PERGAMON

World Patent Information 25 (2003) 195–202

WORLD
PATENT
INFORMATION

www.elsevier.com/locate/worpatin

Markush structure searching over the years

Edlyn S. Simmons

The P&G US Business Services Co., 5299 Spring Grove Avenue, Cincinnati, OH 45217, USA

Abstract

The indexing and retrieval of Markush structures has always been among the most problematic aspects of patent information and the most expensive. Indexing advanced from the simple classification systems of the 1950s to proprietary fragmentation systems, which were followed in the 1980s by topological systems. The cost of access to the latest indexing systems has varied widely over the years. In spite of improvements in indexing and less restrictive access conditions, comprehensive Markush structure searches remain the sole province of well financed organizations.
© 2003 Elsevier Ltd. All rights reserved.

1. Introduction

In 1924, before Eugene Markush won his appeal from the US Patent Office's rejection of the claims in his patent application, the world was a simpler place. Chemists were limited in their ability to synthesize and characterize new compounds and knew relatively little about structure–activity relationships. Information was recorded with pens and manual typewriters. Prior art searches were performed with printed indexes and files of printed cards. Beilstein had been indexing the compounds in the chemical literature by name and chemical formula since 1881, and Chemical Abstracts had begun to do the same thing, introducing a general subject index in 1917, and a chemical formula index in 1920.

Markush was not the first person to attempt covering more than one compound in a patent claim, and his claim (Fig. 1) was relatively simple in comparison with later Markush claims [1]. It had some generic language and a short list of specific compounds. But the Examiner's insistence that a patent claim could not cover alternative compounds resulted in an appeal to the Commissioner of Patents. The Commissioner issued a decision approving of claims with lists of alternatives, provided that the claims were presented in the form used in the Markush claims [2]. This decision, ''in Re Markush'', became precedent; it was cited by other patent applicants, and the name ''Markush group'' was attached to claims reciting chemical fragments ''selected from a group consisting of'' a list of alternatives. Over time, the format known as a Markush claim became standardized as a chemical structure drawing with variable substituents. If Markush had drafted his claim in the format we now know as a Markush structure, it might have looked something like Fig. 2.

The term ''Markush structure'' has grown to designate any chemical structure that contains a required substructure and one or more variable or optional chemical groups. Patent attorneys usually denote Markush groups by the letter R, but other notations are also common. Atoms are often implied in drawings of organic molecules by showing bond angles; the angles are understood as carbon atoms with any empty valences satisfied by hydrogen atoms, but substitution might be permitted if the patent disclosure specifies that the atom is optionally substituted. One feature is essential—groups that are not stated to be either required or optional are forbidden.

As illustrated in Fig. 3, the structure can include variable positions of substitution, variable numbers of substituents, variable bond types, optional substituents, variable chain lengths, and provisos defining combinations of substituents that are forbidden in the claimed genus. The proviso in this structure keeps it from overlapping with the one in Fig. 2 constructed from the Markush patent. A Markush structure represents each of the compounds that can be constructed by combining the variables; the notation used to describe the structure as a whole has no special meaning.

One of the most troubling features of Markush structures is that overlapping Markush structures can be

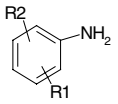
E-mail address: simmons.es@pg.com (E.S. Simmons).

0172-2190/03/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/S0172-2190(03)00073-5

Claim 1. The process for the manufacture of dyes which comprised coupling with a halogen-substituted pyrazolone, a diazotized unsulphonated material selected from the group consisting of aniline, homologues of aniline and halogen substitution products of aniline.

Fig. 1. US Patent 1,506,316 Eugene A. Markush August 26, 1924.

The process for the manufacture of dyes which comprised coupling with a halogen-substituted pyrazolone, a diazotized unsulphonated material prepared from a compound having the formula



wherein R1 and R2 are independently selected from the group consisting of hydrogen, methyl and halogen.

