8. The fact that large batches of questions could be processed against some 150,000 WLNs on an IBM 360/30 demonstrates that retrospective substructure searches are indeed practical with these linear notations and their self-determined bit screens.

STRUCTURE MAKE-UP

The binary screens discussed above have also been used in studying the structure make-up of new compounds being reported in the literature. For example, Figure 9 gives the binary screen statistics for the 153,000 new compounds reported in 1968.

The symbol J is at the top of the list for primary characters. Nearly 60% of the new compounds reported in 1968 contained a ring system other than benzene. Of these, 33% contained carbocyclic ring systems. About 42% of the new compounds contained an unfused benzene ring. About 47% of the new compounds contained a carbonyl group.

Despite extensive use of contraction rules the "1" (for $-CH_3$ or $-CH_2-$) is still the third most frequently used primary character and the most frequent multiple character. Multipliers are used in about 12% of the compounds coded.

A careful review of Figure 9 should be of value to anyone concerned with the make-up of new structures or the design of fragment codes. Studies on screen combinations are planned.

GENERATION TIMES FOR SCREENS

The three-field screens can be generated in less than 5 minutes for the monthly *ICRS* tapes (ca. 17,000 WLNs) and are now provided as part of the *ICRS* system.

AVAILABILITY OF PROGRAMS

The RADIICAL programs referred to above are included in the *ICRS* system provided by the Institute for Scientific Information.

SUMMARY

Substructure searching on files containing hundreds of thousands of compounds is now economically feasible. ISI has developed a series of computer programs for effecting such searches. Part of the efficiency of the programs is directly attributable to the binary screens discussed in this paper.

REFERENCES

(1) Garfield, E., G. S. Revesz, C. E. Granito, H. A. Dorr, M. M. Calderon, and A. Warner, "Index Chemicus Registry System: Pragmatic Approach to Substructure Chemical Retrieval," *J. Chem. Doc.*, **10**, 54–8 (1970).
(2) Smith, E. G., "The Wiswesser Line-Formula Chemical Notation," McGraw-Hill, New York, N. Y., 1968.

Alternatives to Searching Semantic Surrogates of Chemical Structures*

RICHARD I. RUBINSTEIN and ARLENE QAZI
Research and Development Division, BioSciences Information Service of Biological Abstracts, 2100 Arch Street, Philadelphia, Pa. 19103

Received February 22, 1971

Chemical parameters in BIOSIS' data base are derived from edited, author-specified nomenclature, necessitating "semantic synthesis" of keywords and fragments analogous to chemical synthesis in the laboratory. To obviate this complex synthesis, a pilot file of chemical toxicants was created from the biological literature for study of alternate techniques for chemical information handling. Using synonym indexing, CAS Registry Numbers, and Wiswesser Line-Notation, Toxitapes, a computer file of general, industrial, and pharmaceutical toxicology, was initiated.

BIOSIS' HANDLING OF CHEMICAL INFORMATION

Chemical references in BIOSIS' file are in a biological context rather than about chemical or physical properties. Therefore, chemical information is usually of the nature of pharmacology, chemotherapy, biochemistry, or toxicology rather than organic, inorganic, or physical chemistry.

We determined the extent of chemical information in our file by computer. BIOSIS maintains on-line the indexing assignments made to all references announced in its publications since late 1959. Using C.R.O.S.S. (Computer Rearrangement Of Subject Specialties) we obtained a count of all items indexed in the categories Toxicology, Pharmacology, Chemotherapy, Pollution, and Pest Control for *Biological Abstracts*, Volumes 45–51 (1964–70). This did not include purely endogenous biochemistry.

Biological Abstracts has increased by 31% from 107,000 to 140,000 articles per volume during this 7-year period. The coverage of chemical information, however, increased by 72% from 19,024 to 32,764 articles per volume. It now comprises 23.4% of *Biological Abstracts* whereas 7 years ago it was only 17.8%.

* Presented at the Middle Atlantic Regional Meeting, ACS, Baltimore, Md., February 5, 1971.

