for heats of formation derived from utilization of ($\mathbf{v} \times \mathbf{A}_{10}$) indices and Frazer and Herndon *n*-iterative indices, respectively. Methane excluded, the ($\mathbf{v} \times \mathbf{A}_{10}$) index values displayed a linear relationship described by the equation:

$$(-1)\Delta H_f^\circ(g)\ \mathrm{kJ/mol} = 31.084 + 60.512\ (\log\ [\mathbf{v} \times \mathbf{A}_{10}]\ \text{index})$$

[with an r^2 value of 0.983 and F test equal to 2136. The Frazer *n*-iterative (F) index values displayed a linear relationship described by the equation:

$$(-1)\Delta H_f^\circ(g)\ \mathrm{kJ/mol} = 52.948 + 36.4276\ (\log\ [n - (\mathrm{A}_2)]\ \text{F index})$$

[with an r^2 value of 0.992 and F test equal to 4869, methane excluded. The best correlation by a slight margin was exhibited by the Herndon *n*-iterative (H) index values with a linear relationship described by the equation:

$$(-1)\Delta H_f^\circ(g)\ \mathrm{kJ/mol} = 44.749 + 36.7775\ (\log\ [n - (\mathrm{A}_2)]\ \text{H index})$$

[with an r^2 value of 0.993 and F test equal to 5341, methane excluded. The above relatively simple one-parameter modes of estimating heats of formation compare reasonably well with the seven structural parameters utilized by Kalb et al.[11] to approximate such alkane heats of formation to within ±1.5 kJ/mol.

CONCLUSION

Transformation of the base-2 adjacency matrix of an alkane to the base-10 vector enabled a series of matrix computations to be performed that led to several kinds of topological indices for each alkane. One of these, the index calculated by summing the vector elements of the product of the degree vector and the base-10 adjacency vector, gave an index that appeared to be unique, single-sum, and with other indices formed a monotonic series. The above indices were not invariant; hence, a canonical numbering system was used for each graph. There was additionally derived from the iterative procedures that served as the source of the canonical graph numbers a series of invariant molecular topological indices, which appeared to be unique, single-sum, and constituted monotonic series of values.

REFERENCES AND NOTES

(1) Paper 2 of this series: Schultz, H. P.; Schultz, E. B.; Schultz, T. P. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 27-29.

(2) Rouvray, D. H. In *Chemical Applications of Topology and Graph Theory*; King, R. B., Ed.; Elsevier: Amsterdam, 1983; pp 159-177.

(3) Cahn, R. S.; Ingold, C.; Prelog, V. *Angew. Chem., Int. Ed. Engl.* **1965**, 385-415.

(4) Frazer, R. A.; Duncan, W. J.; Collar, A. R. *Elementary Matrices*; Macmillan: New York, 1946; p 142.

(5) Herndon, W. C. In *Chemical Applications of Topology and Graph Theory*; King, R. B., Ed.; Elsevier: Amsterdam, 1983; pp 231-242.

(6) Hansen, P. J.; Jurs, P. C. *J. Chem. Educ.* **1988**, *65*, 574-580.

(7) Trinajstić, N. *Chemical Graph Theory*; CRC: Boca Raton, FL, 1983; Vol. 1, pp 31-41.

(8) Randić, M. *J. Chem. Phys.* **1974**, *60*, 3920-3928; **1975**, *62*, 309-310. Randić, M. *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 171-180.

(9) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*; Academic: New York, 1976; p 32.

(10) Bertz, S. H. In *Chemical Applications of Topology and Graph Theory*; King, R. B., Ed.; Elsevier: Amsterdam, 1983; p 208.

(11) Kalb, A. J.; Chung, A. L. H.; Allen, T. L. *J. Am. Chem. Soc.* **1966**, *88*, 2938-2942.

