From specific to generic
From DARC to Markush DARC

The structural search for generic patents

Bernard Marx

Résumé **Du spécifique au générique, de DARC à Markush DARC : la recherche structurale pour les brevets génériques**

Les nouveaux composés et les nouvelles méthodes de synthèse chimique ont toujours représenté des innovations donnant lieu à de nombreux brevets, en particulier à partir du milieu du XIX^{e} siècle. Les revendications définissent alors des structures bien précises, dites structures spécifiques. Avec le développement de la chimie organique et la complexité structurale de nouveaux composés, des brevets génériques sont acceptés à partir du premier quart du XX^{e} siècle. Dans les années 50, les développements de l'informatique permettaient d'indexer et de chercher les structures spécifiques par des codes fragmentaires, linéaires puis topologiques. L'étape suivante est l'application de ces codes aux structures génériques de brevets. Les premières recherches ont lieu dans les années 1975-1976 et le système Markush DARC (Documentation et Automatisation des Recherches de Corrélations) permet cette recherche générique en ligne à partir de janvier 1989. À partir de cette date, les trois partenaires du projet (l'Institut National de la Propriété Industrielle et Questel en France, Derwent Information Services au Royaume-Uni) ont apporté des améliorations continues au système de recherche.

Mots-clés **Brevets chimiques, recherche structurale, système Markush DARC.**

Abstract New compounds and new chemical synthetic methods, particularly since the latter part of the XIXth century, have traditionally given rise to numerous patents. The patent claims for the compounds they described generally defined very precise "specific structures". As organic chemistry developed, new compounds with increasingly complex structures were synthesized and in the first quarter of the XXth century the National Patent Offices began to issue "generic" patents. In the 1950s, the development of informatics enabled the indexing and searching of specific structures, first by fragment-based, then linear, and finally topological codes. The next step was the development of code to search generic structural patents. Initial research began in 1975 and the first commercial release of the Markush DARC (Documentation and Automated Research of Correlations) system, which supported online generic patent searching was in January 1989. At this stage, a partnership between the French Patent and Trademark Office, Questel, a French online supplier, and Derwent Information Services in the UK was established and has since worked to improve the quality of the indexing and searching features of the system

Keywords **Chemical patents, structural search, Markush DARC system.**

Two independent elements: generic chemical patents and specific compound search

Generic chemical patents

On January 9th, 1923, Eugene A. Markush solicited a patent from the US Patent Office, now the US Patent and Trademark Office. The resulting patent, US 1506316, was granted on August 26th, 1924.

In US 1506316 a process was claimed for the manufacture of dyes which comprised the coupling products of a halogen-substituted pyrazolone with a diazotised unsulfonated material selected from the group consisting of aniline, its homologues and its halogen substitution derivatives [1].

This was not the first generic claim in a patent, but since then, "Markush-type" patents have proliferated and "Markush structures" have come to designate a group of specific compounds, namely a molecular skeleton bearing one or more variable substructures with a list of alternative definitions for the variable portions of the molecule [2] (*figure 1*).

These patents can be inconsistent with the legal definition of a patent in cases where *the various specific compounds designated by the generic structures encompass a very wide structural space and a very small number of examples are provided.* Notwithstanding this, generic patents were not only granted, but their number and the complexity of the structures they included increased dramatically.

Development of new compounds or new processes is a long and costly undertaking. A clear benefit of the use of such generic structures is its lowering of the cost of pharmaceutical and chemical research. Broad coverage in a pharmaceutical patent is important for predictable activity of a homologous series, less expensive than filing different patents and makes it difficult for a rival company to file for a patent presenting only minor differences.

Figure 1 - An example of a Markush structure: in this case, a benzenic cycle with variable substructures on the cycle (X, Y, or Z) and on the carbon chain (R1 and R2).

There are also, however, significant disadvantages to using broad generic structures. The predictable activity will not be reliable if the initial generic structures translate into too many specific compounds. Representing and searching these complex structures can be particularly difficult. When researchers from the Chemical Abstracts Service (CAS) [3] mention "*the generic and prophetic substances represented by the Markush structures presented in chemical patents*", there is clearly a contradiction between a "prophetic" structure and a patent, the latter being meant to recognize such elements as novelty, utility, "non-obviousness" or inventiveness, and instances of specific embodiments of the invention. When a patent contains hundreds of pages but only one exemplified compound, is it still really a patent? National patent rules can also have a bearing on this question. General claims can be made as to the original work for the international or regional patenting body and then made more specific for the national authority involved. Another distinction lies in the different rules between the US and much of the rest of the world: in the US, only granted patents are published; elsewhere, unexamined applications are also published. Application date and content can become almost as relevant as patent issuance.

