which will analyze the amount of vaporization and liquid volume that will result for certain petroleum components when pressure and temperature are varied.

Several data-processing applications are also included in SMART. An example of this type application is the Manpower Retrieval Program. Given a set of "requirements" (in terms of education, experience, skill, job level, and other characteristics), this program will search the entire manpower inventory file and select those employees fulfilling all of the requirements.

SMART manuals have been prepared containing the necessary operating instructions and detailed write-ups for every production applications. At last count well over 400 such manuals have been requested by company employees throughout the world.

SYSTEM STATISTICS

To evaluate system performance in terms of dollars and cents, SMART automatically maintains utilization statistics by remote location and application. Also maintained is the actual elapsed computer time to produce a solution.

Although solution times may vary from a fraction of a second to several minutes, using 1964 statistics the average solution time, for the 20,000 problems submitted during that year, required 12 seconds on the 1410 with a resulting company cost of 35 cents. This amount is particularly significant when viewed in comparison to the several man-hours or man-days required to arrive at comparable solutions by manual methods.

CONCLUSION

The primary advantages of SMART may be summarized as follows:

A. Operating Efficiency
   1. Total job time has been significantly reduced.
   2. Maximum utilization of the computer is achieved since regularly scheduled jobs may be running concurrently.
   3. Computer operator errors and job setup time are virtually eliminated.

B. Remote Capabilities
   1. The large scale computing capabilities of the New York Headquarters have been linked, through existing telephone facilities, to any point in the worldwide Mobil organization.
   2. As many as 64 different company locations may use these facilities simultaneously.

C. Time Current Information
   1. Up-to-date operating information is maintained on the computer's mass memory and obtained by company management within seconds.

It is for these reasons, as well as the direct system savings, that we are convinced the SMART system has contributed to our ultimate goal—company profitability.

Chemical Substructure Searching with Linear Notations*

BEATRICE A. MARRON, GLORIA R. BOLOTSKY, and STEPHEN J. TAUBER
Center for Computer Sciences and Technology, Institute for Applied Technology,
National Bureau of Standards, Washington, D.C. 20234

Received February 8, 1966

A technique for doing chemical substructure searching directly on linear notations, an input format, is described. Some of the implications and limitations are discussed.

Among the requirements for a comprehensive chemical information storage and retrieval system is the ability to search a file of chemical structures both for individual compounds and for classes of compounds defined by any desired substructure. There is presently greater emphasis on searching chemical structures than related information both because of the very basic nature of structure information, and because it lends itself to mechanical searches.

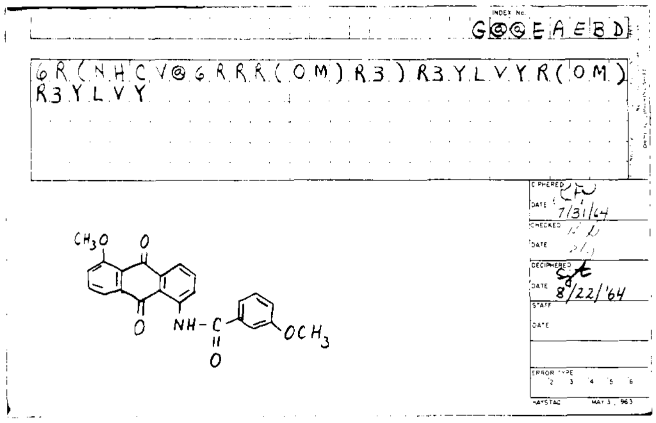
Chemical structures are usually shown in written communications as networks of atoms and bonds, but linear notations have an advantage for computer input because they consist of sequences of symbols. Here is described an experimental system developed at the National Bureau of Standards to perform chemical substructure searches on Hayward linear notations. The general method should be applicable to other linear notations.

INPUTS

The basic unit of information in the system is the chemical structure. Each structure as it enters the system is assigned an eight-digit identification number.

* Presented in part at the 1965 Congress, International Federation for Documentation, Washington, D.C., Oct. 12, 1965.

(Hexadecimal numbers are used for convenience in interfacing with a binary machine. The symbol "@" represents zero, and the letters "A" to "O" represent the digits 1 to 15 respectively.) Figure 1, a card, shows the original hard-copy input to the system, with the identification number in the upper right-hand corner. The lower half of the card contains the structure diagram, universally intelligible to chemists but not readily convertible to machine language. In the boxes there is recorded the Hayward linear notation (*1*), a sequence of symbols which uniquely and unambiguously defines the structure of the compound. Trained chemists are not needed for enciphering most compounds; it has been shown that college students can learn and apply notation rules within two weeks (*2*).



