Pergamon

World Patent Information, Vol. 20, No. 1, pp. 21–27, 1998

Printed in Great Britain
0172-2190/98 $19.00 + 0.00

PII: S0172-2190(98)00009-X

Patent information: are the traditional suppliers as doomed as the dinosaurs?*

Ursula Schoch-Grübler, BASF AG, D-67056 Ludwigshafen, Germany

Introduction

For a long time there has been much speculation as to why the dinosaurs suddenly became extinct. Theories abound: were they too big? Were they too slow? Or were they just too stupid? Whilst you can take your pick of whichever theory you prefer, most of them have one thing in common: the dinosaurs died out because they were unable to adapt quickly enough to a changing environment.

In recent years, more and more evidence has been gathered to suggest that the earth was hit by a huge meteorite, which resulted in enormous climatic changes. The dinosaurs were unable to cope and became extinct.

- Will the Internet have the same impact?
- And if it does — who are the dinosaurs anyway?
- Are they the traditional patent information suppliers?
- Or perhaps are we, today's information professionals, the real dinosaurs doomed to extinction?

Dinosaurs apart, I am going to try and take a look at some of the issues in patent information which confront industry today — focusing particularly on the chemical industry. What are our real problems and how can we tackle them?

Internet and intranets

Let me begin by taking a closer look at our meteorite — the Internet.

The first obvious thing is that my comparison is not particularly appropriate.

- The Internet certainly did not hit the earth and change everything overnight.
- And it has not really changed the climate — at least not yet.
- And it certainly has not wiped out any species — at least not yet.

In fact, without any doubt, the impact so far has already been extremely beneficial to the scientific community as a whole.

Intranets

When we talk about the impact of the Internet on industry we are not just talking about superhighways with access to vast quantities of data, or of a new medium for communication, marketing, advertising or transactions, we are talking about the impact of the technology itself.

Here we have, so to speak, off-the-shelf technology which can be used to create intranets which, among other things, enable companies to market services internally or provide users with access to in-house databases. The huge advantage lies in standardisation, i.e. that they have a uniform client software and that they are not tied to a particular medium, data supplier or location.

However, despite the unstoppable success of intranet systems, several aspects are already clear:

- intranets only provide a front-end to existing different IT systems already in place: these will still need to be retained and maintained;
- the intellectual input required to set up the useful links between tools is an enormous and ongoing task;
- regular updating of the additional information in intranet systems requires extra IT support;
- many of the necessary gateways between systems do not yet exist which makes even mid-term planning difficult, to say the least.

Internet

Let us concentrate now on chemistry in the patent sector. On the Internet we find firstly that there are, for example,

* This paper was presented at the 1997 Chemical Information Conference in Nîmes and published by Infonortics in the Proceedings of that Conference. It is reproduced by kind permission of Mr. H. Collier.

full-text US-patent files available from the USPTO, Questel-Orbit, IBM and Micropatents. Although in some cases only a small time-range is covered, these full-text files permit, among other things, searching for trade names or other aspects buried in the text which are not picked up by the traditional databases. They are generally either free of charge or they demand only relatively small fees. The question is: how long will this situation last?

In addition, there are also some less famous patent offices, such as Malaysia or Brazil, who offer the content of the first page of the patent. We can expect more from other offices in the future. Interestingly, Malaysia, with data going back as far as 1951 and despite being one of the "Asian tigers", is not covered by Derwent or the other major database producers. Surely if the material is already available electronically at the Malaysian Patent Office, why were the traditional database suppliers not able to get it?

Is this a case of a fast cat leaving the dinosaurs behind?

The field of non-patent, scientific literature is often relevant for novelty searches and the technical background to patents. Here the Internet offers a huge number of different systems, even if many of them are still in an experimental state, such as Science Direct from Elsevier or Springer-Link from Springer-Verlag. In addition, the Internet offers unique full-text-searchable sources which are invaluable for digging out hidden prior art:

- material from scientific conferences,
- plenary lectures and posters,
- government information and statistics,
- press releases and other grey literature.