For many years, patents containing generic structures, Markush patents, were accepted in all patent offices and the main concern was establishing the appropriate limit to their broadness. Not surprisingly, industry patentees and patent offices have widely differing views on the matter.

Industry patentees prefer the broader claims for the advantages mentioned above; any curtailment or regulation is viewed as an infringement on their rights and on the freedom of innovation and inventiveness. Patent offices, faced with the onerous task of indexing and searching cleave to a different perspective. Overly broad claims allow claimants to put anything they wish in a patent application, thereby making it very difficult to prove the novelty and utility factors necessary to granting a patent. As Mike Dixon, from Derwent Information Services, concluded in 1990: "*The broad Markush claim cannot be everything to all men. Some people are very happy with it, others are dissatisfied with it*" [4].

Specific compound search

Three different codes are used to tackle the description, input, storage and searching of chemical information: fragment, linear and topological codes.

Fragment codes

Several fragment-based codes of chemical compounds have been proposed. In these programs, the molecules in the database are broken into fragments. No precise information is retained concerning the links between the fragments and the compounds from which they derived. As a result, compounds can only be retrieved by means of the fragments they contain. The problem here is that two different structures can lead to the same code. The two most popular fragmentation codes used in the past years were Gremas and the Ring Code. Gremas was developed in 1957 by Hoechst and BASF in Germany and used by IDC (Internationale Dokumentations Gesellschaft für Chemie). It used 3 000 fragments and made it possible to code compounds, patents and reactions. Fragment codes are extracted automatically from the connectivity matrix of each structure in the database.

The Ring Code was used by Documentation Ring, an association of chemical industries, and commercialised by Derwent Information Service. The system based on Ring Code can index and search both specific compounds and Markush structures. The first step of the search uses 324 fragments.

Linear codes

Linear codes look like the semi-developed formulas used by chemists. The best known of these linear codes, the Wiswesser Line Notation, was developed in 1954 by William Wiswesser of the US Army in collaboration with a number of chemical companies. With some 300 rules for the description of chemicals, only specific compounds were initially described, but subsequent developments extended the applicability of the code to Markush structures.

Topological codes

Topological codes are the most current method for the description of molecules. They can specify the existence and the nature of the atoms as well as the bonds between them. The graph which is used for the input and for the search is equivalent to a structural formula. Two such topological codes are in use today: CAS (Chemical Abstracts Service) and DARC (Description, Acquisition, Retrieval, Correlations).

The CAS code has been used by Chemical Abstract Services since 1965. It is based on a connectivity matrix which shows the links between the atoms, the identity of the atoms and the bond types.

The DARC system was developed in 1965 by Jacques-Émile Dubois, Daniel Laurent and Henri Viellard [5-9] and stands today as a major contribution to chemical informatics from Dubois' group. The DARC system is a documentation system for description and search, but beyond this, it permits links between structure and properties such as reactivity. In DARC, the canonical description of a chemical structure results from the progressive generation of a concentric environment around a focus. The focus may be an atom, chosen freely and independently and used for input or for search. A progressive ordered generation process by ELCO (Environment which is Limited, Concentric and Ordered) propagation around the focus results in the assignment of a linear order label. Each descriptor derived from an ELCO contains the

following independent topological information: a descriptor of the existence of the nodes which unambiguously expresses the topology of the hydrogen-suppressed graph, a descriptor of links which gives the multiplicity order of each bond, and a descriptor of the nature of the nodes which includes the atomic number of each atom.

The DARC system's most important innovation is its generation of the environment by progressive substitution. With this concentric organisation, the program groups atoms with similar positions relative to the focus. This correlation is very important because it reflects the chemical relationships between such pairs of atoms. The characteristics of DARC led to three important applications: property prediction, derivation of synthetic pathways and identification of reaction paths.

