$H_2C = C = CH_2$

E3 MULTIPUNCH POSITIONS:

04&	C_{1-6} chain
042	Unbranched
050	C_{3-4} chain
058	Not a polyvalent chain
069	Hydrocarbon
530	>1 double bond, not conjugated
59&, 60-, 62-, 620	No rings
717	Aliphatic molecule

1972-mid '76, 286 references

Figure 8. Simple compound multipunch retrieval.

ceptably low relevance. The multipunch code may be very effective for complicated molecules, but it is relatively ineffective for simple ones.

Other retrieval problems stem from the fact that products are indexed, but not starting materials, again reflecting the interests of the end-product-oriented pharmaceutical and agricultural industries. It is very difficult to retrieve information on the utilization of a given starting material. Catalyst retrieval, too, is sketchy at the moment. For example, an attempt to isolate rhodium catalysis in the oxo test search gave only middling results. Hopefully the new catalysis manual codes introduced this year will help.

We saw how effective title keywords were for retrieval in the oxo example. They would have been even more effective had it not been for a number of typographical errors: hydroformalation, hydroformulation, and hydro-formylation. The misspelling rate in Derwent titles is far too high, and the unstandardized use of various punctuation marks causes one to miss relevant items. New hyphenation rules promise to help the punctuation problem. Users should certainly try to truncate terms wherever possible, to allow for variant endings, misspellings near the end of the word, and hyphenated combinations. Use the "Neighbor" command liberally, too. But of course misspellings at the start of words are going to be missed no matter what you do, and there is a great need for Derwent to tighten up on the proofreading of their abstract headings. One more point for searchers: remember to use both American and British spellings.

Tests are underway on the addition of indexing keywords to Plasdoc, and, while this is welcome, it would seem more important to add this feature in other sections first, since Plasdoc retrieval facilities are strong in comparison with those for most other sections. In particular, Chemdoc needs better retrieval of starting materials, processes, and simple compounds.

One of the problems in retrieval stems from the fact that changes in indexing have occurred over the years. Improved indexing is, of course, beneficial to users, but the searcher often has to devise different search strategies for different time periods.

Given a system as complex as the CPI, one of the greatest needs for users is detailed instructional materials. In fact the telegraphic style of Derwent headings extends to their instruction manuals, and leaves much unexplained. Recent communiques from Derwent have acknowledged that fact and promised more detailed manuals for the future. These will be very welcome, as will a promised listing of past multipunch codings, which will be an invaluable aid in devising multipunch search strategy.

Retrieval, then, presents more problems than does alerting. Nevertheless, there is considerable retrieval capability, and it has been greatly enhanced by the advent of the on-line file. There is evidence that titles, always a Derwent strong point, have improved and will continue to be improved. Hopefully the keywording experiment will succeed and will be extended to additional sections of CPI. In the meantime, CPI searchers are advised strongly to make use of all of the available, complementary, search parameters.

This paper has not attempted to touch on every aspect of the Derwent products, much less explore them all in detail. Its aim has been to present a balanced picture of the most important strengths and limitations of the system. If there has been a stress on some of those limitations, it has been in the spirit of striving for improvements in a powerful but imperfect information resource. Let there be no misunderstanding: based on extensive experience we regard the Derwent CPI–WPI system as an invaluable information asset, indispensible for anyone who needs to know about the new technology of the present and the recent past.

On-Line Retrieval of Chemical Patent Information. An Overview and a Brief Comparison of Three Major Files[†]

ROGER GRANT SMITH,* LOUISE P. ANDERSON, and SUSAN K. JACKSON
Merck & Co., Inc., Rahway, New Jersey 07065

Received April 21, 1977

Databases which contain patent information and are available to the public via on-line computer networks are listed. A comparison of document and retrieval parameters is made for those on-line databases which provide comprehensive coverage of the chemical patent literature, i.e., CHEMCON, CLAIMS^{TM}, and WPI. Examples of representative searches run in each of these three databases are presented and discussed. The conclusion is reached that no one database may be relied upon to provide all possible relevant answers.

INTRODUCTION

The ability to retrieve patents by means of on-line access to computers is a relatively recent phenomenon. Producers of machine-readable patent databases have been slow to capitalize on the advantages of on-line retrieval for obvious economic reasons: comprehensive patent databases are large and therefore expensive to maintain on-line, while their appeal has always been limited.

Considering the cost factors, it would seem unlikely that many organizations would care to sustain the entire cost of operating an in-house on-line retrieval system for a large file of patents. One company that did perform pioneering work

[†] Presented in the symposium on "Meeting the Challenges of the Changing Patent Literature", Division of Chemical Information, 173rd National Meeting of the American Chemical Society, New Orleans, La., March 21, 1977, and 11th Middle Atlantic Regional Meeting of the American Chemical Society, Newark, Del., April 21, 1977.

*Author to whom inquires should be sent.

Table I. On-Line Databases Which Provide Access to Patents

ON-LINE DATA BASES WHICH PROVIDE ACCESS TO PATENTS

ACRONYM OR SHORT NAME	FULL NAME	PRODUCER
APIPAT	American Petroleum Institute Patent Index[a]	American Petroleum Institute New York, New York
APTIC	Air Pollution Technical Information Center	Environmental Protection Agency Research Triangle Park, N.C.
CHEMCON	Chemical Abstracts Condensates[b]	Chemical Abstracts Service Columbus, Ohio
CA PATENT CONCORDANCE	Chemical Abstracts Patent Concordance	Chemical Abstracts Service Columbus, Ohio
CASIA/CHEMNAME	Chemical Abstracts Subject Index Alert[d]	Chemical Abstracts Service Columbus, Ohio
CBAC	Chemical-Biological Activities	Chemical Abstracts Service Columbus, Ohio
CLAIMS^{TM}/CHEM	Class Code, Assignee, Index, Method Search/Chemistry	IFI/Plenum Data Company Arlington, Virginia
CLAIMS^{TM}/GEM	Class Code, Assignee, Index, Method Search/General, Electrical, Mechanical	IFI/Plenum Data Company Arlington, Virginia
CLAIMS^{TM}/CLASS	Class Code, Assignee, Index, Method Search/U.S. Patent Classification[e]	IFI/Plenum Data Company Arlington, Virginia
ERA	Energy Research Abstracts[a]	Energy Research And Development Administration, Washington, D.C.
THE INFORMATION BANK	The Information Bank	New York Times Information Services Parsippany, New Jersey
INSPEC	International Information Services in Physics, Electrotechnology, Computers And Control[c]	Institute of Electrical Engineers London, England
MRIS	Maritime Research Information Service	Maritime Transportation Research Board, Washington, D.C.
NTIS	National Technical Information Service Bibliographic Data File	National Technical Information Service Springfield, Virginia
PAPERCHEM	Paper Chemistry Information	Institute Of Paper Chemistry Appleton, Wisconsin
PREDICASTS F&S INDEX	Predicasts F&S Index[a]	Predicasts, Inc. Cleveland, Ohio
PREDICASTS MARKET ABSTRACTS	Predicasts Market Abstracts[a]	Predicasts, Inc. Cleveland, Ohio
PNI	Pharmaceutical News Index	Data Courier, Inc. Louisville, Kentucky
POLLUTION ABSTRACTS	Pollution Abstracts	Data Courier, Inc. Louisville, Kentucky
STAR	Scientific And Technical Aerospace Reports	U.S. National Aeronautics And Space Administration, Washington, D.C.
STI	Specialized Textile Information Service[a]	Shirley Institute Manchester, England
SWRA	Selected Water Resources Abstracts	Water Resources Scientific Information Center, Washington, D.C.
TOXIC MATERIALS	Toxic Materials Information Center General File[a]	Toxic Materials Information Center Oak Ridge, Tennessee
TULSA	Petroleum Abstracts	University Of Tulsa Information Services Department, Tulsa, Oklahoma
WAA	World Aluminum Abstracts	American Society For Metals Metals Park, Ohio (For Aluminum Association Of America)
WPI	World Patents Index[a]	Derwent Publications, Ltd. London, U.K.