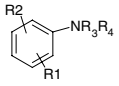
Fig. 2. Claim 1 of US Patent 1,506,316 if drawn in Markush format.

nearly unrecognizable as such. This can happen because the patentees use differing notation, but it happens more often because one patent *requires* a substructure that is optional in the other. This situation is illustrated in Figs. 3–5. Figs. 3 and 4 show hypothetical claims to compounds defined by Markush structures. Fig. 3 shows a genus of aniline derivatives in which the amino moiety may form a heterocyclic ring. Fig. 4 shows azepine derivatives with an optional phenyl substituent. Fig. 5 shows the areas of overlap between the apparently dissimilar Markush structures. When R3 and R4 in the left Markush structure form a 6-carbon alkylene chain, and n in the right structure is zero, these structures overlap.

If a chemist of the early-20th century wanted to find out whether any of the members of a new genus of compounds had been made before, how could he or she find out? This has always been important to industrial chemists, whose companies needed to know whether their new compounds were claimed in a patent or whether they were patentable. It was also important to companies that wanted to monitor the new compounds invented by their competitors. Chemical patents were classified on the basis of structural features by the national patent offices. The chemist could send a representative to the public search room in the patent office and have the classified files searched by hand. For current awareness, he or she could order the gazettes and bulletins issued by the patent offices and scan through the abstracts of the week's patents to see what had been issued during the week. The subscriptions to multiple patent gazettes and the manpower needed for scanning each week's patents were expensive. Only large companies could afford it.

There is no way to search for a generic structure in a name or formula index, but the specific compounds in the examples of patents that were abstracted were retrievable from indexes, by their chemical names and by their molecular formulas. Looking for each of the embodiments of a new generic structure could identify any indexed publications that mentioned one of the compounds, particularly the compounds listed in the indexes of Chemical Abstracts and Beilstein. In Markush structures with predictable chemical names, this was reasonably easy, but many Markush structures include compounds like the embodiments of Figs. 3 and 4 whose systematic names are not clustered together alphabetically. Before the standardization of chemical nomenclature, searching for compounds by name was more difficult than it is in modern indices, for example an

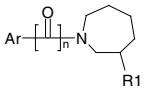
A compound of the formula



wherein R1 and R2 are independently selected from the group consisting of hydrogen, halogen and straight or branched alkyl of from one to four carbon atoms; and R3 and R4 are independently selected from the group consisting of hydrogen and straight or branched alkyl of from one to four carbon atoms; or R3 and R4, taken together, form an alkylene chain of from four to six carbon atoms, with the proviso that when R1 and R2 are hydrogen, methyl or halogen, R3 and R4 are not both hydrogen

Fig. 3. A typical Markush structure.

A compound of the formula



Wherein R1 is selected from the group consisting of hydrogen, C1-C2 alkyl, C1-C2 alkoxy, halogen and trifluoromethyl; n is an integer between zero and 2; and Ar is phenyl or naphthalene, optionally substituted by from one to three groups selected from C1-C4 alkyl, C1-C4 alkoxy, halogen and trifluoromethyl.

Fig. 4. An overlapping Markush structure.

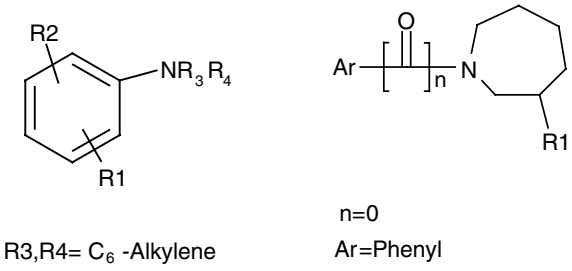


Fig. 5. Areas of overlap.

aminobenzene derivative could have been named as a derivative of aniline and alphabetized accordingly. This left the formula indexes to be searched by calculating the empirical formula of each of the embodiments of the Markush structure and searching each formula in each volume of the formula index. It was tedious, but it was effective, and it is still the only way to search for all the embodiments of a generic structure in the first five collective index periods of Chemical Abstracts.

