ELSEVIER

World Patent Information 27 (2005) 309–315

WORLD
PATENT
INFORMATION

www.elsevier.com/locate/worpatin

Patents: A unique source for scientific technical information in chemistry related industry?

Mervyn Bregonje *

CobidocBV, Kabelweg 37, 1014 BA Amsterdam, The Netherlands

Abstract

A study was done to find the source of first publication and the number of non-patent publications from those whose first publication is in patent literature. Patent information is not always used as source for scientific technical information; however information disclosed in patents does not always appear in non-patent literature later. The latest studies published on the uniqueness of information disclosed in patents were Liebesny [Liebesny F et al. The scientific and technical information contained in patent specifications. The extent and time factors of its publication in other forms of literature. Inform Scientist 1974;8(4):165–77 [Reprint]] and Terapane [Terapane JF. A unique source of information. Chemtech 1978;8:272–76]. This study was undertaken to have current data on what we are missing when patent information is not used in the area of chemistry information. Looking at a sample of substance information reported in both patent and journal literature represented in the CAS databases suggests that a significant amount of this information does appear in patents before it is reported elsewhere; in some cases the disclosure in patents may truly be the only report.

Keywords: Patent disclosures; Journals disclosures; Literature disclosures; Information systems; Chemical industry; Polymers; Chemical compounds

1. Background

A patent is not just a legal document but also needs to be a technical description of an invention. Because the patent application needs to be applied for before the first publication (there are some exemptions) of an invention it is often said that scientific or technical information disclosed in a patent (application) is not disclosed elsewhere later. The proposition is that 80% of information disclosed in patents is not disclosed elsewhere. Terapane et al. [1] concluded, after he carried out a random sample of 435 mainly US patents, that 70% of new technology was not disclosed later elsewhere, 13% partly and 16% was completely disclosed in non-patent literature later. Liebesny et al. [2] concluded that 5.77% of technical information in patent specifications is published in non-patent literature later. Berks [3] concluded that attempts to quantify the uniqueness of data in patents, compared to other literature sources, are difficult. His intention was to demonstrate that important molecules are often disclosed in patents months or years before their disclosure elsewhere. If verified, this would provide evidence that unique data is present in patents that is not available, or is delayed, in other literature sources.

It is generally acknowledged that in many organizations scientists are expected to keep up to date with the literature of interest and when they do literature research they use mainly non-patent information sources. This is especially likely when scientists use different information systems for patent and non-patent literature. Certainly, scientists are aware of the importance of patent information but look at that more as a legal matter then as a source of scientific information. On

* Tel.: +31 (0)20 688 03 33; fax: +31 (0)20 681 86 26.
E-mail address: mervyn.bregonje@cobidoc.nl

0172-2190/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.wpi.2005.05.003

the other hand we see that patent information specialists focus more on patent information sources, though they are aware of the relevance of non-patent literature as prior art for novelty issues.

It is interesting to know for which technologies which source is of more importance. In a survey published by Derwent Information (now Thomson Scientific) in 2000 [4], it was concluded that small-medium sized chemical companies used the patent system more than the cross-industry average, but not as much as their counterparts in the pharmaceutical industry. Across all sectors researched, a majority of SMEs not using patents believes that they are not relevant to their line of business, and this reason was primarily given by companies in the chemical sector. If people think patents are not relevant they probably will not use them as a source of information. Also very few companies carry out searches themselves. Many (65.9% of patenting SMEs and 48% of non-patenting SMEs) rely on patent agents. If patent information is used it is mainly for checking on existing patents or patent infringements rather then as a source of technical or commercial information.

If the above-mentioned propositions are true they imply that patent literature is an under utilized resource for scientific and technical information and not adequately used as such. If patent information is not being used enough, it would be interesting to measure how much is being missed and if this is valuable. Accordingly the objective of this study is to research the position in chemistry related industry to determine the source of first publication and the number of non-patent publications from those whose first publication is in the patent literature, and thus provide current data on what we are missing when patent information is not used in the area of chemistry information.