[a] Restricted access.

[b] Offered as separate files, e.g., 1970-1971 and 1972-present by SDC;
1970-1971, 1972-1976, 1977-present by Lockheed.

[c] Offered as two separate files: (1) Physics and (2) Electronics & Computers.

[d] For use with CHEMCON.

[e] For use with CLAIMS^{TM}/CHEM and CLAIMS^{TM}/GEM

Table II. Details on On-Line Databases Containing at Least 5% Patent Information (as of March 21, 1976)

	APIPAT	APTIC	CA PATENT CONCORDANCE	CASIA/CHEMNAME	CHEMCON	CLAIMS™/CHEM	CLAIMS™/CLASS
APPROXIMATE NUMBER OF PATENTS IN ON-LINE FILE	100,000	5,600	500,000	via CHEMCON	409,000	400,000	500,000 (via CLAIMS/CHEM and CLAIMS/GEM)
APPROXIMATE PER CENT OF FILE WHICH IS PATENT INFORMATION	100%	7%	100%	via CHEMCON	19%	100%	100%
COUNTRIES COVERED[a]	BE, CA, FR, DT, GB, JA, NL, ZA, US	US (primarily 1969-1975) JA (primarily 1970-1974) plus others not systematically covered.	See CHEMCON	See CHEMCON	AU, BE, BR, CA, CH, CS, DK, DL, DT, ES, FR, GB, HU, IL, IN, JA, NL, NO, OE, PO, RU, SF, SU, SW, US, ZA[c] (IT discontinued)	US plus foreign equivalents from BE, DT, FR, GB and NL	See CLAIMS/CHEM and CLAIMS/GEM

SUBJECT AREAS COVERED	Petroleum Refining And Major Petrochemicals	Air Quality;
		Air Pollution;
		Prevention and Control;
		Air Analysis
			See CHEMCON	See CHEMCON	Chemistry and Chemical Engineering	Chemical and Chemically Related Inventions	See CLAIMS/CHEM and CLAIMS/GEM

YEAR RANGE COVERED	1964-Present	1966-Present	1972-Present	1976-Present	1970-Present	1950-Present	Current U.S. Patent Office Classifications
BROKERS	SDC	Lockheed	Lockheed	Lockheed	Lockheed, BRS, SDC	Lockheed	Lockheed
ON-LINE COST	$65/hour	$35/hour	$45/hour	$60/hour	$35/hour (Lockheed) $60/hour (SDC) (BRS see note d)	$150/hour	$150/hour
OFF-LINE PRINT COST	$0.11/citation	$0.10/citation	$0.08/citation	$0.12/citation	$0.08/citation (Lockheed) $0.12/citation (SDC)	$0.10/citation	$0.10/citation
RESTRICTIONS IMPOSED BY PRODUCER	Subscriber use only	None	None	None	None	None	None

Notes: [a] ICIREPAT Country Codes are used, i.e., AR = Argentina, AU = Australia, BE = Belgium, BR = Brazil, CA = Canada, CH = Switzerland, CS = Czechoslovakia, DK = Denmark, DL = East Germany, DT = West Germany, EI = Ireland, ES = Spain, FR = France, GB = Great Britain, HU = Hungary, IL = Israel, IN = India, IT = Italy, JA = Japan, NL = Netherlands, NO = Norway, OE = Austria, PO = Poland, PT = Portugal, RU = Romania, SF = Finland, SU = Soviet Union, SW = Sweden, US = United States, ZA = South Africa.

[b] Pharmaceuticals 1963-Present; Agriculturals 1965-Present; Polymers 1966-Present; All Chemistry 1970-Present; All other patents 1974-Present.

in on-line retrieval of patent information was G. D. Searle Co. which developed the SOLD System for on-line retrieval from a large portion of Derwent's Central Patents Index file in 1973.[1] For the most part, however, it remained for the on-line information "brokers", such as Lockheed Missiles & Space Co. (Lockheed), System Development Corp. (SDC), and Bibliographic Retrieval Services (BRS), to make on-line retrieval of patent information available to anyone with a computer terminal and sufficient money to pay the hourly usage fee.

Currently, there are over 80 databases of all types on-line. The fact that some of these databases contain patent information is of little significance to the brokers. Their main concern is that each database generates sufficient income to justify its continued on-line availability.

ON-LINE DATABASES CONTAINING PATENT INFORMATION

At the present time there are 26 on-line databases which contain patent information. These are listed in Table I. More details are given in Table II for those databases in which patent information constitutes at least 5% of the total file. Additional information on all of these databases may be readily obtained from either the producers, the on-line brokers,[2,3] or the Directory of Computer Readable Bibliographic Data Bases compiled by Williams and Rouse.[4]

THREE ON-LINE DATABASES PROVIDING COMPREHENSIVE COVERAGE OF CHEMICAL PATENTS

Of the databases listed in Table II, there are only three which attempt to provide comprehensive coverage of the chemical patent literature. These three are Chemical Abstracts Condensates (CHEMCON), Class Code Assignee Index Method Search/Chemistry (CLAIMS), and World Patents Index (WPI).

