nonnegative integers each. These table entries describe the structure of a molecule or compound in gradually increasing detail, such that the complete CCT is interpreted as a hierarchy of subtables, linked together implicity by substituent counts and locant values.

For organic substances, most carbon atoms and their common bonds are represented implicitly, by the coding of complete rings and chains in single table entries, thus making the CCT particularly concise. Other atoms and bonds may be represented explicity, but provided that for some cases, particularly inorganics, further appropriate values are defined for bond modification entries.

It is widely recognized that the IUPAC nomenclature is nonunique, and this quality is often decried in relation to the needs of information storage and retrieval activities. However, it should not be overlooked that chemical nomenclatures are used in a wide range of human circumstances other than (indeed, perhaps more often than) in pure research and its related fields. For example, in general trade and commerce, together with its regulatory legislation, uniqueness is of a lower priority than the ever-essential nonambiguity, and particularly familiarity and, hence, relative ease of use over other structural representations are most important. It is here that the versatility of the IUPAC nomenclature is of great benefit and equally that the ability to verify and convert IUPAC names to other forms is of potential use. In this respect the CCT acts as an intermediary stage in a computer-transformation process just as do all connection tables in their own systems.

The CCT as so far defined has been used successfully to code the structural information contained within a variety of IUPAC nomenclatural terms, and complete tables have been generated algorithmically from complete names. Such tables, as exemplified throughout the text, are self-evidently concise in comparison with conventional atom-by-atom connectivity tables for structures of any great size, through the CCT representation of many atoms and bonds by a single table entry.

The expansion of the CCT to a full atom-by-atom connection table has been briefly investigated and shown to be feasible, thus indicating a potential route from the IUPAC nomenclature to many existing information systems based on such connection tables. The CCT is also suitable for further processing for the display of structural diagrams. Algorithms have been developed on the basis of a character-cell technique[8] to display unscaled line segment representations of rings and chains, given the corresponding CCT entries.

REFERENCES AND NOTES

(1) Rayner, J. D. "Grammar Based Translation by Computer of the IUPAC Systematic Chemical Nomenclature". Ph.D. Thesis, University of Hull, Hull, U.K., 1983.
(2) International Union of Pure and Applied Chemistry "Nomenclature of Organic Chemistry," 1979 ed.; Pergamon Press: Oxford, U.K., 1979.
(3) International Union of Pure and Applied Chemistry "Nomenclature of Inorganic Chemistry," 2nd ed.; Butterworths: London, U.K., 1971.
(4) Ash, J. E. In "Chemical Information Systems"; Ash, J. E.; Hyde, E., Eds.; Ellis Horwood: Chichester, U.K., 1975; Chapter 11, pp 156-176.
(5) Smith, E. G.; Baker, P. A. "The Wiswesser Line Formula Chemical Notation," 3rd ed.; Chemical Information Management Inc.: Cherry Hill, NJ, 1975.
(6) See pp 7 and 380 of reference 2.
(7) See pp 20-27 of reference 2.
(8) Rayner, J. D.; Milward, S.; Kirby, G. H. "A Character Set for Molecular Structure Display" *J. Mol. Graphics* **1983**, *1* (4), 107-110.
(9) The unusual spelling of "TIPE" arises historically through the use of the Pascal programming language for the coding of algorithms. The more preferable term "TYPE" is a reserved word in that language and is therefore unavailable as a field identifier.

Comparative Efficiency of Searching Abstract Text in the Chemical Abstracts Service Database†

M. HERZ, H. K. KAINDL, A. A. SALIB, and R. WARSZAWSKI

BASIC, Basel Information Center for Chemistry (Documentation Center of Ciba-Geigy Ltd., F. Hoffmann-La Roche & Co. Ltd., and Sandoz Ltd.), CH-4002 Basel, Switzerland

Received September 5, 1984

Text retrievals were carried out by utilization of Chemical-Biological Activities (CBAC) and Polymer Science & Technology (POST) tapes, which contained in computer-readable form all *Chemical Abstracts* data element fields. The subjects of the queries were randomly selected, and a natural language vocabulary was used for the text profiles. An evaluation of the contributions of all data elements to the retrievals showed that when searchable abstract text was added to a combination of the remaining fields, the recall was greatly improved and the probability of retrieving concepts not contained in the indexes was increased.

