably from the experience gained with the system that it replaced, a system operated for many years under contract to the NCI by two organizations: Chemical Abstracts Service (CAS), which dealt with the chemical information, and VSE Corp., which handled the biological data.² But, partly because that system was not well integrated—its chemistry ran on one computer, its biology on another, and the linkage between the two was never adequate—the NCI decided in 1981 to design and develop the Drug Information System (DIS). The predecessor system and template for the DIS became the Structure and Nomenclature Search System (SANSS), a system which, for a number of years, had been operated jointly by the NIH and the EPA.³ A summary of the relationships between the SANSS and the