when sufficient information about their composition is available.

This paper was designed to call attention to some of the major revisions in the CAS handling of chemical substance index nomenclature. A more detailed (120-page) account is included in the Volume 76 CA Index Guide, Section IV (Selection of Index Names for Chemical Substances).

The complete new documentation used by CAS staff in naming all classes of chemical substances is also available in typewritten form from the CAS Marketing Department. It comprises nearly 2000 pages, including a comprehensive index and alphabetical list of substituent radical names, and is titled "Chemical Substance Name Selection Manual for the Ninth Collective Index Period (1972–1976)."

LITERATURE CITED

(1) Rowlett, R. J., Jr., and Tate, F. A., *J. Chem. Doc.* **12,** 125 (1972); CA Vol. 76 Index Guide (1972), para 13, p. 101.
(2) CA Volume 76 (1972) Index Guide, Section IV, paras 296–300.
(3) *Chem. Eng. News* **30,** 4515 (1952).
(4) Patterson, A. M., Capell, T., and Walker, D. F., "The Ring Index," 2nd ed., Washington, D. C., 1960; Suppl. I-III (1963, 1964, 1965).
(5) *Enzyme nomenclature:* Recommendations of the International Union of Pure and Applied Chemistry and the International Union of Biochemistry. Elsevier, Amsterdam, 1973. This is a revision of the recommendations (1964) of IUB.

The CA Integrated Subject File. II. Evaluation of Alternative Data Base Organizations*

WILLIAM C. ZIPPERER, MARGARET K. PARK,** and JAMES L. CARMON
Office of Computing Activities, University of Georgia, Athens, Ga. 30602

Received October 9, 1973

The relative retrieval performances of the *CA Integrated Subject File* (CAISF), *CA Condensates,* and a Merged File created from these two data bases have been measured. Retrieval performance is reported in terms of recall and precision values as well as costs. The precision and recall retrieval failures—i.e., irrelevant documents and missed documents—have been analyzed for each data base and characterized according to the five major types of failures: index language, indexing, searching, clerical, and miscellaneous. Over-all analysis of the performance suggests that an effective data base can be created by augmenting the *CA Condensates* data base with Registry Numbers and some representation of the CAISF General Subject concept headings, which results in a file approximately half the size of the corresponding CAISF data base and is suitable for search using existing retrieval system software.

A research study was designed to investigate alternative methods of using the *CA Integrated Subject File* (CAISF) (the acronym ISF was used in the previous paper in this series[4]) for computer-based bibliographic retrieval and to compare these approaches to retrieval from *CA Condensates.* Both of these data bases are produced and distributed by Chemical Abstracts Service (CAS). The CAISF contains the index entries in the printed *CA Subject Index,* plus molecular formulas, the CA section numbers, and the CA abstract numbers. It is distributed in two parts for each six-month volume, the Chemical Substance (CS) segment containing the index entries for specific chemical substances and the General Subject (GS) segment containing the conceptual index entries. *CA Condensates* corresponds to *Chemical Abstracts* and includes the bibliographic citation data, keywords, and the CA abstract number.

The purpose of this study was to determine quantitatively the relative retrieval performances of these three data bases and of others constructed from them in terms of both cost and retrieval capacity. The CAISF contains largely controlled indexing information while *CA Condensates* contains bibliographic information and some uncontrolled descriptive information, such as titles and keywords. Both files represent the same set of documents. The comparison, then, includes these data bases organized as distributed by CAS as well as other combinations which can be created through computer processing. The major tasks included the measurement of recall and precision values for several file organizations, comparison of natural language and controlled vocabularies, evaluation of alternative search strategies, and projection of implementation and operational costs for the alternative approaches.

RESEARCH DESIGN

The approach which was taken to evaluate alternative data content and file organizations for the CA-related files was the construction and use of files as near opposite ends of the retrieval continuum as could reasonably be derived from the resources available—that is, a file containing natural language vocabulary and a file with controlled vocabulary; a file designed for postcoordinate index term retrieval *vs.* one with a highly structured, hierarchical, precoordinate retrieval structure; and a single file which collects both chemical compound and general subject entries at a single point suitable for Boolean logic operations *vs.* split files which require postcorrelation of document refer-

* Presented in part before the Division of Chemical Literature, 166th Meeting, ACS, Chicago, Ill., Aug. 28, 1973.
** To whom correspondence should be addressed.

ences. The differences between the data bases allow investigation of a number of characteristics of the vocabulary and file organizations. The indexing vocabulary ranges from the natural language of the titles and keywords in *CA Condensates* to the multi-level, controlled (but open-ended) vocabulary of the CAISF, the latter in turn supplemented by the natural language text modifications. It was not expected that any single one of these data bases would necessarily prove optimum, but analysis of the results was expected to provide quantitative measurements of the magnitude of improvements which could be expected by proposed search algorithms and various information contents.

The study was designed such that three sets of randomly selected questions were searched against each of four logical data bases and the retrieval results analyzed for relevance and recall. Upon completion of the first evaluation, each of the profiles was revised to improve precision based on the results of the first search run. That is, an effort was made to revise the profile to eliminate "false hits" which were occurring for that question in each specific data base. Searches were again run, and the results re-examined for relevance. A third revision was then made to all the profiles to improve the recall based on the results which were obtained for any given question from all four data bases. Thus, the results from search on one data base could be used to improve the results on another data base by examining the "missed" documents. Based on this design, the results give three points on the relevance-recall curve for each question and for each data base. The extent of the differences between subsequent search runs suggests the effects iteration would have on search results if, for example, the system were operating in an interactive mode.

One-hundred and three (103) questions which had been searched against the retrospective *CA Condensates* collection over a consecutive four-month period as part of routine processing in the Georgia Center were classified into three major groups, based on a predetermined set of characteristics. A hypothesis of the study was that performance could potentially vary according to the type of question, and the three generic types selected to test this hypothesis reflected the nature of the CAISF segments: Type I questions required general concept information only, Type II questions required chemical substance information only, and Type III questions required both general concept and chemical substance information. Classification of the questions gave 36 (35%) Type I, 37 (36%) Type II, and 30 (29%) Type III questions. Ten questions from each set were randomly selected to give a proportional design for statistical analysis of the results. The Type II and III questions were searched against four data bases on each of the three runs; the Type I (concept only) questions were searched against three of the four data bases on all three runs, the Chemical Substance data base being omitted. Thus, the design is 30 questions—three sets of ten each—searched against a maximum of four data bases (with results calculated for a fifth data base) in three separate search runs for a total of 330 individual search results.