Chemical Abstracts Online: A Study of the Quality of Controlled Terms

SABINE MARTIN and GÜNTER BERGERHOFF*

Institute for Inorganic Chemistry, University of Bonn, Gerhard-Domagk-Strasse 1, D-5300 Bonn 1, FRG

Received August 27, 1990

Controlled vocabulary is said to be the key to facts covered in the vast amount of literature. An analysis of a certain fraction of CAS controlled vocabulary from the CA File shows a number of errors and inconsistencies. This leads to conclusions for the retrieval process, and more use of the basic index is recommended until the controlled vocabulary has been revised.

INTRODUCTION

CAS ONLINE, being the largest database in chemistry worldwide, demands a high degree of familiarity with its rules of indexing. This leads to unsatisfactory results, especially when used by the bench chemist. As the chemist himself is most familiar with his research subject, the classical documentalist expects that the scientist himself would search the literature for the solution to his problems. Therefore, modern documentation systems should be conceived such that their undeniable advantages in rapidity and flexibility are not offset by overly complicated user instructions.

The search for chemical compounds is a problem of general importance in chemical databases. It has been dealt with in various ways. Whether it is the introduction of registry numbers (the concise and unambiguous representation of substances in most cases), the description of structural characteristics by means of easily-machine-readable topological codes, or the possibility of graphical entry of the structures in question—the effort always aims to make the search as easy and successful as possible (see STN Express, MOLKICK).

But CAS ONLINE users have to cope with difficulties in fact-searching. The problems may be of a formal nature (typographic errors, parameters for headings, length of terms, use of special characters, admissibility of abbreviations) or of a logical nature (homonyms, synonyms, combination of terms, etc.). Right from the beginning CA has provided a controlled vocabulary (CV) for assistance especially in this type of search. The CV applies also to the database version. However, the CV used from the 9th Collective Index (9CI) on, beginning with 1972, differs somewhat from that for the 8CI (1967-1971) and differs clearly from that in the CA File.

The controlled vocabulary is said to be the key to the solution of the problems cited above. Looking for possible improvements we began with an examination of the CV in the

0095-2338/91/1631-0147$02.50/0 © 1991 American Chemical Society

CA File. All entries in the controlled vocabulary beginning with an "A" were analyzed. The letter "A" covers the largest amount of entries relating to a single letter, covering about one-twelfth of the word-material. It can therefore be assumed representative for the vocabulary as a whole.

The presentation of the faults and insufficiencies which we found will be followed by reflections on how to improve the situation. A comprehensive study including a detailed and commented list of index terms is available.[1]

RESULTS OF EXAMINATION

Amount of Vocabulary. The CAS ONLINE searcher was first confronted with the following obstacle when calling the relevant portion of the CV index. According to the printed index of all valid index terms (*CA Headings List of 1985*[2]*—HL in the following*), approximately 6450 terms beginning with "A" were to be expected (85% of which are names of living organisms). However, when calling the online index on April 22–23, 1988, the first index term beginning with "B" appeared only after 11 244 entries! The following attempts of explanation left aside, this fact is problematic for two reasons:

> The search of such a voluminous online index is unreasonably expensive (online time!); the search cannot be accelerated because there is no other access to specific terms than through the alphabetical structure. This fact cancels the advantage of a controlled vocabulary which is the increased predictability of terms used and to be used. Why not try your chance with basic index terms right from the start? This would avoid learning (useless) indexing rules.
>
> There is nothing more annoying to a cost-conscious searcher who has neatly prepared his database call with tools available offline than to find that his efforts have been in vain or that his confidence in the correspondence between online and offline indexes is rewarded with a high rate of missed references.

Origins of the Disparity between Online and Offline Vocabulary. A first survey confirmed that the enormous disparity between the offline and online versions are not due to a planned increase of the vocabulary since the publishing of HL in 1985. Its origin lies rather in a multitude of faulty entries and in terms which were deliberately not included in the printed vocabulary since they were valid for 8CI only.

After elimination of all names of living organisms, the exact comparison of the remaining 2554 terms from the online index with the printed index produced the following results:

841 terms (32.9%) occur in both indexes
1713 terms (67.1%) occur only in the online index

Terms occurring only in the online index can be subdivided according to their provenance:

161 (9.4%) misspelled terms (in the printed index one single mistype was found—that term does not appear on the online list)

117 (6.8%) terms differing in number (singular/plural) from HL

75 (4.4%) names of single substances (CA offers registry numbers for single substances. Hence there is no need to include them in the controlled vocabulary. This is not the case for classes of compounds, which is a special problem.)

