Improvements in Cost-Effectiveness in On-Line Searching. II. File Structure, Searchable Fields, and Software Contributions to Cost-Effectiveness in Searching Commercial Data Bases for U.S. Patents†

J. ROBERT ALMOND* and CHARLES H. NELSON
ICI Americas Inc., Wilmington, Delaware 19897

Received February 20, 1979

The advantages and disadvantages of file structure, searchable fields, and software requirements on searching three groups of commercial data bases for U.S. patents are described. The three groups of data bases analyzed are: Chemical Abstracts Condensates group in SDC and Lockheed, the CLAIMS® group in Lockheed, and the WPI group in SDC. The ease of use and general effects on search cost are examined for the following elements: (1) organization of files, (2) number of access points, (3) assignee designation, (4) classification system, (5) free language searching, (6) equivalents, (7) software considerations, (8) data-base relevance. The overall effect of selection of a given data base and command language on search cost depends on the particular search question.

Costs associated with on-line searching may be reduced by two types of procedures: (1) reduction of costs per unit time (e.g., equipment costs, personnel costs, data-base connect costs, communications network costs); (2) reduction of total connect time (e.g., I/O speeds, access times, keystrokes (I/O), data-base size, frequency and degree of truncation, the number of iterations of a single strategy). However, often a decision of procedure 1 affects the economics of procedure 2. For example, one data base may cost less per connect hour than another, but may require three to four times the number of input keystrokes to define a given concept. Analysis of component costs leads to development of search procedures to minimize the overall costs without affecting retrieval. As a result of our analyses, ICI Americas has reduced current awareness costs by more than 60% over a three-year period and concurrently produces a more timely, highly customized output. This results from the ability to run batch on-line update searches entered at machine speed using Dialog and Orbit. It is interesting to note that our costs in such searching have been consistently less than the SDI service costs offered by the vendors, and this procedure is applicable to all the bibliographic files as opposed to a select few.

The key to such analyses is not simply cost. Obviously results and benefits are equally important. One can measure costs easily and reliably but benefits are much harder to quantify. Accordingly, this paper is not directed at cost/benefit relationships but rather at cost effectiveness (in which one need only assume that the same information retrieved by a variety of techniques has a constant value to a requestor at a single point in time).

When comparing the cost effectiveness of searches among a variety of data bases, relevance and recall are primary considerations at a given cost level, and cost is the primary consideration at a given level of recall and relevance. Therefore, cost effectiveness can be measured in dollars expended for retrieval of 100% of relevant citations or % relevant citations retrieved at a given cost. It is difficult to predetermine the absolute number of relevant citations to a given search, but it is almost impossible to design a search strategy that will cost a predetermined amount. Furthermore, there appears to be no direct proportionality between dollars expended and relevant citations retrieved, particularly at extremes of retrieval. Therefore, percent or number of relevant citations retrieved/dollar expended cannot be extrapolated to 100% retrieval.

† Presented before the Division of Chemical Information, 176th National Meeting of the American Chemical Society, Miami Beach, Fla., Sept 11–17, 1978.

Table I

	CA	WPI	CLAIMS®
Subject	chemical	chemical & chemically related, general, electrical, mechanical	chemical general electrical mechanical
Time frame	1970-present***	1963-present*	1950-present**
Country coverage	(27) AU, OE, BE, BR, GB, CA, CS, DK, SF, FR, DT, DL, HU, IN, IL, IT, JA, NL, NO, PO, RO, ZA, ES, SW, CH, SU, US	(12) BE, CA, CH, DL, DT, FR, GB, JA, NL, SU, US, ZA, SW, European, and Research Disclosures	(1) US only

*coverage of pre-1970 material is fragmentary
**searchable fields vary over the time frame
***material included in CA from 1967-present is available through Lockheed

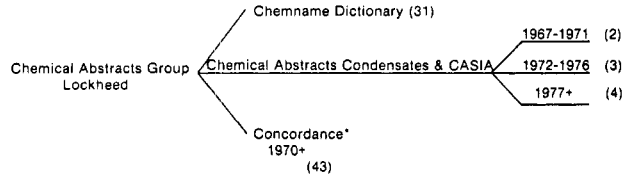
Accordingly we have chosen to use the unit of dollars expended for retrieval of 100% of relevant citations available in a data base as our basis for comparing cost effectiveness. This figure may then be reduced to dollars expended per relevant citation retrieved. For this figure to have any meaning in comparison of data bases, all of the elements of the desired final set must be present in each of the data bases. This need not be the case in commercial data bases drawn from different sources, e.g., World Patent Index from Derwent, CLAIMS® data bases from IFI Plenum, and *Chemical Abstracts* data bases from CAS. Therefore, a careful selection process must be used in framing test questions to ensure that the complete answer set is available in each data-base group to be compared.