Other variables which normally must be considered in the evaluation and comparison of data bases are profile coding differences, search strategies, relevance judgments, and the retrieval systems used. Variations owing to differences between profile coders were minimized by dividing the randomly selected questions equally between two information specialists (both chemists as well) for initial coding, then exchanging the sets of profiles for complete review and resolution of any differences in interpretation of the question or the construction of the search profile. Search strategy variations were minimized through use of a common vocabulary for the uncontrolled language search terms, regardless of whether they were being searched against titles and keywords or text modifications with every reasonable effort being made to make that common vocabulary a superset of the conventions used in these three data elements. Efforts were made to obtain the equivalent controlled indexing term (concept heading) for the General Subject data base through use of CA printed indexes (but not the volume index corresponding to the study file) and the *CA Index Guide*, but this was not always possible. The use of chemical nomenclature as search terms, especially for structural classes, is rather unique and has no direct correlation with the uncontrolled vocabulary search strategies, but it was generally treated in the same manner as were the concept heading terms. The analysis of the retrieval failures (both precision and recall) indicate the extent to which search strategy variations influenced the results. The third variable which was controlled was the judgment of relevance. Since the principal purpose of this study was to determine the relative retrieval performance of the various data bases, not the absolute performance, the relevance of all documents was determined by one person on the research staff, an information scientist with a doctorate in chemistry, on the basis of relevance to the question as stated. Judgments of relevance were made largely on the basis of the index entry or the title and keywords, with consultation of abstracts only in exceptional circumstances. This procedure provides a consistent basis on which to compare the alternative data bases, but it does not measure the performance of the system with respect to the users' needs. The last variable, the search system, was controlled by using the same computer-based search system on all the data bases, thus removing variations owing to search algorithms, programmer coding techniques and/or efficiencies, and computer configuration differences which are particularly pertinent in comparing costs. While one given text search system is not necessarily the "best" system for all of the file organizations, control of this variable puts the results on a common basis which, in conjunction with the search results and failure analysis, can provide a base line against which to measure proposed improvements. The search system used was the UGA Text Search System which is routinely used in the Georgia Center for bibliographic retrieval.[1]

The analysis of the search results from each of the files raises the question of definitions to be used for precision and recall. The precision measurements require a judgment as to the relevance of the document retrieved, and the recall measurements require that the total number of relevant documents in the data base(s) be known. As noted previously, all relevance judgments were made by one person, thus assuring consistency though not necessarily an absolute value with respect to the users' judgment. The size of the files used in the study (131,000 documents and 637,000 index entries) made the determination of an accurate recall base for a large number of questions difficult. The recall base was taken as the number of relevant documents in the union set from all searches of all data bases for each question. This is a practical definition which allows relative comparison of the results from each data base but one which is probably not accurate as far as true evaluation of retrieval performance for the end user is concerned.

DATA BASE CHARACTERISTICS

The data bases used in this study were *CA Condensates* (CA), CAISF Chemical Substance (ISF-CS), CAISF General Subject (ISF-GS), and a Merged File created from the preceding three files. All corresponded to Volume 71 of *Chemical Abstracts*. The results are also given for a

fifth data base, the merged CAISF, which is essentially the combination of the two CAISF segments (ISF-GS and ISF-CS). Results for this data base were obtained by coalescence of the results of the two CAISF segments to obtain the union set of answers, rather than by physically creating and searching a merged CAISF data base. Technical specifications for the *CA Condensates* and CAISF data bases can be found in the CAS documentation.[2,3] Detailed characteristics of all the files as converted for use in this study are reported in the preceding paper of this series.[4]

The principal data elements which occur in these data bases are illustrated in Figure 1. *CA Condensates* contains the title of the article, patent, etc.; keywords which are assigned primarily from the title and abstract; bibliographic information such as the journal title, volume, issue, pagination, and names of authors; and the CA abstract number. This file is a sequential file with each logical document record corresponding to one physical record in the UGA format. The CAISF Chemical Substance data base contains one record per index entry with the major data elements being the CAS Registry Number, the CA Preferred Index name, text modifications for approximately 66% of the entries, and the CA abstract number. Additional data elements of importance to retrieval are the qualifiers and functional categories. The General Subject segment also has one record for each index entry and includes the Concept Heading (which corresponds to the bold-faced heading in the *CA Subject Index*), an associated numeric code assigned to the concept heading and any associated qualifier or functional category, text modifications for approximately 99% of the entries, and the CA abstract number. Both of the CAISF segments are in inverted order, sequenced in the alphabetical order of the *CA Subject Index*. The Merged File, then, contains the Registry Number (with qualifiers or functional categories appended) from the ISF-CS file, the concept heading codes from the ISF-GS file, and the complete *CA Condensates* record. The CAISF data elements, which could be transferred to the Merged File, were limited to the coded representations mentioned because of file size considerations and constraints on interrelationships between multiple occurrences of the same data element in the UGA Text Search System. The CA abstract number ties all of the information for a given document together, as illustrated, where the CS and GS examples are two of the eight specific index entries for the document shown in the Merged File.