CHEMCON is produced by Chemical Abstracts Service of Columbus, Ohio. CLAIMS is produced by IFI/Plenum Data Company of Arlington, Virginia. WPI is produced by Derwent Publications Limited of London, England. CHEMCON is carried by three brokers, Lockheed, SDC, and BRS, while CLAIMS is carried exclusively by Lockheed, and WPI is carried exclusively by SDC. It is important to note that the last two of these, CLAIMS and WPI, are pure patent databases and thus geared exclusively toward patent collection and indexing. On the other hand, CHEMCON is a literature service which looks upon patents as simply another form of technical literature. CHEMCON may be used in conjunction with its companion files CA PATENT CONCORDANCE, CASIA, and CHEMNAME, which are also produced by Chemical Abstracts and are currently offered only by Lockheed. CASIA and CHEMNAME do not refer directly to citations but provide additional search terms which can be used either alone or in conjunction with search terms from CHEMCON to find patents in CHEMCON. CLAIMS also has a companion file called CLAIMS/CLASS which enables the searcher to determine appropriate U.S. Patent Office Classification Numbers for use in finding patents in CLAIMS.

The balance of this paper will deal exclusively with a comparison between CHEMCON (including its companion files), CLAIMS, and WPI.

PREVIOUS COMPARISONS

There is extremely little research literature comparing the on-line retrieval afforded by different chemical patent da-

CLAIMS^{TM}/GEM	PAPERCHEM	POLLUTION	STI	TULSA	WAA	WPI
100,000	45,000	2,000	15,000	54,000	7,600	1,200,000
100%	43%	5%	40%	30%	16%	100%
US	CA, DT, FR, GB, JA, SU, US, directly; others via secondary sources.	All countries	GB, US	All Industrialized	?	BE, BR, CA, CH, CS, DK, DL, DT, FR, GB, HU, IL, JA, NL, NO, OE, PT, RU, SF, SU, SW, US, ZA (AR, AU, EI discontinued)
General, Electrical, Mechanical Inventions	Paper, Pulp and Board Manufacturing	All Aspects Of Pollution	Textile Science and Technology	Oil Exploration and Production	Aluminum and Light Metals	All Patents - Chemical, Electrical, Mechanical (Chemical only for Japan)
1975-Present	July, 1969-Present	1970-Present	1970-Present	1965-Present	1968-Present	1963-Present[b]
Lockheed	SDC	SDC, QL Systems, Lockheed	Shirley Institute	SDC	Lockheed	SDC
$90/hour	$110/hour (non-subscriber rate) $80/hour (academic rate)	$65/hour (SDC & Lockheed)		$125/hour (non-subscriber rate)	$50/hour	$150/hour (non-subscriber rate)
$0.10/citation	$0.15/citation (non-subscriber rate)	$0.10/citation (SDC & Lockheed)		$0.50/citation (non-subscriber rate)	$0.10/citation	$0.15/citation (non-subscriber rate)
None	None	None		None	None	In-depth Indexing Terms Limited To Subscribers Only

[c] Currently only resident patentees or patentees from countries other than the 26 listed are covered in AU, CH, CS, DK, DL, ES, HU, IN, IL, NO, PO, RU, SF, SU, SW.

[d] Depends upon contracted hours of usage.

tabases. This is due to the short period that these databases have been available in the on-line mode. For example, WPI, which contains the largest number of patents, did not go on-line until February 1976. The first comparative study, as far as the authors are aware, was presented by Hare in March 1976.[5] His study of the results of two searches conducted in CHEMCON and WPI indicated that WPI was noticeably superior in the specific fields he searched, namely, quinoxaline *N*-oxides and the fermentation production of citric acid by Japanese companies. Hare concluded that WPI was weaker than CHEMCON in term searching but made up for this lack with its unique chemical fragment codes.

COMPARISON METHODOLOGY

The methodology of this comparison will involve the use of two basic approaches: comparison of database coverage and comparison of parallel search results from each of the three files.

Coverage will be analyzed in two ways: first, in terms of the documents which are included in each database, and, second, in terms of the retrieval parameters—both subject and non-subject—which may be accessed in each database.

The parallel search results will be analyzed in terms of total relevant answers retrieved and a concept which the authors call "exclusive relevant answers". The answers which were missed by each system will also be analyzed and their causes discussed.

Recall and precision calculations will not be made for reasons which will be apparent in the discussion of the search results. Other comparative criteria, such as response time, user effort, output form, and cost will not be discussed in this paper. Discussion of the first three would be meaningless at this point for a variety of reasons; e.g., significant degrees of difference exist between the searchers in their experience with on-line systems and their familiarity with the databases. Thus, reproducible measurement of these factors was not possible. The last-mentioned criteria, output form, is basically similar for all three databases, i.e., a printout of bibliographic citations obtained either on-line or off-line. Typical citations include titles, patent numbers, database accession numbers, patentee names or codes, patent filing information, and classification codes. The format of the citations can be tailored to the user's needs. Neither abstracts nor full text are currently available from any of the three databases.

The parallel searches which are studied will be "parallel" in the sense that the same questions will be asked of each system. The results will thus depend on the document coverage and retrieval parameters available in each database, as well as the strategy employed by the search specialist. The searches run in CHEMCON have the further qualification that only the active file, from 1972 to present, was used. CHEM7071 was excluded for reasons of expediency: the keyword indexing for the 1970–1971 period tends to be less detailed than the indexing used in the current file and the possibility of using both left-hand and right-hand truncation in the SDC file would require a separate search strategy to have been devised, executed, and analyzed for each of the sample questions.

Because of the small number of sample searches this comparison is not intended to be a complete evaluation of the capabilities of each database. The conclusions which are drawn should be viewed in the context of the sample searches only.

DOCUMENT COVERAGE

A tabular comparison of the document coverage of each of the three databases is shown in Table III, arranged by country.

Table III. Document Coverage

Countries	CHEMCON[a]	WPI	CLAIMS-CHEM
Argentina	----	1975-1976	----
Australia	1970-Present[c]	1963-1969	----
Austria	1970-Present	1975-Present	----
Belgium	1970-Present	1963-Present	1950-Present[b]
Brazil	1976-Present	1976-Present	----
Canada	1970-Present	1963-Present	----
Czechoslovakia	1970-Present[c]	1975-Present	----
Denmark	1970-Present[c]	1974-Present	----
Finland	1970-Present[c]	1974-Present	----
France	1970-Present	1963-Present	1950-Present[b]
Germany, East	1970-Present[c]	1963-Present	----
Germany, West	1970-Present	1963-Present	1950-Present[b]
Great Britain	1970-Present	1963-Present	1950-Present[b]
Hungary	1970-Present[c]	1975-Present	----
India	1970-Present[c]	----	----
Ireland	----	1963-1969	----
Israel	1970-Present[c]	1975-Present	----
Italy	1970-1976	----	----
Japan	1970-Present	1963-Present	----
Netherlands	1970-Present	1963-Present	1950-Present[b]
Norway	1970-Present[c]	1974-Present	----
Poland	1970-Present[c]	----	----
Portugal	----	1974-Present	----
Romania	1970-Present[c]	1975-Present	----
South Africa	1970-Present	1963-Present	----
Spain	1970-Present[c]	----	----
Sweden	1970-Present[c]	1974-Present	----
Switzerland	1970-Present[c]	1963-Present	----
United States	1970-Present	1963-Present	1950-Present
U.S.S.R.	1970-Present[c]	1963-Present	----

[a] Includes CHEM7071 Coverage.