Among the considerations necessary to properly design a test question to compare cost effectiveness among the commercial patent data bases are subject coverage, time frame, and country of patent. Table I compares these areas for each of the data base groups.

For a patent question to be equally applicable to all three data-base groups, it must concern U.S. chemical patents included in the sources from 1970 to present. In a case where all relevant art available from on-line sources must be uncovered, i.e., a state-of-the-art search, then all data bases possibly containing the information must be searched in an attempt to guarantee complete retrieval.

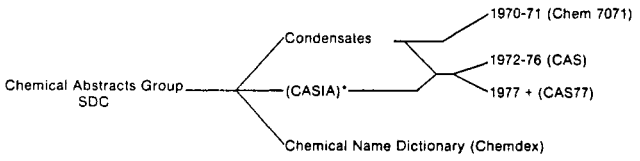
If, on the other hand, the requestor has limited funds available for the search or does not need more than a predefined amount of relevant art (as in questions of English language equivalents, infringement searching, or direct anticipation searching), the information scientist must choose the file and software with which to begin. If a search request is received in which each of a number of files should contain all the relevant citations, then how should the information

0095-2338/79/1619-0222$01.00/0



*The separation of equivalent patent numbers from the remainder of the condensate is relatively inefficient in searching for English language equivalents to patents retrieved by some means other than patent number. Combination of the contents of the Concordance and Condensates files would show increased efficiency similar to that demonstrated by combining the CASIA and Condensates files. Barring a file merger, assigning a Dialog accession number equivalent to the accession number of the basic citation would greatly increase efficiency.

Figure 1.



*The recent addition of the CASIA access points to the CAS record at SDC finally results in complete coverage. The absence of the fragmentation system and concordance database however may tempt many users to continue using the Dialog system.

Figure 2.

scientist choose between accessing files A, B, or C using command language X or Y? The answer is relatively simple.

The decision should be made on the basis of projected costs (not on the basis of ease of usage or lowest connect hour costs). Searches of all the data bases can be made if absolute thoroughness is required. However, if the cost of retrieving any additional material beyond that retrieved with the first search exceeds the value of that information to the user, an ultimate loss of business will occur.

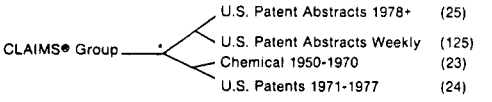
In a previous study we provided a general formula for prediction of costs in searching any given data base.[1] However, often a data-base group is subdivided into component files which must be searched individually. In this case, cost projections for each of the component files must be summed for comparison purposes. The reader is reminded that projections for a particular search are applicable only to that search, and attempts to generalize such projections as to representing an overall or generic measure of data-base cost effectiveness are *not* intended and can be very misleading.

I. FILE ORGANIZATION—EFFECTS ON COST

At times a file may grow to such a size that its efficiency is reduced in retrieval of references. At this point the vendor must decide whether or not to restructure the file. Often a new file is created to handle the overcapacity. When this leads to increased efficiency, it is desirable. However, at times separate files may be created because the vendor does not combine the contents of a new file with the old one for any number of reasons (e.g., cost, disruption of routine searching, differing search parameters, or staffing considerations). In this case to achieve the results desired, the searcher may be forced to search each of the individual files at an increased total cost.

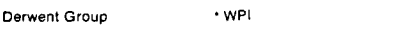
The organizations of each of the data base groups compared in this study are presented in Figures 1–4. Where this structure places a seemingly unnecessary burden on the information scientist, some indication of this is made.

To compare properly the projected costs of any one search using each of these data-base groups, the user must add the projected costs for searching each of the component files and then compare the total projected costs (see eq 1). Note that S_x may not equal the book cps rating because of buffers required for carriage returns and line feeds. Hence S_x may



*The division of the chemical patents into several files (if necessary) might have been more beneficial to the searcher if one file contained the most recent 17 years and the second covered earlier material, in spite of the difference in searchable fields among the files as presently constructed.

Figure 3.



*It might be useful to subdivide this very large data base (perhaps along subject lines, e.g. CPI and non-CPI material) with in-depth indexing (available to all users) at a premium price. Subscribers could recover (some or all) of this premium through reduced connect hour costs or subscription credits.

Figure 4.