The vocabularies represented in the data bases range from completely uncontrolled, natural language text to controlled and systematic indexing and various combinations of these extremes. *CA Condensates* contains the natural language titles and keywords which are uncontrolled except for the application of preferred abbreviations in the keywords. The search strategy normally employed with the *CA Condensates* file is to construct a search profile containing all words and their related word forms (e.g., alternate spellings, abbreviations, synonyms, etc.) which might reasonably appear in titles and keywords, linking them as appropriate with Boolean logic operators. The CAISF index entries, on the other hand, have a hierarchical structure, portions of which are controlled and portions of which are uncontrolled. The Concept Headings reflect a controlled but open-ended vocabularly since the headings normally remain consistent from volume to volume, especially within a five-year collective indexing period—e.g., Spectra. The qualifiers representing modifiers to the concept headings tend to be consistent across time because of their associations with specific concept headings. The use of "infrared" and "Raman" in the boldface headings "Spectra, infrared" and "Spectra, Raman" illustrate the use of qualifiers. Functional categories represent

CA CONDENSATES

TITLE	PROOF AND THIN LAYER CHROMATOGRAPHIC SEPARATION OF ANTIOXIDANTS
KEYWORDS	FATS DETN PHENOLIC ANTIOXIDANTS OILS CHROMATOG
CITATION	QUAL. PLANT. MATER. VEG., 16 (1-4), 292-296, 1968
ABSTRACT NO.	004742G

CHEMICAL SUBSTANCE FILE

REGISTRY NO.	000149917
NOMENCLATURE	GALLIC ACID
QUALIFIER	ANALYSIS
TEXT MODIFN.	DETN. OF ANTIOXIDANT, IN FATS
ABSTRACT NO.	004742G

GENERAL SUBJECT FILE

CONCEPT HDG.	ANTIOXIDANTS
QUALIFIER	ANALYSIS
HDG. CODE	000855
TEXT MODIFN.	PHENOLS, CHROMATOG. OF, IN FATS
ABSTRACT NO.	004742G

MERGED FILE

TITLE	PROOF AND THIN LAYER CHROMATOGRAPHIC SEPARATION OF PHENOLIC ANTIOXIDANTS
KEYWORDS	FATS DETN PHENOLIC ANTIOXIDANTS OILS CHROMATOG
CITATION	QUAL. PLANT. MATER. VEG., 16 (1-4), 292-296, 1968
REG. NOS.	000149917, ANALYSIS: 000128370; 025013165; 000500389
HDG. CODES	000855; 003438; 006644; 007227
ABSTRACT NO.	004742G

Figure 1. Principal data elements in the four data bases

a special type of qualifier in that they are assigned to certain nomenclature entries (and a few concept headings) which frequently have a large number of abstract postings; these are controlled by an authority list of the 17 allowed functional category terms or phrases—e.g., Polymers. Text modifications are uncontrolled language statements which describe the relationship of the controlled portion of the index entry to the document, as for example, the text modification "hydrolysis of" under a specific chemical substance name. Text modifications are normally included in the index entries except for synthetic studies or similar papers of general interest. The search strategy in the two CAISF segments is to combine the controlled and uncontrolled terms with appropriate Boolean logic operators as they would normally occur *in a given index entry*. Each concept or significant term which occurs in the question is treated in turn as the controlled index entry point, with the remaining concepts added as the uncontrolled information which could be expected to occur in the text modifications. As a result, several index entries for a given document will ordinarily be retrieved in response to each question. Reduction of the index entry answers to the set of unique documents requires a second processing step. The Merged File combines the uncontrolled language of *CA Condensates* with the controlled indexing of the CAISF segments, the latter represented as numeric codes which include (or append) the qualifiers or functional categories which occur in the CAISF entries. Thus, this data base might be considered either an "enriched" *CA Condensates* with added Registry Numbers and concept heading codes or a combined CAISF data base in which titles and keywords replace the text modifications. The search strategy for this data base is similar to that of *CA Condensates* except that Registry Numbers and the UGA-assigned concept heading codes can be added as search terms.

Table I. Precision and Recall Results

Data Base	Run 1		Run 2		Run 3		Means	
	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision
Merged File	65.7	50.5	67.9	76.6	89.0	39.1	74.2	55.4
CAISF (Both Files)	60.8	59.4	62.8	77.3	86.9	22.9	70.2	53.2
Chemical Substance File	60.8	64.3	64.4	83.4	76.5	45.9	67.2	64.5
CA Condensates	48.6	51.7	52.8	80.1	73.6	39.8	58.3	57.2
General Subject File	32.9	54.3	34.3	69.9	67.1	26.1	44.7	50.1

SEARCH RESULTS—PRECISION-RECALL RESULTS

Table I gives the precision and recall means obtained for each of the data bases over all runs. Analysis of variance shows significant differences occurring in both the over-all recall means and the over-all precision means. Precision and recall were calculated according to standard definitions, using the assumptions discussed earlier. Ratios for which either the numerator or the denominator was zero were assigned values of zero.[5] Two of the 30 questions retrieved no relevant answers in any of the searches and were disregarded in the subsequent data analysis, rather than adopting a convention for a numerator of zero. The recall and precision averages reported throughout this study were calculated according to the "average of ratios" method as defined by Lancaster in order to retain the inter-question variability for statistical analysis.[6]

Looking at the recall values in Table I, the CAISF (the union set of the results of the two CAISF segments) and the Merged File are significantly better than the remaining three data bases, but they do not differ between themselves. The recall means of the three other data bases do differ significantly from each, decreasing in the order shown. For the precision means, the Chemical Substance file is significantly better than the remaining four data bases, among which there is no difference.

Figure 2 shows the operating ranges of the three data bases which represent the entire document collection: the Merged File, the CAISF (both segments), and *CA Condensates*. In this and subsequent plots, the point in the center is the initial search run result, the point to the right shows the result of the precision-oriented second run, and the left-hand point is the third, or recall-oriented, search run. There was no statistically significant difference in the mean precision values of these three files as averaged over the three runs for each. Their precision means were within five percentage points of each other. However, there was a greater spread between the mean recall values. The recall means of the Merged File and the CAISF file did not differ from each other significantly, but both had recall values which were significantly higher than that of *CA Condensates*. As far as data content is concerned, the meaningful difference between the Merged File and the CAISF is that the Merged File has titles and keywords where the CAISF has text modifications. The difference in mean recall values for these two files is 4%, and the difference in precision values is 3%. Neither of these differences is statistically significant, which suggests that the natural language representations in the two data bases are approximately equivalent in terms of retrieval capabilities.

Figure 3 compares the performances of the individual CAISF segments with their cumulative performance. The Chemical Substance file shows the highest mean precision and the second highest mean recall of the data bases studied. The General Subject file has the lowest recall values. Since this file contains the index entries for chemical compound classes (e.g., Alkaloids or Phenols), all three types of test questions containing chemical substance requirements were searched against the General Subject file. It was, therefore, of interest to determine if the low recall performance of the General Subject file was related to the question type. Figure 4 shows that the questions requiring chemical substance data are detrimental to the over-all performance of the General Subject file. The Type I questions (concepts only) have mean recall and precision values comparable to those obtained in *CA Condensates*. As far as the over-all performance of the CAISF is concerned, search of the chemical substance-containing questions in the General Subject segment of the CAISF is adding some to the questions' recall but at a heavy cost to precision.