INDEX No. G@@EAEBD

6R(NHCV@6RRR(OM)R3)R3YLVYR(OM)
R3YLVY

CIPHERED CFU
DATE 7/31/64
CHECKED
DATE
DECIPHERED Sgt
DATE 8/22/'64
STAFF
DATE
ERROR TYPE 2 3 4 5 6
HAYSTAC MAY 3, 963

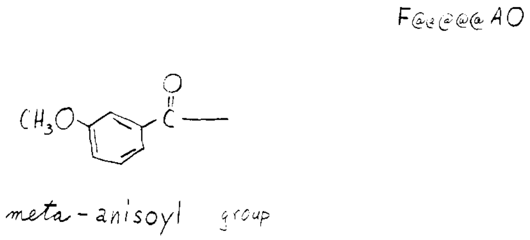
CH3O
O
O
NH-C
O
OCH3

Figure 1. Hard-copy disclosure input card.

The identification information and linear notation from such cards were punched in a prescribed format on an eight-channel punching typewriter in ASCII (*3*). The format included some nonprinting symbols for error-checking purposes, delimiting the start and finish of each entry and requiring exactly eight digits in the serial number. Other nonprinting symbols indicated deletion of any desired number of symbols or of an entire entry. The rate of human error in punching long strings of code is understandably high, and a single punching error invalidates the entire entry.

A loading routine checked the entries for proper punching format (concerning proper notation format *cf.* below) and packed them for economical storage in the computer memory. They were then dumped on magnetic wire cartridges. Each card thus became a variable length record, and a test file of over 2000 organic structures from *Index Chemicus* (*4*) was prepared.

A similar question file was prepared for common substructures from a list of organic substructure names (*5*). New questions can be input from the keyboard during any run. A typical substructure in the file is shown in Figure 2. In contrast to complete structures, however, a given substructure has a cipher which depends on what it is attached to. (Basically, it is a matter of where you may start and what paths you may take through all the atoms.) But there is still only a specific number of ways for any given substructure to be ciphered. In the present example, there are five forms of the cipher. The ciphers for substructures do not use the abbreviations of the notation system (see below). Provision exists for interruption of the symbol string by arbitrary substrings and for a special type of permutation associated with cyclic substructures.



F@@@@@AO

Figure 2. Substructure question card.

THE SEARCH SYSTEM

The PILOT research computer facility at the National Bureau of Standards was used. This includes a three-address, fixed word-length central processor with 32,768 words of core storage, 72 bits per word.

In a search, an attempt must be made to match each disclosure compound with all alternate forms of the question in turn, until either a match is found, or the list of forms is exhausted. The concept is straightforward, but the housekeeping becomes tedious.

Figure 3 is a gross flow chart for the system control routine. The programming is on a modular basis. If a question is already in the file, its number is entered from the console; new questions must first be entered into the file. The routine prints the question number and its Hayward notation; then it deals with each entry in the disclosure file, in turn unpacking the entry into individual

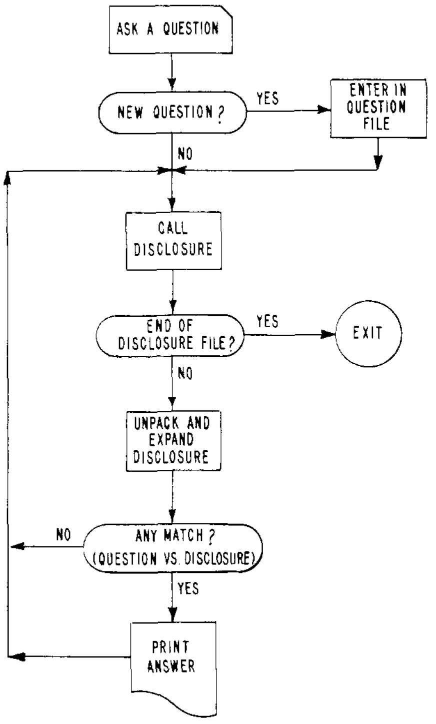


Figure 3. Gross flow chart for system control routine.

characters, expanding abbreviations inherent in the cipher system, checking for internal consistencies, and then matching the resultant character string against each of the alternate strings of the question. For each disclosure that is found to match a question string, its number and cipher are printed.

The cipher abbreviations which occur are (1) "Q", "V", and "Z" for structure units which would otherwise require more symbols, (2) designation of repetition by multiplying numerals, and (3) omission of parentheses from simple substituent designators. The internal consistencies which were required are (1) valid notation symbols, (2) matching parentheses, brackets, etc., (3) an acyclic structure if the notation begins with a multiplied substituent, (4) a cyclic main path if there is a cyclic substituent, (5) valid multiplying numerals, and (6) at least one organic ion in a salt.

