Analysis of Terminology in Various CAS Data Files as Access Points for Retrieval†

MARTHA E. WILLIAMS
Information Retrieval Research Laboratory, University of Illinois, Urbana, Illinois 61801

Received November 22, 1976

An analysis was made of the CASIA data base, Condensate data base, and the keyword indexes in the printed issues of CA. Individual terms and phrases (excluding chemical nomenclature) were compared to determine which source file and data elements or mixtures of data elements from more than one source would provide maximum retrieval capability. The percent overlap for terms in titles, keywords, keyword phrases, CASIA concept headings, and CASIA text modification are given.

INTRODUCTION

Since the advent of computer searching of machine-readable data bases, there has been controversy as to which data elements are most satisfactory for retrieval purposes. Are titles more useful than keywords? Are index terms more useful than keywords? There is no single answer to such questions. The usefulness of such elements, as access points, varies considerably from subject area to subject area and from data base to data base. It is highly dependent on the precision of the terminology in the subject area and on the depth and quality (accuracy and consistency) of indexing and keywording that is provided by the data-base generator. And, in the case of titles, there is seldom any vocabulary control exercised; we are subject to the individuality and whim of authors.

Most studies that have evaluated access points, with respect to their utility for retrieval, have done so by providing various levels and types of search results to panels of users who evaluate the retrieved items as being relevant, nonrelevant, or of marginal interest to their areas of interest or search queries. Additionally, when possible, recall values may be determined by manually searching the source set of documents and comparing the results of the two methods of searching to determine the percentage of the relevant set in the data base that is retrieved by the computer search. Such evaluative procedures are necessarily somewhat subjective as they are dependent on the judgments of human evaluators and people are not completely consistent in their judgments over time. The human evaluator represents a variable that is impossible to control completely in evaluative schemes.

The analysis described in this paper is an attempt to evaluate various types of data elements as access points for retrieval. It is an analysis of data content only and therefore is objective. No search queries were used, no searches were conducted, and no search output was generated for user evaluation. I have looked at only the vocabulary content of files. Five different data elements were analyzed. They are found in three different data bases all of which represent one basic file, namely *Chemical Abstracts* (CA).

All elements were broken down into single terms; e.g., titles were broken down into the single terms that comprise them. Similarly, keyword phrases from *CA Condensates* tapes, Keyword Index (KI) phrases from CA issues, and concept headings and text modification phrases from CASIA (*Chemical Abstracts Subject Index Alert*) tapes were reduced to their single term components. The objective was not to evaluate terms in context but to determine the amount of content information or information-bearing vocabulary that was present in each of the five data elements in order to assess the retrieval capability of each element.

Comparisons were made of the subject vocabulary found in various elements of two machine-readable and one hard-copy of the Chemical Abstracts Service (CAS) generated retrieval tools. The three were *CA Condensates*, CASIA, and the Keyword Index (KI) phrases found in the Keyword (phrase) Index portion of the hardcopy *Chemical Abstracts*. Selected elements from these sources were analyzed and compared. The purpose was to evaluate the single term retrieval capability of individual resources as well as individual elements found within a given resource. The term comparisons resulted in the production of data regarding term occurrences, term redundancy or duplication of terms within an element, overlap of terms between elements, and nonoverlap of terms between elements.

Every attempt was made to treat files as if they were to be searched by computer on a keyword or single term basis. Where the word "term" is used it refers to a unique string of characters in a data element associated with a bibliographic citation or reference.

The problem of the occurrence of variant forms of terms was considered both inclusively and exclusively so that judgments regarding retrieval capability could be made either with or without the use of truncation. Variant forms of terms are considered to be those containing the same stem or left string of characters. The problem of synonyms or acronyms was not treated. Thus, the two terms "nicotinic" and "nicotinate" were treated both as variants and as individual terms. On the other hand, the terms "monoamine oxidase" and "MAO" were treated as individual terms only.

