Both MARPAT and M-DARC have advantages and disadvantages. The choice of which system to use will depend on the particular problem at hand. An important component in this decision is reported on in this issue by Kathy Cloutier;[12] and that component is database content and indexing policies.

It seems to be a part of the American personality trait to ask but which is the best? If I were pressed to decide which of the systems represents the "cutting edge", the conclusion seems inescapable that, at the present time MARPAT is the more mature product from a software viewpoint. This judgement is based primarily on the user-specifiable selective query translation, and MARPAT's integration with the other STN structure and bibliographic files.

REFERENCES AND NOTES

(1) Meurling, A. CAS Online and DARC: A comparison. *Database* **1990**, *Feb*, 54-63.

(2) Warr, W. A.; Wilkins, M. P. Graphics front ends for chemical searching and a look at Chemtalk Plus. *Online* **1990**, *May*, 50-4.

(3) Brueggman, P. Creating chemical structures for online searching with Molkick. *Database Searcher* **1989**, *May*, 22-7.

(4) *Workshop Manual*: *WPI*, *Markush DARC*, Feb 1989 ed.; Derwent Publications Limited: London, 1989.

(5) Fisanick, W. The Chemical Abstract's Service generic chemical (Markush) structure storage and retrieval capability. 1. Basic concepts. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 145-54.

(6) Dubois, J. E.; Laurent, D.; Viellard, H. DARC system. Polymatrix structural description. Writing of formal matrices. *C. R. Hebd. Seances Acad. Sci., Ser. C* **1966**, *263*, 1245-8.

(7) Dubois, J. E.; Laurent, D.; Viellard, H. System of documentation and automation of correlation researches. General principles. *C. R. Hebd. Seances Acad. Sci., Ser. C* **1966**, *263*, 764-7.

(8) Dubois, J. E.; Laurent, D. DARC (documentation and automation of correlation research) system. Population-correlation theory, organization, and description. *C. R. Acad. Sci., Paris, Ser. C* **1968**, *266*, 943-5.

(9) Dubois, J. E.; Mathieu, G.; Peguet, P.; Panaye, A.; Doucet, J. P. Simulation of infrared spectra: an infrared spectral simulation program (SIRS) which uses DARC topological substructures. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 290-302.

(10) Attias, R.; Dubois, J. E. Substructure systems: concepts and classifications. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 2-7.

(11) Dittmar, P. G.; Farmer, N. A.; Fisanick, W.; Haines, R. C.; Mockus, J. The CAS ONLINE search system. 1. General system design and selection, generation, and use of search screens. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 93-102.

(12) Cloutier, K. A Comparison of Three Markush Databases. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 40-44.

The PHARMSEARCH Database[†]

MICHAEL P. O'HARA*

Questel, Inc., 5201 Leesburg Pike, Suite 603, Falls Church, Virginia 22041

CATHERINE PAGIS

INPI, 26 Bis rue de Leningrad, 75008 Paris, France

Received December 19, 1990

PHARMSEARCH, a database produced by the French Patent and Trademark Office (INPI), covers pharmaceutical patents issued by the European, French, and United States patent offices from November 1986 onward. PHARMSEARCH is composed of MPHARM, a structure file searchable using Markush DARC software, and PHARM, the companion bibliographic file. Markush structures claimed in the patent documents are entered into the database as variable generic structures. Specific structures are also included in the database, when they are not part of a Markush structure in the patent document. Chemical index terms describe all moieties of the structure. Indexing also describes the therapeutic activities and preparation processes for the compounds. The indexing policies used in the production of this database are described.

INTRODUCTION

PHARMSEARCH is a pharmaceutical patents database which covers pharmaceutical patents issued by the European, French, and United States patent offices. PHARMSEARCH actually consists of two files: MPHARM, which is a Markush structure file, and PHARM, which is a companion bibliographic file. PHARMSEARCH originally began as a publication in paper form. In 1983, it was acquired by INPI, the French Patent and Trademark Office. After acquiring PHARMSEARCH, INPI began a study to develop automated processes for the production of the publication. As a result of this study, INPI decided to use a graphics structure database to index and record the chemical structures which occur in the pharmaceutical patents. In 1984, INPI joined the Markush DARC development that was already under way at Telesystèmes, the parent of Questel. The actual building of the graphics database began in 1986. The online database was opened to the public in January 1989, making PHARMSEARCH the first Markush structure database to become available for online searching. The searching methodology for Markush structure systems is described in the paper by John Barnard in this issue.[1] This paper focuses on the indexing considerations used in the preparation of this unique database.