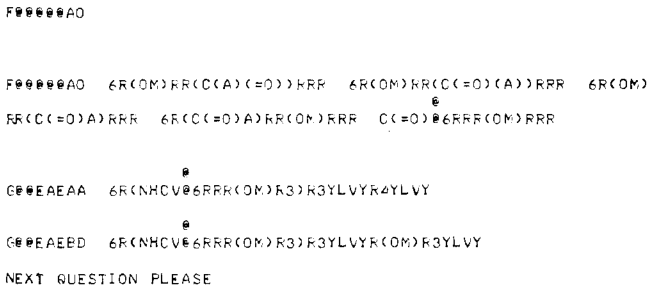
Comparison of strings proceeds character by character, starting in turn with each character of the disclosure string. When the "end of cipher" flag is reached in the disclosure string this not only causes a mismatch but also causes the next alternate question string to be brought up for matching. If one of the alternate question strings matches a substring of a disclosure, then the answer is printed, the next disclosure is immediately brought up, and the first question alternative is brought back. If the last alternate does not match, then the process is repeated with the next disclosure.

Positions within the question strings where any arbitrary substring within a disclosure is to be accepted as matching are signaled by the symbol "A", which does not occur in the disclosures. Such arbitrary substrings are terminated by the first right parenthesis, ")", found.

A refinement has been devised, but not implemented, to generate alternate strings by a substring-by-substring circular permutation of question strings. The beginning of each substring is recognized by certain cipher characters and automatically marked; as many permutations exist as there are substrings. If none of the permutations produces a match, then the string is recopied, substring by substring, from the last substring backwards; each permutation of the reversed string is then a further alternate question string. The question string must be flagged to indicate that permutation is desired. Certain characters which behave anomalously under permutation are recognized in the disclosures and either only counted, position disregarded, or else suppressed entirely—*e.g.*, ring size numerals.

The routine as it has now been implemented is applicable to seeking substructures for which it is practical to write out all alternate symbol strings, with interruption by arbitrary substrings permitted at the end of substituents—*i.e.*, immediately before a right parenthesis. A pragmatic limit is set by the number of alternate strings one is willing to write down. Inclusion of the permutation provision would extend the applicability to a class of substructures for which the number of alternate substrings may be very large: fused ring systems and substituted monocyclic systems.

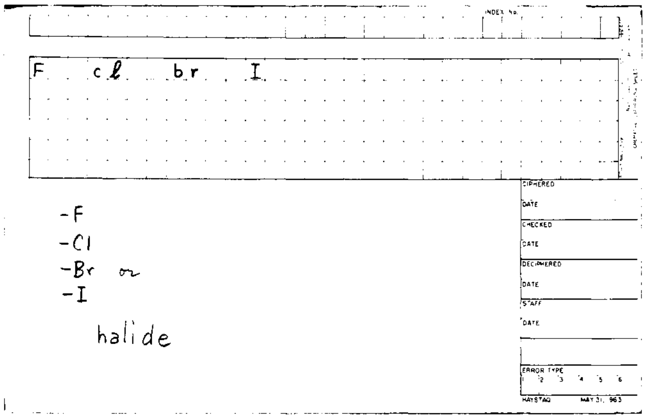
Figure 4 is a sample output sheet. Only the question number in the first line was punched by the operator; the rest is computer printout. This particular question is the *meta*-anisoyl fragment of Figure 2. The five forms of the question are separated by double spaces. Note the "A" used for arbitrary substrings. Here two disclosures were found to match (including the disclosure shown in Figure 1). The notations printed out could be deciphered by anyone who knows the notation, or else hard copy cards could be extracted from files on the basis of the identification numbers. These cards show the structure diagrams for the compounds and could also contain related information.



```
F@@@@@AO


F@@@@@AO  6R(OM)RR(C(A)(=O))RRR  6R(OM)RR(C(=O)(A))RRR  6R(OM)
                                         @
RR(C(=O)A)RRR  6R(C(=O)A)RR(OM)RRR  C(=O)@6RRR(OM)RRR


                 @
G@@EAEAA  6R(NHCV@6RRR(OM)R3)R3YLVYR4YLVY
                 @
G@@EAEBD  6R(NHCV@6RRR(OM)R3)R3YLVYR(OM)R3YLVY

NEXT QUESTION PLEASE
```

Figure 4. Sample output sheet.

Since the treatment of alternate strings representing a substructure is really an application of the Boolean "or", the same technique is applicable to alternate substructures. Figure 5 is a very simple example of this; the question represented is to find any halides in the file.