PROJECTING COSTS

$$\$ = \sum_{y=1}^{n} \left[R_p(P_{off}) + R_y \sum_{x=1}^{m} N_x\left(T_x + \frac{C_x}{S_x}\right) \right] \quad (1)$$

R_p = cost per offline print in file y
P_{off} = number of offline prints
R_Y = composite rate for database y per connect hour
N_X = number of operators of type x
T_X = average turnaround time for an operator of type x
C_X = average number of characters per operator of type x
S_X = effective terminal speed (in characters per second)

vary with the length and number of lines to be printed, as well as with traffic levels on the network and search system.

II. NUMBER OF ACCESS POINTS (SEARCHABLE ELEMENTS)

In designing a search strategy to retrieve a set of desired citations, an information scientist selects the most specific available access points to the citations (provided their use will not exclude relevant material). For example, in searching *Chemical Abstracts* on Lockheed for citations on *tert*-butyl alcohol (RN=75-65-0), the searcher might select:

t(F)butyl(F)alcohol
t(F)butanol
tert(F)butyl(F)alcohol
tert(F)butanol
tertiary(F)butyl(F)alcohol
tertiary(F)butanol
RN=75-65-0

If the registry number were not included as an access point, the searcher would be restricted to the results of free language searching alone. In this instance he would either lose those citations in which the free language selectors were absent (ca. 75% in this case) or sacrifice relevance by using broader selectors, e.g., butyl and alcohol, or butanol.

Access points included in each of the data-base groups compared in this study are listed in Tables IIA–B as source, time, subject, country, or equivalency identifiers. For a more detailed treatment of file scope the reader is referred to an excellent article by Kaback.[2]

A greater number of access points only provides a searcher with an opportunity to use them if the search request provides the necessary input. Many times this is not the case. Although the subject area of a patent search is often precisely defined, it is infrequent, for example, that information identifying a particular inventor, accession number, or update code is contained in a search request. Therefore, these are normally blind access points, i.e., unable to be used by the searcher.

Four common types of patent questions are:

Table IIA

	CA Lockheed	CA SDC	CLAIMS®	WPI
Source	Author Assignee	Author Assignee	Assignee Code Author* Assignee Name	Company Code Author*
Subject	Free Language Terms	Free Language Terms	Free Language Terms	Free Language Terms
	CA Section/ Subsection	CA Section/ Subsection	US Classification Group	Intl. Patent Class
	Patent Classification	Patent Classification	US Classification Code	Derwent Class
	Registry Number	Registry Number*		Manual Codes
	Index Terms	Index Terms*		Multipunch
				Ring Index Numbers

*Authors are available in CLAIMS® only from 1971-date although IFI reports imminent addition of authors to the backfile. In WPI authors are available only when listed on the basic patent (1978-date). Registry numbers and index terms are only present in CASIA containing files.

Table IIB

	CA Lockheed	CA SDC	CLAIMS®	WPI
Equivalency	Accession Number	Accession Number	CLAIMS® Accession No.	Accession Number
	Equivalent Number	Patent Number	Equivalent Number	Equivalent Number
	Patent Number		Patent Number	Patent Number
			CA Accession No.	Application Number
				Application Dates
Time	Update Code	Update Code	Update Code	Accession Year
	Limit by Accession No.	Publication Year	Limit by Accession No. Issue Date	Update Code
Issuing Country	CA Basic	CA Basic	US only	Basic and Equivalent
	CA Equivalents (through Concordance)			

Table III

CA Lockheed	CA SDC	CLAIMS®	WPI
Assignee	Assignee	Assignee Code	Company Code
Free Language	Free Language	Free Language	Free Language
Patent Class	Patent Class	U.S. Classification	Intl. Patent Class
Registry No.	Publication Year	Accession # Limit	Derwent Class
CA Section/ Subsection	CA Section/Subsection	Equivalent Number	Manual Codes
Index Terms	Registry Number	5	Multipunch
Accession # Limit	Index Terms		Ring Index Numbers
Equivalent Number	7		Accession Year
8			Equivalent Number
			Application Date
			Application Number
			11

(1) *Subject questions*: What patents, if any, have issued in a particular subject area (direct anticipation or state of the art)?

(2) *Corporate intelligence questions*: What patents in a given area have been assigned to a particular company?

(3) *Equivalency questions*: Does a particular foreign patent have a U.S. equivalent?

(4) *Infringement questions*: Has a patent been granted in the last 17 years which claims the following invention?

Of the access points in Table II those which we have found useful in answering any of the above types of questions are designated in Table III.