Figure 5 shows the performance of the two types of questions containing chemical substance information in the Chemical Substance File (the solid lines) and in the Merged File (the dashed lines). The Type II questions, shown in the upper right-hand corner, are the chemical substance-only questions. Although the Chemical Substance file appears slightly better than the Merged File, there is no significant difference in the means. In so far as this set of questions is concerned, the Chemical Substance file contains nomenclature and Registry Numbers while the Merged File contains only the Registry Numbers. Thus, the nomenclature adds but a marginal improvement over Registry Numbers alone for this set of questions. One file was not used as the source of Registry Numbers for searching another file; all numbers were obtained manually from available reference sources. The differences in performance for the two data bases are much greater with the Type III questions which include both chemical substance and concept requirements. First, the relative order of the two files reverses on recall from that of the Type II questions, with the Merged File about twenty percent better than the Chemical Substance File.

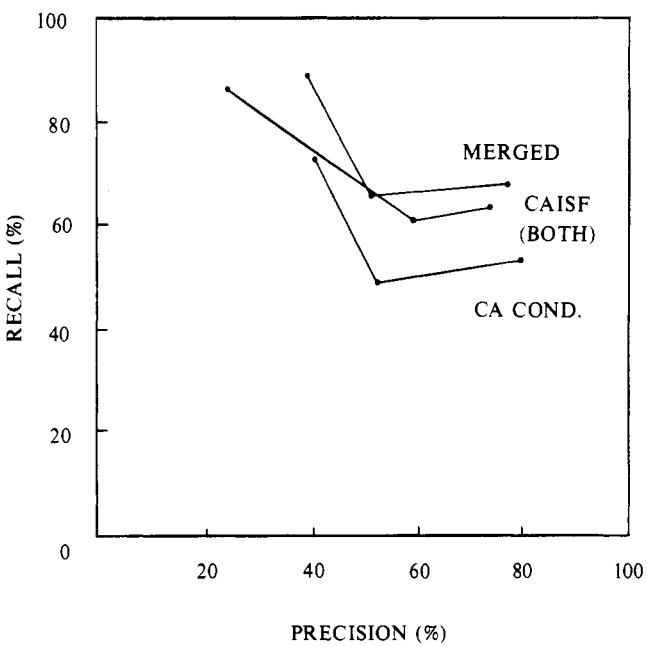


Figure 2. Performance ranges for the Merged File, the CAISF File (both segments) and CA Condensates

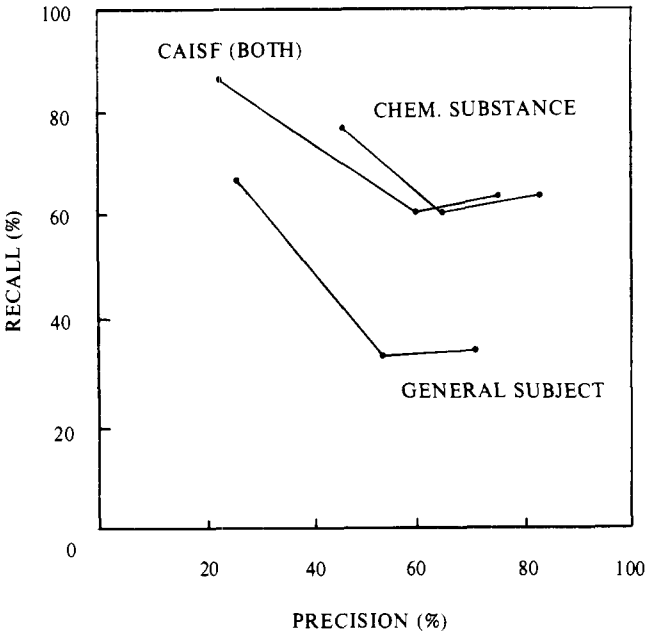


Figure 3. Performance ranges for the individual CAISF files

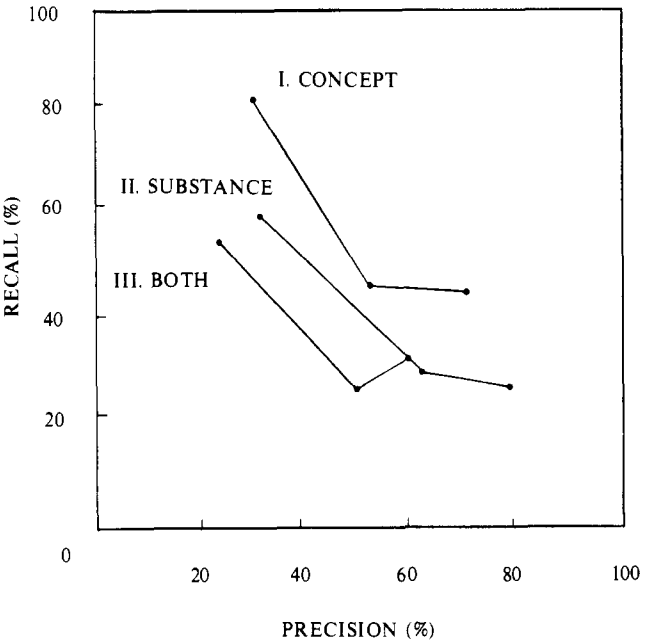


Figure 4. General Subject file performance by question type

If it is assumed that the chemical substance portions of these questions are performing essentially the same as are the Type II questions previously discussed, then the general concept portion of the question has caused the degradation in both precision and recall. The conceptual information is found in the text modifications of the Chemical Substance file and in the titles, keywords, and concept headings of the Merged File. If it is assumed that the approximate equivalence of text modifications and titles-with-keywords holds and that the concept entries perform no better than these natural language sources, as was indicated on the gross data base comparisons, the degradation could be caused either by the absence of text modifications for 34% of the Chemical Substance file entries or by insufficient indexing within the text modifications.