PATENT INDEXING CONSIDERATIONS

When preparing indexing rules for patents databases there are two considerations for which the rules must account: the levels of information which are contained in the actual patent documents and the patent law requirements. There are three levels of information in patent documents: a general level, a preferred level, and a specific level. In the actual patent documents, the general level of information is contained in the description and in the independent claims; the preferred level and the specific level of information are described in the description and in the dependent claims. Of these three levels of information in the patent documents, patent law requires consideration of the general and the specific levels of information.

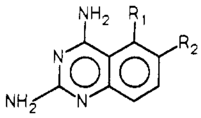
Pharmaceutical patents generally are searched for two basic types of information: the chemical compounds involved and

[†] Presented at the Division of Chemical Information, 200th National Meeting of the American Chemical Society, Aug 29, 1990.

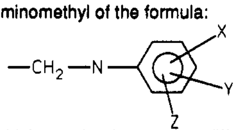
[*] Author to whom correspondence should be addressed.

0095-2338/91/1631-0059$02.50/0 © 1991 American Chemical Society

General Information (Claim 1)



R1 is alkyl of from one to three carbon atoms and R2 is bromine, chlorine, cyano or arylaminomethyl of the formula:



X, Y and Z, which may be the same or different, are hydrogen, halogen, lower alkoxy or trifluoromethyl.

Specifically described compounds in the patent document:

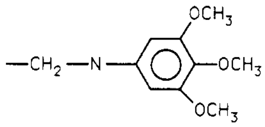
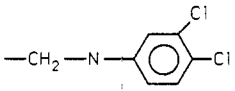
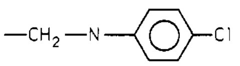
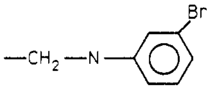
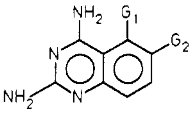
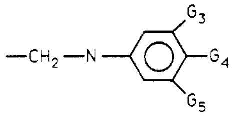
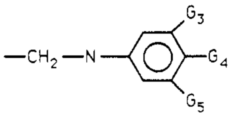
Compound Number	R_1	R_2
1	CH_3	CN
2	CH_3	Br
3	CH_3	Cl
4	CH_3	
5	CH_3	
6	CH_3	
7	CH_3	

Figure 1. Structure information in EP 253396.

Main Group (Group 0) Indexing



Variable Group Indexing

Group	Specific Values	Generic Values
G1	C (1 to 7)	CHK
G2	CN, Br, Cl,  (1) (2) (3)	
G3	OCH3, Cl, Br, H (4) (5) (7) (6)	- O - CHK, HAL
G4	OCH3, Cl, H (4) (5) (7)	- O - CHK, HAL
G5	OCH3, H (4) (5,6,7)	- O - CHK

Attributes

G1 CHK(1) LO

Text Notes

G1 CHK(1) = C1-3

Figure 2. Structure indexing in EP 253396.

the pharmaceutical aspects related to these compounds. The indexing in a pharmaceutical patents database must, therefore, consider these two information types. The indexing for PHARMSEARCH is performed by a staff of 30 people. In order to ensure knowledge of the subject area, INPI uses only trained pharmacists or pharmaceutical chemists for indexing. All of the indexing staff have also received extensive training at INPI to develop a thorough intellectual property orientation.