On top of this, Internet technology now offers, besides the standard features such as Boolean and proximity operators and so on, new retrieval auxiliaries which employ fuzzy logic or intelligent search agents. Over and above these, there are also software packets for data-mining, statistical analysis and clustering of results like those developed by IBM to help exploit the information retrieved. And all is easily accessible without needing to learn peculiar command languages or Byzantine coding systems.

Turning to the producers of the traditional databases, they have always gone to great efforts to devise indexing systems of varying levels of sophistication and quality to enable us selectively to search the vast volumes of information available in the literature. The production of this indexing is labour-intensive and correspondingly expensive. Bearing in mind what the Internet now offers we should ask ourselves: do we still really need the in-depth level of indexing offered by the classical database producers?

And if we consider that in future we may also have the patent applicants' abstracts available from most of the patent offices, will we really need all those other expensive patent abstracts? Are the suppliers of primary information — patent offices and publishers — now going to take over the information market? Are the value-added databases to be added to the list of endangered species? Is this the beginning of the end for the traditional suppliers?

Indexing and value-added databases

Well, we in industry certainly do ask ourselves these questions and we watch carefully how the information market is developing.

To get a clearer picture of the search possibilities on the Internet we made comparison searches of "real life" cases in which we compared the results from the Internet searches with those in the traditional databases. As an example we looked for patents reporting the preparation of retinol.

Searching the traditional files CAplus, MARPAT, WPIDS and WPIM resulted in each case in between 27

Table 1. Patents for Chemical Preparation of Retinol

Search Criteria	Database (Host)	Number of Answers	Relevant Hits
	Traditional Databases		
68-26-8p	CAplus (STN)	51	33 (65%)
11103-57-4p	CAplus (STN)	35	6 (17%)
Structure	Marpat (STN)	27	7 (26%)
0282-P	WPIDS (STN)	50	32 (64%)
Structure	WPIM (Questel)	30	10 (33%)
Chemical Coding	WPIDS (STN)	94	45 (48%)
	Internet		
Retinol, Vitamin A Alcohol	QPAT-US	90	2 (2%)
Retinol or Vitamin A + Prepn.+ Patents	HOTBOT	1883	Not examined
Retinol or Vitamin A + Prepn. + Patents	EXCITE	1251	Not examined
Retinol + Patents, Claims	HOTBOT	111	None
Retinol + Patents, ClaimsClaims	EXCITE	99	None

and 94 answers, of which between 6 and 45 were identified as being highly relevant. In other words our precision varied between 17 and 65 per cent.

A search in QPAT-US gave 90 answers but only two of them were relevant. This is a precision of only 2 per cent. The search for retinol in combination with patents using the two popular search engines HotBot and Excite resulted in well over 1000 answers each at our first attempt. These were not further investigated because it would have required at least two days to download them. In a second attempt with a new search strategy we retrieved 111 and 99 answers respectively, but none of them were relevant.

Leaving aside the problems of security and reliability, the results brutally exposed the limitations of the current Internet options. We conclude from our experiences that:

- Patent searching is almost exclusively restricted to the US-patents.
- Searching the full text of patents leads, depending on the search profile, either to far too many lost answers or far too many non-relevant hits.
- Neither chemical nor Markush structure searching is possible.
- It is far more time-consuming to scan through unstructured material than the corresponding, standardised abstracts.
- Getting reasonable results — as opposed to any old result — requires experience and searching skills comparable to those needed for the classical databases.

So much for click and go.

All told, the results simply confirm that you get what you pay for. The Internet was nominally very cheap but could not deliver the goods. The yield in terms of relevant documents per hour of the searcher's time was miserably low and once this is included in the bill, the Internet search becomes expensive and poor value for money.

So are the dinosaurs in for a reprieve? Not necessarily.

The Internet is still in its pioneering phase and today's results are not necessarily tomorrow's. It is evolving quickly and many of the current deficiencies will no doubt be remedied in the near future.

Traditional suppliers—endangered species?