During the 1950s, the patent literature underwent a major change. Industry was booming all around the world; more and more patent applications were being filed. There were backlogs of patent applications waiting to be examined by the patent offices, so many countries changed their patent laws. They began publishing unexamined patent applications 18 months after their priority filing date. Because the applications were published whether or not they actually claimed a patentable invention, many thousands of applications that would eventually be rejected for lack of novelty were published. Some of the early publication countries even omitted the steps of printing and distributing copies of the applications; they simply displayed a copy in the public search room. The chore of searching manually grew much larger—to see the latest patents, it was necessary to send a representative to the Dutch and Belgian patent offices to read the published specifications. An English chemist, Montague Hyams, recognized that there was a need for quick access to the 18-month publications, and began commuting from his home in London and writing abstracts of published chemical patents. He offered them for sale by subscription to chemical and pharmaceutical companies and named his new company for his house—Derwent Publications Ltd. [3].

2. Markush databases, 1950s

Printed indexes were available to every chemist with access to a good library, and the public search rooms were available to anyone who could afford to travel to the Patent Office or hire someone to do it for him, but prosperous companies were trying to beat out the competition, and they wanted more. Better searches would provide a valuable competitive advantage, and the companies were willing to pay for them. Many companies established their own abstracting and indexing departments; most of them recorded the identity of compounds in card files. The proprietary card decks could be used to index both internal records and published literature and patents. Many companies developed their own notations and coding. And many searched the coded data with the computers that were then available, most of which were IBM card sorting machines. In 1961, the National Academy of Sciences/National Research Council surveyed companies and other institutions in the United States that had internal chemical-structure retrieval systems [4]. At the institutions they visited they found 49 indexing systems, 32 of which were in active use. Most used fragment codes of one kind or another, and a few were using topological codes of some type. Only large companies could afford the cost of indexing the compounds, punching the IBM cards, and searching the card decks. The authors of the report comment that the IBM machines needed for the more advanced indexing systems occupied an entire room, and there was a rental charge of least $1000 a year for a card punch and low-speed sorter alone. In 1961 that was real money!

Although the proprietary indexing systems were designed for specific compounds, some of the companies used them to index patents as well. This turned out to be relatively simple to do with a fragment code. One simply used the technique of ''overcoding'', that is, one indexed all of the structural variables on a single index card. For example, if a patent allowed chlorine and bromine as alternatives, one punched the card in both the position that signified chlorine and the position that signified bromine. This allowed the searcher to retrieve the record whenever he searched for any combination of allowed variables. This produced more false correlations than would have been encountered in a file of specific compounds—these were literally false drops. But it was not a serious problem when there was a small deck of cards to search, and the survey showed that most of the organizations had files of less than 50,000 compounds. As time went on, however, the absolute number of records that had to be screened out grew to be a burden.

3. Markush databases, 1963–1969

The cost of running an in-house indexing system was burdensome, too, even for large, prosperous companies. Now that the searchers in the companies were accustomed to searching decks of coded records, it seemed natural for suppliers to code chemical structures and sell the indexed records along with their abstracts. Monty Hyams' customers urged him to provide structural indexing for the Derwent Fine Chemicals abstracts series. After checking the depth of the customers' pockets, Mr. Hyams decided to create a new service for pharmaceutical companies. He named the new service Farmdoc and hired one of the subscriber representatives, Peter Norton, to create a suitable fragmentation code for the Markush structures in the patents. Farmdoc was introduced in 1963 and the code was used for both specific compounds and Markush structures. The Farmdoc service was available only by subscription. It was followed by two additional subscription patent databases, Agdoc, covering agricultural and veterinary patents, and Plasdoc, covering polymer patents. The new Derwent

indexing systems relieved the subscribing companies of the need to use their in-house experts to index patents, but the retrieval process still required that each company maintain an IBM card sorter to search the Derwent record.

One should note that there is a basic difference between a fragmentation code and a topological indexing system [5,6]. A topological system maps the shape of a molecule by identifying the spatial relationships among the atoms of the molecule. A traditional topological indexing system like the Chemical Abstracts Registry file requires that all of the connections in a molecule be known, and searching the connection table requires a more sophisticated computer than a card sorter. A fragmentation code breaks the molecule into identifiable pieces—rings, chains and functional groups, each of which is identified by a symbol. Only structural units that have codes can be indexed. When the fragmentation code was based on an IBM punch card, the chemical groups were designated by a row and column of the IBM card, and the code was known as a multipunch code. The cards had 12 rows of 80 columns, which meant that any coding system could index less than 960 defined fragments, as some of the positions were needed to identify the document itself and others might be needed to describe non-chemical aspects of the document. A typical fragmentation code has terms for the fragments that are present in some or all of the compounds in an indexed document and a separate set of terms signifying fragments that are required in every compound in the indexed structure. Those ''essential group'' or ''must'' codes can be used in a search strategy by negating them from the answer set, thus limiting the results to compounds where the negated fragment is forbidden in every compound in the record.