[b] Covered as equivalents to U.S. only.

[c] Currently covers patents issued to resident patentees or patentees from countries other than the 26 covered by CAS.

Table IV. Retrieval Parameters–Nonsubject

Parameter	CHEMCON (SDC)	CHEMCON (Lockheed)	CASIA	WPI	CLAIMS™ CHEM
C.A. Accession Number	x	x	-	-	x
Derwent Accession Number	-	-	-	x	-
Accession Year	-	x[a]	x[a]	x[d]	x[a]
Company Code	-	-	-	x	x
Company Name or Patent Assignee	x[b]	x[b]	-	x[c]	x
Author or Inventor	x	x	-	-	-
Patent Number	x	x	-	x	x
Patent Country	x	x	-	x	x
Priority Date	x[c]	-	-	x[f]	-
Priority Country	x	x	-	x	-
Priority Number	x[c]	-	-	x	-
Claim Type (C, P, M)	-	-	-	-	x[e]

Notes: [a] Through limit command using accession number ranges

[b] By /CS- each term is indexed separately, e.g., Syntex/CS, Lilly/CS and Eli/CS

[c] By Stringsearch

[d] Only available 1970-Present

[e] Only available 1972-Present through limit command.

[f] Only available 1973-Present?

It should be noted that Derwent's coverage of chemistry developed stepwise beginning with pharmaceuticals in 1963 and followed by agricultural chemicals in 1965, polymers in 1966, and finally full coverage of chemistry in 1970. It should also be kept in mind that each database employs a slightly different definition of what constitutes a chemical patent. Derwent, for example, uses a somewhat broader definition than Chemical Abstracts, and for that reason picks up a large number of "fringe" patents.

Table V. Retrieval Parameters–Subject

Parameter	Chemcon (SDC)	Chemcon (Lockheed)	CASIA	WPI	CLAIMS CHEM
Title Terms	x	x	-	x	x
International Patent Classification	x	x	-	x[a]	-
Index Terms	x	x	x	-	-
C.A. Publication Section	x	x	-	-	-
U.S. Patent Classification (Original)	x	x	-	-	x
U.S. Patent Classification (Cross Reference)	-	-	-	-	x
Classification Code	-	-	-	-	x
Classification Group	-	-	-	-	x
Derwent Manual Code	-	-	-	x	-
Derwent Class	-	-	-	x[a]	-
Derwent Punch Code	-	-	-	x	-
Ring Index Number	-	-	-	x[b]	-
Rare Fragment Numbers	-	-	-	x[c]	-

[a] 1970-Present

[b] 1972-Present

[c] 1972-1975

Table VI. Searches Used in Study

SEARCH NUMBER	SEARCH TYPE	TITLE
1	SPECIFIC COMPOUND	BLEOMYCIN, ITS DERIVATIVES AND RADIOISOTOPES
2	COMPANY	BRITISH ALUMINIUM
3	CHEMICAL STRUCTURE	12-PHENYL(1,3)OXAZINO(3,2-d)-(1,4)-BENZODIAZEPINES
4	CHEMICAL STRUCTURE	2-ANILINONICOTINIC ACIDS, INCLUDING NIFLUMIC ACID
5	COMPANY/ ACTIVITY	E. MERCK TRANQUILIZERS
6	EQUIVALENTS	U.S. 3,769,282 PATENT FAMILY MEMBERS

RETRIEVAL PARAMETERS

The retrieval parameters which are accessible in each of the three databases are shown in Tables IV and V. CHEMCON offered by SDC and CHEMCON offered by Lockheed are listed separately because the Chemical Abstracts Condensates files can be accessed differently owing to differences in the retrieval programs of the two brokers. CASIA is also listed separately to illustrate that in itself it contains no patent information but provides additional index entry points to data in CHEMCON. The footnotes should be read carefully, since they indicate restrictions on how certain parameters may be used.

It is apparent from Table V that the only subject parameter common to all three databases is the title terms category. Furthermore, the significance of this overlap should be minimized since the titles are frequently enhanced by the different database producers and consist of uncontrolled vocabulary. In essence, there is no subject retrieval parameter which appears consistently the same in all three databases.

Of the three files only WPI makes available on-line all of the retrieval parameters which are available in the original manual or batch searchable database. Both CHEMCON and CLAIMS are designed to provide quick, nonexhaustive answers. These answers may then be used as aids in searching for additional answers in the printed Chemical Abstracts Subject Indexes or the batch-searchable IFI comprehensive Database of Patents.