Table I. Evolution of Chemical Nomenclature Editing

Years	*BA* vol =	Edited title
1959	34	DICHLOROPHENOXYACETIC ACID
1960–62	35–41	DICHLORO PHENOXY ACETIC ACID
1963–67	42–48	DICHLOROPHENOXY ACETIC-ACID
1968–71	49–52	DI CHLOROPHENOXY ACETIC-ACID

Table II. Synonyms for Phenethylamine[11]

β-Aminoethylbenzene
1-Amino-2-phenylethane
Ethylamine, 2-phenyl
Phenethylamine
β-Phenethylamine
Phenylethylamine
β-Phenylethylamine
ω-Phenylethylamine
2-Phenylethylamine

Having determined that nearly one quarter of *Biological Abstracts*' documents are concerned with exogenous chemicals and their effects on living organisms, we then examined the interest in this file and the means for retrieving information using chemical parameters.

Chemical Searching at BIOSIS. In the 5 years that BIOSIS Search Service has been in operation,[1] it has performed several thousand search requests for both contract and one-time users. During the year 1970, more than 20% (144/677) of these were oriented toward an exogenous chemical affecting living organisms. In addition, 87% of search requests from participants in the Toxicology Information Program (PHS Contract PH43-68-1329) were directed toward chemical information. Of the 183 search queries submitted by research toxicologists, 52% of all searches involved a single chemical element or compound; while 24% concerned a structural (generic) class as phenols or aldehydes; and 11.5% were oriented toward functional attributes of chemicals (dyes, pesticides).

In view of the number of chemical searches submitted at BIOSIS, we examined the scope and magnitude of chemical parameters in BIOSIS' file.

Vocabulary and File Organization. BIOSIS' file is derived principally from edited, annotated titles of articles, specifically the "keywords" appearing in the indexing position in *B.A.S.I.C.* (*Biological Abstracts Subjects In Context*). This file now represents over 11 years of *Biological Abstracts* and sister publications and therefore exhibits evolutionary changes of philosophy in title editing.[2] This is especially noticeable in chemical nomenclature because authors may specify individual chemicals by trivial or systematic names or may refer to structural class or some functional attribute. An example of the evolution in chemical nomenclature editing is 2,4-dichlorophenoxyacetic acid (2,4-D). The changes in policy can be seen in the file entries for the following years in Table I.

Semantic Synthesis for Computer Chemical Searching. Just as a chemical is synthesized in the laboratory, so the description of the chemical must be "semantically synthesized" in the search strategy. The raw materials for this are the character strings present in the file, which in this case are not preassembled, cross-referenced, or group indexed to any structural or functional class.[3] There has been a consolidation of frequently occurring surrogates that are similar in both spelling and meaning.

Table III. Basic Strategies for the "Semantic Synthesis" of Phenethylamine

	Strategy	Retrieved	Relevant
1.	PHEN(YL) AND ETHYL AND AMINE	14	3
2.	PHEN(YL)ETHYL AND AMINE	11	2
3.	PHEN(YL) AND ETHYLAMINE	28	18
4.	PHEN(YL)ETHYLAMINE	100	24
5.	PHENETHYLAMINE (1, 2, or 3 parameters)	91	18
6.	PHENYLETHYLAMINE (1, 2, or 3 parameters)	64	29
7.	AMINO AND ETHYL AND BENZENE	1	0
8.	AMINO AND PHENYL AND ETHANE	1	0
9.	PHEN PHENYL AND ETHANE AND AMINE BENZENE ETHYL AMINO	176	44
10.	CONSOLIDATED STRATEGIES	180	47

The means for this synthesis is the computer search program. The program currently operating at BIOSIS coordinates character strings from the inverted file using the Boolean operators AND, OR, and NOT. A valuable aid in demonstrating the results of our "semantic synthesis" is our "Strategy consolidation and analysis program,"[4] which enables us to compare the sensitivity (Se), specificity (Sp), and precision (Pr) for multiple strategies.[5]

Sensitivity is the ability to retrieve relevant documents. Se = 0.73 indicates that 73% of all possible relevant documents have been retrieved. Specificity is the ability to suppress nonrelevant documents. Sp = 0.73 indicates that 73% of all nonrelevant documents have not been retrieved. Precision is the ratio of relevant documents to the total number of documents retrieved. Pr = 0.73 signifies that 73% of the retrieved documents were relevant. For the purpose of statistics, we are assuming 100% recall.

We may now examine how BIOSIS search techniques cope with the various nomenclature problems encountered, particularly synonymy, ambiguity, and generic class conceptualization.