In deciding which data base to choose of those with similar contents, e.g., *Chemical Abstracts* from Lockheed or SDC, the searcher should first decide if any of the access points are particularly suited to the question. Only after this decision has been made (and both data bases are found to contain all useful access points) should any consideration be given to differences in software. In the case of data bases containing different unit records, complete strategies should be prepared for each and then subjected to cost projections to determine the optimum strategy.

```
"PA=Dow
• • •
E7  PA=Dow Badische Co. ..... 9
E8  PA=Dow Chem. ..... 3
E9  PA=Dow Chemical (Nederland) B.V. ..... 2
E10 PA=Dow Chemical Co. ..... 1964
E11 PA=Dow Chemical Co. Ltd. ..... 6
E12 PA=Dow Chemical Investment and Finance Corp. ..... 1
E13 PA=Dow Chemical of Canada, Ltd. ..... 1
E14 PA=Dow Corning A.-G. ..... 1
E15 PA=Dow Corning Corp. ..... 302
                                -more-
? Page
Ref Index-term                          Type Items RT
E16 PA=Dow Corning G.M.B.H. ..... 3
E17 PA=Dow Corning Ltd. ..... 56
E18 PA=Dow-Unquinesa S. A. ..... 6
E19 PA=Dowa Mining Co., Ltd. ..... 71
E20 PA=Dowa Seiko Co., Ltd. ..... 8
E21 PA=Dowa Teppun Kogyo Co., Ltd. ..... 3
E22 PA=DQ Engineering Co. Ltd. ..... 1
```

Figure 5.

```
"PA=Shell
• • •
E5  PA=Sheel Internationale Research Maatschap ..... 1
E6  -PA=Shell .....
E7  PA=Shell Canada Ltd. ..... 1
E8  PA=Shell Chimie ..... 1
E9  PA=Shell Francaise ..... 2
E10 PA=Shell Int. ..... 2
E11 PA=Shell International Research Maatschapp ..... 2
E12 PA=Shell International Research Maatschapp ..... 1
E13 PA=Shell International Research Maatschapp ..... 1
                                -more-
? P
Ref Index-term                          Type Items RT
E14 PA=Shell Internationale Research Maatschap ..... 331
E15 PA=Shell Internationale Research Maatschap ..... 7
E16 PA=Shell Internationale Research Maatschap ..... 426
E17 PA=Shell Internationale Research Maatschap ..... 1
E18 PA=Shell Kagaku K. K. ..... 1
E19 PA=Shell Oil Co. ..... 374
E20 PA=Sheller-Globe Corp. ..... 4
E21 PA=Shemyakin, M. M., Institute of Bioorgan ..... 1

# PA=Shell Internationale Research Maatschap
      1    0  PA=Shell Internationale Re
# PA=Shell Internationale Research Maatschappe N. V.
      2  331  PA=Shell Internationale Re
```

Figure 6.

III. ASSIGNEE DESIGNATION

One important searchable field in patent data bases is the patent assignee field. This field is treated in three different ways by the data bases being compared.

CA. In this group the name of the assignee is spelled out, i.e., is not codified. This presents advantages and disadvantages to the user. It is possible in a highly specific question to select among individual branches of a large corporation. However, its disadvantage is that in order to select in the assignee field, the searcher must always expand (or neighbor) to determine the proper format for the individual assignee name or to determine the multiple entries which must be made for the entire corporate assignments or what truncation (if any) might be suitable. Figures 5 and 6 show examples of such practices. Note the difficulties encountered in the second case.

CLAIMS. In this group the assignees are given a five-digit numerical code. Divisions or branches of the same company all fall under a single code. However, joint ventures receive separate codes. These codes reduce costs by restricting the input characters to nine digits (e.g., #AC=41248) and also reduce the chances of keyboarding errors. As of this writing assignee codes in File 23 are still six-digit codes.

WPI. In this data base assignees are given a four-digit alphabetic code which may be modified depending on previously assigned codes by a following hyphen. Where Derwent recognizes subsidiaries with a common parent, these companies have been assigned the parent company code. This codification presents the same type of cost advantages as the assignee code system employed by the CLAIMS data bases.

IV. PATENT CLASSIFICATION SYSTEM

Two systems are used by the data bases being compared: the International Patent Classification System and the U.S.

BASIC U.S. PATENTS IN CHEMICAL ABSTRACTS CONDENSATES

Sample: U.S. 3,730,000 - U.S. 3,829,999

Basics (from index)		Equivalents (from concordance)	
144	entries per column	141	entries per column
x 83.9	columns	- 43	entries per column corresponding to foreign equivalents of U.S. basics
≈12,092	basic U.S. patents	= 98	entries per column corresponding to U.S. equivalents to foreign basics
		x 122.5	columns
		≈11,995	U.S. equivalent patents

Total Indexed U.S. patents		
	12,092	basics
	11,995	equivalents
	24,087	Total

If U.S. classification code is used as a retrieval requirement, then

$\frac{11,995}{24,087}$ = 49.8% of U.S. patents primarily assigned to that class would be lost (i.e. not retrieved).