2. Study set up

To set up a study to measure the uniqueness of information disclosed in patents we would need a source that discloses the new information disclosed in patent and non-patent literature. CAS provides pathways to published research in the world's journal and patent literature relevant to chemistry, life sciences and a lot of other chemistry related disciplines. CAS Registry^{SM} and CAplus^{SM} databases on STN International® disclose new substances or new information about existing substances in patent as well as in non-patent literature. It would also be interesting to look at data for electrical and mechanical patents but those are not covered by CAS.

The Registry File on STN contains chemical substance records. The substance records contain CAS Registry Numbers, chemical names, structures, stereochemistry, molecular formulas, ring data, biosequence information and classes for polymers. All these data may be displayed and searched in variety of ways. The Registry Number is a unique automatically assigned serial number that contains no chemical intelligence but provides a link to a substance record for each unique substance. When a substance, e.g. an organic substance, an enzyme or a polymer, is published for the first time in public literature it will be registered by CAS with a unique CAS Registry Number®.

The CAplus File on STN is a comprehensive chemistry related bibliographic database. CAplus covers international journals, patents, patent families, technical disclosures, technical reports, books, conference proceedings, and dissertations from all areas of chemistry, biochemistry, chemical engineering, and related sciences from 1907 to the present. Records contain bibliographic information, abstract, indexing terms and CAS Registry Numbers. When new or novel information is disclosed for a new or existing compound an index entry with this CAS Registry Number is added.

With CAS files it is possible to do a date search for new registered compounds in Registry^{SM} and cross them over to the CAplus^{SM} file and find out how many of these compounds are disclosed in patent literature and later disclosed in non-patent literature. To analyze trends for different kind of substances different groups of compounds were studied and samples were taken for different time periods (1980, 1990 and 1999).

3. Searching for the source of first publication

The following are the groups of compounds and the search strategy considered. The groups of compounds are arbitrary and are chosen because they reflect different chemistry related disciplines, e.g. coatings, fibers, optics, pharmaceuticals, photochemistry and plastics.

Polymers: used polymers as class identifier (PMS/CI).
Cyclopentadiene metallocenes: used a generic chemical structure (Fig. 1a).
Alloys: used alloys as class identifier (AYS/CI).
Quinolinone derivatives: used a generic chemical structure (Fig. 1b).
C12–30 unsaturated fatty acids: used a generic chemical structure (Fig. 1c).
Piperidine dervitatives: used the ring identifier (46.156.1/RID) for the piperidene ring system (Fig. 1d).

Fig. 2 represents the search and analysis strategy with the polymer example. Step 1 is selecting the appropriate Registry Numbers for that compound class for that year in CAS Registry file. Samples were taken for the years 1980 (Registry Numbers 72467-42-6 to 76081-80-6), 1990 (Registry Numbers 124508-13-0 to 131232-32-1)

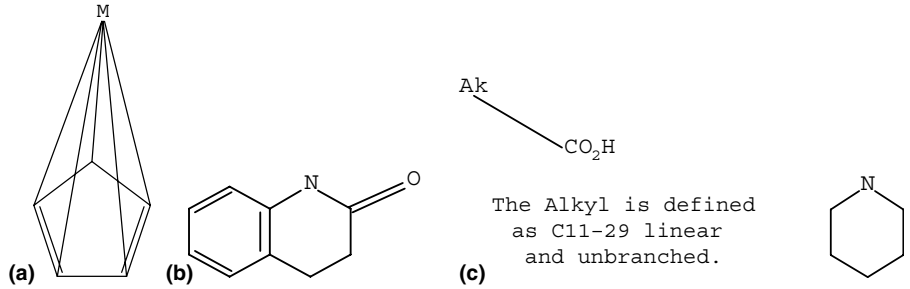

Fig. 1. (a) Cyclopentadiene metallocenes structure query, (b) 2-quinolinone structure query, (c) C12–30 unsaturated fatty acids structure query, (d) piperidine structure.

Step 1.