The following elements were included:

(1) *Title terms* excluding stop words (the titles found in the hard-copy *Chemical Abstracts* and on the *CA Condensates* tapes are identical)

(2) *Keywords* found in keyword phrases on *CA Condensates* tapes

(3) *Keyword Index* (KI) terms found as initial terms in phrases in the Keyword Indexes of the hard-copy *Chemical Abstracts*. Only initial terms in phrases were handled because it is only through initial terms that one can manually locate term phrases in the alphabetically sorted Keyword Index (KI) phrases. One should bear in mind that although most terms in phrases do occur as initial terms, not all terms do; therefore, those that are not initial terms cannot be searched directly. For example, when a phrase of five terms is specified for inclusion in the Keyword Index, that phrase may appear in only three of its five possible linear permutations

(4) *Concept Headings* (CH) found on the CASIA tapes

(5) *Text Modifications* (TM) found on the CASIA tapes

Data elements dealing with chemical nomenclature were omitted from the analysis on the assumption that eventually the use of CAS Registry Numbers and molecular formulas would provide adequate access points for retrieving references related to specific chemicals. While specific chemical nomenclature data elements were omitted from the study,

† This paper represents work carried out on a consulting basis for the Information Sciences section of IIT Research Institute under NSF Grant No. GN-40772. Presented before the Division of Chemical Information, 170th National Meeting of the American Chemical Society, Chicago, Ill., Aug 1975.

Table I. The Type:Token Ratio for Condensates Keywords in the Keyword Phrases

Type:Token	1:2.6
Distribution	2-5
2 occurrences	60.3%
3 occurrences	21.9%
4 occurrences	11.0%
5 occurrences	6.8%

Table II. Variant Forms of Terms in Titles, KI's, and Keywords

Percentage of all unique terms, in titles and KI's combined, that are found in variant forms	7.5%
Percentage of title terms found in variant forms in KI's	9.7%
Percentage of KI terms found in variant forms in titles	17.8%
Percentage of unique terms, found in titles and Condensates keywords combined, that occur in variant forms	8.0%
Percentage of title terms found in variant forms in keywords	12.0%
Percentage of keywords found in variant forms in titles	13.9%

chemical nomenclature terms occurring in *other* data elements, e.g., titles, do appear. The study was restricted to the "concept" type of terminology which is often less specific and therefore more problematic.

Note that the overlap between *CA Condensates* keywords and CBAC (*Chemical-Biological Activities*) keywords is 100%; therefore, any findings for Condensates keywords can be applied directly to CBAC keywords. Naturally, titles associated with the same reference are the same for all sources. The initial terms found in the Keyword Indexes of *Chemical Abstracts* are a subset of terms derived from *CA Condensates* keyword phrases. Thus, not all of the keywords in *CA Condensates* tapes appear as initial terms in the Keyword Indexes in the hard-copy *Chemical Abstracts*.

In all cases the strings that comprise titles, keyword phrases, KI phrases, concept headings, and text modifications were broken into individual terms; thus the analysis is based on individual terms rather than terms in context.

The test file consisted of 50 references selected at random from Volume 78. A second sample of 50 more references was randomly selected from the same issue and was used for spot checking and verification of the first sample. Because the analysis was done manually, the sample was quite small. However, the data should provide a rough indication of the trends and utility of the various data elements in the various data bases, both independently and in relation to each other. It should also provide a basis or model for a more extensive analysis.

DISCUSSION

Duplicate Terms Found in Condensates Keywords. Wherever the expression "unique term" is used, it refers to a unique string of characters or a "type". Multiple occurrences or "token" data within data elements were removed for all analyses except where specifically indicated, as in this case. Token data or duplications of the same term within an element of a record are relatively frequent (Table I).

This phenomenon is to be expected because of the permuted phrase selection procedure used by CAS in assigning keyword phrases. Most tape processors remove the redundancy when inverting files or reprocessing files, but some leave the phrases intact in order to provide "context" or adjacency searching.

Variant Forms of Terms. In looking at terms found in titles, keywords, and KI's, one finds terms that contain common left strings; these are called variant forms of terms. This analysis points out the occurrence of variant forms. Since these forms can be retrieved by means of right truncation, the fact that the right-hand portion of strings differs would not impede computer retrieval (Table II).

Table III. Variant Forms of Terms in Text Modification

Percentage of TM terms that occur in variant forms within the same citation	13.0%
Average number, per citation, of variant forms of TM terms	1
Percentage of citations containing variant forms of TM's	69.0%
Percentage of citations containing zero variant form TM's	31.0%

Table IV. Percentage of Unique Terms That Occurred as Title Terms and Keywords

Titles only	42.4%
Keyword phrases only	33.8%
Both titles and keyword phrases	15.8%
Variant forms in titles and keyword phrases	8.0%
	100.0%

Table V. Percentage of Terms That Occurred as Title Terms and KI Terms

Titles only	57.8%
KI only	22.6%
Both titles and KI	12.1%
Variant forms in titles and KI	7.5%
	100.0%

Within the CASIA Text Modification (TM) element, there are a number of variant forms of terms (e.g., *polymer* and *polymeric*) that occur in the same citation (Table III).