The DARC system is also used for the representation of cyclic compounds [10]. When the notion was developed in DARC of undetermined infra FREL (Fragment Reduced to a Limited Environment) [11], it immediately enabled the coding of and search for generic Markush structures.

Merging the two elements: structural search for generic patents

The generic nature of Markush structure patents makes them difficult to index and thus to search. Indexing each individual embodiment of a Markush structure is impossible. A more subtle approach was clearly necessary. Research into this problem proceeded *via* the following stages.

The earliest technique for indexing chemical compounds in patents was the classification of each patent on the basis of the most important structural parts claimed in the patent. This method is highly practical from the point of view of patent office examiners because it limits the number of cases they must consider. The difficulty lies in finding the pre-existing codes corresponding to new and large number of possible structures.

Half a century ago, the use of mechanical card-sorting devices followed by the advent of computers made it possible to overcome the limitations of single-substructure-based classification by indexing all the substructures in a Markush structure [12]. Fragmentation codes were *used to search specific compounds, and also patents.* There is a very real *consistency* between a fragmentation code and chemical fragments obtained from a developed Markush structure: all the different fragments are linked in a Boolean OR operation. Such a search, however, may be irrelevant because the fragments are not necessarily connected the way they were within the original structure.

A fragmentation code was used by IFI/Plenum Data, a US company, for its CLAIMS Uniterm Database of US patents. The limitation of this system lay in the large numbers of false drops that were retrieved because of the large number of patents with common chemical functions. IFI/Plenum also developed a more specific fragmentation code for its Comprehensive Database with occurrence counts and specific operators such as "MUST" and "POSSIBLE".

The Chemical Patent Index (CPI) was developed by the Derwent Information Service. This was a fragmentation code defining not only the functional group but also its position of attachment to the molecule. Between 1970 and 1980, CPI's potential was enhanced, but its use for Markush structures encountered severe limitations: the code proved to be difficult to learn and use. Like the IFI/Plenum code, it resulted in too many false drops and it did not allow reconstruction from the coded record of the complete structure of the indexed compounds.

Thus there was clearly a need for complete and pertinent searches using topological codes. Previously developed for specific compounds, these were now applied to generic patent searching.

Markush DARC: a groundbreaking application

The extension of the DARC system from substructure searching for specific compounds to search and retrieval of patent generic structures began in 1975 in Jacques-Émile Dubois's laboratory with the development of a graphical interface for fragment search and an atom by atom search to obtain the exact number of structures corresponding to the generic structure query.

As of 1982, Dubois joined forces with the French Patent and Trademark Office (INPI) and the French Department of Industry to further develop the DARC system for chemical generic patent search on Telesystemes, the online computation service. This allowed access to Chemical Abstracts Service databases both by text search and by DARC structural search.

In 1983, a contract was entered between the Department of Industry and INPI to "*extend DARC software to the management of structural data of patents*". During the same year, INPI and Telesystemes signed a general agreement whose aims included the development of DARC for its application to patents. The French Department of Research and Industry also sponsored a contract for the development of DARC for Pharmsearch, a pharmaceutical structural database produced by INPI. In 1984, INPI and Derwent Publications Ltd signed a general agreement to cooperate with Telesystemes-Questel to develop their patent databases. In the same year, the Department of Research and Industry asked INPI to produce an integrated search system for patents using the DARC software and Questel computers. In 1985, Derwent and INPI agreed to a 50/50 partnership to finance the development of DARC software for patents. The new software was to handle specific groups, superatoms, free-text and dictionary terms. At the same time, INPI worked to develop Pharmsearch and a contract was signed between INPI and the European Economic Community with the same objective. From 1986 on, contracts were signed between INPI and Telesystemes-Questel to develop DARC for generic patents, and in January 1989, Markush DARC became the first system in the world providing online users with structural access to generic patents.

Markush DARC today

What is it?

Markush DARC is an enhancement of the generic query capabilities of the substructure DARC search system. One of the major problems with Markush structures is matching the representation of a generic group *with the specific structures that derive from it.* To represent these groups, Markush DARC uses a set of 22 "superatoms", each representing a different type of group, such as acyclic hydrocarbons, cyclic systems, metal and others. There is no hierarchical link between superatoms. Some of them are given special "attributes" specifying, as an example, the length of a carbon chain.