INTRODUCTION

Our investigation dates back to 1976 and was based on searches of Chemical-Biological Activities (CBAC) and Polymer Science & Technology (POST) tapes, two services available at that time containing all *Chemical Abstracts* (CA) data element fields in computer-readable form, including the abstract text.[1] As a result of improvements in its editorial process, Chemical Abstracts Service (CAS) began producing the full abstract text from all 80 CA sections in computer-readable form in 1975.[2] Prior to that time, the abstracts were only available in computer-readable form in CBAC and POST. CAS subsequently included abstract text in searches for some topics in its *CA Selects* current awarenesss service[3] and optionally for all searches in the *Individual Search Service* (ISS) batch current awareness service.[4] Although abstract text is the most voluminous data element in the CA text file (e.g., in 1976 CA text comprised about 21% of the complete database), we thought having the capability to search it would greatly improve the recall for certain queries. In an earlier study, Barker et al.[5] compared the search performance of the abstract text with that of the other free-text data elements (keywords and titles), which at that time were accessible in computer-readable services from CAS, and found that abstract text greatly improved recall. In addition, several studies have been devoted to searching the abstract text in other data bases.[6]

† Address correspondence to Dr. M. Schellenbaum, BASIC, Ciba-Geigy Ltd., CH-4002 Basel, Switzerland.

0095-2338/85/1625-0111$01.50/0 © 1985 American Chemical Society

Our research consisted of conducting text retrievals to find and select the most efficient combination of data elements appropriate for an in-house batch system. We had established that a controlled vocabulary, i.e., index entries, was indispensable for retrievals in which the target query is a chemical substance in conjuction with text. Our next goal was to evaluate the efficiency of CA data elements when the queries were formulated in a natural language vocabulary—the method a chemist might use in constructing a search profile. Chemical professionals, especially laboratory workers, are quite often infrequent users of online search systems and have limited search experience. They are likely to favor a natural language vocabulary when formulating queries. This is especially true of the nonchemist (the biologist, for example), who most probably is not familiar with CAS indexing but who may want to search CA (a) for subjects that are not covered in the index, or are covered in the index by other terms, and (b) with a strategy already utilized in other databases.

With the widespread online availability of CA CONDENSATES, CASIA, and later CA SEARCH, our plans to develop our own text search system were subsequently abandoned. However, we view the results of this study with renewed interest now that CAS intends to make the currently printable abstract text also searchable in its CAS ONLINE service and in light of the current research being conducted into the capabilities of full-text searching of ACS primary journals.[7,8]

METHOD

Prior to beginning our study, we reformated the CBAC and POST tapes of Volume 82 of CA (January through June 1975), which included all data element fields, for searching in our batch system. The latter offered, among other features, the advantage of left- and right-hand truncation of terms and made possible searching of all nomenclature fields (i.e., Heading Parent, Substituent, Name and Text Modification) of the *Chemical Substance Index*, as well as all fields of the *General Subject Index* (Concept Heading, Functional Category, etc.). We selected 10 queries at random and prepared 10 corresponding search profiles, eight for CBAC and two for POST (see list of queries in Table V), using a natural language vocabulary. Since the subjects of the queries were also randomly selected, neither they nor the vocabulary used in the profiles was intentionally adjusted to conform to CA index vocabulary. However, if such conformance was coincidental, no effort was made to avoid it. For example, if a commonly used chemical term such as "benzodiazepine" happened to agree with the CA index nomenclature or with a term within a General Subject index entry, it was not removed. Naturally, no CAS Registry Numbers were used.