The expanded Markush structure in Fig. 2, which was constructed from the Markush patent, would have the codes shown in Fig. 6 in the original version of the Derwent fragmentation code. The codes Fig. 6 do not describe the structure in complete detail; optional groups are weighted equally with required groups and the linkages between fragments are ignored. All of the codes represent groups that are present in a great many organic compounds. This record would be retrieved by many searches for compounds outside the scope of its disclosure, but this was not considered to be a serious problem, as all users expected to screen the results of any search. No one expected the code to be easy to learn and use, either. None of the coding schemes in use at the time were easy enough for use by non-experts, and Derwent was able to provide manuals, training and coding sheets to the few people at each company who would become retrieval experts.

Column/Row	Fragment
37/3	Benzene
37/8	One aromatic group present
46/11	Amine linked to aromatic ring
52/7	Halogen
71/5	Aromatic (non-heterocyclic) compounds
31/3	Pyrrolidine
31/6	Piperidine
31/8	Azepine
71/4	Mononuclear heterocyclic compounds
46/6	Ring tertiary amine, one present

Fig. 6. Derwent fragment codes, 1963.

Derwent's coded records were extremely popular among pharmaceutical companies. Very soon, other companies wanted access to patent information similar to what Derwent was providing to pharmaceutical companies. Agdoc was introduced with its own version of the fragmentation code and Plasdoc was introduced with a specialized polymer code. By 1970 Derwent had expanded coverage to all fields of chemistry, created modifications of the fragmentation code for general chemistry and dyes, and named the expanded service the Central Patents Index. Only Sections A (Plasdoc), B (Farmdoc), C (Agdoc) and E (Chemdoc) of the Central Patents Index received chemical coding. Remaining technologies were indexed only with a simple system of Manual Codes [7] and were priced accordingly. In 1974 non-chemical patents were added, and the service became the World Patents Index, and the name of the 13 chemical sections of WPI became the Chemical Patents Index. When computers using magnetic tapes began to supplant card sorters, code records on tape were provided as an alternative to punched cards. The concept of punched card records was continued, however, even as subscribers migrated to magnetic tape for retrieval.

As the card decks grew and screening out false drop took more and more time, users asked for refinements of the fragmentation code. More fragment codes were added in 1970 and 1972, defining essential groups for negation, the characteristics of linkages between rings, and the details of carbon chains. Rare ring codes based on the Patterson Ring Index [8] were introduced as a separate retrieval service. By the time the World Patents Index went online in 1976, it had five heavily overlapping code systems plus non-overlapping codes for ring systems, steroid molecules and polymers [9,10]. In the online environment, searchers became aware that the division of the fragmentation codes into several different, overlapping codes with 960 terms apiece was meaningless.

In 1978, a subscriber committee was formed to help the Derwent coding division integrate and refine the fragmentation code. In the end, a new four-character alphanumeric notation was introduced; each symbol designated the same fragment in all sections of the database that contained chemical structure indexing. Negation codes, which apply to groups of structurally related fragments, have a two-character format that

Code	Fragment	Added
H1	Amine essentially present	1970
M521	1 mononuclear hetrocyclic ring	1970
M520	No mononuclear heterocyclic ring	1970
M320	No multivalent carbon chains	1972
M210	$C_{1\text{-}6}$ alkyl chain	1972
M270	Alkyl attached to heteroatom	1972
M273	Heteroatom is N	1981
F011	Substitution on 1-position of heterocycle	1981
H641	1 Halogen linked to aromatic ring	1981
H642	2 Halogens linked to aromatic ring	1981

Fig. 7. Additional Derwent fragments 1970, 1972, 1981.