How does it work?

The indexing of a chemical generic patent involves different steps: listing the specific values which correspond to the variables in the original structure, listing the generic terms (superatoms) which correspond to these values, and finally listing any attributes (see *figure 2* for an example of indexation).

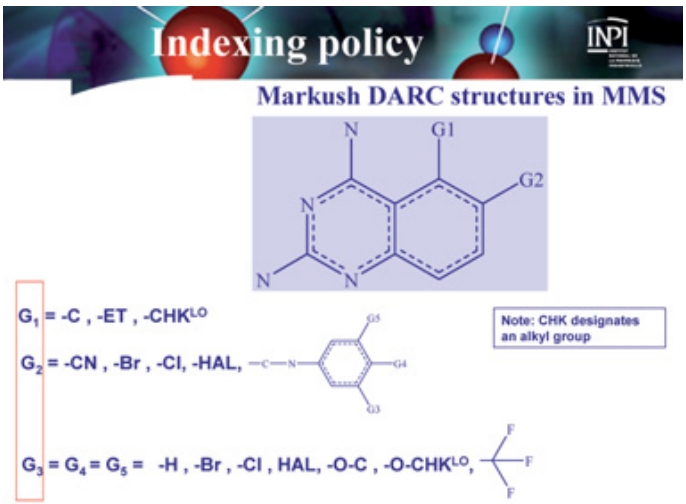

Figure 2 - Indexation of a Markush DARC substructure of the Markush structure seen in *figure 1* in the MMS (Merged Markush Service) database jointly produced by Derwent and the French Patent and Trademark Office, INPI.

The structural search consists of two main steps: the first, the RE process ("recherche par Ecran" or screen search) is the screening process based on FRELs (Fragments Reduced to Limited Environment) developed by Dubois. These are locally limited fragments defined about a central atom, branching out to two levels [13] (*figure 3*).

The second step, the AA ("recherche Atome par Atome" or atom by atom) process, conducted on the results from the first step, is an iterative atom by atom search, an exact search of generic chemical structures (*figure 4*). The structure results can be visualized on the screen in different ways through the step called "VI FO" or Viewing Focus (*figure 5*).

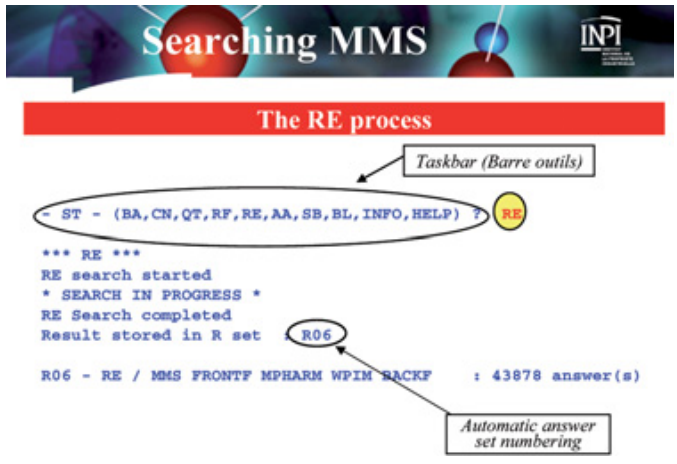

Figure 3 - The first step of a Markush DARC substructure search is the RE ("Recherche par Écran", French for screen search) process, a screening process based on FRELs.

The process is initiated by selecting "RE" (highlighted above) from the taskbar which appears at the top of the DARC screen. In the illustrated screenshot above, the search yielded 43 878 Markush structures as potential answers, with R06 being the number automatically attributed by the computer to designate this particular answer set. Each answer set is limited to a maximum of 1 million candidates and is obtained from the intersection of the query structure with a list of predefined screens. This constitutes the first of a two-step search process. In the second step, the answer set is narrowed down by a more specific atom by atom search (see *figure 4*).

