World Patent Information 44 (2016) 48–52

Contents lists available at ScienceDirect

World Patent Information

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

Indexing chemical structures: Exemplified compound indexing in patents by the vendors Thomson Reuters, Chemical Abstracts and Elsevier – A comparative study by the Patent Documentation Group (PDG)

Mark Ede [a, *], John Endacott [b], Mark Harper [c], David Rees [d]

[a] GSK, David Jack Centre for Research and Development, Park Road, Ware, Hertfordshire, SG12 0DP, United Kingdom
[b] GSK, 980 Great West Road, Brentford, Middlesex, TW8 9GS, United Kingdom
[c] IP Intelligence, Sanofi R&D, 91385, Chilly Mazarin, France
[d] Elanco Animal Health, c/o Novartis Animal Health Inc., Schwarzwaldallee 215, CH-4058, Basel, Switzerland

ARTICLE INFO

Article history:
Received 13 May 2015
Received in revised form
9 December 2015
Accepted 29 December 2015
Available online 18 February 2016

Keywords:
Patent analysis
Exemplified compounds
DWPI
Reaxys
CAS
Chemical abstracts
CASREACT
SureCHEMBL

ABSTRACT

The following article briefly discusses the history of the PDG and why Taskforce/Working Group efforts have been an important part of PDG activities. The authors describe a methodology for analysing exemplified compound indexing from the patent literature in three databases: Chemical Abstracts (CAPLUS and REGISTRY), Derwent World Patents Index (DWPI) and Reaxys. This article also discusses the post analysis feedback from the database vendors. The feedback and the further discussions with the vendors culminated in some examples of positive changes to their indexing policies.

1. Introduction

The PDG (www.p-d-g.org) was founded in 1957 by 13 European companies seeking to promote the effective and efficient use of patent information. The organisation now comprises 39 multinational companies and has an unrivalled reputation in Europe as an indispensable partner in efforts to enhance the quality and availability of published and indexed patent information. The PDG also strives to promote effective searching and retrieval of all types of patent-relevant information.

Current objectives for the PDG include discussions of new developments and challenges in patent information, gathering opinions from members on shared interests, and communicating formulated proposals and suggestions to third parties in the field of intellectual property. The PDG comprises several active Working Groups. This study was undertaken by members of the Networks and Online Working Group (WG Online) which is a group that focuses on seeking to improve patent information products and services [1,2].

There are publications that comment on the retrieval of chemical information from particular databases [3–7]. There are also some more selective examples of publications where authors aim to compare some of these databases with the purpose of understanding how indexing policy and quality affects the database content [8–15]. This has an impact on searching and retrieval. The work that is described herein is a comparative study of this type. We will now discuss the objectives for this study and the background that led to this study being made.

2. Objectives

This article will make frequent references to “exemplified”

* Corresponding author.
E-mail address: Mark.Ede@gsk.com (M. Ede).

http://dx.doi.org/10.1016/j.wpi.2015.12.003
0172-2190/

compounds.

Exemplified compounds refer to specific compound substances, including salts, complexes and stereochemical forms of specific compounds. The analysis was therefore on "exemplified" compounds and excluded Markush defined structures or variable substructure definitions. Our objective was also to look at "small molecule" exemplified compound cases since polymeric compound and macromolecule cases were outside our area of expertise.

Chemical Information Scientists within the PDG had noticed ad hoc chemical indexing errors over several years but believed that these errors had not been counted or analysed in a coordinated way before. The WG Online was aware that articles had been published that described in detail how to search for chemical structures in patent information sources and the added value indexing available in such systems [3–7]. However, the team was not aware of any previous comparative study that had attempted to determine the vendors indexing policies in significant detail by extracting the compounds that the vendors had indexed in specific patents, determining where in the document the compounds were located and establishing the reasons for why they had been indexed. Therefore if the team conducted a coordinated analysis of exemplified compound indexing from 3 key information providers (Chemical Abstracts, Elsevier and Thomson Reuters) it could yield interesting and valuable data that had not been previously revealed. It was hoped that any findings of the analysis could be shared constructively with the vendors, with a view to influencing future enhancements to their product offerings. It would also help to improve the understanding of PDG member companies in this area.

From henceforth the source Chemical Abstracts will be referred to as CAS whilst the Derwent World Patent Index will be referred to as DWPI.