298 (17.4%) terms followed by subdivisions. According to the introduction to the HL, this procedure dating from the early days of the CA printed edition was used to subdivide too voluminous entries. But there is no hint at the fact that in the database these subdivisions are included in the CV index in contrast to the printed edition. The searcher who uses the original index terms for his query will often lose large numbers of hits. Because the CV terms are bound phrases, the same loss of information must be blamed when one tries to replace the full descriptor by combining the two parts by proximity operators. Only the entry of the entire enlarged term is successful or truncation of the term by a question mark.

1045 (61%) terms which are impossible to find in the HL for various reasons (e.g., index terms from superseded vocabularies; personal names; enzyme-class denominations, which generally do not occur in the HL).

Obviously many of the terms listed in the HL correspond to extraordinarily few documents:

42.5% refer to one document only
67.3% refer to 1–9 documents

Even if there is no formal condition such as a minimum of documents corresponding to an index term, it appears strange to include these rarely used terms in the controlled vocabulary. This leads us to the assumption that many of these terms were used as descriptors without any control. This does not apply to the names of living organisms. In this group of terms 51.8% refer to one document only; 87.7% to 1–49 documents; and 97% to 1–99 documents due to the high specificity of binary/ternary names used in that field.

Faulty Terms. The fault rate of 10.9% in the whole online index corresponds to the average of other examinations.[3] If numerical faults are not taken into consideration, it lies even significantly below: 6.3%. Nevertheless this fault rate is a considerable obstacle to many experienced users: The controlled vocabulary is just the place where you would not expect any fault (remember that the examined portion of the printed HL is fault-free with one exception). Every tenth term is lost by error. In the examined portion this leads to a loss of 1–3756 documents per faulty term (13.241 documents in total). Although many of them correspond to terms which are faulty only as to their number (singular/plural), they must be considered as faulty terms because their very existence is only rarely revealed in the HL or the index guide (IG in the following[4]). One hundred sixty-one terms corresponding to 243 documents suffer from misspelling, wrong punctuation, or straight nonsense.

Abbreviations. Things are similar in the case of abbreviated terms: HL and IG are full of references from abbreviations to (permitted) terms. For instance, when advised "At. nuclei—use Atomic nuclei (9.11.CI)", who shall deduce from this instruction that in the online search both terms must be used in order to (hopefully) obtain all the relevant information?

The examined material includes 27 (1.6%) terms abbreviated according to CA rules and marked with a point. The rules are to be found in the leaflet: Standard Abbreviations, Acronyms, Special Characters and Symbols—224 documents correspond to these terms. We also found 30 (1.8%) entries which were simply cut for overlength—this manner of "shortening" is not mentioned in any manual, and without prior consultation of the online CV list, it is impossible to find the corrupted entries.

Length of Terms. Abbreviations are a consequence of overlength of index terms. The increase of term length has presumably even worse consequences such as the danger of decreasing the predictability of the current vocabulary and the further complication of query formulas. Compared with the overall mean of 19 characters in the index of textbooks in chemistry,[5,6] the statistics shown in Table I occurred.

The high proportion of long terms can be interpreted as a sign of quality since a term's specifity is generally linked to

Table I

terms for	short terms (≤19 chars)	long terms (>19 chars)
classes of compds	58.9%	41.1%
facts	27.3%	72.2%
organisms	63.8%	36.2%

Table II

terms for	single-part terms	multi-part terms
classes of compds	31.1%	68.9%
facts	22.8%	77.2%
organisms	14.9%	85.1%

its length. Indexers and searchers are given a quite dense grid of terms. But particularly since these long terms do correspond to very few documents, the negative aspects cited above become predominant. It is impossible to dream up index terms like "Air, conditioning or purification of" or "Alkali- and water resistant, anticorrosive coatings" without some inspiration from either an online or offline index, or even to match the exact punctuation and grammar! Given that two-thirds of the index terms actually in use are not listed in the printed index, queries relying on the controlled vocabulary do hardly make any sense, unless you accept the loss of an unknown amount of information or you do not mind the costs and the effort of searching the thousands of terms for just that entry you might be looking for. In this regard the names of living organisms are an honorable exception: their biosystematic formulations provide a high degree of predictability.