Synonymy. The profusion of chemical synonyms in the scientific literature is a result of using trivial names and systematic nomenclature, as Beilstein's *Handbuch der organischen Chemie*, the International Union of Pure and Applied Chemistry (IUPAC), and Chemical Abstracts Service.[6] Many of the trivial names were derived from either Greek or Latin and now serve as the basis for systematic nomenclature. Others are modern day trade names or cryptograms.[7]

We can best illustrate the synonym problem by the semantic synthesis of phenethylamine. Phenethylamine is listed in the "Desk-Top Analysis Tool for the Common Data Base" (DAT)[11] as shown in Table II. As in a chemical synthesis, we must first obtain the reactants. In the laboratory, phenethylamine can be synthesized by the reduction of benzyl cyanide[8] or by the aminoboration of styrene.[9] In the semantic synthesis, the reactants are the character strings in the file which comprise all or part of the chemical rubrics.

The first four strategies for semantic synthesis shown in Table III were formulated from the character strings PHEN, PHENYL, ETHYL, AMINE, and all concatena-

Table IV. Comparison of Sensitivity (Se) for Different "Semantic Syntheses" of Phenethylamine

Search logic	PHEN	PHENYL	Total
1. PHEN(YL) + ETHYL + AMINE	...	0.06	0.06
2. PHEN(YL)ETHYL + AMINE	...	0.04	0.04
3. PHEN(YL) + ETHYLAMINE	...	0.38	0.38
4. PHEN(YL)ETHYLAMINE	0.38	0.13	0.51
ALL STRATEGIES	0.38	0.62	1.00

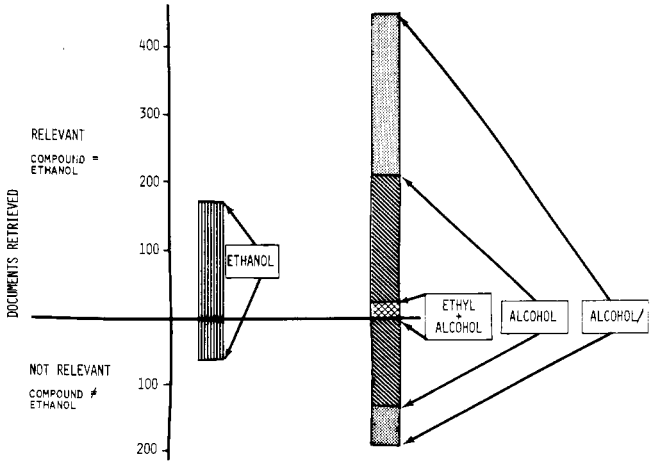

Figure 1. Ambiguity of the rubric ALCOHOL

The sensitivity and specificity for the following strategies are: ALCOHOL/ Se = 0.75, Sp = 0.24; ALCOHOL Se = 0.36, Sp = 0.48; ETHYL + ALCOHOL Se = 0.04, Sp = 0.99; ETHANOL Se = 0.29, Sp = 0.76

tions thereof. (Numeric and Greek letter prefixes are not in the file.) Strategies 5 and 6 demonstrate the difference between using PHEN and PHENYL as starting materials. Because this compound is also listed as β-aminoethylbenzene and 1-amino-2-phenylethane, we used strategies 7 and 8, neither of which yielded any relevant material. Strategy 9 is a composite of ALL character strings used in the 8 other strategies. The consolidated list of 180 references is assumed to be 100% recall for purposes of this type of experiment. A relevant document was defined as one which pertained to phenethylamine and not to one of its derivatives. There were 47 such documents from the 180 documents retrieved (Pr = 0.26).

Note in Table IV the results of 8 semantic syntheses of the compound. PHEN appeared only in the single parameter PHENETHYLAMINE (Se = 0.38). PHENYL (Se = 0.62) was divided among the 4 syntheses as indicated. The preference for PHENYL vs. PHEN directly reflects author preference in naming compounds whereas the number of parameters employed is a function of editorial policy at BIOSIS.

Ambiguity. Many of the nonrelevant documents pertained to substituted phenethylamines, mainly 3,4-dihydroxyphenethylamine (dopamine). This, of course, is an example of ambiguity.