PHARMSEARCH INDEXING

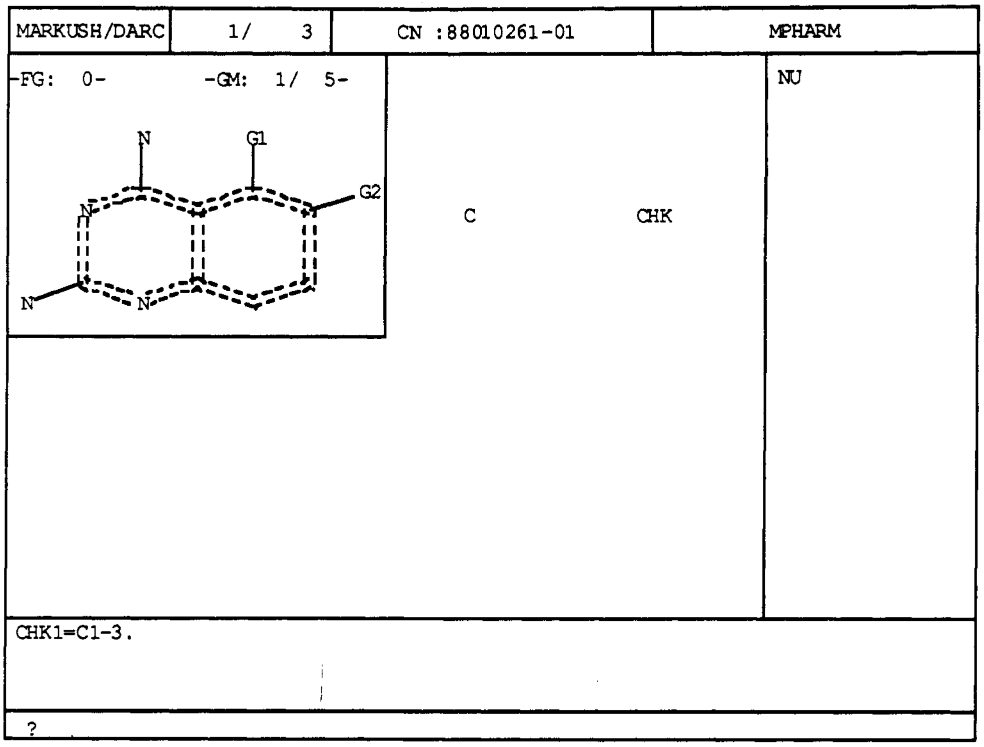
Chemical substances are indexed in PHARMSEARCH when they are claimed and/or described as new therapeutically active compounds, new synthesis intermediates, part of new pharmaceutical compositions, prepared by new processes, or have new therapeutic uses.

When a chemical substance meets these criteria, the structure of the substance is entered into the MPHARM Markush structure file. The chemical substance is also indexed in the companion PHARM bibliographic file. In PHARM, nomenclature terms are used to describe the basic chemical, the chemical functions, the oxidation state, the fixed groups, and the variable groups.

An important PHARMSEARCH indexing consideration for chemical substances is that PHARMSEARCH indexes generic terms, such as alkyl, only when the patent document includes a specific representation corresponding to this generic term. The pharmaceutical aspects of the invention are also covered in PHARM. These pharmaceutical aspects include the composition (oral, tablet, sustained release), the formulation, the process used in the preparation, the claimed effects, any analogous effects, any drug interactions, any toxic and side effects, the therapeutic classes, and the process classes.

CHEMICAL SUBSTANCE INDEXING

Markush structures in the patent documents are entered into the database as variable generic structures. Specific structures are also included in the database when they are not part of a Markush structure in the patent document. The indexing of chemical substances is best understood by looking at an example which illustrates the indexing process. Figure 1 lists the general substance information from claim 1 in EP patent 253396 as well as the seven specifically described substances in the patent document. In all of the specific examples, only methyl was illustrated as a value for group R_1. In the actual indexing process, the indexer first prepares a representation of the basic substance structure with R groups replaced by G groups (the symbol used within Markush DARC to represent variable groups). This representation is illustrated in Figure 2. The indexer then examines the patent document for the specific substances. The indexer then lists the specific values which correspond to the variables in the original structure. These are listed in Figure 2 in the specific value column with the number label for the specific substance as listed in Figure 1. The next step for the indexer is to list the generic terms (superatoms or groups) which correspond to these specific values. These terms are listed in the generic values column of Figure 2. Note only those generic values which have a corresponding specific value are listed. For group G5, for example, HAL (the halogen superatom) is not listed as a generic value since there was no specific example of a halogen at this position in the patent document. The final indexing step is for the indexer to list any attributes or text notes which should be associated with this structure. In this example, the CHK (the alkyl superatom) in group G1 is described to be a 1–3 carbon alkyl. The indexer therefore, lists the attribute of LO (for low carbon count). In order to convey the specific information that the document describes a carbon count of 1–3, the indexer also adds a text note of "C1–3" to this



```
MARKUSH/DARC   1/   3   CN :88010261-01            MPHARM
-FG:  0-        -GM:  1/  5-                       NU

                         C              CHK

CHK1=C1-3.

?
```