ANALYSIS OF RETRIEVAL FAILURES

Each of the retrieval failures was examined to determine a principal cause. The reasons for the failures as used in this report are based on the failure causes given by Lancaster, where full descriptions of each are available.[6] In judging the relevance of the answers, any document judged relevant in any data base (for a given question) was automatically assumed relevant for all other data bases. However, the irrevelant answers and the missed answers were judged independently for each data base in assigning the primary cause of failure. Thus, a given document may have been retrieved for more than one data base and have been assigned a different failure cause for each of the retrievals. Similarly, the recall failure reason for a given document may differ, depending on the data base. Each of these reasons is discussed in the following section in terms of major significant effects associated with data bases, with search runs, and with question types as determined through the Duncan Multiple Range Test. The Interactions between these factors (or the absence of interaction) suggest additional relationships beyond the main effects reported, but discussion of these interactions has been omitted as too detailed for general interest. The failure analysis which follows is organized by type of failure. Summaries by other variables, such as by data base, are available from the authors.

The mean numbers of retrieval failures by reason for each of the data bases are given in Table II where the means have been calculated as the total number of failures for each reason (all three runs) divided by the number of questions and the number of runs.

False Coordination. The first precision failure, false coordination of terms, is the classic failure in which two or more terms associated in an AND relationship appear in the document surrogate, but not in the proper relationship to each other as specified in the question. For example, a request for information on the color of meats using the terms COLOR and MEAT, retrieved an item entitled "Test strip color reaction for quick detection of nitrite in meat products" in which there is actually no direct relationship between the two terms. Analysis of the results indicates there are significantly fewer precision failures for this reason for the General Subject File than for the other three data bases, among which there are no significant differences. The numbers of false coordination failures in *CA Condensates* and the Merged File are not surprising

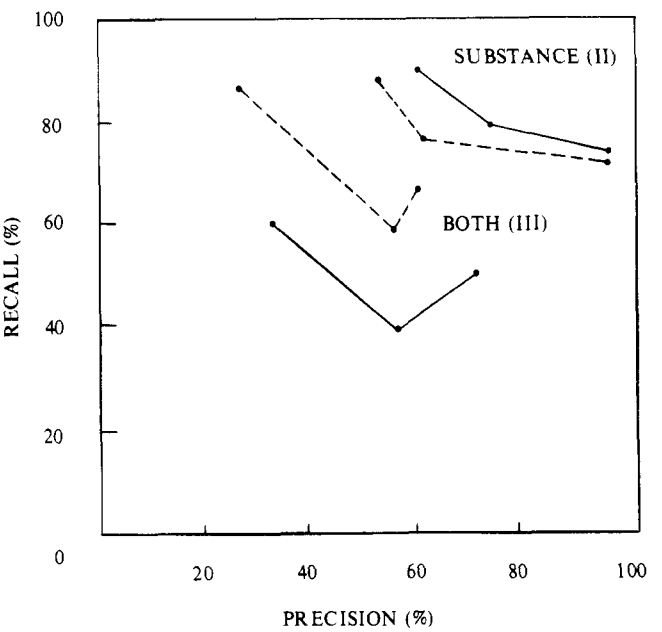


Figure 5. Performance of question types II and III in the Chemical Substance File (——) and the Merged File (- - -)

Table II. Mean Failures by Reason and Data Base

Precision Failures	CS	GS	CA	Merged
(1) False Coordination	16.2	1.0	19.9	13.2
(2) Incorrect Relationship	2.6	1.8	6.2	3.3
(3) Strategy too Broad	17.3	52.3	20.3	41.3
(4) Inevitable Retrieval	0.4	0.2	8.1	1.5
(5) Marginal Relevance	0.3	0.6	1.5	0.4
(6) Clerical Error	127.3	0	0.8	22.7
(7) Lack of Specific Indexing	2.8	13.4	0	11.7
Recall Failures				
	CS	GS	CA	Merged
(8) Inexhaustive Indexing	7.6	10.4	4.8	4.1
(9) Restrictive Search Formulation	1.1	0.1	0.7	0.7
(10) Index Language Insufficient	3.9	23.9	20.0	5.8
(11) Insufficient Term Expansion	5.9	5.8	5.7	4.9
(12) Document not in File	6.9	7.3	0.8	0.8
(13) Misspelled Word in File	0	0.1	0.5	0.1
(14) Clerical Error	14.0	0	0.1	0

since both are essentially free text data bases. In the Chemical Substance file, over two-thirds of the false coordinations occurred within the nomenclature data elements in searches for compound classes. For example, a request for information on N-oxides of some heterocyclic amines (e.g., indole) used a search strategy requiring the cooccurrence of the two terms INDOLE and OXIDE within the nomenclature data element. This strategy retrieved as an irrelevant answer, the compound INDOLE-3-CARBOXALDEHYDE, COMPD. WITH TRIPHENYLPHOSPHINE OXIDE (1:1). This is not a relevant item since the search terms are not related to each other as specified in the original request. When analyzed by run, there are significantly more failures for the recall-oriented third run than for the other two, between which there is no difference, thus indicating some correlation (negative) of this precision failure reason with the recall performance. False coordination failures are also independent of the question type as categorized in this study.

Incorrect Term Relationships. Incorrect term relationship failures are those irrelevant retrievals in which the requisite search terms are related, but in a manner other than that intended by the question. For example, a request for articles on the analysis of vegetable oils retrieved "the use of vegetable oil in the analysis of pesticides." The General Subject file shows significantly fewer failures for this reason than does *CA Condensates*, but it does not differ significantly from the other two data bases nor do the other three data bases differ among themselves. The third run shows significantly more failures owing to this reason than do the first two runs, between which there is no difference. These failures are also independent of question type. Five questions accounted for the majority (72%) of the failures.

Search Strategy Too Broad. Examples in which the search formulation was too broad include the use of broad subject codes—e.g., CA section numbers—or generic terms in which the retrieved answers owing to these terms were either exclusively or predominantly irrelevant. Most of these failures occurred in queries which had only one major concept so that the search strategy made little or no use of the Boolean operators "and" or "not." Analysis of the results indicates that the Chemical Substance file and *CA Condensates* had significantly fewer failures for this reason than did the other two data bases, but the data bases in each of these pairs did not differ between themselves—e.g., CS and CA. The generic level of the concept headings appears to be the reason for a large portion of these failures in the General Subject and Merged Files. Failures for this reason are clearly correlated with the intention of the search strategy, as the precision-oriented run (the second run) has significantly fewer failures for this reason than does the initial run (the first), which in turn has fewer than the third or recall-oriented run. Performance is independent of question type, and three questions accounted for over 50% of the total number of failures.