Precombinations. Long terms do not usually occur because a single word is too long; they are composites of several words or word fragments (see Table II). Precombination leads to two-part terms in most cases, but many considerably more complex descriptors exist, mostly resulting from the use of conjunctions in term construction. A multitude of precombined terms undergo further permutation, which explains the high amount of commas, by far the most frequent special character (see below). Whether it is advantageous or even necessary to the database version of *Chemical Abstracts* to construct terms in this way shall not be discussed in this paper. The point is that the procedure is not applied consistently. Among terms permuted in many different ways you also find combined terms in a normal sequence. The following examples (terms with number of hits) reflect the mixture of irregular term modeling.

1	Absorption-dementalation petroleum refining catalysts
1	Absorption-dementalation petroleum refining
49	Activation energy of electron exchange reactions
42 421	Air pollution
2 075	Air, pollution of
1	Air, polution of
1	Alcohols, c11-14-secondary
1	Alcohols, c16 and c18-unsatd.
1	Alcs., polyols
1 295	Alkaline earth compounds
1	Alk. earth compds.
1	Arom. polyamide-polyanhydride-polyesters
440	Ash
13 447	Ashes
12 421	Ashes (residues)
1	Ashes (residues), residues
2	Ashes, es
15 954	Atmosphere
1	Atmosphere environmental
1	Atmosphere, enviromental
3 457	Atmosphere, environmental
10 064	Atomic nuclei
1	At. nuclei

Table III

special character	terms for classes of compds	terms for facts	names of organisms
comma	432	321	
hyphen	199	33	52
bracket	22	14	25
period	44	29	3
apostrophe	1	7	

Number (Singular/Plural). There is another aspect to term construction where precise rules exist. It is thoroughly explained in Appendix III of the IG when plural index terms have to be introduced into the index. Therefore, one should not think that deviation from the HL number-version is responsible for most errors found in the online index. In some cases a term occurs in the singular only and not in the plural or vice versa. Then there is a chance that the user reading the message "0 hits" becomes suspicious and tries the other number-version. But there are lots of terms which occur in both singular and plural. Then there is a great danger of loss of information because the user is not aware of the system's deficiency. The reasons for the discrepancy between the printed and the database index lie in the revisions of the vocabulary establishing new rules for number attribution. Number rules are confusing, not only for searchers but also for indexers. This explains why many rare terms are attributed the incorrect number (e.g., why are denominations of tools, materials, and other objects plural but not those of animals, plants, and their parts? Who knows that one is supposed to choose the plural form although the singular would be correct when dealing with "terms which appeared as Latin plurals in previous indexes"?).

Special Characters. The situation is also confusing in the domain of special characters, which account for 16.2% of all faults found in the examined material. The fact that the exact sequence of the terms must be obeyed when searching with controlled vocabulary increases the problem of special characters: more rules to learn, more mistakes to make. The problems for indexers and searchers increase automatically with the amount of special characters. The fault rate illustrates that indexers cannot really handle the set of special characters, and the searcher's problems also result from the fact that neither IG nor database manuals provide clear information on the use of special characters. An additional handicap for the searcher is that, in particular, descriptors with special characters do not appear in the HL. Table III shows the special characters that appear in the examined material: (Number of terms in which the special character occurs at least once. Hence a single term can be accounted for several times.)

The use of special characters is closely related to other modes of term construction. Commas replace conjunctions in enumerations or separate parts of permuted terms. Brackets embrace (permitted) terms specifying homonyms and others. Hyphens often express a close relationship between two of several partial terms. Eventual modifications of the use of special characters in the controlled vocabulary therefore are linked to the discussion of other modes of term construction and the fundamental question of whether extensive specification expressed in terms carrying special characters is necessary at all.