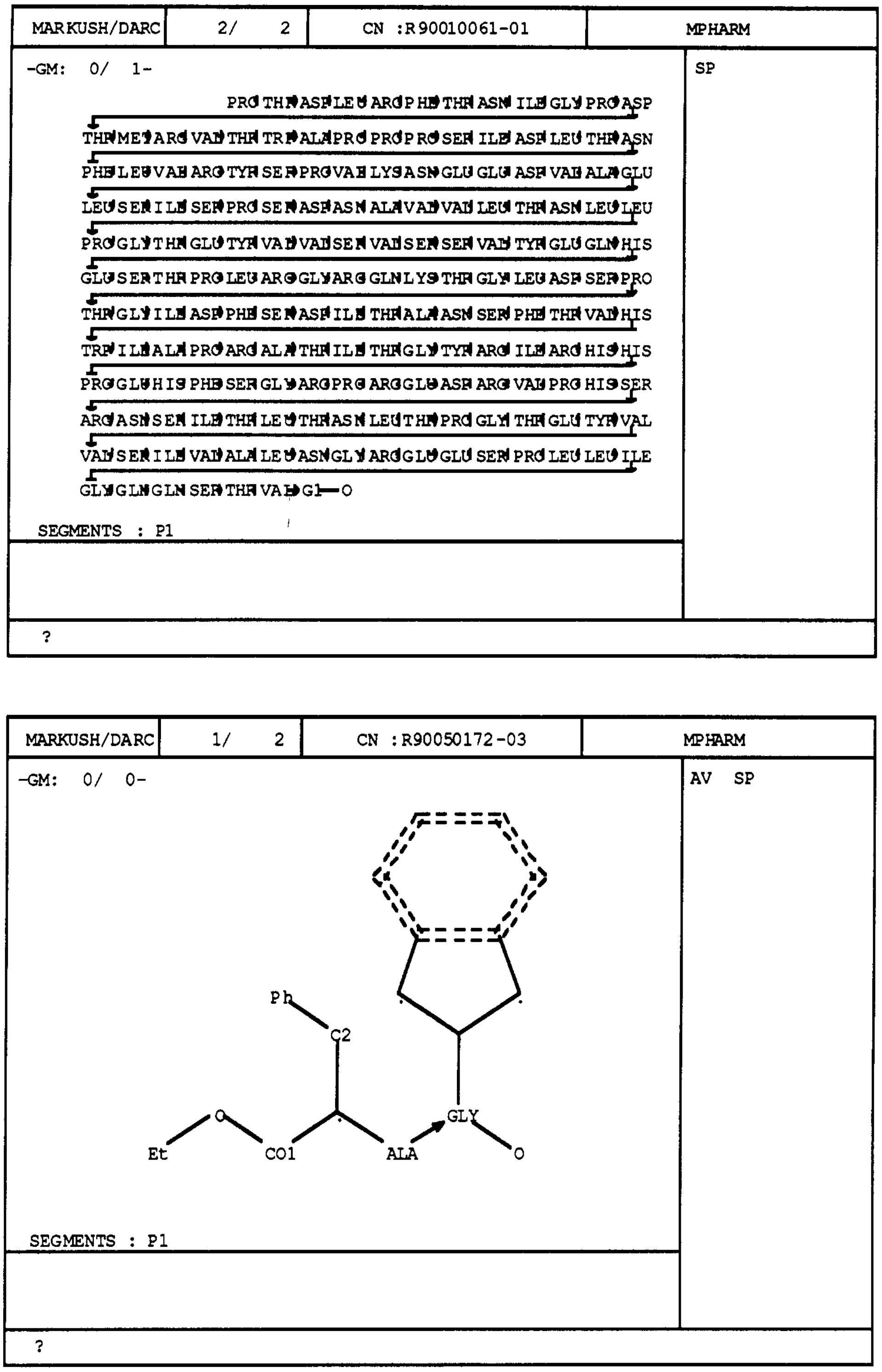
Figure 3. MPHARM structure index record for EP 253396.

1/1 - (C) INPI
AN - 88010261
CN - 88010261-01; 88010261-02; 88010261-03
PN - ***EP253396*** - 880120
AP - EP87110310 870716
PR - US88646386 860717
PA - WARNER-LAMBERT COMPANY
PAC - US
IC - C07D-239/95; C07C-121/52; C07C-121/54
EAB - Process for preparing 5-alkyl-6-(bromo, chloro, cyano or phenylaminomethyl)-2,4-quinazolinediamine derivatives which comprises: a) nitrating 2,6-(dibromo or dichloro)-1-alkylbenzene derivatives, b) converting the obtained 2,6-dihalo-3-nitrobenzenes by the action of copper cyanide into the 2-mono or 2,6-dicyano compounds, c) reducing them into the corresponding amines and finally d) reacting the last products with chloroformamidine hydrochloride at a temperature of between 150C and 250C, followed by isolation and purification of the desired compounds. The intermediates of formula 1-(bromo, chloro or cyano)-2-alkyl-3-cyano-4-(amino or nitro)benzene are also disclosed