Another case of ambiguity is that of alcohol. A consolidation of several strategies revealed 593 documents relevant to the compound ethyl alcohol. The total number of documents retrieved was 845 (Pr = 0.70). Note in Figure 1 that the surrogate ALCOHOL appeared in the annotated title of 342 documents, 210 of which pertained to ethyl alcohol (Pr = 0.61). The sensitivity was 0.36, or 210 relevant documents out of a possible 593. Of these 210, however, only in 22 did the surrogate ETHYL co-occur. The sensitivity was increased by using the truncated stem ALCOHOL/—i.e., ALCOHOLS, ALCOHOLIC, ALCOHOLISM, etc. This strategy retrieved 638 documents, 446 of them relevant to studies on ethyl alcohol. The sensitivity of this strategy was 0.75, 446 relevant out of the possible 593. Using the surrogate ETHANOL results in a sensitivity of 0.29, with a 4% overlap with the ALCOHOL/ strategy. The specificity, that is, the ability to suppress nonrelevant documents for the surrogates ALCOHOL, ALCOHOL/, and ETHANOL were 0.48, 0.24, and 0.76, respectively.

Generic Chemical Searching. The most difficult type of semantic synthesis is that of the chemical generic. One of these was an actual request, "Toxicology and Metabolism of Hindered Phenols (those used as antioxidants)." Ignoring the nonchemical parameters, toxicology and metabolism, we can consider this query to be both functionally and structurally oriented. Consequently we formulated several strategies based on semantic synthesis. The precision of the over-all synthesis of strategies was 116/427 or 0.27. In Table V, note the first strategy is functional, coordinating the surrogates suggesting antioxidant with PHENOL, PHENOLIC. The latter 2 surrogates have been compiled from other files representing phenols and phenolics and represent over 1400 occurrences in the file.

Only 15% of the 116 relevant articles were retrieved using this strategy. The specificity was also poor, retrieving nearly 50% nonrelevant articles. This was due mainly to the coordination yielding oxidation of phenols by the enzyme phenol oxidase.

Strategy #2 was based on steric hinderance and synonyms describing this phenomenon. This yielded 14 relevant hits (Se = 0.12) and a specificity of (Sp = 0.92) signifying a high rejection of irrelevant material.

Strategy #3 was a structural semantic synthesis using rubrics signifying alkyl and aryl groups and descriptors as ORTHO and TERTIARY. The yield from this synthesis had a sensitivity (Se = 0.15) 17/116 and a relatively high specificity (Sp = 0.70).

These three strategies, based on the semantics of the search request, accounted for only 43/116 documents (Se = 0.37).

Considering that all phenols are not always comprised of the rubric phenol, we formulated a strategy using other phenols coordinated with the alkyl parameter described in strategy #3. Note in Table V that this strategy #4 resulted in four more relevant listings (Se = 0.034). Going one step further, we semantically synthesized the concept of a phenol—i.e., hydroxy and aromatic. In strategy #5 this synthesis, coordinated with the alkyl parameter, yielded 48 relevant articles out of a possible 116 (Se = 0.42). All of these were references to either butylated hydroxy anisole or butylated hydroxy toluene. An additional 17 references were found on these compounds using cryptograms BHA and BHT and commercial names as IONOL or IONOX (Se = 0.15). In all, 54/116 articles (Se = 0.465) specified butylated hydroxy anisole or toluene or their cryptograms, rather than the rubrics, namely phenol, suggested by the inquiry.

Table V. Toxicology and Metabolism of Hindered Phenols
(Phenolic Antioxidants)