For a proper assessment of their retrieval efficiencies, the searches were run separately in the following *individual* data element fields: (a) Title and Keyword Phrases (considering the content of the CA CONDENSATES tapes, Title and Keyword Phrases have been treated as a single data element field)—TK; (b) Index (Chemical Substance and General Subject fields)—I; (c) Abstract—A. Thus, no hits could result from the intersection (Boolean combination) of terms from different fields. Relevance with respect to a query was judged in the entire context of a retrieved record. A "relevant" or "irrelevant" reference was defined as such in all data element fields of such a record. The total of all relevant references for a given query was defined as that retrieved by searching all fields.

EXAMPLES

The query "benzodiazepines in anesthesiology", one of the 10 chosen for our study, can be used to illustrate our procedure and the results obtained. Figure 1 lists the search strategies we developed for this query. The terms within each of the Search Statements 1 and 2 were combined using the Boolean "OR". In Search Statement 3, the search results of both previous statements were combined by using "AND". Table I illustrates the results of this search.

```
Search strategy:

Search Statement 1    :BENZODIAZEPIN:          :EPAM:
                      :AZEPATE:                :EPAN:
                      :DIAZEPOXIDE:            :OLAM:

Search Statement 2    :ANESTHE:                :SADDLE BLOCK:
                      :PREMEDIC:               :SADDLE-BLOCK:
                      :PREOPERAT:              :LYTIC COCKTAIL:
                      :OPERATION:              :LYTIC-COCKTAIL:
                      :SURGER:                 :HALOTHAN:
                      :SURGIC:                 :THIOPENT:
                      :DENT:                   :THIAMYLAL:
                      :ODONT:                  :METHOHEXITAL:
                      :PAIN:                   :METHITURAL:
                      :NEUROLEPTANALG:         :FLURAN:
                      :NEUROMUSCULAR BLOCK:    :NITROUS OXIDE:
                                               :N2O:

Search Statement 3    Search Statement 1 AND Search Statement 2

: = truncation
```

Figure 1. Search strategies for the query "benzodiazepines in anesthesiology".

Table I. Results of Searching the Query "Benzodiazepines in Anesthesiology"[a]

CAN[b]	TK	I	A
82:000241	+		+
82:025830	+		
82:038611	+		
82:068285	+	+	+
82:080584	+ (F)	+ (F)	+ (F)
82:144982	+	+	+
82:038651		+ (F)	
82:093143		+	+
82:106282		+	
82:164862		+	+
82:011204			+
82:051497			+
82:051638			+ (F)
82:064536			+
82:093286			+ (F)
82:175185			+ (F)

[a] (+) = all hits; (F) = false drops. [b] CAN = Chemical Abstracts Number (abstract number).

Table II. Retrieval Efficiencies according to Individual Data Element Groups

	TK + I + A	TK + I	A	contribution of A exclusively
relevant references	11	8	8	3
false drops	5	2	4	3
total hits	16	10	12	6
precision (%)	68.8	80.0	66.7	50.0
recall (%)	100.0	72.7	72.7	27.3

The number of hits (retrieved relevant references and false drops) for groups of data elements and for those unique references retrieved from searching the abstract exclusively are summarized in Table II. This table also shows the retrieval efficiencies for groups of data elements and particularly for the abstract text. For purposes of this paper, we have defined precision and recall to be the following:

$$\text{precision} = \frac{\text{number of relevant references retrieved} \times 100}{\text{total number of references retrieved}}$$

$$\text{recall} = \frac{\text{number of relevant references retrieved} \times 100}{\text{total of relevant references retrieved}}$$

Column "TK + I + A" shows the results (relevant references, false drops, total hits) from all data element fields searched. Column "TK + I" shows the results from searches of the following two groups: (1) titles and keywords and (2) index entries (recall that no hits could result from the inter-

Table III. Results of Searching the Query "LD50 of Compounds Containing Chlorine"[a]