Table VII. Search Strategies and Performance Dates

SEARCH	CHEMCON*/CASIA SEARCH STRATEGY AND DATE PERFORMED	CLAIMS SEARCH STRATEGY AND DATE PERFORMED	WPI SEARCH STRATEGY AND DATE PERFORMED
1. BLEOMYCIN, ITS DERIVATIVES AND RADIOISOTOPES	CHEMCON: BLEOMYCIN: AND P (UC) DATE PERFORMED: 2/4/77	BLEOMYCIN OR BLEOMYCINIC OR BLEOMYCINS from EXPAND BLEOM DATE PERFORMED: 1/14/77	{[B02-B(MC) OR C02-B(MC)] OR [011/B,B1(MP)] OR [011/C,C1(MP)]} AND STRINGSEARCH :BLEO: (TI) DATE PERFORMED: 1/7/77
2. BRITISH ALUMINIUM	CHEMCON: BRITISH/CS AND ALUMINIUM/CS AND P (UC) DATE PERFORMED: 1/3/77	AC = 011400 DATE PERFORMED: 1/14/77	BRHA (PC) DATE PERFORMED: 1/7/77
3. 12-PHENYL(1,3)OXAZINO (3,2-d)(1,4) BENZODIAZEPINES	CHEMCON: [(PHENYL:OXAZINO:-BENZODIAZEPIN:) OR (OXAZINO:-BENZODIAZEPIN:) OR (OXAZINO AND BENZODIAZEPIN:)] AND P(UC) CASIA: Select: 1,3(W) OXAZINO(4W) BENZODIAZEPINE 1,3(W)OXAZINO (4W)BENZODIAZEPIN } In CHEMNAME File Print Registry Number of retrieval citations. Select Registry Number from CASIA file. Limit /PAT. DATE PERFORMED: 1/5/77	(EXPAND OXAZIN AND EXPAND BENZODIAZEP) OR (EXPAND PHENYLOXAZIN AND EXPAND BENZODIAZEP) (all relevant terms were selected from each expand) DATE PERFORMED: 1/14/77	40388 (RR) OR ({[B06-E05(MC) OR C06-E05(MC) OR E06-E05(MC)] OR 26&/B,B1,B2,C,C1,C2,E3(MP)] OR [C07D-099/04(IC)]} AND {373/B,B1,B2,C,C1,C2,E3(MP)} AND STRINGSEARCH :OXAZINO: (TI) AND :BENZODIAZEPIN: (TI)) DATE PERFORMED: 1/7/77
3. 12-PHENYL(1,3)OXAZINO (3,2-d)(1,4) BENZODIAZEPINES	CHEMCON: [(PHENYL:OXAZINO:-BENZODIAZEPIN:) OR (OXAZINO:-BENZODIAZEPIN:) OR (OXAZINO AND BENZODIAZEPIN:)] AND P(UC) CASIA: Select: 1,3(W) OXAZINO(4W) BENZODIAZEPINE 1,3(W)OXAZINO (4W)BENZODIAZEPIN } In CHEMNAME File Print Registry Number of retrieval citations. Select Registry Number from CASIA file. Limit /PAT. DATE PERFORMED: 1/5/77	(EXPAND OXAZIN AND EXPAND BENZODIAZEP) OR (EXPAND PHENYLOXAZIN AND EXPAND BENZODIAZEP) (all relevant terms were selected from each expand) DATE PERFORMED: 1/14/77	40388 (RR) OR ({[B06-E05(MC) OR C06-E05(MC) OR E06-E05(MC)] OR 26&/B,B1,B2,C,C1,C2,E3(MP)] OR [C07D-099/04(IC)]} AND {373/B,B1,B2,C,C1,C2,E3(MP)} AND STRINGSEARCH :OXAZINO: (TI) AND :BENZODIAZEPIN: (TI)) DATE PERFORMED: 1/7/77
4. 2-ANILINO NICOTINIC ACIDS, INCLUDING NIFLUMIC ACID	CHEMCON: [ANILINO:NICOTIN: OR **ANILINO AND NICOTIN:** OR PHENYLAMINO:-NICOTIN: OR PHENYLAMINO AND NICOTIN: OR PHENYLAMINO:-PYRIDINECARBOXYLIC OR PHENYLAMINO AND PYRIDINECARBOXYLIC OR ANILINOPYRIDINECARBOXYLIC OR ANILINO AND PYRIDINECARBOXYLIC OR NIFLUM:] AND P(UC) CASIA: (1) S HP = 3-PYRIDINECARBOXYLIC AND (2) S HP = 3-PYRIDINE CARBOXAMIDE (3) S PHENYLAMINO (4) 3 AND (1 OR 2) Printed Registry Numbers (RN's) S RN = in CASIA file. Limit PAT. DATE PERFORMED: 1/5/77	(EXPAND ANILINONICOTIN) OR (EXPAND PHENYLAMINONICOTIN) OR (EXPAND TRIFLUOROMETHYLANILINONICOTIN) OR (EXPAND TRIFLUOROMETHYLPHENYLAMINONICOTIN) OR (NIFLUM?) OR {[PHENYLAMINO OR PHENYL(W)AMINO OR ANILINO OR TRIFLUORO(W)METHYL(W)ANILINO OR TRIFLUORO(W)METHYL(W)PHENYL (W)AMINO OR TRIFLUOROMETHYL(W) ANILINO OR TRIFLUOROMETHYL(W) PHENYLAMINO OR TRIFLUOROMETHYLANILINO] AND [EXPAND NICOTIN OR (EXPAND PYRIDIN AND EXPAND CARBOXYL) OR PYRIDINECARBOXYL?]} (all relevant terms were selected from each expand) DATE PERFORMED: 1/14/77	{[B07-D04 OR C07-D04 OR E07-D04 (MC)] AND [(314 AND 373 AND 47&) /B,B1,B2,C,C1,C2,E3(MP)]} AND [(46& OR 46-)/B,B1,B2,C,C1,C2, E3(MP)] DATE PERFORMED: 1/7/77
5. E. MERCK TRANQUILIZERS	CHEMCON. MERCK/CS AND PATENT/CS AND ALL TRANQUIL: AND P (UC)	(AC = 054144 OR AC = 054146) AND TRANQUIL? DATE PERFORMED: 2/17/77	MERE(PC) AND {[B12-C10(MC) OR C12-C10(MC)] OR [600/B, B1,B2,C,C1,C2(MP)]} DATE PERFORMED: 1/7/77
6. U.S. 3,769,282 PATENT FAMILY MEMBERS	CA PATENT CONCORDANCE: EXPAND PN = US3769282 DATE PERFORMED: 1/24/77	PN=US3769282 DATE PERFORMED: 1/24/77	US3769282 (PN) DATE PERFORMED: 1/24/77

* ALL CHEMCON searches were performed on the SDC file.

Table VIII. Search Results

SEARCH #1: BLEOMYCIN, ITS DERIVATIVES AND RADIOISOTOPES **A**

	CHEMCON/CASIA	CLAIMS	WPI
TOTAL HITS	28	11	24
RELEVANT HITS	27	11	24
RELEVANT INVENTIONS	24	9	24
EXCLUSIVE RELEVANT INVENTIONS	6	0	4
RELEVANT INVENTIONS NOT RETRIEVED	6	21	6
EXPLANATION OF WHY RELEVANT INVENTIONS WERE NOT RETRIEVED	5-Too early for file 1-Outside scope of file	20-Outside scope of file (Not issued in U.S.) 1-Retrieval program limitation (Inability to truncate title term to the left)	3-Missing from file (Japanese unexamined) 3-Term not in title ("BLEO")
RELEVANT U.S., 1972-PRESENT	10 of 12	11 of 12	11 of 12

SEARCH #2: BRITISH ALUMINIUM **B**

	CHEMCON/CASIA	CLAIMS	WPI
TOTAL HITS	15	31	46
RELEVANT HITS	15	31	46
RELEVANT INVENTIONS	14	29	46
EXCLUSIVE RELEVANT INVENTIONS	0	22	29
RELEVANT INVENTIONS NOT RETRIEVED	54	39	22
EXPLANATION OF WHY RELEVANT INVENTIONS WERE NOT RETRIEVED	35-Too early for file 16-Outside scope of file 3-Patentee name misspelled ("BRITISH ALUMINUM")	39-Outside scope of file	21-Too early for file 1-Missing from file ("Z-Number")
RELEVANT U.S., 1972-PRESENT	5 of 10*	7 of 10*	10 of 10*