Inevitable Retrieval. There were only a few answers classified as inevitable retrievals where the retrieved article was correctly indexed, and the document correctly matches the search formulation, but the answer is nevertheless irrelevant. All such examples were due to homographs, and the failures occurred significantly more often in *CA Condensates* than in the other data bases, among which there was no difference. Twenty per cent occur for one Type II (chemical substance only) question, which makes this reason for failure question-dependent in this study for the *CA Condensates* data base only. The particular example was a question on amine oxides which the term N OXIDES was used in the profile. The question subsequently retrieved articles on the nitrogen oxide gases as well as amine oxides because of the use of the element symbol N for nitrogen in the keywords of the early *CA Condensates* volumes.

Marginal Relevance. Because of the design decision to judge relevance with respect to the question rather than the users' interests, very few failures can be attributed to value judgment in this controlled environment. These failures were documents which were only remotely related to the question topic and which, in the judgment of the information scientist, would not have provided sufficient information on the question topic to be of value. There are not significant differences between data bases or between question types. The recall run does cause more failures of this type than do the other two runs.

Clerical Errors. Two types of precision failures occurred owing to clerical or profile coding errors. One of these was the erroneous use of the same search term in two concepts which were related by the "and" operator, thus retrieving all documents containing this term and causing irrelevant retrievals. This type of clerical error occurred only once and this was in the Merged File. All items with the keyword "OXIDN" were retrieved, causing over 1000 irrelevant retrievals. The second source of error occurred with an incorrect, but valid, Registry Number. The exact source of the error is unknown but it may have occurred in one of the listings used to obtain Registry Numbers, in the transcription of the Registry Number to the profile, or in the keying of the numbers for the search. This incorrectly used Registry Number was assigned to a compound with a very high frequency of occurrence (Sodium Chloride), causing a large number of irrelevant retrievals. As is apparent from the table, the Registry Number error is the principal cause of failures for this reason, occurring in the two data bases which include this data element (CS and Merged). All such errors occurred in the third run and for one question.

Lack of Specific Indexing. Following the first two search runs, it was apparent that documents were being missed because of the absence of specific index terms assigned to the documents. The *CA Subject Index* and/or the *Index Guide* includes the term(s) initially used as search terms, but these terms had not been used in indexing the document. The search strategies were subsequently revised for the recall-oriented search run, using additional or broader terms. Precision failures caused by these revised strategies were assigned to reason 7, the lack of

specific indexing pertinent to the question. As would be expected, these failures occur significantly more often in the third run than in the other two runs. Both the General Subject file and the Merged data base have significantly more failures for this reason than does *CA Condensates*. None of the other pairwise comparisons show significant differences. Failures for this reason are independent of question type, but six questions account for over 80% of the total number of failures.

Lack of Exhaustive Indexing. Reason 8, lack of exhaustive indexing, is the first of the recall failures. This reason was assigned to unretrieved documents in which one or more concepts expressed in the document were not covered in the indexing, determined by examining the appropriate document and indexing records which were subsequently retrieved by the CA abstract number. Such failures occurred significantly more often in the General Subject file than in *CA Condensates* or the Merged File, with no significant differences between the Chemical Substance and General Subject files. The Chemical Substance file showed more failures for this reason than did the Merged File, but did not differ from *CA Condensates*. There was no significant difference between *CA Condensates* and the Merged File. An inspection of these document surrogates suggests that the absence of index entries is based on indexing policy, rather than indexing "errors of omission." In the Chemical Substance file, the concepts would have to appear in the text modification, and in this data base, the general policy is to omit the text modification entirely for papers dealing with synthesis (regardless of additional information in the paper) and to generalize the text modification for the remaining entries so as to cover broadly the entire scope of the document with respect to the particular chemical substance being indexed. In the General Subject file, many of the index headings are restricted as to the types of documents for which index entries can be made (e.g., DENSITY-General studies relating to density are indexed at this heading. For studies of density of specific materials or classes of materials, see those specific or class headings.) And, at the specific entries, the text modifications are frequently generalized to cover the full scope of the document as indicated previously. It appears that the titles and keywords of *CA Condensates* (hence the Merged File) may actually provide more searchable concepts per document than do the index entries of the General Subject file, but a detailed investigation has not been made to verify this observation for the full document collection. There are definite question type dependencies for this type of recall failure in the various data bases. The interactions are complex, and while reasons can be postulated, the data are insufficient to either confirm or refute them. The recall failures for this reason are independent of the run, however.

Restrictive Search Formulation. Items which were missed because the search formulation was too exhaustive through a restrictive Boolean logic expression were assigned to reason 9. The Chemical Substance data base showed more failures for this reason than did the General Subject file, but otherwise there were no significant differences between the remaining pairs of data bases. And, as is apparent from the means in Table II, there were very few failures for this reason even in the third run (failures are independent of runs).

Index Language Insufficient. Reason 10 is assigned to recall failures which are attributable to a lack of specificity in the index terms and, for the CAISF only, where indexing has not followed the process described by the entry vocabulary (the *Index Guide*). All of these types of failures represent some flaw in the system vocabulary, but in some cases this was caused by indexers not following the authority list rather than an inherent weakness of the index language. Analysis of the results indicates that the General Subject file and *CA Condensates* have significantly more failures for this reason than do the other two data bases, and there are no differences between the members of the two pairs (e.g., GS and CA). In the General Subject file, these failures tend to occur when documents are indexed under terms other than the ones to which the searcher is directed through the entry vocabulary in the *Index Guide* or when index terms are inconsistently applied to the documents. Consider, for example, the question dealing with the analysis of vegetable oils. The *Index Guide* directs the searcher from *Vegetable oils* to *Oils* through a cross reference and from *Oils, arachis* (also groundnut and peanut) to *Peanut oil.* The document "Fatty acid composition in the oil of groundnut (arachis hypogaea)" was subsequently indexed at *Peanuts* (text modification, fatty acid composition of), rather than at *Peanut oil,* and at *Fatty acids* (text modification, in peanuts). In *CA Condensates,* these failures tend to occur when a degree of specificity is missing from the indexing as represented by the title and keywords. For example; a request for information on metal carbonyl complexes did not retrieve the following relevant item from *CA Condensates:* "Mass spectra of organometallic compounds VIII. Some transition metal organometallic halide derivatives" with keywords/Palladium/Org compds Organometallic/Mass spectra/Molybdenum/Iron/Tungsten. The concept "Carbonyls" is "indexed" only under the very generic terms "Organometallic" and "Org compds" in *CA Condensates* while the more specific indexing of the General Subject File included an index entry at "Carbonyls." In both the General Subject and the *CA Condensates* data bases, the substance only questions (Type 2) show more failures for this reason than do the other two types. Also, the precision-oriented search run shows more recall failures for this reason than does the recall-oriented search where strategies have been broadened.