CAN	TK	I	A	CAN	TK	I	A
82:026809	+		+	82:081422			+
82:026834	+			82:081430			+
82:052359	+			82:081431			+
82:052366	+		+	82:081502			+ (F)
82:081391	+			82:092888			+
82:119681	+			82:093137			+ (F)
82:011275			+	82:093916			+
82:011949			+	82:093933			+
82:011957			+	82:093959			+ (F)
82:011958			+	82:094195			+
82:025653			+	82:106145			+ (F)
82:026048			+ (F)	82:106282			+ (F)
82:026706			+	82:106487			+
82:038624			+	82:106521			+
82:038740			+	82:107149			+
82:039241			+	82:107170			+
82:039601			+ (F)	82:107184			+
82:051346			+	82:107344			+
82:051586			+	82:119481			+
82:051673			+	82:119799			+
82:052302			+	82:120106			+
82:052349			+	82:132776			+
82:052377			+	82:132796			+
82:052582			+ (F)	82:133647			+
82:068177			+ (F)	82:133668			+
82:068250			+	82:133997			+
82:068948			+	82:147184			+
82:069004			+	82:149236			+ (F)
82:069173			+	82:149333			+
82:080343			+ (F)	82:149997			+
82:080352			+	82:164612			+
82:080592			+ (F)	82:164778			+
82:080660			+ (F)	82:165605			+
82:081390			+	82:165836			+ (F)
82:081421			+	82:165858			+

[a] (+) = all hits; (F) = false drops.

Table IV. Retrieval Efficiencies according to Individual Data Element Groups

	TK + I + A	TK + I	A	contribution of A exclusively
relevant references	56	6	52	50
false drops	14	0	14	14
total hits	70	6	66	64
precision (%)	80.0	100.0	78.8	78.1
recall (%)	100.0	10.7	92.9	89.3

section of a keyword term with an index term in our experiment). These data elements are available in the CA SEARCH file, accessible through several vendors (DIALOG, SDC, DATA-STAR, BRS, etc.).

Column "contribution of A exclusively" contains references that were found by searching the abstract text *only* and were not retrieved by searching the remaining data element fields. As shown in Table II, we retrieved three additional unique relevant references (27.3% of all relevant references) by searching the abstract text exclusively.

```
Search Statement 1    :CHLOR:

Search Statement 2    :LD50:          :TL50:
                      :LD 50:         :TL 50:
                      :DL50:          :LETHAL TOX:
                      :DL 50:         :LETHAL DOS:
                                      :ACUTE TOX:

Search Statement 3    Search Statement 1 AND Search Statement 2
```

Figure 2. Search strategies for the query "LD50 of compounds containing chlorine".

The query "LD50 or acute toxicity of compounds containing chlorine", another of the 10 studied, further illustrates our method. We searched this query in the same manner as the previous example with the same three data element groups, TK, I, and A. Figure 2 gives the search strategies for this query. Table III illustrates the results of this search.

Table IV summarizes the total number of hits retrieved by searching the three data element groups with the LD50 query. This table also shows the retrieval efficiencies of the individual data element groups, their precision, and recall.

Searching the abstract text with the LD50 query exclusively contributed 50 out of a total of 56 relevant references retrieved, or 89.3%. This compares to only three unique hits out of a total of 11 relevant references, 27.3%, when the abstract text was searched for the benzodiazepine query. The LD50 query yielded only 10.7% of the relevant references when it was searched against the data element groups TK and I. This clearly demonstrates that, in certain cases, abstract text retrieval can substantially increase the total number of relevant retrievals and the resulting recall.

CONCLUSIONS

By comparing the search result efficiencies of all 10 queries (Table V), we see that information obtained from searching abstract text exclusively added an average of 50.2% more relevant references to those already obtained by searching the remaining combined data elements (TK + I). Abstract text retrieval contributed 80.6% of all relevant references. Thus, from this study it would follow that if searchable abstract text were added to the presently accessible files in free-text retrievals with randomly selected queries, the recall would be nearly doubled. On the other hand, the precision would be lower.