Strategy						Retrieved	Relevant	Se	Sp
1. **PHENOLIC ANTIOXIDANTS**						168	17	0.15	0.52
ANTIOXIDANT	OXIDATION		PHENOL						
ANTIOXIDANTS	OXIDATIVE	AND	PHENOLIC						
OXIDANT	OXIDIZE								
OXIDANTS	OXYGENATION								
2. **HINDERED PHENOLS**						40	14	0.12	0.92
CRYPTO	SUBSTITUENT		PHENOL						
HINDERED	SUBSTITUENTS	AND	PHENOLIC						
STERIC	SUBSTITUTED								
STERICALLY									
3. **ALKYL (ARYL) PHENOL**						112	17	0.15	0.70
ALKYL	DIPHENYL		PHENOL						
AMYL	DIVINYL		PHENOLIC						
ARYL	ISOPROPYL								
BUTYL	ORTHO	AND							
BUTYLATED	PHENYL								
DIALKYL	PROPYL								
DIBUTYL	TERT								
DIISO	TERTIARY								
DIISOPROPYL	TRIPHENYL								
CONSOLIDATION, STRATEGIES 1, 2, & 3						306	43	0.37	0.15
4. **ALKYL (ARYL) PHENOLS**						43	4	0.03	0.88
ALKYL	DIPHENYL	AND	CATECHOL						
AMYL	DIVINYL		CRESOL						
ARYL	ISOPROPYL		DIPHENOL						
BUTYL	ORTHO		GUAIACOL						
BUTYLATED	PHENYL		HYDROQUINONE						
DIALKYL	PROPYL		NAPHTHOL						
DIBUTYL	TERT		PHLOROGLUCINOL						
DIISO	TERTIARY		PYROCATECHOL						
DIISOPROPYL	TRIPHENYL		RESORCINOL						
5. **ALKYL (ARYL) HYDROXY AROMATICS**						60	48	0.42	0.96
ALKYL	DIPHENYL	AND	HYDROXY	AND	AROMATIC				
AMYL	DIVINYL		HYDROXYL		ANISOLE				
ARYL	ISOPROPYL				BENZENE				
BUTYL	ORTHO				CUMENE				
BUTYLATED	PHENYL				MESITYLENE				
DIALKYL	PROPYL				NAPHTHALENE				
DIBUTYL	TERT				PHENANTHRENE				
DIISO	TERTIARY				STYRENE				
DIISOPROPYL	TRIPHENYL				TOLUENE				
					XYLENE				
6. **TRADENAMES and CRYPTOGRAMS**						18	17	0.15	0.99
BHA	IONOL								
BHT	IONOX								
CAO									
7. BUTYLATED HYDROXY ANISOLE/TOLUENE						54	54	0.46	1.00
BUTYL	AND	HYDROXY	AND	ANISOLE					
BUTYLATED				TOLUENE					
OR BHA									
BHT									
TOTAL (PRECISION = 0.27)						427	116		

These semantic syntheses were done by trained search personnel, familiar with chemical nomenclature and with a knowledge of the structure and content of the BIOSIS retrospective file. Based on these and similar experiences, we concluded that this detailed knowledge and experience using this file could not be disseminated to potential users of the Toxicology Information File. Consequently, we investigated alternative techniques to searching semantic surrogates of chemical structures.

ALTERNATE TECHNIQUES OF CHEMICAL INFORMATION HANDLING

Among the stated goals of the Toxicology Information Program were coordination of existing chemical information systems to provide toxicologists and related professionals with an integrated toxicological information network. To achieve these goals, National Library of Medicine (NLM) supported Chemical Abstracts Service (CAS)

Table VI. Comparison of Storage Space for Different Chemical Techniques

Index technique	Average length	# Rubrics/ chemical	Space required	Occurrences of chemical
CAS Registry Number	7	1	7	3.91
Single rubric	17	1	17	
Full text (Common data base)	17	1.37	23	3.91
Full-text (Non-common data base)	17	1.125	19	1.47
DAT Synonyms	17	10.45	178	3.91
Wiswesser Line Notation	19	1	19	

in the development of their Registry System and also initiated a file of chemical toxicants in Wiswesser Line Notation (WLN).[10]

Because of these goals AND the fact that preliminary evaluations of our retrospective toxicology information searches indicated a need for alternative chemical information handling techniques, we have studied the input problems and retrieval capabilities as well as the cost and feasibility for CAS Registry Number, Wiswesser Line Notation, and full abstract text input.

Chemical Abstracts Service Registry Number. The "Desk-Top Analysis Tool for the Common Data Base" (DAT)[11] served as the primary source of CAS Registry Numbers. If a chemical was not listed in DAT, we sent all available information on that chemical through NLM to CAS for assignment or identification of a Registry Number. In addition to receiving Registry Numbers, we also served as an input for NLM's Toxicology Information Network.

Wiswesser Line Notation. The majority of notations for chemicals listed in DAT was obtained from NLM. Other notations were encoded in-house, based on availability of structure or systematic nomenclature. We encoded according to the rules in Smith's text "The Wiswesser Line Formula Notation"[12] and subsequent tutorials. Two extremely valuable references for encoding cyclic compounds were "The Ring Index"[13] and "Wiswesser Line-Notation Corresponding to Ring Index Structures."[14] Use of uncertainty codes enabled us to encode certain chemicals for which structural knowledge was either incomplete or uncertain.

Full-Text Abstract Input. In the first phase of the Toxicology Information File, we keyboarded the full abstract and citation for all documents in our experimental file. This provided a means of measuring the effectiveness of any sophisticated information system, because it involved no intellectual decisions. Every character string is retrievable; therefore, this input served as a control for both chemical and biological indexing, allowing study of whether full-text input is practical and effective.

