World Patent Information 53 (2018) 49–57

Contents lists available at ScienceDirect

World Patent Information

journal homepage: www.elsevier.com/locate/worpatin

A deep analysis of chemical structure-based patent searching in the Derwent index space

Andreas Barth[1]

FIZ Karlsruhe, Leibniz Institute for Information Infrastructure, Hermann-von-Helmholtz-Platz 1, D-76344, Eggenstein-Leopoldshafen, Germany

ARTICLE INFO

Keywords:
Patent searching
Structure searching
Markush
Database crossover
Indicators
Chemical information

ABSTRACT

Structure-based patent searching with high precision and recall is the most important information request in chemistry. The different parts of chemical patent information are stored in different databases and the challenge is to bring the different parts together in a consistent manner and retrieve a complete set of structures and patents. With the application of set theory, it is possible to develop a method which enables the user to obtain a group of corresponding, consistent answer sets after subsequent searching in the different databases and utilizing the special features of both structure and concept searching. A second goal of this paper is to demonstrate the relevance of Markush structures for patent searches in a quantitative manner. Another goal is the development of a set of simple indicators which can be applied to the search results of structure-based patent searches. Four indicators have been chosen with special focus on the relevance of Markush structures for patent searches. The searches and the analysis were carried out with ten pharmaceutical examples. The results show in a quantitative manner, that a search in the Markush database yields a number of inventions (patent families) that cannot be found via searches in specific structure databases alone.

1. Introduction

Chemical information retrieval with high precision and recall is of central importance both for industrial and for academic researchers. Patent searches, especially when freedom-to-operate is requested, require utmost comprehensiveness of search results. In order to fulfill these demands it is necessary that content providers index their databases with high quality. Information retrieval systems, on the other hand, must support sophisticated retrieval functions for information searches. Since chemistry is based to a large extent on structures and reactions, the corresponding structure databases play a central role in any kind of chemical information search. However, in most cases the structure is not the final goal and therefore it is necessary that the structure databases are linked to the corresponding literature and patent databases.

In Chemistry, there are several organizations that provide high quality chemical databases based on careful manual curation of information with additional indexing of controlled terms and special classifications. The most important providers of chemical patent databases are Chemical Abstracts Services (CAS) [1] and Clarivate Analytics (formerly Thomson Reuters and Derwent) [2]. CAS has the mission to document all chemical relevant information worldwide and to offer the most comprehensive databases in chemistry. The CAS system consists of CAplus with more than 47 million research records, mostly from scientific literature and patents, of Registry with approximately 140 million unique organic and inorganic substances (excluding sequences), and of MARPAT with more than 1.16 million Markush structures from patents. In addition, there is also the CASREACT database of chemical reactions. All databases from CAS are interconnected via the CAS accession number and the CAS registry number.

Clarivate Analytics, on the other hand, is focusing on the provision of comprehensive patent information in general, including chemical information. Patents from different authorities worldwide are collected and grouped as patent families (inventions) in the Derwent World Patent Index (DWPI). Specific chemical structures are indexed in the Derwent Chemistry Resource (DCR) and Markush structures are stored in the Derwent Markush Resource (DWPIM). Indexing of DWPIM began in 1961 while DCR began in 1999. Before topological searching was introduced, both Markush and specific structures were indexed by fragmentation codes. In 2017 DWPI contained about 3.2 million chemical inventions (patent families) corresponding to 2.8 million chemical compounds in DCR and more than 2 million Markush structures in DWPIM. The Derwent databases are linked via the Derwent accession number, the DCR registry number, and the DWPIM number.

Since comprehensiveness of patent information is both a goal and a challenge in most patent searches, several authors have compared the

E-mail address: ab.barth@online.de.

[1] Present address: Blumenstr. 2, D-76297 Stutensee, Germany.

https://doi.org/10.1016/j.wpi.2018.05.005
Received 13 October 2017; Received in revised form 11 April 2018; Accepted 15 May 2018

various databases. For example, Ede et al. [3] have studied the quality of chemical structure indexing with a few exemplified searches for the providers Chemical Abstracts Service, Clarivate Analytics, and Elsevier. Markush databases and Markush structure searching with different information systems has been analyzed by several authors, e.g. by Simmons [4], Schmuff [5], Barnard [6], Berks et al. [7], Berks [8], and Geyer [9]. Another set of studies looks at open access public databases (e.g. Kim et al. [10], Papadatos et al. [11]) and/or chemical information extraction from patent full-text databases (e.g. Habibi et al. [12], Krallinger et al. [13]). Also first level patent data have been compared with value-added patent information by Emmerich [14].

This paper focusses not on individual databases and their quality but on the relationship and the relevance of inter-database information retrieval. The study comprises the formal interrelations between specific structures (chemical compounds), Markush structures, and the corresponding patent information in the Derwent information system. With the application of set theory it is possible to develop a method which enables the user to obtain a group of corresponding, consistent answer sets after subsequent searching in the different databases, and utilizing the special features of both structure and concept searching. In practice this is done by successively searching in the different, but interrelated databases. A simple crossover between the databases makes it possible to use the answer sets of the previous searches in the complimentary database. In every search step it is possible to combine the answer set from the previous search with additional parameters and to narrow the results down until the final outcome is obtained. The search procedure may start from a structure search (structure-based) in a specific structure and/or Markush database or from a patent search (concept-based) in a patent database.