In view of the very small number of queries involved, some of them have contributed much more to these overall statistics than others; if a large number of queries were used, the "average" recall might change considerably.

The usefulness of abstract text searching seems to be dependent on the nature of the query and on the selectivity of

Table V. Comparison of Hits and Efficiency of Search Results for All 10 Queries according to Individual Data Elements[a]

query subjects	relevant hits through TK + I + A	relevant hits through TK + I	relevant hits through A	relevant hits through A exclusively
fixation of dyes (P)	45 (77)[b]	21 (27)[b]	39 (70)[b]	24 (50)[b]
homopolystyrene and carbon black (P)	10 (24)	5 (7)	9 (21)	5 (17)
pyrimidines as growth hormones (C)	7 (12)	5 (10)	3 (3)	2 (2)
terpenes as perfumes (C)	19 (22)	10 (10)	16 (19)	9 (12)
vaccines (1 parameter) (C)	47 (47)	34 (34)	32 (32)	13 (13)
benzodiazepines in anesthesiology C)	11 (16)	8 (10)	8 (12)	3 (6)
LD50 or acute toxicity of barbiturates (C)	2 (11)	0 (0)	2 (11)	2 (11)
LD50 or acute toxicity of compounds containing chlorine (C)	56 (70)	6 (6)	52 (66)	50 (64)
acetanilides as herbicides (C)	32 (35)	24 (26)	25 (28)	8 (9)
photosynthesis of plants (C)	50 (107)	26 (32)	39 (90)	24 (75)
total	279 (421)	139 (162)	225 (352)	140 (259)
precision (%)	66.3	85.8	63.9	54.1
recall (%)	100.0	49.8	80.6	50.2

[a] (P) = POST; (C) = CBAC. [b] Total hits (relevant references + false drops) are indicated in parentheses.

the vocabulary used. Queries containing terms seldom, if ever, found in the indexes, for example, biological data and numerical parameters like "LD50", are particularly suited. In some other cases abstract text searching might be of great disadvantage. Especially in those where the topics are already well covered in the indexes, any possible advantages of such an additional search must be weighed against the loss of precision incurred. This would be of special importance if, instead of the small file we used, the *entire* CA text file (all years and all sections) were to be searched.

It would be inaccurate to compare the results we obtained in searching the combination of fields TK + I with those which might be obtained with CA SEARCH (alone and in conjunction with the abstract). We have searched and retrieved hits within the separate fields only and did not retrieve references if terms satisfying the profile were in different fields. CA SEARCH in its current online usage includes all index entry terms, titles, and keywords in the Basic Index, so such references would probably be retrieved.

We have illustrated that a very simple free-text search technique is sufficient to achieve a substantially increased recall, to provide the nonspecialist with a higher retrieval capability, and, possibly, to attract additional user circles to access the CA text file online.

ACKNOWLEDGMENT

We thank CAS staff for their continued technical support and interest in the results of this study and, in particular, Patricia S. Wilson as well as others who, in addition to reviewing and correcting the draft of this paper, have contributed many valuable suggestions. We also thank the management of BASIC for their encouragement and for providing resources and especially H. Kniess from the Ciba-Geigy Information Center for fruitful discussions and valuable assistance in preparing this paper.