Comparison of Title Terms and Keywords. Of all unique terms found in either or both titles and keyword phrases on Condensates tapes, some occurred in titles only, some in KI's only, some in both, and some in variant forms in both titles and KI's (Table IV).

On a *citation* basis it was found that in 15.7% of the citations there was no overlap between titles and keyword phrases. Therefore, an exclusive search of either titles or keyword phrases would result in a loss of 15.7% of the terms. The relative merits of titles and keywords differ by 18.4%; i.e., 66.0% of the unique terms found in titles and/or keywords are found in titles and 47.6% of the unique terms are found in keywords.

Comparison of Title Terms and KI Terms. Of all unique terms found in either or both titles and KI's, some occurred in titles only, some in KI only, some in both, and some in variant forms in both titles and KI's. Each of these categories is treated exclusively (Table V).

These data represent *terms* found in both portions of the records and indicate term overlap and nonoverlap for the two elements of records. On a *citation* basis it was found that in 23.5% of the citations there was no overlap between title terms and KI terms. If one were restricted to the use of either title terms in *CA Condensates* (or CBAC or any other subset data base), or the initial terms in the permuted phrases in the hard copy *Chemical Abstracts*, there would be more retrieval capability found in the titles since the titles alone accounted for 57% of the unique terms in both elements. If one adds to that the 12.1% that overlapped and the 7.5% that occurred in variant forms, 76.6% of the terms appear in the titles.

Analysis of KI Phrases. In the analysis of KI phrases, duplicate word pairs were eliminated. The numbers that relate to usefulness of word pairs were based on the judgments of two chemical information specialists, each of whom has had more than ten years experience in searching. Their judgments indicate that the word pairs, as pairs, found in the KI phrases are of little utility as search terms. In fact, the pairs are far less useful than the single terms. This does not mean, however,

Table VI. Analysis of KI Phrases

Average number of KI phrases per citation; range 2-7	4.14
Average number, per citation, of distinct contiguous word pairs embedded within the KI's; range 3-11	7.47
Percentage of total embedded word pairs judged to be useful	10.5%
Average number of useful word pairs, per citation	0.78
Average number, per citation, of terms in useful pairs that also appear in titles	0.2
Average number, per citation, of terms in useful pairs that do not appear in titles	0.65

Table VII. Occurrence of Concept Heading Terms per Citation

Percentage of citations in which CH's occurred	88.0%
Percentage of citations in which no CH occurred	12.0%
Average number, per citation, of unique CH's for all citations	4.43
Average number, per citation, of unique CH's for all citations in which CH's occurred	5.02

Table VIII. Overlap between Concept Heading and Title Terms

Percentage, of citations with overlap between CH's and titles	49.0%
Percentage of citations with no overlap between CH's and titles	51.0%
Average number of overlapping terms per citation for all citations	0.86
Average number of overlapping terms per citation for citations in which overlap occurred	1.76

Table IX. Overlap between Concept Heading Terms and Keywords

Percentage of citations with overlap between CH's and Keywords	49.0%
Percentage of citations with no overlap between CH's and Keywords	51.0%
Average number of overlapping terms per citation for all citations	0.63
Average number of overlapping terms per citation in which overlap occurred	1.29

that the phrases are not useful. It appears that the single terms found in the phrases would be useful for searching, and the phrases would be useful for postsearch screening to determine context. The phrases or pairs treated as bound terms, however, are too restrictive for searching. They are not controlled strings and hence are unlikely combinations for a searcher to select as search terms (see Table VI).

Term Occurrence in CASIA Concept Headings. CASIA concept heading terms (CH's) occur in a large percentage of the citations (Table VII). (Stopwords and duplicate occurrences of the same term within a citation were omitted.)

Comparison of CASIA Concept Heading Terms and Title Terms. A comparison of the individual terms (stop words omitted) of CASIA concept headings and titles indicates that approximately half of the terms overlap. Thus, a keyword type of search of titles alone would provide only half of the retrieval capability that can be had by using concept headings also (Table VIII).