Code	Fragment
34701	Benzene possible
34700	Benzene required
34193	1 Halogen possible
34194	2 Halogens possible
34598	Pyrrole ring possible
34667	Pyridine or piperidine possible
34711	Azepine or hexamethylenimine possible
33001	1 Tertiary amine possible
33000	Tertiary amine required

Fig. 8. CLAIMS CDB fragment codes.

distinguishes them from codes for groups that are present in a structure. Preexisting codes were converted to the new notation, and additional codes were added to refine the definition of chemical structures. Additional codes that would have been applied to the chemical structure in Fig. 2 are illustrated in their four-character format in Fig. 7. As before, polymers had an indexing system of their own, which will not be addressed here. The World Patents Index database, now universally accessed through reloaded files on online search services, had been freed from the constraints of the IBM card. Although the World Patents Index database has since been opened up to non-subscribers, Derwent still restricts access to the fragmentation code to subscribers.

The proprietary databases designed by individual companies did not completely disappear when Derwent's indexing system became available. By 1964, the indexing staff at DuPont had created an Index to United States Patents with a well-developed fragmentation system for organic and inorganic molecules and a separate fragmentation system for polymers. In 1971 the database and the indexing staff were transferred to IFI/Plenum Data Co., the producer of the Uniterm Index to United States Patents [11,12]. The merger of the Uniterm file with aspects of the DuPont indexing system produced a system that could be used for indexing and searching Markush structures, but without the full power of the DuPont indexing system. The complete DuPont database was made available in a second IFI file, the comprehensive database (CDB). Both systems were sold by subscription as magnetic tape databases, which made them attractive only to the well-financed companies that could afford subscription fees, computers, and an expert staff to encode queries and operate the computers.

The DuPont/IFI structure code consists of an open-ended vocabulary of five-digit terms The indexing and retrieval system continues to reflect its development in an era when computer resources were tightly controlled. Only new compounds and generic structures are indexed directly with the fragmentation code; common specific compounds are indexed in a companion registry file and are searched with their own code numbers. In the CDB, but not in the Uniterm database, retrieval can be limited by using negation codes and by codes for roles that are linked to the indexing of the compounds. A large number of "must" codes exist; they are assigned by indexers to fragments that are required in a chemical structure, while "possible" codes are assigned to fragments that are either required or optional. Searchers negate the codes for required fragments when those fragments are not within the scope of the query structure. Some of the CDB codes for the structure in Fig. 2 are shown in Fig. 8. The roles for compounds other than polymers allow the searcher to distinguish among patents covering a compound as the product of a reaction, a reactant in a chemical reaction, and a compound present in a product. The Uniterm and CDB were searched in-house with a weighting algorithm that allowed retrieval of imperfect matches as well as patents with all the codes present, an early form of relevance ranking. Online, searching is done with ordinary Boolean logic and proximity operators. Like the Derwent codes, access to the CDB coding is limited to subscribers, but non-subscribers can search the Uniterm codes for a limited total connect time each year. Non-subscribers were allowed to search the CDB briefly to introduce the file to a broader audience, but access was later withdrawn [13].

Perhaps the most effective and the most expensive fragmentation code was the GREMAS code originally developed by one of Germany's largest chemical companies, Hoechst, in 1959 [14,15]. Shortly thereafter, file building and system development became a joint project with BASF and Bayer. In 1967 an organization called Internationale Dokumentationsgesellschaft für Chemie m.b.H., IDC, was founded for the purpose of indexing organic chemistry from the journal and patent literature, and access to the database was made available to other companies. There were never more than a handful of subscribers—in 1990 only eight companies shared the cost of IDC among them. The GREMAS system ran on a mainframe computer and, as early as 1967, the structures of specific compounds were entered graphically and the code terms were posted by a conversion program. Not surprisingly, indexing for Markush structures

was entered by hand. Although searches with the GREMAS code required an enormous investment, all of the papers I have read [16,17] assure the reader that it was capable of such remarkable precision that the saving in time screening out false drop made up for the original cost.