Number of Documents per Term. In a large database it is important to compare the amount of controlled vocabulary and the number of documents attributed to single index terms. The denser the available set of index terms, the more precisely the content of a document can be described by single index terms, and hence the number of unwanted hits in the answer sets is reduced. But it also increases the vocabulary the searcher must be familiar with in order to find all his target documents. Given the volume of almost 10 million documents, in many domains the temptation may be great to always think up new index terms. As the closer examination below shows, the result is a hardly manageable "controlled vocabulary".

Even in the printed index 15% of terms correspond to a single document and almost 19% correspond to 2–9 documents, hence every third index term admitted refers to less than 10 documents (10 documents represent about 0.000001% of all documents!). For the online index these parameters rise to 43% and 25%, respectively; hence two out of three of these terms were used for the indexing of less than 10 documents! Vice versa, among the descriptors in use online and offline only 21.3% refer to more than 1000 documents; in the online index this parameter is even as low as 6.6%.

The following groups of terms are remarkable when considering the vocabulary in detail:

Terms referring to names correspond to 1.25 documents on average.

Among the terms referring to institutions only two correspond to more than 5 documents.

Although the general descriptor "Awards" exists, there are many extremely special denominations of medals and awards which refer on average to 1.4 documents only.

The case is similar with the group of "activation energy of" terms: beside the general term "activation energy" there are almost 300 more specific terms, a fourth of which refer to a single document and half of which refer to less than 10 documents.

On the other hand there are many groups in the medical domain, for instance, or the denominations of apparatuses and tools, which refer to 1000 documents on average.

Many terms formed with "Standard Qualifiers" are found corresponding to high numbers of documents. Terms with the addition "preparation", "reactions", "analysis", and "uses and miscellaneous" correspond to 1000 documents on average. Terms with the addition "properties" reach even 1800. (A comparatively inhomogeneous picture is found for the terms for classes of compounds; we do not develop this further in the present paper.)

Theoretically a large number of index terms seem to permit a highly specified description of the document's content. But it is not practical. Having lost the control over the available vocabulary the indexer tends to index rather generally when there is a doubt, or he is tempted to create new specificiations. This leads to a concentration of the mass of documents under a relatively small range of index terms and to a large range of index terms with fairly low occurrence. This situation leads to a first-class obstacle for the searcher: When searching a relatively specific question he must choose the search term from an extremely wide range of terms. There is a high risk that the searcher uses a different reasoning when choosing a term than the indexer did when indexing that document. Paradoxically this is true when searching with the very exact term defining the subject of interest. In practice, documents are indexed rather generally. On the other hand, when searching a more general subject, a multitude of very specific terms must be introduced in order to cover the whole domain. The possibility to assemble automatically special terms for a more general search does not exist.

Synonyms and Quasi-Synonyms. A fundamental cause for the multitude of terms referring to extremely few documents lies in the details of natural speech as well as of many specialized terminologies which offer various expressions for the very same thing (or for very similar things). When the database producer disregards strict control of the vocabulary, the user may find his preferred term in the vocabulary, but of course this concession to user-friendliness results in a vocabulary of unmanageable dimensions without a discernible structure and saturated with descriptors of the same or only slightly different meanings.

The following types of synonyms are predominant:

abbreviation/plain version
("Alk. earth compds."/"Alkaline earth compounds")
different versions of the same personal name
("Alikhanov"/"Alikhanov, Abram Isaakovich")
varieties of spelling
("Alnico VS55"/"Alnico VS 55")

Quasi-synonyms occur as terms of very similar or overlapping meaning:

("Animal growth substances"/"Animal growth regulators").

IG and HL cannot help because they sometimes lack references to existing similar terms. Frequently IG and HL indicate supposedly invalid terms and their preferable terms, but in fact the "invalid" terms do exist. These are causes (in addition to the more fundamental causes described above) for loss of information and/or increased search costs.