Comparison of CASIA Concept Heading Terms and *CA Condensates* Keywords. Approximately half of the CASIA concept heading (CH) terms are also found as *CA Condensates* keywords. (Stop words and duplicate occurrences of the same word within a citation were omitted.) Obviously, CH's provide retrieval capability that is not found in *CA Condensates* keywords (see Table IX).

Comparison of CASIA Concept Heading Terms and KI Terms. The overlap between CASIA concept heading terms and KI terms is only 29.0% (see Table X).

Table X. Overlap between Concept Heading Terms and KI Terms

Percentage of citations with overlap between CH's and KI's	29.0%
Percentage of citations with no overlap between CH's and KI's	71.0%
Average number of overlapping terms per citation for all citations	0.35
Average number of overlapping terms per citation for citations in which overlap occurred	1.20

Table XI. Overlap between Concept Heading Terms and Text Modification Terms

Percentage of citations with overlap between CH's and TM's	71.0%
Percentage of citations with no overlap between CH's and TM's	29.0%
Average number of overlapping terms per citation for all citations	1.69
Average number of overlapping terms per citation for all citations in which overlap occurred	2.39

Table XII. Duplication of Text Modification Terms within Citations

Percentage of citations in which TM's occurred	100.0%
Percentage of citations containing no TM's	0.0%
Average number of TM's per citation; range 0-20	7.68

Table XIII. Duplication of Text Modification Terms

Average number of duplicate TM terms per citation; range 0-20	5.14
Percentage of citations containing duplicate TM terms	91.0%
Percentage of citations containing zero duplicate TM terms	9.0%

Table XIV. Overlap between Text Modification Terms and Title Terms

Percentage, per citation, of total unique TM terms that overlap with title terms	24.0%
Average number, per citation, of TM's that overlap with title terms; range 0-6	1.88
Percentage of citations containing TM's that overlap with title terms	80.0%
Percentage of citations containing no TM that overlaps with title terms	20.0%

Comparison of CASIA Concept Heading Terms and CASIA Text Modification Terms. Although CASIA concept headings (CH) and text modifications (TM) perform different functions, when analyzed on a term-by-term basis there is considerable overlap of terminology (stop words omitted); see Table XI.

Analysis of CASIA Text Modification Terms. In analyzing text modification (TM) terms, stopwords and multiple occurrences (duplicates) of the same word were omitted, but variant forms (discrete strings) of the same word were included (Table XII).

Duplicate CASIA Text Modification Terms Associated with a Given Citation. Duplicate terms found in text modifications within a given citation are largely due to the fact that it is quite common (and necessary) for multiple TM's associated with a given reference to be identical except for one word. Duplication data are presented here, but in the comparisons that are given in later sections of this paper the duplicates were removed; see Table XIII.

Comparison of CASIA Text Modification Terms and Title Terms. CASIA text modification (TM) terms were compared with title terms. It was found that 80% of the citations contain one or more TM terms that overlap with one or more title terms; however, on a per term basis, only 24% of the TM's

Table XV. Overlap between Text Modification Terms and Keywords

Percentage of citations in which one or more TM terms overlapped with keywords	73.0%
Percentage of citations in which no TM overlapped with keywords	27.0%
Average number, per citation, of TM terms that overlapped with keywords for all citations	1.29
Average number, per citation, of TM terms that overlapped with keywords, for citations in which overlap occurred	1.78

Table XVI. Overlap between Text Modification Terms and KI Terms

Percentage of the total unique TM terms that overlap with KI terms	9.0%
Average number of TM's per citation that overlap with KI terms; range 0–3	0.66
Percentage of citations containing TM's that overlap with KI's	50.0%
Percentage of citations containing no TM that overlaps with KI's	50.0%

Table XVII. Nonoverlap of Title Terms

Percentage of citations containing one or more title terms that did not overlap	88.0%
Percentage of citations in which all title terms overlapped with terms in one or more of the other elements	12.0%
Average number of title terms (stop word and duplicates omitted) per citation	6.67
Average number of nonoverlapping title terms per citation for all citations	2.0
Average number of nonoverlapping title terms per citation for citations in which nonoverlap occurred	2.27

overlapped with title terms (Table XIV).

Comparison of CASIA Text Modification Terms and *CA Condensates* Keywords. The average number of CASIA text modification (TM) terms, per citation, for the sample was 7.68. Of these, a considerable number overlapped with Condensates keywords (Table XV).