Nevertheless, before we decide to junk our traditional suppliers let us look what they currently offer compared to the Internet:

- They are technically rapid and reliable.
- They have a single language: the abstracts are in English and you only need to search in English.
- They are one-stop shops. They bring together numerous sources of original literature within a single database and numerous databases within a single search system.
- They offer multi-file searching and clustering of databases with duplicate detection and elimination.
- For patents there is only one record per patent family — retrieving equivalents is avoided.
- The backfiles contain much more relevant prior art from the past thirty years than the Internet can offer.
- The levels of precision and recall achieved through intellectual indexing remain far superior to those achieved with Internet sources no matter how many bells and whistles have been added to the search engine.
- For chemists there is the added bonus of high precision searching of chemical structures and generic structures.

When you think about it, it is incredible that the Internet is hogging the headlines at all when the traditional suppliers have so many "goodies" on offer. In my opinion, the Internet with all its new features has not materialised as the huge meteorite which caused the climatic changes.

Unfortunately, we have not actually needed any meteorite to change the climate; the dinosaurs have managed to do it themselves already. This may be news to you, but there is another theory about the extinction of the dinosaurs which we have not mentioned so far- that the dinosaurs committed mass suicide:

- Some of the traditional suppliers behave as if they were exclusive clubs — closed shops for members only. They have high barriers to entry: any newcomers are expected to stump up hefty, up-front subscriptions without having any real idea of the cost — effectiveness of the files. They are expected to buy a pig in a poke.
- Over and above this, the marketing and training of many suppliers was mainly geared to introducing a product to new users. But if they really want to educate the users to become "professionals" who get the most out of a product, the following issues need to be addressed:
    - If the suppliers do not understand the real uses of patent information and how industry employs their products, how can they convince non-users of its potential value and sell them the products?

- If they are not familiar with the competitors' products, how will they recognise their own competitive advantages and sell their products? How can they provide feedback to their own companies to improve their products?

- How can trainers expect to be able to teach effective searching techniques unless they fully understand the indexing systems? Conversely, how can they demonstrate the optimum use of indexing if they themselves have too little experience of searching?

I have really never been able to understand why suppliers do not recruit more of the missing skills from the user community to market their products.

That said, what I find much more difficult to understand are some of their marketing strategies.

- Instead of trying to lure newcomers to take a ride on their flagships — the value-added, deep-indexed files — some suppliers fob them off with raw patent material which, for what it is worth, is already easily available elsewhere. Take Derwent, for example, who instead of trying to promote the advantages of the World Patent Index, dumps the Patent Explorer on to the potential users. What sort of marketing strategy is that? One that keeps the best assets hidden! Do suppliers really believe that by stuffing new customers full of junkfood, they will get hungry for lobster? Now there is no way that I condone drug pushers, but they could sure give suppliers a lesson in marketing. How can anyone who has never been treated to the "luxury" of a decent set of search results ever become hooked on indexing. Let them try the real thing and they will be coming back for more.

- Even patent offices have fallen into the same sort of trap. They have invested very heavily in technology to make enormous volumes of data readily accessible to the public. This investment will be a waste of resources if the result is simply that people drown in data rather than make creative use of it.

- One might have expected that all the traditional suppliers would see the ready availability of these data as a golden opportunity to rationalise their production processes, avoiding the costly duplication of data input, and invest the savings achieved in increasing the content of intellectually added value in their products. Instead, however, some prefer to try to make a quick buck by taking raw data and putting a fancy label on it.

- By and large, the traditional suppliers have always been eager to pass on any increased costs to their customers as price increases. This creates a vicious circle where each price increase drives the customers who can least afford it to seek cheaper alternatives. In the end only a limited circle of users remain to bear a multi-million dollar burden on their shoulders. Sooner or later, they will question whether it is not wiser to invest their money in creating the files which fulfil their specific needs — even if it means the rebirth of private clubs like the IDC.

The writing has been on the wall for years now. For too long suppliers have been happy to scoop up the profits and to assign reinvestment a low priority — a good recipe for extinction. Compared with the strides made in information technology, the suppliers have been crawling along at a snail's pace. And the investment that there has been, was not always wisely spent. It often went to fund peripheral products rather than enhancing the core activities.