Grid of Terms. The most obvious cause for the enormous amount of vocabulary and for the multitude of terms corresponding to extremely few documents is the use of highly specialized terms. These are used even when the broader term corresponds to few documents in turn.

Thirty or more narrower terms for one broader term are not an exception. Top of the chart is the heading "Activation energy" (3313 documents) with 297 more specified terms "Activation energy of ...", followed by "Alkali metals, compounds" (2367 documents) with 121 narrower terms and "Aminotransferases" (2621 documents) with 119 narrower terms.

On the other hand there are many terms without any subdivision referring to considerable numbers of documents (e.g., "Algorithm" refers to 6248 documents, "Animal cell" refers to 18 085 documents). There are even cases where terms referring to a fair number of documents are joined to another term, creating a new term (e.g., "Appetite depressants and antiobesity agents" referring to 250 documents, the parts of the term referring to 485 and 41, respectively).

As a result, the grid of terms is very inhomogeneous. In places it is extremely dense, elsewhere it is astonishingly wide. For this reason the predictability of the success of a query relying on basic index terms is the same as when relying on the controlled vocabulary without prior consultation of the online index.

Organization of the Vocabulary. A clearer response to the question "what information can I get from CA using the controlled vocabulary?" is given by a further examination of the vocabulary as to its content. This was done partially by examination of typical queries and partially by classification of the material considered.

Of the whole material, 8960 terms are names of living organisms. The high proportion of bio terms in the vocabulary of a database in chemistry may be surprising. Terms for substances, fitting better into a chemistry database, account for 13.2%, and only 8.6% of the vocabulary is dedicated to facts. The 2554 terms remaining after elimination of the bio terms can be classified as follows.

58.2%	terms for classes of compounds
38.1%	terms for facts
3.7%	names of persons, institutions, awards

Among the examined terms for facts (=100%) we found

40.2%	(natural) phenomena
28.1%	reactions/procedures

16.1% medical terms
6.1% properties/characteristics
3.7% apparatuses
2.8% constants/measures
1.9% specific domains
0.9% theories/models/(natural-)laws

Given the fairly high proportion of terms for substances, one can say that the vocabulary is very substance oriented. This is underlined by the fact that the groups established above include a very high proportion of substance-oriented "Standard Qualifiers". Fifty percent of terms of the group "reaction/procedures" and 42.5% of the group "properties/characters" become substance oriented by standard qualifiers. Leaving aside the terms "activation energy of" (which account for more than 90%), more than one-third of the terms of the group "phenomena" are substance oriented. Hence the number of terms available for fact search is even smaller than the statistics above might suggest. This means that in your search strategy you should choose to ask for substance-oriented formulations rather than for facts. For instance, when you are interested in "Abrasiveness" (1 document) you also want to check "Abrasives" (2749 documents).

The general difficulty of defining a subject and expressing it in terms convenient for fact searching and the small number of terms actually available for facts make a practical vocabulary very desirable. This vocabulary should be subdivided in term categories, define the possible relationships between terms, and give a precise definition of the term in use. The very first attempts have been made with the CA Hierarchies—only available offline so far—and the database NUMERIGUIDE designed as a guide to the numeric databases of STN.

RECOMMENDATIONS FOR REFORM

Of course a fundamental reform of the controlled vocabulary is desirable for several reasons. Just think of the revolution in the area of chemical information in the last years. In particular the ever-growing importance of databases and the central role CA could play therein owing to the registry numbers for substances already in use and especially owing to a new and appropriate vocabulary yet to come. But many of the faults and insufficiencies described above could be cured immediately without taking such a drastic step as a fundamental reform.

Organization of the Index. By grouping bio terms (three-fourths of the whole material) in a separate online index, the survey of available index terms would be rendered much clearer. This has already been done for the printed index which is published in two volumes.

As a measure of further improvement the elimination of terms for classes of compounds should be taken into consideration. The file REGISTRY offers the tools needed for substance-oriented search anyway.