Comparison of CASIA Text Modification Terms and KI Terms. CASIA text modification (TM) terms were compared with KWIC terms. On a term-by-term basis, only 9% of the TM terms (stop words and duplicates omitted) overlapped with KI terms. TM terms, therefore, provide considerably more retrieval capability than KI terms (Table XVI).

Nonoverlap. Nonoverlap statistics are perhaps more meaningful than bilateral comparisons between terms found in various data elements. The bilateral comparisons were done to assist data-base processors who may wish to consider alternative methods of retrieval, e.g., titles alone or keywords alone, etc. The nonoverlap statistics for terms found in each element demonstrate the unique retrieval capability provided by that element. *CA Condensates* keywords are compared with terms in KI phrases, titles, CASIA concept headings, and CASIA text modifications. Each of the terms in each of the five elements is compared with the terms in the other four elements.

Title Term Nonoverlap. Title terms were compared with KWIC terms, *CA Condensates* keywords, CASIA concept headings, and CASIA text modifications. Of all the title terms in the citations, an average of two terms per citation was found not to occur in the other elements; that is, nearly 30% of all title terms are not found in the other elements. Eighty-eight percent of all the citations contained one or more terms that did not overlap with terms in the other elements. Titles, therefore, can be seen to contribute significantly to retrieval capability (see Table XVII).

Table XVIII. Nonoverlap of *CA Condensates* Keywords

Percentage of citations containing one or more keyword terms that did not overlap	47.0%
Percentage of citations in which all keywords overlapped with terms in one or more of the other elements	53.0%
Average number of keywords (stop words and duplicates omitted) per citation	4.37
Average number of nonoverlapping keywords per citation	.76
Average number of nonoverlapping keywords per citation in citations in which nonoverlap occurred	1.63

Table XIX. Nonoverlap of CASIA Concept Heading Terms

Percentage of citations containing one or more CH terms that did not overlap	59.0%
Percentage of citations in which all CH terms overlapped with terms in one or more other data elements	41.0%
Average number of CH terms (stop words and duplicates omitted) per citation	4.43
Average number of nonoverlapping CH terms per citation for all citations	2.27
Average number of nonoverlapping CH terms per citation for citations in which non-overlap occurred	3.87

Table XX. Nonoverlap of CASIA Text Modification Terms

Percentage of citations containing one or more TM terms that did not overlap	90.0%
Percentage of citations in which all TM terms overlapped with terms in one or more of the other elements	10.0%
Average number of TM terms (stop words and duplicates omitted) per citation	7.68
Average number of nonoverlapping TM terms per citation for all citations	5.20
Average number of nonoverlapping TM terms per citation for citations in which nonoverlap occurred	5.76

Keyword Nonoverlap. *CA Condensates* keywords were compared with title terms, KI terms, CASIA concept heading terms, and CASIA text modification terms. Of all the keywords in the citations, an average of 0.76 per citation was found not to occur in the other elements. On a citation basis, almost half the citations were found to contain one or more keyword terms that did not overlap with terms in the other elements (see Table XVIII).

CASIA Concept Heading Term Nonoverlap. CASIA concept heading terms were compared with title terms, KI terms, *CA Condensates* keywords, and CASIA text modification terms. Of all the concept heading terms associated with the citations, an average of 2.27 per citation was found to occur in other elements for the same citations. Thus 51.0% of the unique CH terms, per citation, are not found in the other elements (see Table XIX).

CASIA Text Modification Term Nonoverlap. CASIA text modification terms were compared with title terms, KI terms, *CA Condensates* keywords, and CASIA concept heading terms. Citations in the test contained an average of 7.68 text modification terms; of these an average of 5.20 terms did not occur in the other elements. Thus, 68% of the unique TM terms per citation are not found in other elements (see Table XX).

CONCLUSION

While the analysis involved an admittedly small sample, and even if the resulting data were to vary by 10%, one could still conclude that the CASIA data-base concept heading terms

and text modification terms provide considerably more retrieval capability, based on single term occurrence, than can be found in titles alone, keywords alone, or KI phrases alone. It also provides retrieval capability beyond that which is available using both titles and keywords.

Beyond the question of objective retrieval capability, since the CASIA file contains phrases, it is possible to use this context information to screen or narrow down preliminary search results. If this file were to be made available on-line, it would greatly enhance the access to *Chemical Abstracts* that is now available through *CA Condensates* with its titles and keywords.