Figure 7.

Patent Classification System. Each of these systems (to be treated separately) provides codes indicating areas in which claims of the patent reside. It would be of great use to have classification codes from *both* systems on *all* patent records in a data base, but this is not presently the case.

A. U.S. Classification System. In the U.S. patent classification system, each patent is assigned at least one code based on the area most relevant to its claims. This code is composed of a three-digit class (which refers to a broad area, e.g., Class 260—Chemistry of Carbon Compounds or Class 104—Railways) and a subclass of up to five digits (two decimals) with optionally one or two alphabetic characters ("unofficial" subclass); e.g., Class 260, Subclass 326.83, refers to heterocycles containing a five-membered ring of four carbons and one nitrogen—specifically pyrrolidines containing a sulfur atom in a thiocarbamyl group. The classification is hierarchical, but the coding is not; i.e., it does not necessarily correspond in levels to changes in equivalent levels in the code. For example, amino-substituted dibenzanthrones is a fifth-level subclass.

CLASS 260, CHEMISTRY, CARBON COMPOUNDS

Subclass	350	Carbocyclic or Acyclic
	351	. Anthrones or Anthranols
	352	. . Benzanthrones
	353	. . . Dibenzanthrones or Isodibenzanthrones
	354	. . . . Amino
	355	. . . . Oxy (non leuco)
	356	. . . . Halogen
	357	. . . . Dibenzanthrone
	358	. . . . Isodibenzanthrone
	359	. . . Anthanthrones

In the example four levels of classification are covered by a code change in the units level. As a result, the searcher cannot readily truncate this code to generate all amino-, oxy-, or halogen-substituted dibenzanthrones (other fifth-level subclasses). Indeed, if the code were truncated, one would get not only dibenzanthrones but anthanthrones and otherwise unspecified benzanthrones, anthrones, or anthranols, and carbocyclic or acyclic carbon compounds as well. However, the searcher can use a ranged select command, e.g., #-CL=260354000:CL=260358000 to obtain these fifth-level subclasses in a single step. The U.S. patent office has recently (1972) begun issuing patents in "unofficial" subclasses, e.g., Class 260, subclass 615B, and these "unofficial" codes are not presently searchable with any reliability in the on-line chemical patent files with the exception of CLAIMS 1971 date.

In the *Chemical Abstracts* data bases the U.S. Classification Code is provided only if the basic patent (the patent abstracted first from a family of equivalents) is a U.S. patent. In this case the code is complete, i.e., containing full class, subclass, and unofficial subclass of issue. If one required a U.S. class code to be present for retrieval, our estimates (shown in Figure 7) indicate one would lose an average of nearly 50% of the relevant material primarily assigned to that class by the U.S. patent office because that percentage of U.S. filings issue first in a foreign country. In addition, CA further limits the utility of this field by including only the original classification code and excluding all cross-referenced classes and subclasses.

In the CLAIMS–CHEM data bases the U.S. Classification Codes are provided for all patents. These data bases include not only the primary classification code but all cross-referenced codes as well. In addition, CLAIMS provides access to the new or changed classifications as they are introduced by Patent Office reclassification via annual update. At this time unofficial subclasses are not available in the back file.

B. International Patent Classification System. In the International Patent Classification System, each patent is assigned at least one code based on the area which best encompasses its claims. This code consists of three parts (e.g., *C07c 143/66*). The first of these is four characters in length (a leading and terminal alphabetic character with two imbedded numeric characters) representing the class. The next digits which follow the class and precede the slash represent the main group within the class. The digits following the slash represent the subclass. This system is also hierarchical and

CLASS		**C07C ACYCLIC AND CARBOCYCLIC COMPOUNDS**
Main Group	143	Sulphonic acids, their esters, halides, and amides; Sulphonated fats or fatty oils
Subclass	50	. . . . Low molecular weight water soluble sulphonated phenol aldehyde condensation products
	52	. . . Carboxylic sulphonic acids
	525	. . . . with esterified carboxyl groups
	53	. . . . Carboxamides thereof
	54	. . . . Hydroxy carboxylic sulphonic acids
	55	. . . Nitro sulphonic acids
	56	. . . Amino sulphonic acids
	58	. . . . of benzene
	60	. . . . of naphthalene
	62	. . . . of stilbene
	63	. . . . containing halogen
	64	. . . . Hydroxy amino sulphonic acids
	66	. . . . . of naphthalene or its derivatives
	665	. . . . Amino anthraquinone sulphonic acids
	67	. . . . Amino carboxylic sulphonic acids
	675	. . . . N-acylated amino sulphonic acids
	68	. Esters of sulphonic acids

may not be readily truncated at the subclass level for reasons similar to those given in the above description of the U.S. Classification System. Truncation following the complete class, or complete main group codes, should cause no problems.