Insufficient Term Expansion. Failures owing to insufficient search term expansion occurred when synonyms and appropriate closely related terms were not included in the search formulation. There were no significant differences between data bases or between question types for this reason. The precision-oriented search run showed more failures for this reason than did the recall-oriented search, but there were no differences between the other pairs of runs.

Document Not in File. Recall failures which occurred because documents were not indexed in a particular file were assigned to reason 12. Volume 71 of *CA Condensates* contains references for approximately 130,000 documents; the Chemical Substance file and the General Subject file contain index entries for approximately 70% and 88% of the documents, respectively. There are also some 1282 documents for which index entries occur in one or both of the CAISF Segments which do not occur in the Merged File used in this study. As would be expected, the two CAISF segments show significantly more failures for this reason than do the other two files, and the failures are independent of run. For the General Subject file, the failures are also question dependent with Type 2 questions showing more failures than Type 3, and Type 3 showing more failures than Type 1. This is consistent with the subdivision of the index on the basis of type of index entry where the chemical substance components of the Type 2 and 3 questions fail more frequently in the General Subject file than do the questions related to the conceptual content.

Misspelled Words. Recall failures owing to misspelled words in the data base show no significant differences either by data base or by run, and over-all, this reason accounts for a very small number of missed references. The redundancy which is inherent in all these data bases ob-

viously compensates for the mispellings which were observed in the answers.

Clerical Errors. Clerical errors causing recall failures were due to two misspelled search terms which caused 14 recall failures. Such errors occurred only in the Chemical Substance file and in *CA Condensates* and in each case in one question.

Precision and Recall Failure Groupings. The reasons for precision and recall failures were grouped according to the five classes given in Table III. The first two reasons for precision failures, false coordination and improper term relationships, are normally considered index language failures, while the seventh reason, lack of specific indexing, is an indexing failure. These two types of failures are associated specifically with the characteristics of the data base. The third and sixth reasons, search strategy too broad and clerical or profile coding error, are searching and clerical failures, respectively, associated with the searching system. Both the fourth and fifth reasons have been assigned to the general category "Other," but it may be more appropriate to consider the inevitable retrievals (reason 4) as index language failures as all of these failures were due to homographs. Recall failure reasons 8 and 10 are indexing and index language failures, respectively. Reasons 9 and 11 are search system failures. The clerical errors (14) were grouped together in the Clerical category, but can generally be considered search system failures. Reasons 12 and 13 are grouped under "Other" in the summary table.

In an information center environment, the only variables which are really under the control of the center, as opposed to the data base supplier, are the Searching and Clerical factors. In this study, both the precision and recall searching failure groupings were dependent on the runs, with the recall failures occurring significantly more often in the precision-oriented run and the precision searching failures occurring significantly more often in the recall-oriented runs. This would suggest that the only general form of optimization which center staff could apply would be on an individual case basis, optimizing the search results to either precision or recall, but not to both at the same time. The types of searching failures reflect the classic inverse relationship for these two performance variables and are directly related to the index language and/or the indexing which are characteristics of the data base. Most of the clerical failures observed in this study are failures of the search systems, rather than the data bases. They occur infrequently, but can be catastrophic in nature when they do occur. The principal benefit to be gained in this area from the analysis is the identification of the types of clerical errors which do occur and the emphasis of manual checking to eliminate as many as possible. None of the error types was amenable to automatic checking procedures which might have been included in the system per se. The other two principal sources of failures—index language and indexing—are characteristics of the data bases and under the control of the data base supplier. Within the two CAISF segments, the inverse relationship between precision and recall tends to hold for each of these variables with the possible exception of indexing in the General Subject File. In this data base, improvement could be obtained by closer adherence to the indexing authority (the *Index Guide*). Other improvements, however, imply a change in indexing policies or procedures and many of them would necessarily affect the balance between precision and recall.

Table III. Precision and Recall Failure Means by Class of Reason

Precision Failures (Means)

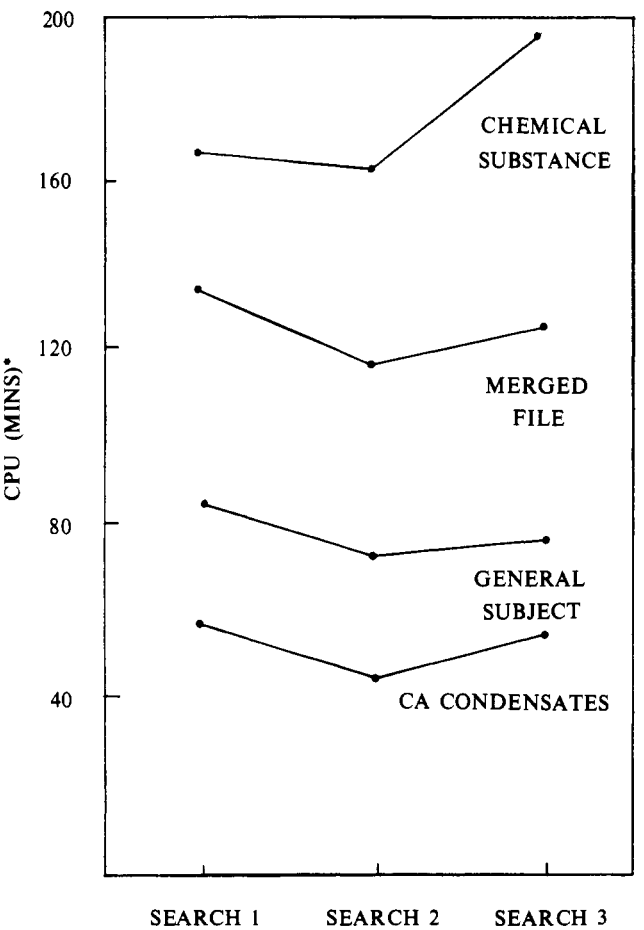
	CS	GS	CA	Merged
(1) Index language	18.8	2.8	26.1	16.5
(2) Indexing	2.8	13.4	0	11.7
(3) Searching	17.3	52.3	20.3	41.3
(4) Clerical	127.3	0	0.8	22.7
(5) Other	0.7	0.8	9.6	1.9