Comparison of the Retrieval Effectiveness of *CA Condensates* (CACon) and *CA Subject Index Alert* (CASIA)†

D. L. DAYTON,* M. J. FLETCHER, C. W. MOULTON, J. J. POLLOCK, and A. ZAMORA

Chemical Abstracts Service, P. O. Box 3012, Columbus, Ohio 43210

Received November 23, 1976

CA Condensates (CACon) is the computer-readable file corresponding to *Chemical Abstracts* (CA) containing bibliographic information and keywords. *CA Subject Index Alert* (CASIA) is the computer-readable file containing index entries for the Chemical Substance, General Subject, and Formula Indexes of the CA Volume Indexes. This paper studies the vocabulary characteristics of the two files and the relative retrieval effectiveness of CACon and CASIA for a range of general subject and chemical substance topics. A vocabulary study quantifies the unique substantive word content of each of the files, the distribution of this vocabulary over citations, and the overlap of the file vocabularies. The results demonstrate that chemical substance nomenclature is a distinctive content feature of CASIA. A search study shows that CASIA gives significantly better recall than CACon for chemical substance topics and that both files give comparable recall for general subject topics when appropriate methodology is used.

INTRODUCTION

Chemical Abstracts Service (CAS) started distributing computer-readable chemical information files in 1965 with the introduction of *Chemical Titles* on magnetic tape. Since then, CAS has introduced a variety of computer-readable chemical data bases which are aimed at meeting various information needs of the chemical community. Many computer-readable files distributed by CAS are now available on-line, making it possible for organizations without computer facilities to use these files. The increased number of searchers using the CAS data base without direct contact with CAS has created a need for disseminating information detailing the characteristics of these files. The purpose of this paper is to report on our research into the retrieval effectiveness of the computer-readable files *CA Condensates* (CACon) and *Chemical Abstracts Subject Index Alert* (CASIA) based on demonstration searches, and then to correlate this with file vocabulary characteristics.

The CAS Data Base[1] consists of a large body of bibliographic information, abstract text, and index data which is supported by several authority files, such as the Registry Files. Computer-readable files like CASIA and CACon provide access to selected portions of the CAS Data Base. CACon, introduced by CAS in 1969 and covering the period from July 1968 to the present, corresponds to the documents identified in the printed issues of *Chemical Abstracts* (CA). It includes names and affiliations of authors, patentees, and patent assignees; titles of papers, books, patents, and conference proceedings; and source document bibliographic citations, including CA section and subsection numbers and CA reference numbers. The keyword phrases used to create the *Chemical Abstracts* Keyword Index for each issue comprise a major search component of the file. These keyword phrases do not contain the CAS systematic nomenclature used in the CA Volume Subject Indexes. Substances are indexed using author terminology or compound class names.

Soon after CACon was introduced, CAS experimented with a computer-readable version of the *Chemical Abstracts* Volume Index called the *Chemical Abstracts* Integrated Subject File. This file was divided into Chemical Substance Index entries and General Subject Index entries and was in index entry order. This file was the subject of a study by Zipperer et al.[2,3] With the emphasis on current awareness and document-oriented retrieval, CAS introduced CASIA in order to supply these volume index entries on a more timely basis.

CASIA files are issued every two weeks and contain the index entries for the Chemical Substance and General Subject Indexes of the *Chemical Abstracts* Volume Indexes sorted by CA publication citations. The ability to perform retrospective searching is also available in this new packaging format because document-ordered CASIA files have been made available to cover the period from January 1967 to the present. The General Subject Index entries basically consist of concept headings selected from a controlled vocabulary, with uncontrolled-vocabulary text modifications which provide the context of the subject of the index heading in the original document. Each Chemical Substance Index entry contains a substance name and, usually, a text modification which indicates the context in which the substance was mentioned in the original document. In addition, Chemical Substance Index entries in CASIA contain molecular formulas and CAS Registry Numbers.

At the outset of our research we had two main goals: to investigate data-base vocabulary characteristics to assist users of computer-readable files in understanding the substantive content and retrieval capability differences between CASIA and CACon, and to search CASIA and CACon to demonstrate the relative retrieval performances which can be expected for different question types using each file. O'Donohue[4] and Prewitt[5] have stressed the importance of using search aids for effective profiling, and we also feel that such methodology is important to enable other users of CAS files to duplicate our

† Presented, in part, before the Division of Chemical Information, 172nd National Meeting of the American Chemical Society, San Francisco, Calif., Aug 31, 1976.

* Author to whom correspondence should be addressed.