=>FIL REGISTRY

=> set ran=72467-42-6-76081-80-6
SET COMMAND COMPLETED

=> s pms/ci not (rr/fa or ar/fa or dr/fa)
L1 13455 PMS/CI NOT (RR/FA OR AR/FA or DR/FA)

In this step we excluded replaced (rr/fa), alternate (ar/fa) and deleted (dr/fa) Registry Numbers

=> s L1 not RELATED POLYMERS/FA
L2 11755 L1 NOT RELATED POLYMERS/FA

In this step we excluded related polymers. Those polymers would also be reflected by other Registry Numbers

Step 2.

=> FIL HCAplus

=> set ran=1980
SET COMMAND COMPLETED`

=> s L2
L3 4179 L2

=> S L3 and patent/dt
L4 2367 L3 AND P/DT

=> S L3 and journal/dt
L5 1674 L3 AND J/DT

Step 3.

=> S L3 (L) PREP/RL
L6 1143 L3 (L) PREP/RL

=> S L6 AND PATENT/DT
L7 502 L6 AND PATENT/DT

=> S L6 AND JOURNAL/DT
L8 628 L6 AND JOURNAL/DT

Step 4.

=> SET RAN=ALL
SET COMMAND COMPLETED

Fig. 2. Search strategy for the polymer search.

```
=> TRANSFER L4 HIT RN
ENTER ANSWER NUMBERS, RANGES (1-), OR ?:1-
L9      TRANSFER L4 1- RN HIT :   5434 TERMS
L10     7487 L9
ALL TERMS IN L22 RETRIEVED.

=> S L10 not L4
L11     5120 L10 NOT L4

=> S L11 NOT PATENT/DT
L12     1144 L11 NOT PATENT/DT

=> ANALYZE L12 HIT RN 1-
L13       ANALYZE L12 1- RN HIT :    403 TERMS

=> DIS L13 TOP 10 D
L13       ANALYZE L12 1- RN HIT :    403 TERMS
```

TERM #	# OCC	# DOC	% DOC	RN	*PY of 1st Non-Patent Disclosure*
1	53	53	6.13	74359-03-8 *	*1981*
2	33	33	3.82	74433-64-0	*1982*
3	33	33	3.82	75797-33-0	*1995*
4	24	24	2.78	72980-71-3	*1980*
5	20	20	2.31	73379-95-0	*1980*
6	20	20	2.31	75835-87-9	*1981*
7	18	18	2.08	73708-02-8	*1989*
8	15	15	1.74	72662-97-6	*1981*
9	13	12	1.39	75503-70-7	*1982*
10	12	12	1.39	73989-24-9	*1983*
11	12	12	1.39	75268-90-5	*1983*

```
* Compound with Registry Number 74359-03-8 is 53 times reported (# DOC=53) in non-patent
literature, after its first publication in a patent.
The Italic written publication year is manually added. This is the first year of publication of non-patent
disclosure.

Step 5.

=> ANALYZE L12 PY
L14       ANALYZE L12 1- PY :     26 TERMS

=>DIS L14 ALPH D    (Alphabetically in descending order)
```

Fig. 2 (*continued*)

and 1999 (Registry Numbers 216431-13-9 to 252063-50-6). The first Registry Number of each range is the first added to CAS Registry file in that year and the second the first of the successive year (use help RNYEAR online in STN). Registry numbers within this range were added in that year and thus reflect new compounds.

Step 2 searches for the source of first publication by refining with the document type for patents and journals in CAplus[SM] . The same date ranges were used as in Registry[SM] (1980, 1990 and 1999). We have to be aware that a compound can be as well published in patent as in non-patent literature in the same year. From Steps 4 and 5 we will have a better idea of the number of compounds republished in non-patent literature at a later time.

Step 3 differentiates for the role the compound plays in its first publication. CAS Roles are since 1994 intellectually assigned by CAS analysts, 1967–1994 CAplus[SM] records are indexed with a computer algorithm. CAS Roles are 3 or 4 letter codes that describe new or novel information reported about a compound. Interesting is what kind of new or novel information is reported with new compounds. CAS Roles used are: preparation (prep/rl), use (uses/rl) and process (proc/rl). CAS Roles information will tell us what type of information is more likely to be disclosed in patent or journal literature.