. Final products are folic acid antagonists useful for treating

malaria and tumor
IT - INTERMEDIATE; QUINAZOLINE; AMINOQUINAZOLINE; DIAMINOQUINAZOLINE; METHYLQUINAZOLINE; BROMOQUINAZOLINE; CHLOROQUINAZOLINE; CYANOQUINAZOLINE; AMINOMETHYLQUINAZOLINE; PHENYLAMINOMETHYLQUINAZOLINE; METHOXYPHENYLAMINOMETHYLQUINAZOLINE; TRIMETHOXYPHENYLAMINOMETHYLQUINAZOLINE; CHLOROPHENYLAMINOMETHYLQUINAZOLINE; DICHLOROPHENYLAMINOMETHYLQUINAZOLINE; BROMOPHENYLAMINOMETHYLQUINAZOLINE; CYANOBENZENE;DICYANOBENZENE; AMINOBENZENE; TOLUENE; NITROBENZENE; CHLOROBENZENE; BROMOBENZENE; TRIMETREXATE
PROC- SYNTHESIS PROCESS; SYN
EFF - CLAIMED EFFECT; CLEF; FOLIC ACID ANTAGONIST; MALARIA; TUMOR; PALUDISM; CANCER
PHC - 21; 03; 15

Figure 4. Bibliographic record for EP 253396.



```
MARKUSH/DARC   2/   2   CN :R90010061-01   MPHARM
-GM:  0/  1-                                SP
          PRO THR ASP LEU ARG PHE THR ASN ILE GLY PRO ASP
  THR MET ARG VAL THR TRP ALA PRO PRO PRO SER ILE ASP LEU THR ASN
  PHE LEU VAL ARG TYR SER PRO VAL LYS ASN GLU GLU ASP VAL ALA GLU
  LEU SER ILE SER PRO SER ASP ASN ALA VAL VAL LEU THR ASN LEU LEU
  PRO GLY THR GLU TYR VAL VAL SER VAL SER SER VAL TYR GLU GLN HIS
  GLU SER THR PRO LEU ARG GLY ARG GLN LYS THR GLY LEU ASP SER PRO
  THR GLY ILE ASP PHE SER ASP ILE THR ALA ASN SER PHE THR VAL HIS
  TRP ILE ALA PRO ARG ALA THR ILE THR GLY TYR ARG ILE ARG HIS HIS
  PRO GLU HIS PHE SER GLY ARG PRO ARG GLU ASP ARG VAL PRO HIS SER
  ARG ASN SER ILE THR LEU THR ASN LEU THR PRO GLY THR GLU TYR VAL
  VAL SER ILE VAL ALA LEU ASN GLY ARG GLU GLU SER PRO LEU LEU ILE
  GLY GLN GLN SER THR VAL G1—O
SEGMENTS  : P1
?
```

```
MARKUSH/DARC   1/   2   CN :R90050172-03   MPHARM
-GM:  0/  0-                                AV  SP
SEGMENTS  : P1
?
```

Figure 5. Peptide structure indexing in MPHARM.

structure record. Figure 3 is the actual Markush DARC structure record for group G1 for this substance.

The chemical substance information is also indexed by using nomenclature terms in the IT (Index Terms) field of the PHARM bibliographic file. Figure 4 is the PHARM record for this patent document.

PEPTIDE INDEXING

Peptides containing two or more peptide linkages are indexed in MPHARM by using 30 peptide superatoms. The peptide superatoms are symbols such as ALA for alanine, Gly for glycine, etc. The peptide superatoms are linked by oriented peptide "bonds". The peptide superatoms may also be substituted as shown in Figure 5.

PROCESS AND COMPOSITION INDEXING

The process and composition information given in the patent document are indexed in natural language terms in the Index Terms field of the PHARM file.

For example, in EP 266042, claim 1 states the following: "A process for the preparation of a desired substantially pure (6R or 6S) diastereoisomer of a derivative of tetrahydrofolic acid or a salt or ester thereof which process comprises the steps of ..." and claim 4 reads "A process according to any one of claims 1 to 3, wherein the derivative of tetrahydrofolic acid or salt or ester thereof obtained is a substantially pure (6S) diastereoisomer of leucovorin (5-formyltetrahydrofolic acid) or salt or ester thereof". To convey this information, the PHARM indexing for this patent document includes the terms

Table I. Patent Coverage in PHARMSEARCH (as of Nov 29, 1990)

year	structures (week)	peptides
1986	8644–8652	none
1987	8701–8754	none
1988	8801–8852	none
1989	8901–8952	none
1990	9001–9043	9001–9043
BSM	1961	1961

"leucovorin; purification; optical resolution and tetrahydrofolic".