Correction of Errors. A first and efficient step to reduce the amount of entries in the online CV index would be the elimination of misspellings. This appears to be pretty easy using the sophisticated error-detection tools which already exist according to Zamora.[7]

The generation of complete identity between the online and offline index demands also the purge of all invalid terms from the online list as well as reconsidering correct terms dating from before the 9CI period when establishing a new HL.

Minimum Amount of Documents per Term. In practice 100 hits can easily be checked for relevance by the display command. Hence, let us set the condition that terms included in the vocabulary refer to at least 100 documents. This means the elimination of 50–80% of all terms. The relevant documents have to be attributed to more general terms, most of which already exist. For instance, the term "Alpha chi sigma award in chemical engineering" gives way to "Awards" and "Acyltransferases, *O*-1-alkenylglycero-3-phosphorylcho" gives way to "Acyltransferases".

Surely this measure diminishes the (very theoretical) precision of retrieval. It is nevertheless justified by the fact that it is the easiest way of assuring a complete set of responses covering all documents of interest. In order to cope with the problems described above, in particular the minimal number of documents condition is appropriate for fact terms. It would also be convenient for terms for classes of compounds as long as rules for nomenclature are not reliable. But the condition should not be applied to bioterms because their unambiguous rules of nomenclature provide a high predictability, even in the long term.

Length of Terms. Having implemented the above recommendations, a condition as to the length of terms becomes superfluous because most long terms would have been eliminated anyway. Abbreviations due to the lack of space will vanish as well, except for the bio index. In order to facilitate the entry of these terms, the possibility could be considered to permit biosystematic codes like those used in the reference journal "biological abstracts"/"BIOSIS" (analogous to registry numbers for substances). Given the already established cooperation between the producers of CAS ONLINE and BIOSIS (database BIOCAS) this, in principle, should be possible.

An uncomplicated way of reducing the length of certain terms for facts and classes of compounds could be the suppression of conjunctions. This would also eliminate several subdivisions such as "Disease or disorder" or "Uses and miscellaneous". Such a convention seems reasonable also because it would avoid the need for masking the terms which the system would otherwise recognize as Boolean operators. The resulting partial revision of the vocabulary could lead to the modification of terms (e.g., by using general terms: "Nonbiological effects" instead of "Chemical and physical effects") or to the splitting of terms into parts which can be searched jointly if the user wishes to do so. Often these terms already exist in the vocabulary.

A general rejection of the principle of precombination is not recommended because it is the only way to express complex facts with a precision sufficient for a database of the dimension of CAS ONLINE. Problems occur only when the accumulation of increasingly precombined index terms is supposed to create an ever increasingly dense grid of terms. Instead, a manageable and controlled vocabulary should be combined according to an appropriate and defined grammar (see Fugmann[8]).

Special Characters. The use of special characters in precombined descriptors could be restrained although they are legitimate or even necessary, such as the comma in permuted terms and the hyphen in respect to English spelling and chemical nomenclature. This applies especially to the period. Its importance is heavily reduced when avoiding abbreviations and simply marking Greek letters with an additional period. In the examined portion periods remained only in overly specified terms which could easily be modified. The same is true for the apostrophe, which mainly occurs in English genitive constructions. In order to reduce the legal set of special characters even further, and hence to make it handier, brackets should be abolished. They sometimes have been used to mark explanatory items—the latter could also be abolished.

Number (Singular/Plural). More clarity is desirable in regard to the number-version of terms. Certainly the distinction between singular and plural adds certain information to the term. But surely the searcher prefers a comprehensive answer when looking for the appropriate search term? The most practical solution would be to adopt the convention used by encyclopedias, namely to always use the singular term. The plural may be used for terms for classes of compounds.

Synonyms and Quasi-Synonyms. In order to make the handling of the controlled vocabulary easier, the modification of the treatment of synonyms and quasi-synonyms is recommended. Since we dispose of tools like the IG which refers from invalid terms to terms included in the CV, all synonyms except the preferential term could be declared invalid and listed in the IG. Ideally one could think of including the invalid terms in the database with a device which automatically changes all synonyms into the corresponding preferential terms.