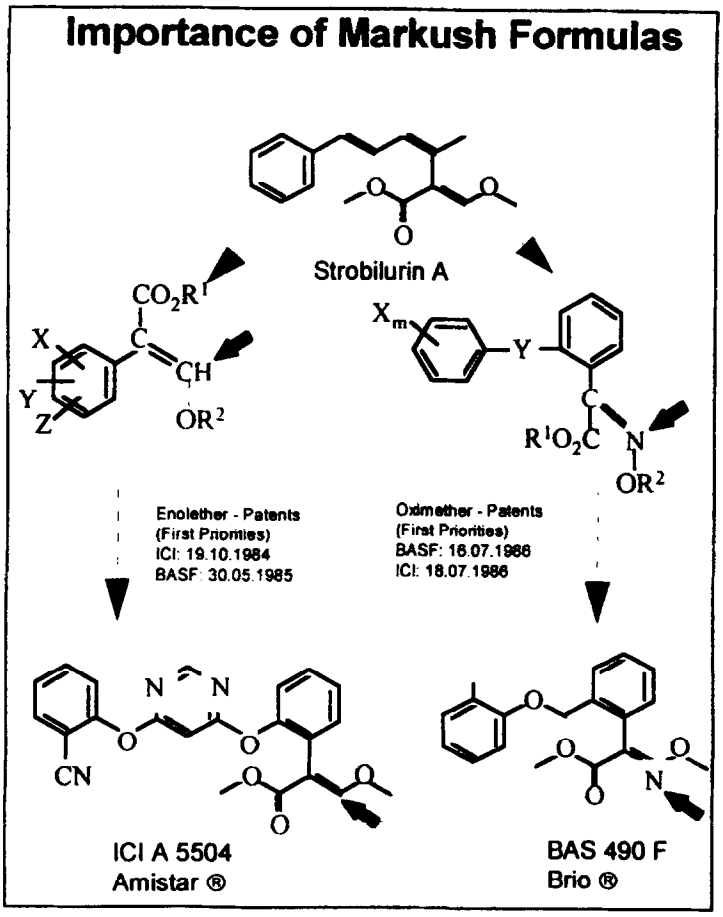
At the same time really innovative new products are mushrooming all over the place — except from the traditional suppliers. Most of them, however, live and die as niche products because, no matter how good they are, they are always too limited in their application to achieve wide acceptance in industry. These products are rarely developed to their full potential because they lack the financial backing and promotion to make the big time. It is painful to witness so much creative talent going to waste. Why do the database producers not exploit this situation and take on board those products which can be integrated into their portfolios so that they can refit their flagships to be seaworthy for the 21st century?

Markush files

I think perhaps it is worthwhile now to examine why industry expends so much effort in searching for good patent information. I can assure you we do not spend millions on searching just to keep our professional searchers happy and contented. The database producers do not see the world much differently: good indexing is such a difficult and time-consuming task that I can guarantee to you that nobody does it just for fun. So why do we do it?

A key factor of survival in the chemical and pharmaceutical industries is the effective protection of innovations by patents. An idea of just how hot the competition can become is demonstrated by the example of two fungicides of the strobilurine type which were developed simultaneously by ICI and BASF. Both the production and use of them is protected by patents in which the molecules are described using huge Markush formulas.

The real difference within these structures centres on a single nitrogen atom. The applications for patent protection from the two companies were within days of each other. In one case ICI was first and in the other BASF won the race. Each of the patented products, Amistar from ICI and Brio from BASF, are expected to achieve sales of more than $100 million per year.



There is much at stake and a prerequisite for investment in such projects is a knowledge of what has already been patented — right down to the finest details of the Markush structures. To get this information fast, or even at all, deep-indexed patent databases are essential.

At the moment there are three different systems on the market:

- MARPAT from Chemical Abstracts Services available on STN
- WPIM from Derwent available via Questel-Orbit's Markush Darc
- Pharmsearch from INPI which is also available via Markush Darc but has different search features.