REFERENCES AND NOTES

(1) "Information Tools"; Chemical Abstracts Service: Columbus, OH, 1976. CBAC is presently accessible online via the National Library of Medicine (NLM) System as a component of TOXLINE.
(2) Buntrock, R. E. "Searching Chemical Abstracts vs. CA Condensates". *J. Chem. Inf. Comput. Sci.* **1975**, *15*, 174.
(3) Blake, J. E.; Mathias, V. J.; Patton, J. "*CA Selects*—A Specialized Current Awareness Service". *J. Chem. Inf. Comput. Sci.* **1978**, *18*, 187. Blake, J. E.; Ebe, T. "Abstract Text Searching for *CA Selects*". Presented at the 2nd Chemical Congress of the North American Continent, 180th National Meeting of the American Chemical Society, Las Vegas, Nevada, 26 Aug 1980.
(4) "CAS ONLINE & Search Services Catalog"; Chemical Abstracts Service: Columbus, OH, 1984; p 13.
(5) Barker, F. H.; Veal, D. C.; Wyatt, B. K. "Comparative Efficiency of Searching Titles, Abstracts, and Index Terms in a Free-Text Data Base". *J. Doc.* **1972**, *28*, 22.
(6) Wagers, R. "Effective Searching in Database Abstracts". *Online* **1983**, *7* (5), 60.
(7) Durkin, K.; Egeland, J.; Garson, L. R.; Terrant, S. W. "An Experiment to Study the Online Use of a Full-Text Primary Journal Database". Presented at the 4th International Online Information Meeting, 9-11 Dec 1980, London, England.
(8) Cohen, S. M.; Schermer, C. A.; Garson, L. R. "Experimental Program for Online Access to ACS Primary Documents". *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 247.

An Algorithm for Chemical Superstructure Searching

PETER WILLETT

Department of Information Studies, University of Sheffield, Western Bank, Sheffield S10 2TN, United Kingdom

Received October 29, 1984

Chemical superstructure searching involves the identification of those molecules that are contained within a given query structure. This paper presents an algorithm for carrying out such searches that may involve fewer accesses to backing storage than a previously described algorithm.

SUBSTRUCTURE AND SUPERSTRUCTURE SEARCH

An important facility in computer-based chemical information systems is the ability to carry out chemical *substructure* searching.[1] Given a query structure, Q, and a set of N molecules $\{M_j\}$ $(1 \leq j \leq N)$, a substructure search results in the identification of those molecules that contain Q as a substructure. Substructure searching is a special case of the more general subgraph isomorphism problem that involves determining whether one graph is a subgraph of another; this problem has been studied extensively and is known to be NP complete.[2] Because of this, substructure search systems[3,4] operate in two stages, with a simple and rapid initial search mechanism being used to eliminate the great majority of the molecules in the file; only those few compounds that pass this initial screening search then undergo the computationally demanding atom-by-atom substructure search.

The screening search is effected by defining a set of features, called screens, that are used to characterize the molecules in the file and the query structure Q; a wide range of substructural features may be used for this purpose, as is illustrated by the screening mechanisms used in the CAS ONLINE system.[5] For some molecule, M_j, to be a *possible hit* in the substructure search, i.e., one that needs to be processed by the atom-by-atom algorithm, it must contain all of the screens that have been assigned to Q: such screens will be referred to here as *query screens*. The screening search may be performed efficiently by setting up an inverted file that contains a series of lists, one for each of the possible screens, with the ith list containing the identifiers of those molecules to which the ith screen has been assigned. The intersection of the lists corresponding to the query screens results in a list containing the identifiers for all of the possible hits that must be processed in the atom-by-atom search.

Wipke and Rogers[6] have recently discussed the inverse of chemical substructure searching, which they call *superstructure* searching. A superstructure search results in the identification of those molecules that are substructures of Q, a search facility that is of importance in computer-aided synthesis design programs. Efficiency of operation is again achieved by the use of an initial screening search based on an inverted file; however, the inverted file is used in a quite different manner from that employed for substructure search. Whereas the latter involves taking the intersection of the query screen lists to identify possible hits, superstructure searching involves taking the union of the lists corresponding to the screens that have not been assigned to Q to identify all of the *definite nonhits*: the atom-by-atom superstructure search is thus restricted to those molecules not eliminated by the

0095-2338/85/1625-0114$01.50/0 © 1985 American Chemical Society