RECOMMENDATIONS TO CAS ONLINE USERS

The controlled vocabulary being available only in the form described previously the user must adopt retrieval habits which can cope with the faults and insufficiencies inherent in the system.

Very keen users spending hours searching for the exact term which ought to describe the document they want to find should bear in mind that the indexer spends only a fraction of this time on the same question. Hence he cannot always find the most appropriate terms in the mass of terms of CV. So instead of a single precise term use a set of more or less precise ones. The IG with its list of nonvalid and superseded terms can be a helpful source of inspiration.

When very complete answers are to be obtained, the online CV index should be consulted before retrieval, besides IG and HL. Use the expand command to get a list of CV terms around your choice and transfer the appropriate E-numbers to the search command. It does not necessarily drastically prolong the online time, but you will get more certainty.

When scrolling the online index, be aware that permutations of the term you have in mind may exist. Interesting index terms may therefore occur at various places on the list.

In any case you should be prepared for errors. Depending on where in the term the misspelling occurred, you would have to look at more or less extensive portions of the list to rediscover the accidentally modified term. The most striking example in the examined material was "acd solutions", to be found 429 entries distant from "acid solutions".

When supposedly appropriate descriptors lead to no or unsatisfactory results, and other possible terms cannot be found in the CV, an additional search in the basic index should be undertaken. If you have at least one hit, display the index terms (IT) or the sample format which displays title (TI), keywords (KW), and IT in order to get some inspiration for further descriptors.

OUTLOOK

"The use of controlled vocabulary is essential to comprehensive and precise results".[9] Given the great importance of CV as expressed in this statement, it is depressing and hardly understandable that this potentially helpful tool is in fact creating obstacles to efficient retrieval, many of which the producers of CA could remove without much effort. The user, resigned to his fate, must try to overcome this. He worries about unreliable results and increasing costs. Since 1988 fees depend on the number and type of searched terms. Not only is the customer offered an inadequate service, he also is asked to pay for his effort to cope with it.[10]

The enormous amount of information produced in the domain of chemistry cannot be handled without an efficient database system. Therefore the obvious insufficiencies of the CAS ONLINE system do not only concern database producers conscious of their reputation and the demand for their product. Improvements in the domain of CAS ONLINE are of interest for everybody working in chemistry and related fields.

REFERENCES

(1) Martin, S. Untersuchung des kontrollierten Vokabulars in Chemical Abstrats Online. Ein Beitrag zur Gestaltung eines idealen Informationssystems in der Chemie. Dissertation, Bonn, 1990.

(2) Chemical Abstracts Service. *CA Headings List, General Subjects. CA Heading List, Plants and Animals.* CAS: Columbus, 1985.

(3) Bourne, C. P. Frequency and impact of spelling errors in bibliographic data bases. *Inf. Process. Manage.* **1977**, *13* (1), 1.

(4) Chemical Abstracts Service. *Chemical Abstracts, Index Guide 1982–1986*; CAS: Columbus, 1987.

(5) Hendrickson, J. B.; Cram, D. J.; Hammond, G. S. *Organic chemistry*, 3rd ed.; McGraw Hill: New York, 1970.

(6) Durrant, P. J.; Durrant, B. *Introduction to advanced inorganic chemistry*, 2nd ed.; Longmans: London, 1970.

(7) Zamora, A. Control of spelling errors in large databases. In *The information age in perspective*, Proceedings of the 41st ASIS Annual Meeting, New York, Nov 13–17, 1978; Brenner, E. H., Ed.; White Plains, 1978.

(8) Fugmann, R. Zur Frage der Vereinheitlichung Indexierens. *Nachr. Dok.* **1970**, *29*, 121.

(9) Zass, E. DIALOG and ORBIT IV: Ein praxisorientierter Vergleich zweier Datenbanksysteme am Beispiel von CA SEARCH ("Chemical Abstracts Service"). *Nachr. Dok.* **1982**, *33*, 129.

(10) Müller-Lorentz, M. Heilmittel oder Gift? *Cogito* **1988**, *4*, 30.