This classification system is employed by both Derwent and *Chemical Abstracts* in preparing their data bases.

Chemical Abstracts. CA indexes only the first patent in any series of equivalents. This patent often issues in a "quick-issuing" country such as West Germany, Belgium, or the Netherlands. Depending on the policy of the issuing country, the classification may be specified only as deeply as the class (excluding the main group and subclass). Since CA indexes only the first patent of a family, further definition of the classification code which might be provided by later issuing equivalents is not included.

WPI. Derwent also indexes the basic patent (the first issuing from a major country) and includes whatever class codes are available. For basic U.S. patents, often only an abbreviated international class is included. However, as equivalents issue, this code is amplified and expanded based on additional classification present from the equivalents. Since Derwent does not include U.S. classification codes in their data base, U.S. patents indexed by Derwent will often be difficult to recall using classification codes, while maintaining the desired level of relevance.

From this discussion we can see that for maximum relevance one should employ full classification codes. However, if one attempts to retrieve on classification codes alone, it is essential that this be done only in CLAIMS. Although classification codes exist in Derwent and CA records, for complete recall the selection must be made on abbreviated codes supplemented by free language terms. In CA it is essential that equivalent U.S. codes be selected along with the international codes and these sets be merged before refinement.

V. FREE LANGUAGE SEARCHING

Although free language searching is probably the most frequently used access point in any search system, it can be one of the least efficient means of searching. The reliance of the searcher on the author's inclusion (in the patent title of any of the searcher's specific terms) lowers the probability of complete recall to an incredibly low level. However, the relevance of patents retrieved using this type of strategy should be high, if the searcher uses precise entry forms.

In WPI and CLAIMS the patent titles or expanded versions thereof were the only access points available to the free language searcher, and these often contained fewer than four significant terms. IFI has improved this situation by providing searchable abstracts (1978+), and Derwent is increasing its level of title enhancement. In CA two additional levels of vocabulary exist (keywords and index terms). Keywords often represent modifications of title terms with a few controlled entry terms. Index terms consist of General Subject Index Headings (controlled) and modifiers (controlled).

A unique fragmentation algorithm provided in Lockheed's CA data bases generates significant chemical name fragments as index terms. However, with improper use or a poor understanding of the algorithm, extremely noisy outputs may be obtained. For example, the term cyanate will retrieve not only citations relevant to compounds with CNO groups but also isocyanates (-NCO), isothiocyanates (-NCS), thiocyanates (-CNS), etc. A skilled searcher can avoid this problem by specifying that the desired term be a free term (not a fragment), e.g., # cyanate/FF. Additionally a list of fragments, stop terms, and fragmentation rules can be obtained from *Chemical Abstracts*.

One additional problem faced by free language searchers in WPI is their spelling variants. A controlled vocabulary has reduced this problem. For all terms with spelling variants the searcher previously had to enter the British variant as well, e.g., colour/color, litre/liter, sulphonyl/sulfonyl, or polymerisation/polymerization. In the last case and many similar instances, the searcher could simplify his selection by use of the variable-space character #. Hawkins gives an interesting breakout of spelling variants in data bases from different sources.[3] However, because of the generality of terms in the controlled vocabulary, the problem reoccurs in stringsearching the print record. Such stringsearching should increase in frequency, for now both coloured and colourless (for example) are indexed under the same term.

It is advisable for the free language searcher to enter and merge all possible synonyms for a given term to maximize retrieval. Obviously each of these select commands takes a certain amount of entry and processing time. It is, therefore, necessary to compare the costs of entry and output for all of the potentially useful terms as opposed to only those which are actually useful in obtaining the citations.

Truncation is another complex decision. It is cost effective to truncate when both of the following conditions are met: (1) All nonrelevant items selected by the use of the word stem will be removed from the final set by intersection with other sets. (2) The increased turnaround time caused by selection of noise terms is offset by the decreased input-output time from elimination of the need to enter each and every variable at the terminal. Therefore, the decision to truncate should be based on the length and number of synonyms eliminated, the terminal input speed, the number and postings level of irrelevant terms, and the remaining logic of the search.