Bearing in mind the example I presented, one would expect that there is a large, genuine demand for these three databases. However, their usage has remained persistently low and now, to boost profits, the producers are questioning their futures. Why should it have come to this?

- All three files compete with one another. The bitter pill for us to swallow is that although the files are based on the same source material and thus theoretically might be expected to be redundant, searching them often leads to different results. This is due not only to different indexing philosophies and different search systems, but mainly because they contain different errors! Hence one needs to search two or even all three files.
- Despite this, most users' online budgets will no longer allow searches in more than one of the files. So if you know that none of them are reliable, why bother to search any of them? A poor reputation, once earned, is hard to shake off.
- Searching these files requires expert knowledge, but early retirement and staff reduction in industry have meant that the pioneers have gone and the novices have never been "broken in". To compound this misery, the suppliers' training is hopelessly inadequate to take on this task.
- Patent offices, themselves a potentially large user group, are not using Markush files much, because they focus only on the specific compounds in their novelty searches. They pass the burden of infringement and opposition searching back to industry.

I find it amazing that the three database suppliers have not already sat down together to work out how they might co-operate to improve their own situations. Even if it is not possible to get a perfect Markush file, then at least all three existing files should be available with the same search features under one and the same improved search system.

Document delivery

Now just imagine that you have struggled with the Markush files, or even worse, a set of fragmentation codes, and then you have happily scanned through the abstracts sorting them for relevance and then....crash. You are suddenly stopped short at the next barrier, namely, getting hold of the patent specifications.

You now have a wonderful set of choices for getting the full text of the original document:

- via Internet (World Wide Wait)
- via a reading room of the patent office (book me a flight please)
- via a conventional document supplier (where's the cheque book?)
- via a library in-house or a CD-ROM collection (if you are lucky)

What I am sure you would like to do — what I would definitely like to do — what it makes economic sense to do — is to sit at the desk and call up the documents electronically. Just at the click of a mouse.

For the privileged few who work at the EPO or in some large industrial companies this dream has already become reality. But is there any reason why such a service as this should not be available via the public networks with access for everyone, big or small?

The reasons lie in a confused perception of competitive advantage. Every owner of data wants to bring their products directly to the users without any middle man. Yet if we look at the most effective information retrieval chain, it consists of

- searching indexed information
- evaluating abstracts
- examining the full text of the original
  all in one unified system.

Like it or not, the natural middle man in this chain is a host.

However, instead of seeking co-operation with hosts, many publishers and patent offices are investing heavily in electronic publishing technology — trying to reinvent the wheel instead of focusing on topics where they have more expertise. To their astonishment their square wheels have not proved anywhere near as popular as they imagined they would. By trying to go it alone, the earnings have been small or even negative and progress has been awfully slow. In the meantime users are confused and frustrated at what might have been achieved at only a fraction of the cost. Let us hope that enlightenment is forthcoming and that the suppliers of primary literature recognise that it is wiser to integrate into the existing systems and by doing so put into place the final link in the information chain.

Approaches such as the inclusion of patent full-text files such as EUROPATFULL and USPATFULL with easy crossfile possibilities to other files at STN or the AUTODOC project at FIZ-Karlsruhe are clearly both steps in the right direction; let us have some more!

The battle for survival

So are the traditional suppliers as doomed as the dinosaurs?

Perhaps the dinosaurs are starting to feel the cold now. The weather is certainly becoming a little rougher because the economic climate has changed. Information budgets are being put under ever increasing pressure by the flood of information, the effects of which have been amplified by the emergence of new media. No one can take their market share for granted any longer, for there are plenty of newcomers looking for a slice of the cake. There is no more room for slack management, silly pricing policies or egoism. Management policies directed towards more efficient production, broadening the customer base and delivering what the customer really wants are the orders of the day. Those who look to our needs can count on the support of industry, but there will be no rescue for those who cannot deliver the goods.

Consequently, it is essential that the traditional patent information suppliers concentrate on improving their core activities. They must ensure:

- the coverage and completeness of the information,
- the timeliness with which it is delivered
- the consistency with which it is processed.