Step 4 selects all the HIT Registry Numbers (HIT RN) from the patent publications and searches them again in the whole CAplus file then refines for non-patent literature (not patent/dt). An analysis of the HIT Registry Numbers of this non-patent literature answer set will tell us how many compounds from patent literature are reported later in non-patent literature.

Step 5 analyzes the publication year for the non-patent literature answer set. This was done to see the time-spread over which compounds reported in patents show up in non-patent literature.

4. Analyzing the uniqueness of information disclosed in patents

Fig. 3 shows that there are different trends for the different compounds. Polymers are likely to be disclosed more in patent literature whereas metallocenes are more frequently disclosed in journal literature. This might say something about the commercial value of the kind of compounds. Most compound groups show that the percent of disclosures in patent versus journal literature increases over time, except for the quinolinones.

The result shows us that for all compound groups both sources are relevant. Of course we want to know what percentage is unique. Almost by definition a patent should be the first disclosure of an invention. However, the publication of a patent (application) can be delayed and in the meantime the invention could be partly disclosed elsewhere.

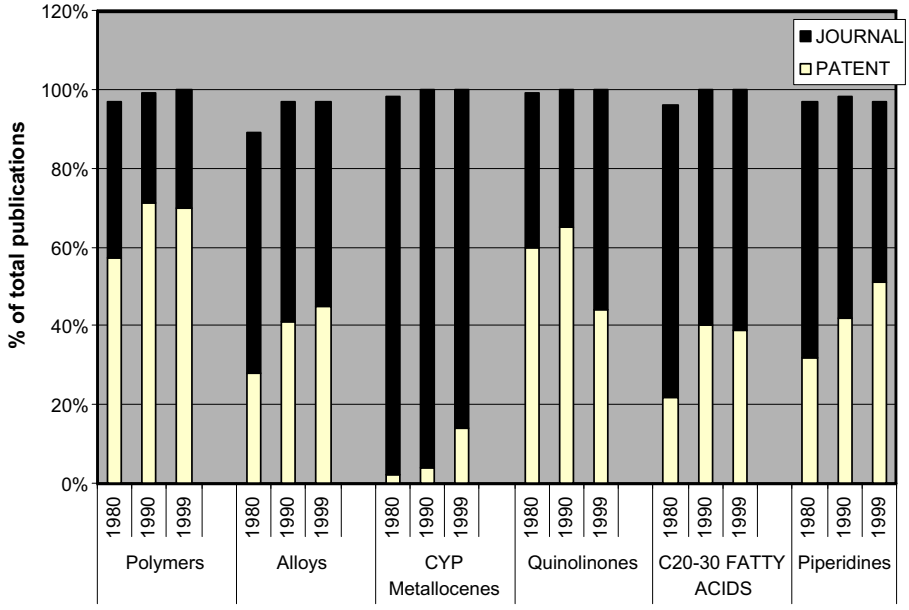

Fig. 3. First source of publication.