SEARCH #3: 12-PHENYL(1,3)OXAZINO(3,2-d)(1,4)BENZODIAZEPINES **C**

	CHEMCON/CASIA	CLAIMS	WPI
TOTAL HITS	11	4	12
RELEVANT HITS	6	3	11
RELEVANT INVENTIONS	5	3	11
EXCLUSIVE RELEVENT INVENTIONS	1	1	6
RELEVANT INVENTIONS NOT RETRIEVED	8	10	2
EXPLANATION OF WHY RELEVANT INVENTIONS WERE NOT RETRIEVED	4-Indexed elsewhere in file (OXAZOLOBENZODIAZEPINES, BENZODIAZEPINE, DIAZEPINE) 2-Too early for file 1-Missing from file 1-Misspelling of index term ("OXAMINOBENZODIAZEPINE")	6-Outside scope of file (Not issued in U.S.) 4-Terms not in title ("OXAZINO" and/or "BENZODIAZPIN")	2-Term not in title ("OXAZINO")
RELEVANT U.S., 1972-PRESENT	3 of 5	2 of 5	4 of 5

SEARCH #4: 2-ANILINONICOTINIC ACIDS, INCLUDING NIFLUMIC ACID **D**

	CHEMCON/CASIA	CLAIMS	WPI
TOTAL HITS	32	13	203
RELEVANT HITS	28	11	34
RELEVANT INVENTIONS	26	11	34
EXCLUSIVE RELEVANT INVENTIONS	11	2	20
RELEVANT INVENTIONS NOT RETRIEVED	25	40	17
EXPLANATION OF WHY RELEVANT INVENTIONS WERE NOT RETRIEVED	17-Too early for file 5-Indexed elsewhere (ANILINOPYRIDINE, PYRIDINE, AZOPYRIDINE, SALICYLATE, ANTHRANILATE) 2-Outside scope of file 1-Retrieval program limitation	33-Outside scope of file (Not issued in U.S.) 6-Terms searched not in title 1-Retrieval program limitation	13-Indexed elsewhere (Primarily as derivatives) 2-Missing from file 2-Outside scope of file (Spain not covered by WPI)
RELEVANT U.S., 1972-PRESENT	8 of 11	7 of 11	6 of 11

SEARCH #5: E. MERCK TRANQUILIZERS **E**

	CHEMCON/CASIA	CLAIMS	WPI
TOTAL HITS	4	3	33
RELEVANT HITS	3	3	32
RELEVANT INVENTIONS	3	3	32
EXCLUSIVE RELEVENT INVENTIONS	2	1	29
RELEVANT INVENTIONS NOT RETRIEVED	32	32	3
EXPLANATION OF WHY RELEVANT INVENTIONS WERE NOT RETRIEVED	20-Indexed elsewhere in file (Not indexed to tranquilizers) 12-Too early for file	27-Outside scope of file (Not issued in U.S.) 5-Terms not in title (Tranquilizer not in title terms)	3-Indexed elsewhere (Not indexed to tranquilizers)
RELEVANT U.S., 1972-PRESENT	0 of 6	3 of 6	5 of 6

SEARCH #6: U.S. 3,769,282 PATENT FAMILY **F**

	CA PATENT CONCORDANCE	CLAIMS	WPI
TOTAL HITS	6	4	10
RELEVANT HITS	6	4	10
EXCLUSIVE RELEVANT HITS	0	0	4
RELEVANT HITS NOT RETRIEVED	4	6	0
EXPLANATION OF WHY RELEVANT HITS WERE NOT RETRIEVED	4-Outside scope of file	4-Outside scope of file (Countries not covered) 2-Missing from file	

* 3 of 10 possible U.S. are non-chemical

DESCRIPTION OF PARALLEL SEARCHES

A total of six questions was employed in the study (see Table VI). The questions were hypothetical ones which the authors considered to be representive of the types of questions which are posed in the industrial/pharmaceutical environment in which they work. This environmental bias was considered an unavoidable factor since it enabled the authors to draw upon their areas of expertise in preparing search strategies and in analyzing the results.

SEARCH STRATEGIES AND PERFORMANCE DATES

The search strategies used to answer each of the six questions in the three databases are presented in Table VII. The Boolean logic used in these searches is typical of the search systems provided by most on-line brokers. Search strategies are not necessarily exhaustive; the object of the searches, as with all on-line searches, was to be cost effective.

At the end of each strategy, the date the search was performed is indicated. Since each of the databases is regularly updated, the performance date has a direct bearing on the size and content of the file at the time of the search.

The use made of CHEMCON's companion files, CASIA and CA Patent Concordance, is noted in the CHEMCON column of Table VII. Hereafter, when reference is made to the CHEMCON search results, it will be assumed that the answers contributed by CHEMCON's companion files are included.

SEARCH RESULTS

The bibliographic citations which resulted from the six searches were screened for relevancy by one or more of the authors by means of the corresponding hard copy abstract publications, i.e., CHEMCON via Chemical Abstracts, CLAIMS via the U.S. Patent Office Official Gazette, and WPI via Central Patents Index Basic Abstract Journal. These hard copy sources reflect the different meaning that each database ascribes to the word "citation": CAS and Derwent consider it to be the first-issuing patent in a family of equivalent patents, both U.S. and foreign, whereas IFI/Plenum considers each U.S. patent to be a citation, regardless of whether it is a continuation or divisional of another U.S. patent. These differing interpretations of the concept of a citation are a potential source for confusion which must be kept in mind when analyzing the search results.

The results of the six searches are tabulated in Tables VIII.A through VIII.F. The authors employed the following conventions in totaling the search results in Tables VIII.A–E:

Total hits is simply the total number of bibliographic citations which were produced as search output for a given search by each of the databases.

Relevant hits is the number of citations considered to be relevant, based on a review of the corresponding abstract.

Relevant inventions is the number of relevant hits minus "duplicate" hits. These "duplicate" hits are equivalents (i.e., members of the same patent families) of other hits in the search output.

Exclusive relevant inventions is the number of relevant inventions found in that particular database only.

Relevant inventions not retrieved is calculated by determining the total number of distinct inventions (i.e., individual patent families) found collectively by the three databases and then substracting relevant inventions.

Explanation of why relevant inventions were not retrieved is the authors' determination of the reasons why relevant inventions were not retrieved by the search.