F cℓ br I

-F
-Cl
-Br or
-I

halide

INDEX No.
CIPHERED
DATE
CHECKED
DATE
DECIPHERED
DATE
STAFF
DATE
ERROR TYPE
1 2 3 4 5 6
HAYSTAQ
MAY 31, 963

Figure 5. Halide question card.

This system is intended to be one part of a comprehensive chemical information handling system (*6*) in which encoded structure information is one of the access points to all types of chemical information. Screening techniques would, of course, have to be included in the comprehensive system. The specific computer programs developed apply to the Hayward notation system, but the basic method appears to be applicable to any notation system in which at least some substructures can be recognized by characteristic sequences of symbols.

LITERATURE CITED

(1) Hayward, H. W., "A New Sequential Enumeration and Line Formula Notation System for Organic Compounds," U. S. Patent Office Reaearch and Development Reports, No. 21, U. S. Patent Office, U. S. Department of Commerce, Washing-

ton, 1961. Available from Superintendent of Documents, Washington, D. C. 20402, $1.25. Under revision.

(2) Hayward, H. W., Sneed, H. M. S., Turnipseed, J. H., Tauber, S. J., *J. Chem. Doc.* **5**, 183 (1965).

(3) Proposed Revised American Standard Code for Information Interchange," American Standards Association Committee X3.2, document 206, American Standards Association, New York, January 21, 1965.

(4) Garfield, E., ed., "Encyclopedia Chimicus Internationalis (Cumulative Index Chemicus)," Institute for Scientific Information, Philadelphia, Pa., 1962.

(5) "The Naming and Indexing of Chemical Compounds from Chemical Abstracts," American Chemical Society, Easton, Pa., 1962, p. 87N.

(6) Tauber, S. J., "Digital Handling of Chemical Structures and Associated Information" in "Association for Computing Machinery, Proceedings of the 20th National Conference," Association for Computing Machinery, New York, 1965, p. 206ff.; Marden, E. C., "HAYSTAQ—A Mechanized System for Searching Chemical Information," National Bureau of Standards Technical Note No. 264, National Bureau of Standards, U. S. Department of Commerce, Washington, D.C., 1965, p. 29ff. Available from Superintendent of Documents. Washington, D. C. 20402. 50¢.

Links and Roles in Coordinate Indexing and Searching: An Economic Study of Their Use, and An Evaluation of Their Effect on Relevance and Recall[*]

J. G. VAN OOT[a], J. L. SCHULTZ[a], R. E. McFARLANE[a], F. H. KVALNES[a]
Textile Fibers Department, E.I. du Pont de Nemours
& Company, Inc., Wilmington, Delaware

and A. W. RIESTER
Film Department, E. I. du Pont de Nemours & Company, Inc., Buffalo, New York

Received March 24, 1966

In a two-phase evaluation of links, roles, and type of indexing vocabulary (prescribed terms *vs.* a freer vocabulary), it was found that certain roles, used only on selected terms, can increase relevance of an answer with only minor reduction of recall. Role definitions must be clearly mutually exclusive. Roles are not useful if, fairly frequently, any term is indexed in several roles in one document-link, or searched in several roles in one question. Links also increase relevance, with essentially no reduction in recall, but are economically used only when the indexing is such that repetition of terms from one link to another for the same document is not frequent.

This paper covers a three-year period of part-time work of the five authors plus assistance from other personnel: clerks, computer programmers, consultants, and experts in the various fields of chemistry and textile technology who judged the relevance of the documents retrieved in our test of links and roles discussed in this paper.

Of necessity then, this description must be highly condensed. We present, therefore, just an outline of the experimental procedures used, concentrating on the results; giving a philosophy of using links and roles which we developed from this study of the effectiveness of links and roles in reducing false retrieval, and of the economics of their use.

THE PROBLEM

The specific question which faced us when we began this experiment was whether or not our indexing system should be converted to the use of links and/or roles. By our indexing system, we mean only one of perhaps 10 information centers in the Du Pont Co. which use concept coordination techniques to index internal company reports of proprietary information.

Our system, serving the needs of the Textile Fibers and Film Departments, was started in 1950 (*1*). It was not until eight years later that our colleagues in the Engineering Department reported (*2*) the results of their studies of the use of links and roles in coordinate indexing. We followed with interest the successful establishment in the Plastics Department of a coordinate index to patents, using links and roles (*3*). Other departments started indexing centers during the period 1958–1962 using links and roles in concept coordination techniques.

[*] Presented before the Division of Chemical Literature, 148th National Meeting of the American Chemical Society, Chicago, Ill., Sept. 3, 1964.

[a] Present address: Secretary's Department, Information Systems Division, E. I. du Pont de Nemours & Co., 3211 Centre Road Building, Wilmington, Delaware 19898.