A second goal of this paper is to demonstrate the importance of Markush structures for patent searches in a quantitative manner. The answer sets of chemical compounds and of Markush structures are analyzed with respect to the corresponding answer sets of patent families. An important result of the analysis is the determination of the quantitative impact of Markush structures on the comprehensiveness of patent searching. As a consequence it is indispensable for chemical information retrieval to augment typical searches in databases with specific structures like Registry or DCR with additional searches in the corresponding Markush databases.

Another goal is the development of a set of simple indicators which can be applied to the search results of structure-based patent searches. Four indicators have been chosen with special focus on the relevance of Markush structures for patent searches. These indicators provide quantitative answers to questions like.

- What is the ratio of Markush structures and specific structures? (see Eq. (10), section 5)
- What is the ratio of patents which are only found via Markush structures with respect to the total number of patents? (see Eq. (11), section 5)
- What is the ratio of Markush structures generating new patents and the total number of Markush structures? (see Eq. (12), section 5)
- What is the impact of Markush structures on the patent answer set? (see Eq. (13), section 5)

The searches and the analysis were carried out with ten pharmaceutical examples: Ibuprofen, a nonsteroidal anti-inflammatory drug (NSAID) will be analyzed in full detail. In order to broaden the scope of the analysis nine additional examples have been analyzed and the results are presented together in summary. The additional drugs that have been chosen arbitrarily are Benzodiazepine, Celecoxib, Keterolac, Letrozole, Metformin, Moxifloxacin, Oseltamivir, Pantoprazole, and Propanolol. Although the query structure contains only specific elements we will find all structure variations of the query structures with a substructure search (SSS). The corresponding patent results show, that a search in the Markush database yields a number of inventions (patent families) that cannot be found via searches in specific structure databases alone. The actual contribution may vary from one query structure to the other. Finally, it should be noted that the comparison was done on the new STN platform.

2. Derwent concept of chemical structure and patent databases

In chemical patents the substance-related information is often documented as a generic or Markush structure, supplemented by exemplified specific structures. Markush structures were first introduced by Eugene Markush when he defended his need for variations to describe a set of related compounds [15]. A Markush structure is a generic structure together with a corresponding set of chemical variations and it corresponds to a chemical structure space containing in general a large variety of chemical structures. The concept has been described in a review by Downs and Barnard [16] and the specific concept for the Derwent Markush Resource (DWPIM) was introduced by Barth et al. [17].

In the Derwent universe chemical information is indexed in the following databases: the Derwent World Patent Index (DWPI) contains the patent inventions (patent families), the Derwent Chemistry Resource (DCR) contains the specific chemical structures referenced in DWPI, and the Derwent Markush Resource (DWPIM) contains the Markush structures also referenced in DWPI. DCR and DWPIM are linked to DWPI via their structure keys which are also indexed in DWPI. The relationship between the databases is illustrated in Fig. 1. The structure keys are the primary keys in the structure databases and secondary keys in the patent database. However, the primary keys of the patent database are not indexed in the structure databases.

Since the primary keys of the structure databases DCR (DCR Key) and DWPIM (DWPIM Key) are indexed in the patent database DWPI, it is possible to perform a search in one database and to crossover to the other database. However, the actual procedure depends on the direction of the crossover. When we start with a structure search one obtains an answer set with structure keys and these keys can be searched directly in the patent database. The opposite procedure builds on a search in the patent database. In order to crossover to the structure database it is necessary to extract all structure keys from the patent records of the patent answer set before these structure keys can be searched in the structure database.

Typical search strategies require the combination of these databases in order to obtain a search result which is both as comprehensive and as precise as possible. At the end of this workflow it is important to obtain corresponding and consistent answer sets where the set of structures and/or Markush structures correspond(s) exactly to the bibliographic or patent answer set.

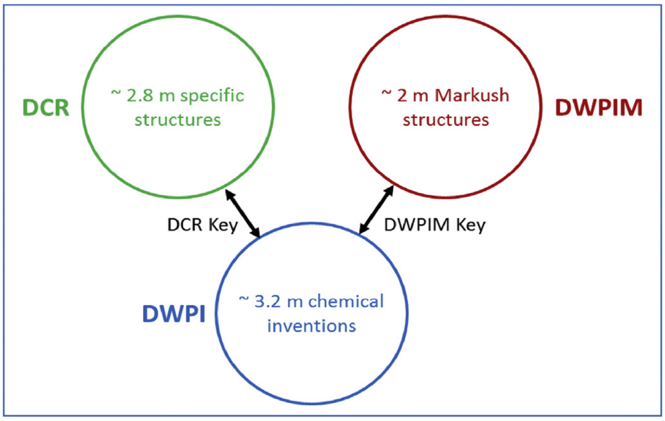

Fig. 1. Relationship between the Derwent databases DCR, DWPIM, and DWPI.