Relevant U.S., 1972–present is calculated for the sole purpose of comparing the only area of document coverage having a high degree of overlap between all three databases. The figures indicate the number of relevant U.S. patents found out of the total number of relevant U.S. patents issued during the period, 1972–present.

The authors employed a different set of conventions in analyzing the answers to the equivalents search (Table VIII.F). In this case, the object of the search was to find as many members of a single patent family as possible. Thus, all answers pertain to one invention only. In this context, "total hits", "relevant hits", and "exclusive relevant hits" refer to family members found while "relevant hits not retrieved" is the number of known family members not found. For the purposes of this comparison, the set of total known patent family members was restricted to patents found only in the three databases under consideration.

DISCUSSION OF SEARCH RESULTS

The obvious fact which emerges from the parallel search results is that no one database found all of the answers all of the time. This is shown clearly in Table IX which focuses on the number of relevant answers found in each database. (Answers in this context mean relevant inventions for searches 1–5 and relevant patent family members for search 6.) In general WPI achieved the best results. This is due to WPI's combination of broad country coverage and high-specificity retrieval parameters such as company codes and chemical fragment codes which provided superior results in these instances. CHEMCON achieved high results in searches 1 and 4 because it contains as index terms the names of specific compounds (BLEOMYCIN) and chemical fragments (NICOTINIC, NIFLUMIC, etc.). CLAIMS failed to surpass its competitors in any of the searches because of its limited country coverage and its heavy reliance on title terms for subject matter retrieval. Search 6, the equivalents search, seems to indicate the superior coverage of WPI, since only WPI found all ten patent family members.

Table IX. Summary of Search Results

	SEARCH #1	SEARCH #2	SEARCH #3	SEARCH #4	SEARCH #5	SEARCH #6
POSSIBLE RELEVANT ANSWERS	30	68	13	51	35	10
CHEMCON	24	14	5	26	3	6
CLAIMS	9	31	3	11	3	4
WPI	24	46	11	34	32	10

Table X. Exclusive Relevant Answers

	SEARCH #1	SEARCH #2	SEARCH #3	SEARCH #4	SEARCH #5	SEARCH #6
CHEMCON	6	0	1	11	2	0
CLAIMS	0	22	1	2	1	0
WPI	4	29	6	20	29	4

Another way of looking at the search results is to focus on the relevant answers which were retrieved by only one of the three on-line databases. These so-called "Exclusive Relevant Answers" are an indication of how many answers would have been missed if that particular database were not searched. A summary of exclusive relevant answers is shown in Table X. Although WPI generally supplies the highest number of exclusive relevant answers, it is readily apparent that CHEMCON and CLAIMS are also quite capable of providing exclusive relevant answers. The ability of each database to provide unique sets of answers is a direct consequence of their differences in document coverage and retrieval parameters.

Precision and recall ratios have not been calculated since to do so requires that a single meaning be applied to the word "citation" so that the results from each database are directly comparable to the others. Whichever definition of "citation" is chosen will necessarily enhance the Precision and Recall figures for one database to the detriment of the others.

Moreover, to present a rigorous comparison of retrieval capability, it would be necessary to restrict the comparison to the only area of overlap between the databases, namely the U.S. Chemical patents, 1972 to present. This area of overlap will be examined below, but to focus on it exclusively would drastically limit the size of the files being compared, and would fail to reflect the coverage advantage enjoyed by each of the systems, e.g., CHEMCON's and WPI's broad foreign coverage and CLAIMS' coverage back to 1950. The authors prefer the approach of highlighting the database differences by discussing the relevant answers which each of the databases failed to retrieve.

A summary of answers which were missed in all six searches is presented categorically in Table XI. Keeping in mind that these figures are directly dependent on the questions which were posed, there are a few obvious inferences which may be drawn. The first category—Too Early For File—reaffirms what was determined by the coverage comparison, i.e., that CHEMCON and WPI are inferior to CLAIMS if older patents are to be included in the search. Similarly, the second category shows how a lack of foreign coverage results in CLAIMS missing many answers which are outside the scope of its coverage. CHEMCON also missed a number of answers owing to its exclusion of patents on the fringe of chemistry. A more serious problem is the third category of answers which ought to have been in the file based on the boundaries of its coverage, but are missing. All three files are guilty of such lapses. The fourth category of answers, those which were missed due to title terms not being present (only CLAIMS and WPI were searched using this parameter), serves to

Table XI. Summary of Missed Answers

CAUSE OF MISSED ANSWERS	ON-LINE FILE		
	CHEMCON (1972-Present)	CLAIMS	WPI
1. Too Early For File	71	0	21
2. Outside Scope Of File	23	129	2
3. Missing From File	1	2	6
4. Terms Not In Title	0	15	5
5. Indexed Elsewhere In File	29	0	16
6. Misspelled Retrieval Parameter	4	0	0
7. Retrieval Program Limitation	1	2	0
TOTALS	129	148	50

Table XII. Search Results–U.S. Patents, 1972–Present Only

	SEARCH #1	SEARCH #2	SEARCH #3	SEARCH #4	SEARCH #5
POSSIBLE RELEVANT U.S. PATENTS	12	10*	5	11	6
CHEMCON	10	3	3	8	0
CLAIMS	11	7	2	7	3
WPI	11	10	4	6	5

* 3 of 10 are non-chemical.

underscore the danger in relying upon title terms for thorough searching. The fifth category reflects the searchers' inability to predict all of the retrieval parameters associated with the relevant hits. (In some of the cases for CHEMCON this was due to incomplete indexing of the chemical compounds.) The sixth category indicates that CHEMCON is the only one of the three files to exhibit the human errors of misspelled retrieval terms. (Obviously, the authors are not convinced on this limited evidence that human errors are absent in CLAIMS and WPI!) The final missed answers result from the fact that left-hand truncation was not available for the CLAIMS and CHEMCON files; otherwise, the terms embedded in the titles might have been used to retrieve the patents which were missed.

DISCUSSION OF SEARCH RESULTS FOR 1972–PRESENT U.S. PATENTS ONLY

Coverage differences can be eliminated theoretically by examining only those search results which fall in the overlap area between the three on-line files. As mentioned above, this overlap area consists of U.S. patents issued since 1972.

Table XII indicates the relevant U.S. patents which were found in each of the first five searches. Search 6 is excluded for the obvious reason that the only relevant U.S. patent was known a priori. The obvious inference which may be drawn from Table XII is that when only U.S. patents are sought CLAIMS is now on a roughly equal footing with the other two on-line files. Except for WPI in search 2, none of the databases are able to supply all of the answers to any of the searches.