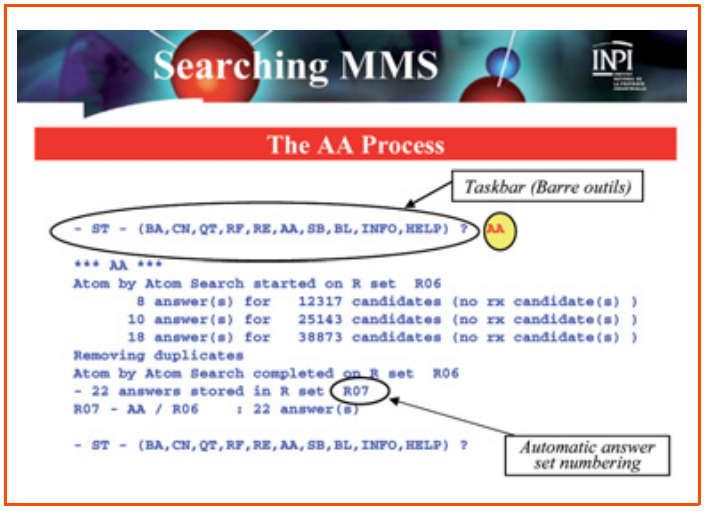

Figure 4 - The second step of a Markush DARC structural search is the iterative atom by atom (AA) search which continues the search started in the RE process (see *figure 3*).

The more detailed atom by atom process begins by selecting "AA" (highlighted above) from the taskbar at the top of the DARC screen. The 43 878 structures in the answer set R06 found in *figure 3* are screened in a step-by-step process in which the computer searches in sections of the RE result set and then removes any duplicates. The AA process in this particular example yields 22 matching structures, which are automatically grouped and numbered as answer set R07.

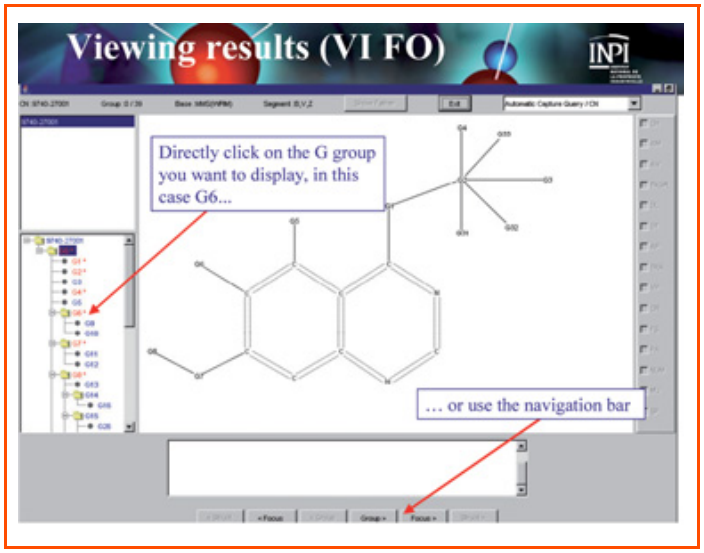

Figure 5 - The VI FO or Viewing Focus is one of the possible visualizations of the DARC display.

The VI FO displayed above is of one of the 22 generic Markush structures found as a result of the RE and AA search processes (see *figures 3* and *4*) performed on the Markush structure shown in *figure 1*. The VI FO shows both the molecular skeleton as well as each of the variable substituents or substuctures of the generic Markush structure. One can see the latter either by clicking on the G group (column to the left of the screen) one wants to display or by using the navigation bar at the bottom of the screen.

Applications

Two structural databases, MPHARM and WPIM, use Markush DARC.

MPHARM (Markush Pharmsearch) is an INPI (National Institute for Industrial Property) database which includes all pharmaceutical patents issued since 1989 by European, French and US patent offices. Patents issued by the Patent Cooperation Treaty (World Intellectual Property Organisation or WO), UK and Germany were added to the database some years later. A companion bibliographic file PHARM is linked to MPHARM to obtain bibliographic information on patents.

WPIM (World Patents Index Markush) is produced by Derwent; its companion bibliographic file is WPI (Derwent World Patent Index).

Together, these databases provide extensive coverage of chemical patents for the pharmaceutical and chemical industries.

Recent developments

From 1990 to 1992, INPI, Derwent and Questel received a grant from the Commission of the European Communities under the Impact Program. The aim was the further development of the Markush DARC software. The two most notable improvements were the superatom translation attributes and the use of variable attachment positions. The translation makes it possible to match both generic terms against specific instances and specific instances against generic terms. The variable attachment positions improvement enables the user to enter searchable variable positions of attachment in the structure query [14-15].