Time was not standing still, and neither was technology. As more powerful computers became available, topological search systems began to appear. Beginning in the late 1970s, the Chemical Abstracts Service's CAS Online service, the DARC system on Telesystemes Questel, and other search software made it possible to search for compounds by drawing or graphing a chemical structure and searching for matching structures in a database. Both CAS Online and the original DARC were designed to search for individual compounds from the Chemical Abstracts Registry database, and both allowed substructure searching as well as searches for fully defined compounds. There was a major difference in the way the search systems interpreted query structures. In CAS Online, any free valence was interpreted as a free site, a position open to any possible substitution. In DARC, following the pattern of the Markush format, any free valence was interpreted as substituted only by hydrogen. Free sites had to be specified by the searcher. As a result, the output from a CAS Online search included any compound in the database that contained the query substructure, while the output from a DARC search consisted only of compounds that matched the query structure. After a few years, both systems had been refined so that generic structures could be searched, but although the searcher could use a Markush structure to define the scope of the search, the database contained only specific compounds, and the output consisted of a group of specifically indexed individual compounds.

4. Introduction of topological Markush search systems

Since it seemed on the surface to be a simple step further to index Markush structures topologically, searchers expressed a desire to replace fragmentation coding with topological search systems. Compared to the new topological search systems, the fragmentation codes were difficult to learn and difficult to use, and the output was imprecise and hard to interpret. Unfortunately, indexing and retrieving Markush structures is by no means straightforward. The human mind has an enormous power to recognize patterns and relationships, and a chemist with a good vocabulary of chemical names and structures is able to read a patent and recognize specific and generic terminology and the relationships between them. He or she will be able to zoom from specific to generic and back, recognizing the equivalence of methyl and alkyl and hydrocarbyl. Translating between specific and generic terms is crucial to understanding a patent. Teaching a computer to do that requires far more than simply drawing a connection table and finding a match in a file of specific molecules [18].

In 1979, Professor Michael Lynch, at the University of Sheffield began an ambitious research program with the aim of developing a topological search system for indexing and searching Markush structures [19]. He obtained funding from, among others, Derwent, IDC and CAS, and, with the assistance of his research group, developed a language called GENSAL, which could be used for input representation of generic structures, as well as a connection table format for representing them. The Sheffield group's approach to handling generic nomenclature terms was based on formal grammar theory, and their algorithm was capable of translation between generic and specific terms. They also developed algorithms for automatic generation of fragments from generic structures, and for matching structures.

During the early 1980s, the major patent indexing organizations began working diligently toward the creation of commercially viable topological Markush structure search systems. All three used the Sheffield work in their development. IDC made direct use of the GENSAL language, and was using it for input to generate GREMAS codes in the early 1990s. Before an actual GENSAL-based retrieval system was implemented, however, the companies that financed IDC decided to close down the organization. At the end of 1992 GREMAS returned to its original status as a corporate database, increasingly out of date but still useful ten years later for retrospective searches in some technologies [20]. CAS surveyed some of its customers to find out what was lacking in its patent coverage and decided to create a new patent database containing Markush structures. Development was done by CAS staff, notably Bill Fisanick, who was granted a patent for his method for storing searchable files of Markush structures [21]. CAS intended the MARPAT file to be used as a supplement to the Registry file—only Markush structures would be indexed. Derwent, Telesystemes Questel, and INPI, the French Patent Office, joined together to develop the Markush DARC system, which was intended for use in topologically searchable patent files. Two databases were developed, a companion file to the World Patents Index database and a companion structural file for INPI's PharmSearch database, which had not previously been commercially available. Derwent's WPIM file was seen as an eventual replacement for fragmentation coding. The parallel and competing development programs of CAS and Derwent, Questel and INPI took on the character of an arms race.

If speed was the goal, Markush DARC won the race. Markush PharmSearch was released on Questel in February, 1989. The Derwent WPIM file and the CAS MARPAT file on STN followed soon thereafter. Each

of the databases had only a few years of data when they were released; nevertheless the producers were surprised that their customers, who had begged for direct topological searching of Markush structures, showed a remarkable lack of interest in using the files. If profit was the goal, nobody won.