Recall Failures (Means)

	CS	GS	CA	Merged
(1) Index language	3.8	23.9	20.0	5.8
(2) Indexing	7.6	10.8	4.8	4.1
(3) Searching	7.0	5.9	6.4	5.6
(4) Clerical	14.0	0	0.1	0
(5) Other	6.9	7.4	1.3	0.9

COSTS

Figure 6 shows the search costs for the various data bases in terms of CPU time on an IBM 360/65. The same search system was used on all searches, so the timings are directly comparable. However, the search system used is designed for sequential document files like *CA Condensates* and the Merged File. It makes no use of the alphabetical ordering hierarchy of the file or the index language hierarchy of the data elements, both of which could be used to advantage to search the Chemical Substance and General Subject files. *CA Condensates* is the least expensive in terms of search time, requiring just under one CPU hour per search for the 30 questions. The Merged File is approximately twice this cost at just over two CPU hours per search. The Chemical Substance file, which is the largest, is also the most expensive at about three hours per search for the smaller set of 20 questions. The time for the full CAISF data base would closely approximate the



*IBM 360/65

Figure 6. Search costs in terms of CPU time

sum of the costs for the two segments, or about 4⅓ CPU hours. Standard statistical methods show the differences between all pairs of files to be significant. It should also be noted that the bibliographic data provided from *CA Condensates* and the Merged File includes the full bibliographic citation while the two CAISF segments provide only the CA abstract numbers directly. An additional processing step would be required to use these CA abstract numbers to retrieve bibliographic citations for display.

CONCLUSIONS

The results of this study indicate that addition of the chemical substance indexing from the Chemical Substance portion of the CAISF (e.g., Registry Numbers) markedly improves the recall performance of *CA Condensates*, as would be expected. Precision, however, is not improved significantly. Also, indications are that the text modifications of the CAISF files provide the same performance levels as do the titles and keywords of *CA Condensates*. Over-all performance suggests that an effective data base can be constructed by augmenting the *CA Condensates* records with the Registry Numbers and some representation of the conceptual index headings, providing a much smaller file than the corresponding inverted CAISF data base with an improvement in both recall and precision over either data base alone.

No direct cost comparison is appropriate as only the search system optimized toward the sequential, document-oriented data bases was used in this study. The Merged File thus takes advantage of existing computer software in this case, but it seems reasonable that comparable timings could be obtained for the inverted file structure with a system designed for this purpose.

ACKNOWLEDGMENT

The original data bases used in this study were obtained from Chemical Abstracts Service under standard lease or license arrangements. Their cooperation in reviewing the results of the study as it progressed is appreciated. Special acknowledgment and thanks go to Ben Pitman, III and Robert E. Stearns, Jr., the senior analyst-programmers who undertook all the programming for creation of the search files and for design and implementation of the automatic reporting system which was used to record and report the retrieval analysis results. Glenn O. Ware, Assistant Professor of Statistics and Computer Science, assisted with the statistical analysis of the results. Margaret C. Caughman assisted with initial profile coding and was consulted on various phases of the project. All funding of this research work was provided by the University of Georgia.

LITERATURE CITED

(1) "UGA Text Search System," 4 Volumes, Office of Computing Activities, University of Georgia, Athens, Ga., January 1971.
(2) "Data Content Specifications for the *CA Integrated Subject File* in Standard Distribution Format," Chemical Abstracts Service, Columbus, Ohio, 1971.
(3) "Data Content Specifications for CA Condensates in Standard Distribution Format," Chemical Abstracts Service, Columbus, Ohio, 1970.
(4) Zipperer, W. C., Stearns, R. E., Jr., and Park, M. K., "*The Integrated Subject File.* I. Data Base Characteristics," *J. Chem. Doc.* 13, 92 (1973).
(5) Lancaster, F. Wilfrid, *Evaluation of the MEDLARS Demand Search Services*, National Library of Medicine, Bethesda, Md., 1968.
(6) Lancaster, F. Wilfrid, "Information Retrieval Systems," Wiley, New York, 1968.

Evaluation of the *ACS Single Article Announcement Service**

SELDON W. TERRANT**
American Chemical Society, 1155 16th St. N.W., Washington, D. C. 20036
WILLIAM H. WEISGERBER
Omnisearch, 632 McLean Ave., Yonkers, N. Y. 10705

Received October 8, 1973

The *ACS SINGLE ARTICLE ANNOUNCEMENT*, issued semimonthly, is a current awareness service which provides the tables of contents for 18 ACS primary journals. Copies of articles cited can be ordered by use of a form provided. The service began on a subscription basis in January 1971. A survey of selected samples of 1971 subscribers (renewal and nonrenewal) was conducted in 1972. The objective was to evaluate the service by identification of factors influencing subscription renewal. The survey methodology and results are presented.

Most users of primary information need access to specific archival literature and also ways to keep abreast of advances in their general fields of knowledge. The first need is usually satisfied via indexes, which provide references to pertinent journal articles. Current awareness can be achieved by the scanning of journals or by the use of alerting services. In an ACS survey[1] of users of the *Journal of Organic Chemistry* conducted in 1969, it was found that over 80% of the respondents cited the title of an article as

* Presented in part before the Division of Chemical Literature, 166th Meeting, ACS, Aug 28, 1973, Chicago, Ill.

** To whom correspondence should be sent.