For chemical retrieval using full-text, only the author's nomenclature is available. We did increase this capability by coordinating all text-specified synonyms for a given chemical. For example, if in one abstract a chemical is called 1,1,1-trichloro-2,2-bis(*p*-chlorophenyl)ethane and DDT in a second abstract, either name will retrieve BOTH abstracts.

The Chemical File. The chemical file being described is that of BIOSIS' Toxicology Information File, a subset of the total data base. There are presently more than 5000 unique chemicals in the file. We are reporting, however, only on the 2900 chemicals encountered in the first phase of file building. The source for this pilot file was approximately 5700 toxicology abstracts in *Biological Abstracts*, Volume 49 (1968), and 100 toxicity reports from product bulletins supplied by Dow Chemical Company and Eastman Kodak Company.

File Dependence. Of the 2900 chemicals in this file, 1500 were in DAT. Therefore, CAS Registry Number, Wiswesser Line Notation, and a full array of synonyms were available for nearly 1300 of these. Approximately 100 of these were mixtures or polymers having MX or PM numbers for which synonyms were available but no WLN.

For nearly 1000 chemicals not in DAT, we were able to obtain 947 Registry Numbers. Due to the turn-around time, only 279 of these were received in time to be added to the file. No additional synonyms were supplied to the file except as appeared in the text or title of any abstract.

Computer Storage. Note in Table VI the comparison in computer storage space for these alternate techniques. From the group of 1500 chemicals in DAT, we found that they occurred in our file an average of 3.9 times. Full-text input provided an average of 1.37 rubrics per chemical. Full synonym input from DAT would have resulted in an average of 10.45 rubrics per chemical. Each chemical in the second group, those not in DAT, occurred an average of 1.47 times. Full-text input provided only one additional rubric for every 8 that uniquely occurred. MX and PM numbers provided 1.25 rubrics per chemical and occurred an average of 3.44 times.

The average and median length of all rubrics is 17 and 14 characters respectively as seen in Figure 2.

CAS Registry Numbers were most economical from the point of storage space, occupying an average of 7 character spaces and occurring only once per chemical.

Wiswesser Line Notation had a median and average length of 13 and 19 characters, respectively. Note in Figure 3 that the inorganic notations have slightly lower median and average than the organic notations. The longer notations represent both complex molecules and

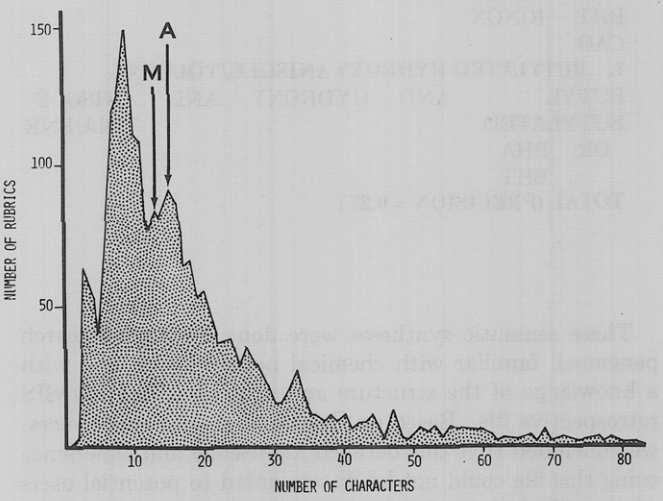

Figure 2. Distribution of length of chemical rubrics
The median (M) and average (A) lengths are 14 and 17 characters, respectively.

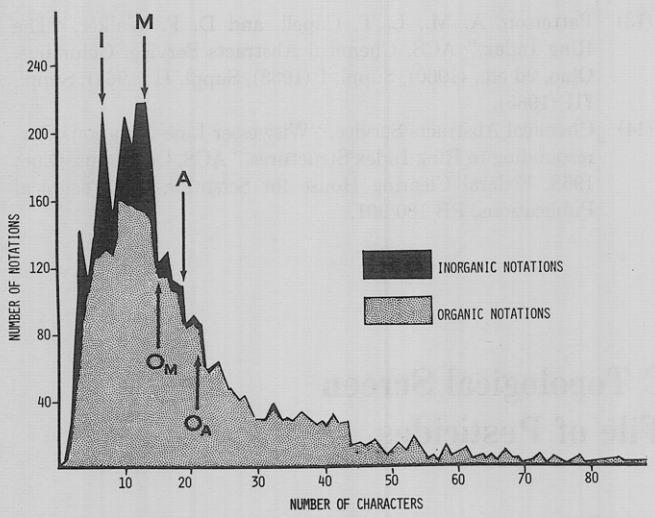

Figure 3. Distribution of length of Wiswesser Line Notation
The Median (M) and average (A) lengths for all notations are 13 and 19 characters, respectively. The median (O_M) and average (O_A) for organic notations are 15 and 21. Inorganic notations (I) have a median and average length of 10 characters

MANTRAP codes. In phase I, 80% of the file was encoded in WLN.