The project commenced with a view to analysing and comparing just CAS and DWPI exemplified compound indexing policies. The team initially chose only two sources since we were uncertain how long the analysis and comparison of a significant sample set would take. It was anticipated that we wanted to analyse CAS and DWPI indexed patent references for accuracy and comprehensiveness, to see how well they handled stereochemical forms, to see if their naming conventions were consistent and to see how well they handled the indexing of complexes. With reference to CAS we also believed that this would be a good opportunity to analyse where it indexed prophetic compounds from within patent references. We understood that CAS made a commitment to index prophetic compounds during 2007. Prophetic compounds represent compounds that are exemplified in patent references, either by name or structure, where no physical data characterisation has been provided for the compound(s) within the patent reference or anywhere else in the published art.

3. Methodology

The methodology was to analyse recent patents that were less than 2 years old from their first publication date. This was so that the analyses would reflect current indexing policies and the current indexing qualities of the vendors. The analyses were done for cases containing a manageable amount of compounds which was determined to be 99 or less for both CAS and DWPI. It was known to the team that if more than 99 compounds satisfied the DWPI indexing policy that Thomson Reuters adopted a selective indexing policy based on structural diversity. We were keen to avoid cases that had adopted this policy so we ideally focused on publications indexing fewer than 99 compounds in each source. The delegation process did allow the team to exercise a small degree of flexibility around the 99 limit and publication date so that suitable patent cases could be quickly sourced. The analysed patent cases were simply patents that the team members would have uncovered during their day-to-day work and considered suitable for analysis. There were team members contributing from three pharmaceutical companies (GSK, Novartis and Sanofi) who naturally chose pharmaceutical patent cases claiming compound structures with medical indications. A single chemical company (Henkel) contributed to the analysis by choosing their own inventive patent cases for chemical compounds that were claimed as hair dyes.

Using STN online command language, compounds were selected and tabulated from equivalent patent records and had to be from the same family member. Within CAS this select command was SEL RN 1- and within DWPI the commands were SEL DCN 1- and SEL DRN 1-. These fields (RN, DCN and DRN) provide the unique identifier numbers for exemplified compounds within patent references (RN represents the Chemical Abstracts Registry Number, DCN represents the Derwent Compound Number and DRN represents the Derwent Registry Number). The selected compounds were displayed online using the STN free format "SCAN", they were then tabulated using the STN Table Tool and finally they were exported to Microsoft Excel for analysis.

The tables were merged in Microsoft Excel and sorted into molecular formula order, at which point an examination was made to see which of the indexed compounds were either duplicates or were unique to each source. The patent specifications were then carefully examined to see where both the unique and duplicated compounds had been indexed from the publications with special attention made to the claims and the preparative example sections.

The team initially examined 14 published cases in a comparison between CAS and DWPI. Some interim conclusions were made at this stage. We were aware that the quality of the indexing can be affected by the quality of the primary information available to the vendors. Allowing for that fact, we observed that DWPI had slightly more erroneous entries than CAS with some of these entries being related to compounds that could have varied enantiomeric or tautomeric interpretations. CAS was more likely to index compounds from the claims, but did not always achieve 100% indexing of all the claimed compounds. CAS also generally indexed more compounds from the same patent document with more indexing of compounds with catalyst, reactant and intermediate roles from the examples.

The team reflected on the interim findings and on the time taken to analyse the cases. It decided after careful consideration to add the resource Reaxys to the evaluation but discount adding the resource CASREACT since this was another Chemical Abstracts resource that would have a high degree of duplication with CAS. Furthermore CASREACT is specialised to capture the indexing of reactions rather than compounds. The team believed that it was covering the 3 most important sources for indexing exemplified compounds in patents. The source SureCHEMBL was also briefly considered as an additional source, however the team wanted to focus on the 3 sources provided by the established commercial vendors that we believed to be the most important. Since the time of the analysis we understand that the SureCHEMBL data has been made publically and freely available [20].

At the time of analysis (2011–2012) the resource Reaxys indexed patent cases from the International Patent Classification (IPC) areas C07 (Organic Chemistry), *A61K (Drugs, Medicinal, Dental, Cosmetic Preparations), A01N (Biocides Agrochemicals, Disinfectants, etc.) and C09B Dyes. *We were aware that if main IPC was A61K, the cases were only indexed if the secondary IPC was in the C07 area. At the time of its launch in 2009 it covered the patent information from the CrossFire Beilstein, CrossFire Gmelin and Patent Chemistry databases. The 3 issuing authorities PCT (WO, 1978 -), European (EP; 1978 -), US (1976 -) were initially indexed and this coverage