In October 1994, an agreement to work together to produce a database was signed by INPI and Derwent. This was an important agreement, with three main objectives: to investigate the advisability of creating a jointly produced database, to merge the data so far compiled by each partner into a database referred to as the "new database", to agree to share the property rights in the new database and to exploit it independently. This agreement rules the technical part of the co-production.

From July 1994 to June 1995, a study was conducted by INPI, Derwent, Questel and CAS to "*explore the feasibility and cost effectiveness of creating a common Markush file and related software with common ownership*". At that point in time, there were two independent systems: STN (Scientific and Technical Information Network) from CAS with Messenger, structural software for Registry and MARPAT (a generic patent database), and Questel with Markush DARC for the MPHARM and WPIM databases. Different scenarios were studied, ranging from partial to complete integration; the latter would have resulted in a World Markush System and a World Markush Database. After a year of meetings, the project was abandoned.

In 1999 however, an agreement was signed by Derwent and INPI to jointly produce a common database called Merged Markush Service (MMS) and to share the task of indexing the generic patents [16-17] (*figure 2*). MMS contains 70% Markush structures and 30% single compounds.

Today, the co-production of Markush DARC by INPI and Derwent covers the specific drug French patents from 1961 to 1973, all pharmaceutical French patents (FR), European patents (EP), Patent Cooperation Treaty (WO), American patents (US) since 1978, British (GB) and German (DE) patents since 1980 and Chinese patents (CN) since 2000. All in all, some 1 600 000 pharmaceutical patent structures are covered, a number that is continually increasing.

References

[1] Markush E.A., Patent Office 1506 316, August 26 1924, Application filed January 9 **1923**, Serial 611 637.
[2] Sibley J.F., *J. Chem. Inf. Sci.*, **1991**, *31*, p. 5.
[3] Ebe T., Sanderson K.A., Wilson P.S., *J. Chem. Inf. Comput. Sci.*, **1991**, *31*, p. 31.
[4] Milne G.W., *J. Chem. Inf. Comput. Sci.*, **1991**, *31*, p. 30.
[5] Dubois J.-E., Laurent D., Viellard H., *C. R. Acad. Sc. Paris*, **1966**, *263(C)*, p. 764.
[6] Dubois J.-E., Laurent D., Viellard H., *C. R. Acad. Sc. Paris*, **1966**, *263(C)*, p. 1245.
[7] Dubois J.-E., Laurent D., Viellard H., *C. R. Acad. Sc. Paris*, **1967**, *264(C)*, p. 348.
[8] Dubois J.-E., Laurent D., Viellard H., *C. R. Acad. Sc. Paris*, **1967**, *264(C)*, p. 1019.
[9] Dubois J.-E., Viellard H., *Bull. Soc. Chim.*, **1968**, *3*, p. 900.
[10] Dubois J.-E., Viellard H., *Bull. Soc. Chim.*, **1971**, *3*, p. 839.
[11] Dubois J.-E., Bonnet J.-C., Goldwaser D., Attias R., The DARC system: a chemical information system *EURIM II*, 23-25 March **1976**, Proceedings edited by W.E Batten, p. 135.
[12] Simmons E.S., *J. Chem. Inf. Comput. Sci.*, **1991**, *31*, p. 45.
[13] Schmuff N.R., *J. Chem. Inf. Comput. Sci.*, **1991**, *31*, p. 53.
[14] O'Hara M.P., Pagis C., *J. Chem. Inf. Comput. Sci.*, **1991**, *31*, p. 59.
[15] Benichou P., Klimczak C., Borne P., *J. Chem. Inf. Comput. Sci.*, **1997**, *37*, p. 43.
[16] Marx B., *La Propriété industrielle, Sources et Ressources d'Information*, Nathan Université, Paris, **2000**, p. 68.
[17] Dickens D., Buffet P., Takashima Y., *CSA Newsletter Autumn 2003*, International Chemical Conference, Nîmes, 22 oct. **2003**.

Bernard Marx
was Former Vice-President of the Documentation and Information Department, INPI (Institut National de la Propriété Industrielle, France)*.

* 49 rue de la Convention, 75015 Paris.
E-mail: bemarx@wanadoo.fr