The important question is whether or not the retrieval capability of a particular technique is proportionate to the amount of storage space and the intellectual and clerical effort in building the file.

CONCLUSIONS

Semantic synthesis is necessary for maximum retrieval in formulating search startegies using an uncontrolled vocabulary as a source of chemical surrogates.

We have experimented with inputting three techniques to reduce the need for semantic synthesis. The retrieval capability of the file prepared for the first phase of this study is currently being evaluated by participants in the Toxicology Information Program. We have observed some characteristics of indexing and comprehensiveness of each technique.

Dependence on outside organizations delays file building. CAS Registry Numbers, requiring an average of 7 characters per chemical, are most economical of computer storage space. However, CAS Registry Numbers were available for only 62% of the total file.

Full-text input, because it involves no intellectual decision, is the most economical input method from the point of view of indexer time. It provided 13% of the total synonyms of DAT. Whether or not these are the most valuable or most frequently used rubrics will be revealed in future tests. Further addition of synonyms requires constant reference to additional compendia, intellectual decisions as to which synonyms to use, and is least economical in terms of computer storage space.

Use of Wiswesser Line Notation requires that the user be familiar with its rules and syntax. However, it requires less storage space than full-text rubric input. Approximately 80% of the chemicals in the file have been encoded. Of the remainder complex mixtures and heteropolymers comprised 10% and for the remaining 10%, no structural information was available. WLN requires the effort of a trained indexer, but, due to its inherent nature, does provide information for substructure searching.

FUTURE PLANS

Based on these observations and our experience, we have begun the next phase of study; preparation of a file called Toxitapes. It consists of citations from *Biological Abstracts*, Volumes 51 and 52 (1970–1) in general, industrial, and pharmaceutical toxicology. In this phase the chemical indexing is being accomplished by CAS Registry Number, Wiswesser Line Notation, all chemical rubrics in text, and complete synonym indexing from DAT. We are presently inviting all interested parties to be participants in the evaluation of these alternative chemical indexing techniques.

ACKNOWLEDGMENTS

We would like to acknowledge the contributions, advice, and encouragement from David W. Fassett, and his staff at the Eastman Kodak Company, Laboratory of Industrial Medicine, and from Mark A. Wolf and his colleagues at the Dow Chemical Company, Biochemical Research Laboratory; Al Weissberg and Charles Rice; and the present management of the Toxicology Information Program at the National Library of Medicine, H. Kissman and B. Vasta.

LITERATURE CITED

(1) Schultz, L., "Problems of Retrospective Searching," presented at the meeting of the Association of Scientific Information Dissemination Centers, Rochester, N. Y., September 24, 1970.

(2) Parkins, P. V., "Approaches to Vocabulary Management in Permuted-Title Indexing of Biological Abstracts," in "Automation and Scientific Communication," pp 27–28, H. P. Luhn, Ed., American Documentation Institute, Washington, D. C., 1963.

(3) Bernier, C. L., "Correlative Indexes IV. Correlative Chemical-Group Indexes," *Amer. Doc.*, 8, 306 (1957).

(4) Junkins, K., and L. Schultz, "An Alternate Strategy to Iterative Searching," *Proc. Amer. Soc. Info. Sci.* 7, 323 (1970).

(5) Saracevic, T., G. Baumanis, M. Bobka, E. Brown, I. Hazelton, L. Rothenberg, J. B. Subramaniam, C. Zull, and A. J. Goldwyn, Comparative Systems Laboratory Final Technical Report, An Inquiry into Testing of Information Retrieval Systems, Part I: Objectives, Methodology, Design, and Controls, and Part II: Analysis of Results. Center for Documentation and Communications Research, School of Library Science, Case Western Reserve University, Cleveland, Ohio, 1968.

(6) Verkade, P. E., "Organic Chemical Nomenclature, Past, Present, and Future," in "Chemical Nomenclature," Advances in Chemistry Series, 8, pp 75–82, ACS, Washington, D. C., 1953.