Release of the databases was followed by a flurry of articles comparing the search systems [22–26]. Both include superatoms, artificial chemical symbols representing generic groups—that is to say, fragment codes—so that generic disclosures could be indexed and generic searches could be done. The graphic representations of the structures in the records were not easy to read—the database conventions look decidedly different from the structures in a printed patent. The initial release of Markush DARC could not translate between the generic groups and specific groups. This forced the searcher to use alkyl and methyl and ethyl and propyl and so on as alternatives, straining the system limits. Alternatively, it forced the searcher to apply a free site and turn the Markush search into a substructure search, generating significant false drop. MARPAT could do translation when it was released, but it translated so freely that false drops were abundant. Neither system permitted direct combinations of structure searches with text searches. Both systems were slow, and the computers were programmed to spend only a limited amount of time attempting to match each record with the query. The computers timed out and left a file of incompletely processed records, which the searcher would have to evaluate manually. That was an especially unpleasant task because those were the most complex Markush structures in the file. Neither system could handle simple Markush structures or structures composed entirely of common substructures. None of the databases was comprehensive—PharmSearch included a limited number of patenting countries, Marpat excluded some of the countries covered by Chemical Abstracts, and WPIM excluded the patents they called "nasties", patents with Markush structures too complex for input into the system.

And, of course, both were expensive. In each file there was a substantial charge for each search, as much as $100, in addition to the connect time needed to run the search. And each of the three databases was an entirely new database, a supplement to the existing databases, not a replacement. The advantage, for searchers without an appropriate Derwent subscription, was that Mpharm and Marpat were available without paying a subscription fee.

WPIM, on the other hand, seemed to many to be entirely dispensable—it was still necessary to use the fragmentation code to retrieve records more than three years old. It was Derwent's intention to discontinue fragmentation coding soon after the Markush DARC file went online. Considering the limitations of Markush DARC, this suggestion was met by a chorus of howls, and the end of fragmentation coding was postponed indefinitely.

5. Improvements in topological Markush search systems

With a few years of additional research, the programmers at Questel and STN overcame many of the original shortcomings of the search software. Markush DARC was taught how to do translation. MARPAT's tendency to over-translate was curbed. Derwent learned how to index "nasties". System limits were expanded. And the files have grown much larger.

Best of all, the cost of searching the topological files has dropped. A discount is provided on STN for MARPAT searches done in conjunction with Registry file searches. In 1998, Derwent and INPI consolidated their indexing efforts and combined their Markush structure files into the Merged Markush Service. The cost of a structure search in the Merged Markush file was reduced to the level of a search of the former Mpharm file, which was always much less expensive than the WPIM file. The cost of searching the Merged Markush Service is even lower than the online fees suggest—the entire file is available to everyone. For the first time, chemical structure searches of the Derwent World Patents Index did not require a subscription! Whether or not this is good news for Derwent subscribers is a subject for another time.

By the beginning of the 21st century, Markush structure searching had stabilized, without improvement in underlying flaws in the search systems [27,28]. The Derwent and IFI fragmentation codes are still in use, but input of coding is now facilitated by the use of computer-based input software. INPI has been increasing the size of the backfile in MMS. Both STN and Questel-Orbit increased their search limits, speeding searches and reducing the number of incompletely processed records, an essential improvement considering the increased size of the MARPAT and MMS databases. In the near future, additional improvements can be expected. Derwent has hinted that they may update the indexing in older fragmentation code records. Questel-Orbit has announced work on a new generation of the MMS service on the Unix platform. It remains to be seen whether these changes will change the way we do Markush structure searches.

Acknowledgements

This article is based on a presentation given by the author at the Division of Chemical Information, 216th National Meeting of the American Chemical Society, New Orleans, August 24, 1999, at the Herman Skolnik Award symposium in honor of Stuart M. Kaback.

References

[1] US 1,506,316. Assigned to Pharma-Chemical Corporation by Eugene A. Markush. Filed January 9, 1923; issued August 26, 1924.

[2] Ex parte Markush. 340 O.G. 839, 1924.

[3] http://www.derwent.com/profile/history.html.

[4] Hunsberger 1M, Frear DEH, Harmon RE, Smith EG. Survey of Chemical Notafion Systems. Washington, DC: National Academy of Sciences-National Research Council; 1964. Publication 1150.