VI. EQUIVALENTS

Each of the data bases employs different methods to store and process equivalent patent numbers. In Lockheed's CA data-base group, a separate data base (CA Patent Concordance) must be used. Searchers must *rekey* their search results into this file to obtain equivalent patent numbers. If accession numbers corresponded between the files, nonrekeyed accession could be achieved (an example of the inefficiency of the rekeying process may be seen in Figure 8). In CLAIMS and WPI the equivalents are an integral part of the primary record, providing low-cost rapid retrieval without the necessity of rekeying search results from one file to another. However, in CLAIMS foreign equivalents from only five countries (i.e., Great Britain, France, Germany, Belgium, and the Netherlands) are included. In comparing the inter-data-base cost effectiveness of obtaining U.S. patent numbers from foreign equivalents, one must include the following cost elements of the search: (1) cost of entry and output of foreign patent number, (2) cost of turnaround time of the selector, and (3) cost of print (off-line or on-line).

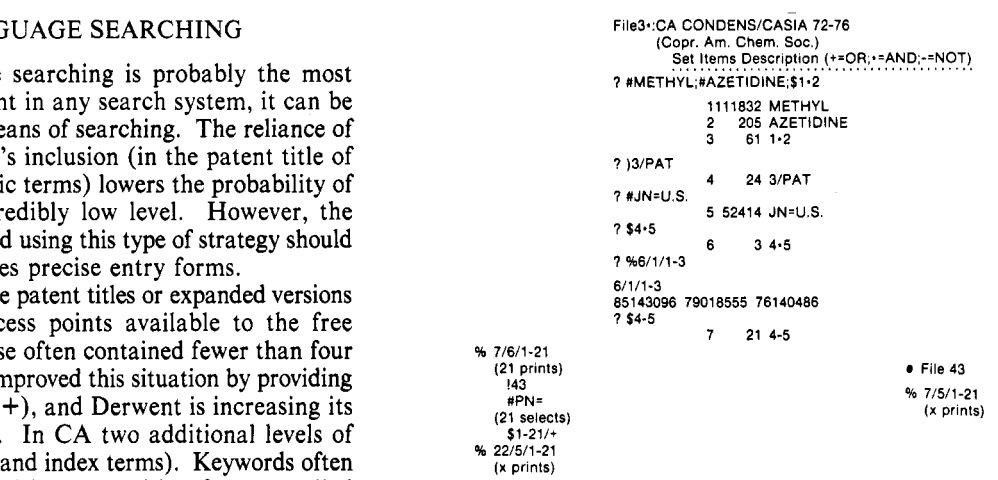


Figure 8.

VII. SOFTWARE CONTRIBUTIONS

A number of features have been incorporated in the Dialog and Orbit[4] command languages which present unique benefits to the searcher. If the relevancy and recall of a search can be enhanced by employing one of the features unique to a software system, then the value of the additional citations should be weighed against the increased cost of utilizing the feature. If the relevancy and recall remain at the same level, then projected costs of each strategy should be compared to maximize cost effectiveness.

A. Features Altering Recall or Relevancy. (1) Stringsearching. The Orbit software package has a unique feature, stringsearching, in which a set of citations may be serially searched for an ordered string of characters. This feature is exceptionally useful in locating fragments of chemical names or embedded conceptual terms such as polymer in polymerization, terpolymer, or copolymerization. This feature may also be used to ensure adjacency and order of terms, e.g., :instruction#manual: to exlude manual instruction. Another use of stringsearching is in reducing connect time; e.g., it may on occasion be less costly to form a small set and stringsearch it for a highly posted term than to build that term into the Boolean logic.

(2) Linking Operators. The Dialog software package has a "full text operator" feature which ensures that two or more terms are found in a particular citation or field or are no farther apart than a specified number of terms. These linking operators may be used with coded as well as free language terms, e.g., # selenium (c) SC=CA022005. Chemical fragments can be restricted to originating from the same word by using a (OW) operator.

However, these operators are often more costly, as they require serial or computational searching rather than simply

searching and matching on an inverted file list.

B. Features Not Altering Recall or Relevancy. (1) Tailored Print Formats. Orbit software allows the searcher to specify which fields are to be included in the printed record. Therefore, the searcher controls the characters per print information necessary to determine relative costs of on-line and off-line prints. In Dialog one must select from a number of predesigned formats.