(7) Jensen, K. A., "Problems of an International Chemical Nomenclature," in "Chemical Nomenclature," Advances in Chemistry Series, 8, pp 38–48, ACS, Washington, D. C., 1953.

(8) The Merck Index, P. G. Stecher, Ed., Merck & Co., Inc., Rahway, N. J., 8th ed. (1968).

(9) Redeuilh, G., and C. Viel, "Hydroxyboration and Aminoboration of Styrenes," *C. R. Acad. Sci.*, Paris, Ser. C., **267** (26) 1858 (1968).

(10) Rice, C. N., "Toward a National Systems Resource in Toxicology," *J. Chem. Doc.*, **9**, 181 (1969).
(11) Chemical Abstracts Service, "Desk-Top Analysis Tool for the Common Data Base," ACS, Columbus, Ohio, 1968, Federal Clearing House for Scientific and Technical Publications, PB 179 900.
(12) Smith, E. J., "The Wiswesser Line-Formula Chemical Notation," McGraw-Hill, New York, N. Y., 1968.
(13) Patterson, A. M., L. T. Capell, and D. F. Walker, "The Ring Index," ACS, Chemical Abstracts Service, Columbus, Ohio, 2d ed., (1960), Suppl. I (1963), Suppl. II (1964), Suppl. III (1965).
(14) Chemical Abstracts Service, "Wiswesser Line-Notations Corresponding to Ring Index Structures," ACS, Columbus, Ohio, 1968, Federal Clearing House for Scientific and Technical Publications, PB 180 901.

Application of the MCC Topological Screen System to a Small File of Pesticides

SAMUEL T. MORNEWECK*
Esso Research and Engineering Co., Linden, N. J. 07036

BRUCE G. HAWTHORNE
Esso Mathematics and Systems, Inc., Florham Park, N. J. 07932

Received August 19, 1970

A file of 8600 compounds, which were tested for pesticide activity, was manually coded and processed by computer to give substructure indexes using the MCC Topological Screen System developed at the University of Pennsylvania. Information is presented on coding rates and coding errors, computer processing times, and substructure searching experience. The indexes are found to be relatively inexpensive to produce and to provide considerable substructure searching capabilities.

The nature of pesticides research requires the ability to search chemical files by substructure. It is frequently necessary to assemble all the compounds containing a particular substructure as a first step in preparing structure-activity relations. On the other hand, the need to recall a particular compound is rather infrequent and can usually be met by intersecting two or three substructure lists if the substructure strings are reasonably long or by maintaining a separate empirical formula index.

A number of manually generated fragment coding systems are known, but an extensive experiment with such a system was unsatisfactory. The system used was developed at Esso and was based on the A.P.I. Thesaurus of Chemical Aspects. The vocabulary was substantially expanded, and the definitions of roles and links were modified to indicate connections between groups of atoms. Problems of thesaurus maintenance and the expense of coding, both of which would be common to all manually generated fragment codes, required that another system be found.

The following criteria were established for a substructure search system:

Reasonably long, descriptive fragments
Minimal cost for coding and processing
Printed indexes so that computer searching was not required for each request.

These criteria seemed to be fulfilled in a system that was being developed by Lefkovitz and coworkers at the University of Pennsylvania, the MCC Topological Screen System (TSS).[1,2] This system uses a nonunique line notation for input, which promised even easier coding than the popular Wiswesser Line Notation, and generates printed indexes of reasonably long fragments, which appear to be easy to search.

With the agreement of the sponsors, the application of this system on a file of 8600 compounds was undertaken. The compounds were intermediates and final products, which were synthesized in a pesticide research program, and selected compounds from other company laboratories, which were tested for pesticidal acitivity. Thus, the file had very few small molecules (<10 non-H atoms) and very few large, polycyclic molecules.

MCC CODING

Compounds were coded from the data sheets submitted by the chemists using the Mechanical Chemical Code (MCC) described in reference 1. The MCC uses standard atomic symbols except for elements that normally have hydrogen attached. Thus

is a, $-CH_2-$ is b, $-CH_3$ is c, $-NH-$ is M, $-NH_2$ is Z, $-OH$ is Q. Certain other contractions are used for common groupings, i.e., $>C=O$ is L, $-NO_2$ is NX,

* To whom inquiries should be addressed; present address, Department of Chemistry, St. Peter's College, Jersey City, N. J. 07306.