Taken together these add up to quality. Quality in these terms cannot be achieved through purely technical means, but always requires intellectual effort to provide added value in converting raw data into structured information. It is precisely this added value that industry is prepared to finance.

The key is to recognise that this is not a game where there is only one winner since no-one has the capacity to do it all alone. Survival and success will depend as much on co-operation as exploiting one's own strengths. All parties — patent offices and publishers, database and software suppliers, hosts and users — need to start to work together constructively — both technically and politically. To gain a win-win situation they need to develop systems which can be integrated in the information chain, independent of their locations or to whom they belong.

Besides what I have proposed already there is plenty of scope for more co-operation, for example:

- To provide quality products at a price we can afford, database producers must streamline their production and strive towards optimum efficiency. Moreover, to reduce production costs the amount of data duplication must be reduced by maximising access to the available machine-readable data. The Patent Documentation Group (PDG) has already requested the WIPO to standardise not only text but also graphical information in machine-readable form.

- Another obvious candidate for co-operation would be for CAS to make the Registry Numbers available to other data suppliers at a fair price and thus make them the universal key to chemistry. This would not only guarantee CAS "immortality", but it is also the only way to avoid the development of parallel registry systems. If new systems do emerge, however, we can be fairly sure that the first to offer open access to all will become tomorrow's standard.

- Users are interested in good information at their fingertips. They do not care whether it comes from external or internal sources or on which medium it is stored. They very much care, however, that they can access it through a standard IT environment. The implementation of the new technologies will only be exploited to the full by those who can achieve the synergies arising out of combining them with the existing technologies. What industry needs are solutions for the whole business process. These can only be achieved if technology from different sources is fully compatible and can be integrated. These solutions must be what the client needs and not what the suppliers dictate.

- User-friendliness and ease of access will be the critical factors in winning over a new generation of users. Why do we have to make it so difficult for newcomers? To give them access to the deep-indexed, high-quality files more intuitive search tools must be developed. Instead of spending weeks on training courses and delving into manuals, intelligent user-interfaces are needed such as:

  - software to convert natural language queries into search queries for codes or controlled vocabulary,

  - easy-to-use interfaces for structure searching-for example by putting SciFinder capabilities onto STN,

  - translation tools to allow searching multilingual files in any language; a first small step might be to offer as an online search tool the International Patent Classifications in the five languages in which they already exist.

- On the one hand we now have data available in plenty from the suppliers and on the other processing tools such as data-mining, clustering and pattern recognition which have been created by talents from outside the traditional information field. Would it not be an excellent idea if they got together to exploit their potential and add value to the traditional files. Through this we could achieve a shift from information management to knowledge management. This would indeed be a classic case of the sum being greater than its parts.

- The time is ripe for patent offices, database producers and hosts, to get together with industrial users to develop a joint program to educate novices in the rites of patent searching. Patent offices should teach the relevant aspects of patent law, the suppliers and hosts should provide the databases and search tools and professionals from industry should contribute their experience and know-how to show how to deal with "real-life" problems.

There is evidence today that at least some dinosaurs do see the need to evolve and that they are already doing so.

For example, over the past few years Chemical Abstracts Services and FIZ-Karlsruhe have become less introverted and have focused on the users. Production has been streamlined and there has been an enormous improvement in the currency of patent coverage. Clearly, familiarity with Darwin's postulates about the survival of the fittest is a good management basis. At the same time they have improved communication software for professionals with STN-Express, and have developed SciFinder and STN-Easy for end-users. They have achieved all these and at the same time have still demonstrated an appreciation of the budget constraints of the users: price increases have been held down to inflation rates.

If CAS, in addition, were to become more co-operative and open-minded with its special friends such as MDL, DIALOG, Derwent and others, it should be more than warmly congratulated.

Coming back to our metaphor for the industry, I am optimistic that at least some of our dinosaurs have a sporting chance of leaving the Jurassic behind and evolving into birds.

Finally, I would like to thank my colleagues Dr Isabella Adams and Dr Geoffrey Fairhurst for their help and advice in preparing this paper.