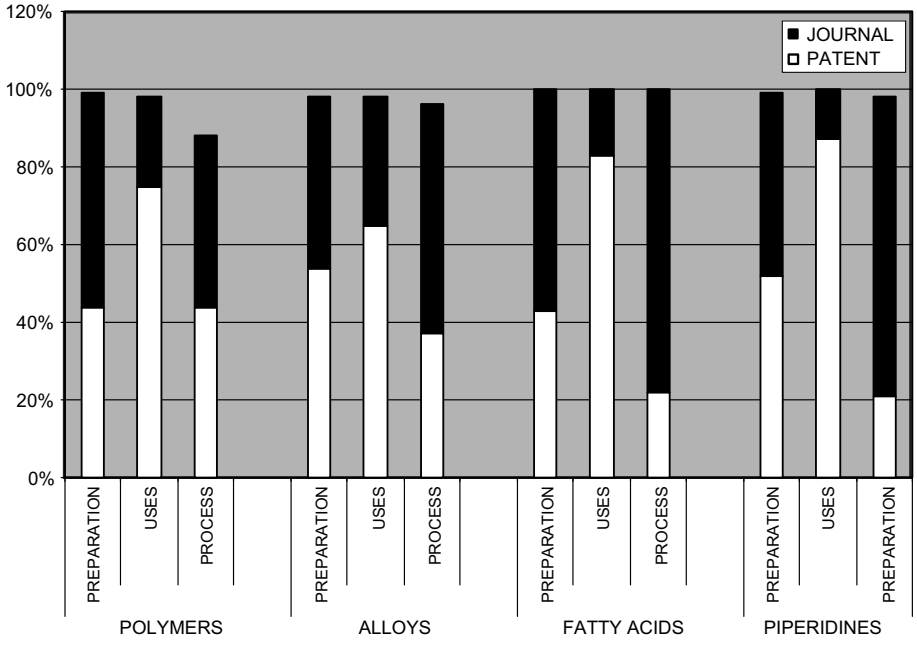

Fig. 4. First publication by role.

As we look at the role of compounds in their first publication we can see from Fig. 4 that its use is more likely to be disclosed in a patent. If one is interested in the use of a compound, patent literature becomes more relevant. Information on processes is less published in patent literature. Also metallocenes are most of the time part of a catalysis process. It is the author's perception that chemistry related industry is less likely to disclose industrial processes in any type of publication.

From the result in Step 4 (Fig. 2) we can see that for polymers 403 compounds out of 5434 (extracted from the patent records) were published elsewhere, which makes 7.4%. From the search result we cannot determine what was published, e.g. partial or complete disclosure. If we look at this as the percentage missed by scientists when they do not take patent publications into account we need to look at the total number of polymers registered that year which is 11,755 (Fig. 2, Step 1). The amount missed is then $5434 - 403 = 5031$ which is 43%. In Table 1 the data is grouped from all compound groups newly registered in 1980.

We did not look at the type of non-patent literature for disclosure of the registered compounds. To have an idea what other sources are relevant we had a closer look with the polymer example. Most of the polymer registrations are disclosed in journal literature and less than 1% in conferences, reports or dissertations. Five per cent of the polymer registrations came from National Chemical Inventories (e.g. TSCA) and did not show up in conventional publication types at all.

When analyzing the later disclosure of compounds by years with the polymer example in Fig. 5 it shows that disclosure is also spread over a time period of more than 20 years! For the 10 most reported Registry Number we looked at the first publication year of non-patent disclosure which can be found in Fig. 2, Step 4 (table with L13). With this top 10 Registry Numbers we found that the majority was published within 4 years but it can take to up to 15 years later for first publication in non-patent literature.

5. Conclusions

Whether a compound is disclosed in patent or journal literature probably depends on its commercial as well as

Table 1
Percentage of newly registered compounds in 1980 published in patents only

Number of ...	Polymers	Alloys	CYP metallocenes	Quinolinones	C20–30 fatty acids	Piperidenes
Compounds	11,755	11,418	3784	685	290	6334
Compounds reported in patents	5434	2789	37	582	72	2710
Compounds reported in non-patent later	403	246	10	57	3	524
% of compounds republished later	7	9	27	10	4	19
% of compounds published in patents only	43	22	1	77	24	35