[5] Almond JR, Welsh HM. Chemical substructure searching—industrial applications and commercial systems. Drexel Library Quart 1982;18(2):84–105.

[6] Simmons ES. The grammar of Markush structure searching: vocabulary vs syntax. J Chem Inf Comp Sci 1991;31(1):45–53.

[7] http://www.derwent.com/cpi-codes/.

[8] Patterson A, Cappell LT. The ring index; a list of ring systems used in organic chemistry. 2nd ed. American Chemical Society; 1960.

[9] Kaback SM. Chemical structure searching in Derwent's World Patents Index. J Chem Inf Comp Sci 1980;20(1):1–6.

[10] Simmons ES. The central patents index chemical code, a user's viewpoint. J Chem Inf Comp Sci 1984;24(1):10–5.

[11] Donovan KM, Wilhide BB. A user's experience with searching the IFI comprehensive database to US chemical patents. J Chem Inf Comp Sci 1977;17(3):139–43.

[12] Kaback SM. The IFI/Plenum chemical indexing system. In: Barnard JM, editor. Computer Handling of Generic Chemical Structures, Proceedings of a Conference organized by the Chemical Structure Association at the University of Sheffield, England, 26–29 March. Aldershot, UK: Gower; 1984. p. 49–57.

[13] Lambert N. How to search the IFI comprehensive database online... tips and techniques. Database 1987;10(6):46–59.

[14] Kolb AG. Topological coding as a basis for the GREMAS file. In: 200th National Meeting of the American Chemical Society Washington, DC. Washington: American Chemical Society; 1990. CINF 23.

[15] Schoch-Grübler U. (Sub)Structure searches in databases containing generic chemical structure representations. Online Rev 1990;14(2):95–108.

[16] Franzreb KH, Hornbach P, Pahde C, Ploss G, Sander J. Structure searches in patent literature: a comparison study between IDC GREMAS and Derwent chemical code. J Chem Inf Comput Sci 1991;31:284–9.

[17] Haxel C. Patent information at Henkel: from documentation and information to collaborative information commerce. World Patent Information 2002;24(1):25–30.

[18] Barnard JM. Substructure searching methods: old and new. J Chem Inf Comput Sci 1993;33:532–8.

[19] Lynch MF. Generic chemical structures in patents (Markush Structures): the research project at the University of Sheffield. World Patent Information 1986;8(2):85–91.

[20] Schoch-Grübler U. Personal communication, August 6, 2002.

[21] Fisanick W. US 4,642,762. Assigned to American Chemical Society, February 10, 1987.

[22] Barnard JM. Online graphical searching of Markush structures in patents. Database 1987;10(3):27–34.

[23] Cloutier KA. Comparison of three online Markush databases. J Chem Inf Comp Sci 1991;31(1):40–4.

[24] Hajime Tokuno. Comparison of Markush structure databases. JCICS 1993;33(6):799–804.

[25] Schmuff NR. A comparison of MARPAT and Markush DARC software. J Chem Inf Comp Sci 1991;31(1):53–9.

[26] Wilke RN. Searching for generic chemical structures. J Chem Inf Comp Sci 1991;31(1):36–40.

[27] Berks AH. Current state of the art of Markush topological search systems. World Patent Information 2001;23(1):5–13.

[28] For additional discussion of the systems discussed herein, see Berks AH. Markush structures in patents. In: Schleyer PVR, Allinger NL, Clark T, Gasteiger J, Kollman PA, Schaefer III HF, Schreiner PR, editors. The Encyclopedia of Computational Chemistry, vol. 3. Chichester: John Wiley & Sons; 1998. p. 1552–9. An updated version of this article is in press for the series "Chemoinformatics—From Data to Knowledge" edited by Johann Gasteiger, expected to be published by Wiley in 2003.

Edlyn S. Simmons serves as Section Manager, Patent Information, at Procter & Gamble Co. in Cincinnati, Ohio. She holds B.S. and M.S. degrees in chemistry and is a registered US patent agent. She is a founding member, past Chair, and Director at Large of PIUG (Patent Information Users Group Inc.) and is the Course Director of the PERI (Pharmaceutical Education and Research Institute) course on Patent Information for Pharma/Biotech.