Another example illustrating formulation information is from EP 221732. In this patent document, claim 1 reads "A sustained release pharmaceutical formulation in tablet unit dosage form which ..." and claim 2 reads "A formulation of claim 1 wherein the active agent is fenoprofen calcium". The PHARM indexing for this document includes the terms "oral; tablet; sustained release; fenoprofen; calcium salt".

EFFECTS INDEXING

During the examination of the patent document, the indexer notes all claimed effects, analogous effects, drug interactions, or toxic and side effects. This information may appear in any part of the patent document. The effects are described in the effects field in the PHARM bibliographic file. Within this field, the type of effect is listed in the spelled out version, followed by a four-letter acronym for the type of effect, followed by the effect as described in the patent document. The inclusion of the four-letter acronym provides a unique search capability. One can, for example, identify all claimed effects for a class of chemical compounds by first performing a structure search in MPHARM and then searching for the presence of the acronym "CLEF" (the acronym for Claimed Effect) in the effects field for all documents which cite compounds from this set. Examples of the indexing of each of the types of effect follows.

Claimed Effects Indexing. In EP 208420, claim 12 reads "A pharmaceutical composition suitable for the therapy of a mammal suffering from diabetes and/or hyperlipemia which contains as the effective component a thiazolidinedione derivative of the general formula ...". For this document, the indexing in the effects field is "Claimed Effect; CLEF; hyperlipemia; diabetes".

Analogous Effects Indexing. In EP 264232, the description of experiment 1, an in vivo test in rat, includes the following statement: "The results were expressed as compared with 100 of the inhibition rate of the same amount of ticlopidine (positive control) as that of the test drug". For this document, the indexing in the effects field is "Similar Effect; ANEF; ticlopidine".

Drug Interaction Indexing. In EP 207193, which is in German, one of the three official languages of the European Patent Office, the title reads: Synergistische kombination van Flupirtin und 4-acetamide-phenol. The abstract includes the statement: "Arzneimittel mit synergistischer wirkung, enthaltend eine kombination des analgetikums Flupirtin mit Paracetamol". For this document, the indexing in the effects field is "Drug Interaction; DINT; synergism; paracetamol; flupirtine".

Toxic and Side-Effects Indexing. In U.S. 4639466, the abstract includes the statement: "Furthermore, these compounds have been found to be potent antidotes of avermectin". The indexing in the effects field for this document reads "Toxic and side effect; TOXI; avermectin detoxification".

TIME COVERAGE AND SCHEDULE

The PHARMSEARCH database covers pharmaceutical patent documents from the European, French, and United States patent offices. Table I lists the time coverage in the database as of November 29, 1990. The weeks noted in the table are the weeks of publication of the patent documents. BSM patents are special medicament patents which were issued by the French patent office. The indexing schedule for this database is such that by the end of the first quarter of 1991, the EPO and French patents will be in the database within 6–7 weeks of publication and the U.S. patents within 14 weeks of publication. Peptides patents issued from the beginning of 1990 are currently in the database. Peptide patents from weeks 8751 to 8835 will be added to the database by the end of 1990. Peptide patents from weeks 8836 to 8903 are planned to be added by mid-1991 and from weeks 8904 to 8952 by the end of 1991. By the end of 1991, the schecule calls for all peptide patents to be covered in the database. This will make PHARMSEARCH a single source for accessing information on generic patents. Once the currency goals are obtained and the backlog of peptide patents has been indexed, INPI plans to begin indexing of the backfile of pharmaceutical patents from the three offices. PHARMSEARCH, when complete, will cover EPO and U.S. pharmaceutical patents from 1978 to the current time and French patents pharmaceutical patents from 1961.

SUMMARY

PHARMSEARCH is a pharmaceutical patent database covering patents from the European, French, and United States patent offices. The database is accessible both by Markush structure and by pharmaceutical indexing. The indexing is done by experts in the field of pharmaceutical chemistry with an intellectual property orientation. The indexing conveys both the general and specific information from the patent documents. The database is one of the most current of the indexed patent databases and provides early access to pharmaceutical patents in a cost-effective manner.

REFERENCES AND NOTES

(1) Barnard, J. M. A Comparison of Different Approaches to Markush Structure Handling. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 64–68.