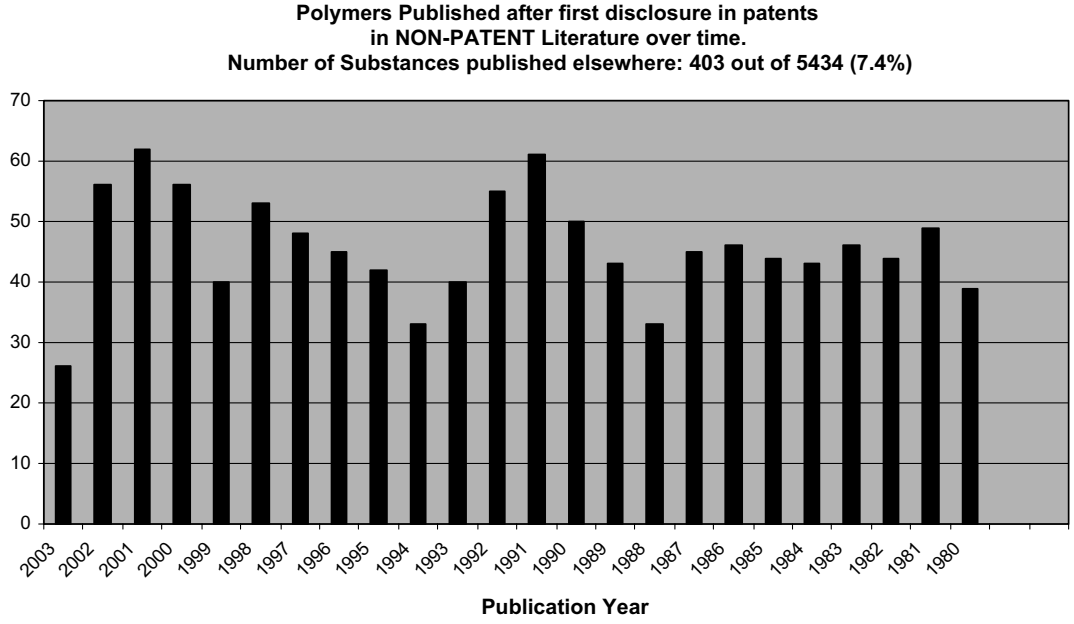

Fig. 5. Polymers published after first disclosure in patents in 1980, in non-patent literature, over time.

scientific value, and governmental regulations (e.g. for pharmaceuticals) may also be a factor. The key findings from this study are that, for scientific technical information in chemistry related industry:

(1) In most areas of chemistry, patents are a rich source of first publication (see, for example, Table 1, Fig. 3).
(2) In most fields of chemistry the proportion of compounds in which patents are the first source of publication is increasing with time (see, for example, Fig. 3).
(3) Patents are most often the first source of information for uses of compounds (see, for example, Fig. 4).
(4) If information is disclosed first in a patent there is a small chance that it is published later again in non-patent literature, i.e. if information is disclosed first in a patent, it is often the only source of that information (see, for example, Table 1).
(5) If the information is republished in non-patent literature, it is often spread over a long time period (see, for example Fig. 5).

For most areas of chemistry we conclude that both journal and patent literature are of major importance if we want to search for novelty or select literature to read in for a new project. Patent information specialists are expected to understand the importance of both journal and patent literature. It is the author's perception that for scientists this need is less obvious, not because literature research is of secondary importance to them; naturally, they want to do new research. Scientists are not expected to be aware of the differences between information systems and their scope. The information profession would do well to show scientists the value of different sources.

It is important to offer sources for scientists and patent information specialists that cover both patent and non-patent literature. There are database producers that do acknowledge the importance of both sources. CAS covers more than 50 patent issuing authorities, 9000 currently published journals and many other sources in the area of chemistry and life sciences. Some databases cover a limited set of patent offices and periodicals for a certain area of interest, e.g. BIOSIS, ENERGY and PAPERCHEM2. However, a broader range of literature and patent coverage is advisable for comprehensive searching.

Acknowledgements

This article has been developed from the author's research for WON, the society of patent information specialists in the Netherlands (Werkgemeenschap Octrooi-informatie Nederland). I gratefully acknowledge CAS staff, especially Eric Shively, for providing help and advice during the preparation of this manuscript.

References

[1] Terapane JF. A unique source of information. Chemtech 1978;8:272–6.
[2] Liebesny F et al. The scientific and technical information contained in patent specifications. The extent and time factors of its publication in other forms of literature. Inform Scientist 1974;8(4):165–77 [Reprint].
[3] Berks AH. Competitive intelligence value of patents vs. other literature sources for drug compounds. Book of Abstracts. 213th ACS National Meeting, San Francisco; April 13–17, 1997.
[4] Dismantling the Barriers: A pan-European survey on the use of patents and patent information by small and medium-sized enterprises (SMEs). Derwent Information, 2000.

Mervyn Bregonje is a manager at Cobidoc BV and is responsible for the scientific and technical information products. Working at various positions in research and sales in the pharmaceutical industry, he started in 1998 with Cobidoc BV. He manages training, sales and marketing for the information products STN International, SciFinder and SciFinder Scholar in the Netherlands. He is a member of the managing committee of the WON (the Dutch society of specialists in patent information).