Table XIII. Exclusive Relevant U.S. Patents, 1972–Present

	SEARCH #1	SEARCH #2	SEARCH #3	SEARCH #4	SEARCH #5
CHEMCON	0	0	1	1	0
CLAIMS	0	0	1	2	1
WPI	0	3*	2	4	3

* Non-chemical

Table XIV. Summary of Missed U.S. Patents Issued since 1972. Searches 1 through 5

CAUSE OF MISSED ANSWERS	ON-LINE FILE		
	CHEMCON	CLAIMS	WPI
1. Terms Not In Title	0	9	2
2. Indexed Elsewhere In File	11	0	6
3. Retrieval Program Limitation	0	2	0
4. U.S. Cited As Equivalent To Pre-1972 Non-U.S. Patents	4	0	0
TOTALS	15	11	8

The number of exclusive answers found by the databases is shown in Table XIII. Although overall there are fewer answers which can be found only in one database, there are still cases (searches 3 and 4) where all three on-line files must be searched to provide a complete list of answers.

A summary of the causes for 1972–present U.S. patents being missed by each database is presented in Table XIV. (These totals cover the first five searches only.) Note that a new category of missed answers has been added to accommodate the indexing practice of CHEMCON and WPI which do not reindex U.S. patents which are equivalent to earlier issued non-U.S. patents, but instead cite them only as equivalents. The inferences which can be drawn from Table XIV are not explained since they follow the same patterns as those drawn in Table XI. The smaller number of missed answers is a natural consequence of focusing on a smaller file.

CONCLUSIONS

The authors have briefly surveyed the patent-information-containing databases which are currently available on-line and have compared the coverage and retrieval capabilities of the three on-line files which offer the broadest coverage of chemistry: CHEMCON, CLAIMS, and WPI. It was found that none of these was superior to the others in all aspects of patent searching. Moreover, as the parallel search results indicate, no one database can provide all the answers all the time. Each of the three is capable of supplying unique relevant answers not found in the other two. The more thorough a search you wish to perform, the more on-line files should be consulted. Thus, the judgment of the user must be exercised when selecting the on-line file (or files) to be searched.

Certainly not to be neglected are the other databases on our list which cover more specialized areas of technology. Users are encouraged to try those which purport to cover the year ranges, countries, and subject areas which are germane to their questions.

A final characteristic of on-line systems which should be emphasized is the speed with which the literature on on-line retrieval becomes outdated. Improvements are taking place so fast (in databases, search programs, and communications techniques and equipment) that there is little which can be said now which will remain unchanged for more than a few months. Users are advised to stay in close contact with the database producers, on-line brokers, and communications vendors to keep abreast of the changes.

LITERATURE CITED

(1) G. V. O'Bleness and G. Cohen, "Presentation of "SOLD" Computer Retrieval System", paper presented at Central Patents Index Subscribers Meeting, Washington, D.C., May 1973.

(2) R. Donati (Lockheed Information Systems), "Overview of Patents Covered by DIALOG Retrieval Service", presented at Session on Patent Literature Systems from the Symposium on Intellectual Property as Sources for Scientific, Technical and Business Information sponsored by the Canadian Patent Office and others, Ottawa, Ontario, Nov 24, 1976.

(3) M. Bonner (System Development Corporation), private communication, Dec 1976.

(4) M. E. Williams and S. H. Rouse, Ed., "Computer-Readable Bibliographic Data Bases - A Directory And Data Sourcebook", American Society for Information Science, Washington, D.C., 1976.

(5) J. B. Hare, "On-Line Searching for Patent Information—A Comparison Of The CHEMCON And WPI Data Bases", paper presented at Pharmaceutical Manufacturers Association Science Information Subsection Meeting, Hot Springs, Va., March 1976.

An Interactive Substructure Search System

R. J. FELDMANN and G. W. A. MILNE

National Institutes of Health, Bethesda, Maryland 20014

S. R. HELLER*

Environmental Protection Agency, Washington, D.C. 20460

A. FEIN, J. A. MILLER, and B. KOCH

Fein-Marquart Associates Inc., Towson, Maryland 21212

Received April 29, 1977

A family of programs for searching on the basis of chemical structure through data bases of chemical information has been assembled and is now publically available on a commercial computer network. The design of and results obtained with these programs are reported, and the status of the system is described and discussed with particular reference to the NIH–EPA Chemical Information System (CIS) and the Toxic Substances Control Act (TSCA).

INTRODUCTION

The ability to use a computer to search for a particular chemical structure or substructure in files of chemical data has for some time been sought after by chemists, and the need for such capability is currently becoming very pressing. A widening interest in the relationships between chemical structure, on the one hand, and various properties, such as toxicity, pharmacological activity, and mutagenicity, on the other, has led in recent years to considerable efforts to generate computer programs which will enable the scientist to locate all occurrences of a given structure or substructure in chemical databases. Further decisive pressure behind these developments has been provided by the enactment, in November 1976, of the Toxic Substances Control Act (Public Law 94-469). This law will require that chemical compounds whose use in commerce is envisaged must first be located within Governmental regulatory files. If they are not in these files, they are deemed "new" and their use in commerce becomes subject to a series of regulations, depending upon their respective toxicities.

In this paper, we describe the NIH–EPA substructure search system, a family of interactive computer programs which allow the user to define a chemical structure or substructure and then to search for occurrences of the structure or substructure in the various databases of the NIH–EPA Chemical Information System.

During the past 25 years, a considerable number of methods of machine representation and handling of chemical structure have been proposed and studied for their utility in manual and automatic data retrieval methods. Some of the better known among these include the German GREMAS system,[1] the British CROSSBOW system,[2] and, in the U.S., the programs developed at the National Cancer Institute,[3] Walter Reed Army Institute of Research,[4] Chemical Abstracts Service,[5] and the Army's Chemical Information Data System.[6]

With the resulting progress in the area of computer-handling of chemical structures, it has become clear that structure records in the form of two-dimensional connection tables are absolutely necessary for structural representation and that both linear notations and chemical nomenclature are at a serious disadvantage vis-â-vis connection tables as far as unambiguity and completeness are concerned. For many years, however, there was no adequate means of screening such connection tables and so, in spite of their intrinsic value, they were not used in any retrieval system.

In the area of structure retrieval, most effort was expended in the development of systems that were designed to fulfill a specific local need. A system of this type that is currently perhaps the most widely used by the chemical industry is the CROSSBOW program,[2] a dozen or so versions of which are in operation around the world. Other systems, such as the GREMAS system[1] or the Walter Reed system[4] require special equipment that is not generally available. While the larger industrial organizations can often afford such luxuries and often also demand in-house facilities of this sort, such systems are of little value to the general chemist. It is this dilemma which led to the development of the NIH nested tree structure searching system, which can operate on a connection table database and which is susceptible to wide dissemination and