(2) Interwoven Selects and Combinations. Orbit software allows the searcher to form a set by selecting and merging or selecting and intersecting component search terms in a single operation. This feature is an apparent convenience to the user, but the time savings is restricted to replacement of program and user cue messages in machine keystrokes. Furthermore, component terms are "dead" as opposed to operationally viable at this point. This feature has recently become available in Lockheed's Dialog and is termed "superselect". In use of Dialog's "superselect" command component sets may be returned "live" if the "#Steps" option is employed.

(3) Command Stacking. Both Dialog and Orbit softwares allow the searcher to perform a number of operations (limited to one line) in response to a single prompt from the computer. By separating each operation with semicolons (in Dialog or Orbit IV), several operations may be performed sequentially to produce that same number of live sets available for further operations.

(4) Nesting Logic. By judicious use of parentheses, the searcher in Dialog may in a single operation execute a complex logical statement. SDC has announced efforts toward making nested logic available on their system in the near future.

(5) Time Slicing. Any given portion of the computer memory and core may be allocated to only one customer at any given instant. In Orbit the processing time is divided into processing units (time slices) which are allocated on a first-come first-serve basis. If, however, a particular operation is not completed in the allotted time slice (a common occurrence in stringsearching—a set containing >100 records), then the user is advised of this overflow and asked if he wishes to continue (i.e., wait for the next time slice). In Dialog a similar situation occurs but, unless interrupted (by pressing the break key), the system assumes that the user wishes to continue the search. The user feels more included in an Orbit time overflow (being prompted every 10–12 seconds for his opinion Y/N). Intuitively, if the Orbit system awaits a customer interaction before proceeding, the turnaround time in Orbit should be increased accordingly.

(6) Saving Searches. If a comprehensive search strategy is designed which is applicable to more than one data base, this strategy may be saved and executed in a second file, thus minimizing operator keystrokes. In Lockheed when the ".Execute" operator is used, only the final set is live (i.e., capable of further use), although two diverse live sets can be obtained by use of the Keep command. Recent innovations at Lockheed have enabled temporary saving of searches (i.e., to be dropped at the end of one day), and selected stepwise execution (".Execute Steps") where each of the constructed sets is returned live. One may also selectively execute part of a saved strategy, but set numbering difficulties prevent effective use of combine commands in such search fragments. In SDC all sets are live. It should be noted, however, that stringsearching and restacking are not permitted in saved searches. In Lockheed there is no charge for a saved search unless it is stored from one billing period to the next. In SDC the saved search strategy (SAVESEARCH) incurs no charge but is stored for one day only. If the search is stored for later use (STORESEARCH), a charge of $0.10/command is incurred. To ensure the integrity of the search strategy against loss of search statements due to lack of postings, the SAVE command is used in the beginning of the SDC search. For retention of any satisfactory series of SDC search sets, the SAVEOLD command is entered upon completion of the search.

(7) Command Length. Dialog has historically allowed use of a symbolic command language with a single character representing the command, and continues this policy. This capability is also possible now in SDC by judicious use of the Rename command. In both of these systems this symbolic approach should be utilized to minimize input of characters, especially in the case of multiple profile searches (on-line SDI current awareness runs).

VIII. DATA-BASE RELEVANCE

The relevance of a given data base to any given search can be expressed in two forms: (a) the likelihood of the data base to contain records of any relevant citations as judged by its time frame and subject coverage, and (b) the amount of extraneous material that needs to be searched to identify those citations. The data base with the broadest general applicability will be more time consuming to search, especially in stringsearching and linked operator searching. To some degree in an inverted file search, data-base size determines the time used, e.g., in combinations of larger sets.

If the information scientist wishes to improve on-line cost effectiveness, he should regularly consider the relationship of the search to the following variables:

(1) File scope (likelihood of the data base to contain the desired search results)
(2) Unique file access points
(3) Organization of component files
(4) Controlled vs. uncontrolled vocabulary
(5) Character requirements for input output messages
(6) System software advantages
(7) File relevance (amount of extraneous material in data base)
(8) Type of search (aimed at a complete or predefined level of retrieval)
(9) Cost projections

However, he should be extremely wary of drawing generalizations concerning data-base or software utility for what may in one search be the ideal combination of system and data base may for a different search be entirely inappropriate.

REFERENCES AND NOTES

(1) J. R. Almond and C. H. Nelson "Improvements in Cost Effectiveness in On-line Searching I", *J. Chem. Inf. Comput. Sci.*, **18**, 13–15 (1978).
(2) S. M. Kaback, "Retrieving Patent Information ONLINE", *ONLINE*, 2 (1), 16–25 (1978).
(3) D. T. Hawkins, "Multiple Database Searching", *ONLINE*, **2** (2), 9–15 (1978).
(4) Orbit software has recently undergone its third major revision resulting in Orbit IV.