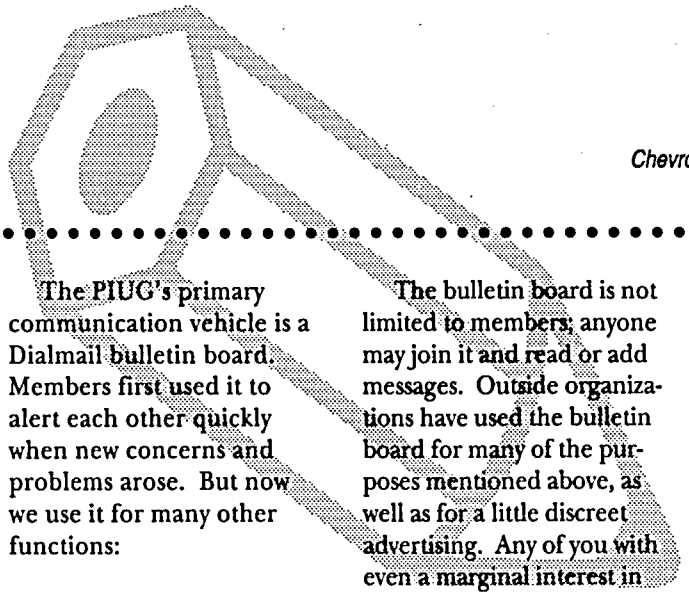
The PIUG Pen



by Nancy Lambert
Chevron Research and Technology Company

First let me introduce myself. I have been a patent searcher for eighteen years, having first studied chemistry (B.S. from Carnegie-Mellon University; two years of graduate work at Princeton University) and information science (M.S., Columbia University School of Library Service). I discovered and fell in love with patent information when I took my first industrial job in 1974, at 3M in Minnesota (a beautiful place from May through October). Since the late 1970s, I have been an active member of the patent information community.

Around seven years ago, I moved from Minnesota to Chevron in California (a beautiful place all year round), where I am currently chief cook and bottle-washer for patent information. At Chevron I have had to expand my horizons to search a lot of sci/tech and a little business literature as well. But my first love, patent information, remains my true love — a fact reflected in most of the talks I've given and papers I've written. This column, while it will cover items of interest in the sci/tech online literature, will probably also reflect my passion for patents, at least as much as the editor lets me get away with.

Next, let me introduce PIUG, the Patent Information Users Group. This informal organization of North American patent information professionals has been around since 1988. Originally formed as a sort of advcacy group where patent searchers could voice their common concerns about developments that affect the patent information world, it has evolved into a strong network of about 115 members.

The PIUG's primary communication vehicle is a Dialmail bulletin board. Members first used it to alert each other quickly when new concerns and problems arose. But now we use it for many other functions:

- ➢discussing new patent products and problems encountered in using them;
- ➢asking for help and answers to questions, and responding to same;
- ➢announcing meetings and patent-related courses;
- ➢reporting on meetings;
- ➢announcing job openings; and
- ➢pointing out interesting or even amusing patents (Yes, patents can be funny.)

One of the most prolific uses in the beginning was as a forum in which PIUG members urged each other to support causes. The current hot topic is computer and printer problems people have encountered trying to use the multitude of CD-ROM patent products newly available.

The bulletin board is not limited to members; anyone may join it and read or add messages. Outside organizations have used the bulletin board for many of the purposes mentioned above, as well as for a little discreet advertising. Any of you with even a marginal interest in patent information should join this bulletin board — and also join the PIUG, for that matter. (Dues are $10 per year. Membership information is available from the PIUG Treasurer, Pat Dorler.)[1]

Originally, this column was conceived with the idea that I, as a PIUG member and Recording Secretary, would lift my pen every other month to discuss PIUG matters. (Hence the name of the column. What?! You read a different meaning into that name? Tsk, tsk!) As I intimated above, this too has expanded. So I will indeed discuss PIUG items of interest, but I will also touch on new developments in all areas of patent and sci/tech information.

My approach to information retrieval has, of course, been strongly influenced by my concentration on patent information. We patent searchers, especially those of us who mostly computer search (as opposed to the unfortunates who spend their time digging through bins of patents at the U.S. Patent Office), are privileged characters. For one thing, our clients — especially patent lawyers — practice a "money-is-no-object" philosophy in searches for important cases; so we don't need to worry about doing the search cheaply. On the other hand, we do worry, very often, about doing it thoroughly. We can't afford to miss relevant patents when, for instance, our clients launch a $10 million business on the assumption that they are not infringing anyone else's patents, based on our search results.

We have to use every trick in the book to search our databases; and we must know the databases well enough to be able to give our customers a realistic idea of how complete a search we have provided for the question at hand. We must be able to judge when to employ more drastic measures, such as a manual search at the Patent Office.

The patent databases available to us also make patent searchers privileged. We don't usually have to resort to hit-or-miss free-text searching. Patent databases reflect a lot of intellectual effort invested in the design and construction. They have some of the finest indexing in the information business.

Major subject-searchable databases exclusively or largely devoted to patents — for example, Derwent World Patents Index (WPI), the IFI Comprehensive Index,

portions of Chemical Abstracts, the American Petroleum Institute Patent file (APIPAT) — provide a variety of subject access points, particularly for chemical patents. These include controlled-vocabulary indexing of subject concepts; registry number indexing of specific chemicals; chemical fragmentation systems; and polymer coding systems. In addition, U.S. and international patent classifications are searchable in various places.

Each database has its own particular strengths and frustrating weaknesses; each suits some questions better than others. Frequently, in a question combining several Boolean sets — for example, a chemical composition and its application — some databases suit searching one set better than another. As a result, patent searchers have all the usual problems when we must search several databases for one question.

And, of course, we patent searchers normally search all possible databases for a particular question, in order to avoid missing relevant patents. This practice has become rather easier as online hosts improve cross-file searching capabilities, letting us move sets of patents between databases and merge search results from different databases. We can even start a search in one database, transfer a large set of discovered patent numbers to another database, and finish the search there. Most online hosts currently provide the ability to search multiple databases at the same time, an ability useful in patent files. Hosts that emphasize patent information will soon provide duplicate detection in patent records, most likely based on priority filing information.

But we haven't achieved Nirvana just yet. All the current and upcoming bells and whistles still leave us searching Chemical Abstracts with only CA search parameters, Derwent with only the indexing in Derwent records, and so on. Simultaneous multifile searching with duplicate detection and removal will only give us a faster way of merging results from separate database searches.

Let me digress a bit before I finally get to the main point of this month's column. The hottest news in the patent information community these days is FIZ Karlsruhe's recent announcement that it will mount the Derwent files, most notably the World Patent Index, on STN International, the CAS (Chemical Abstracts Service) online host. For the first time, patent searchers will have access to Chemical Abstracts in all its glory (with abstracts and STN's structure-searching capability) and the Derwent World Patents Index on the same host — thus partly defusing complaints long heard in the patent information world. Whatever the motivation behind this move, patent searchers enjoy seeing cooperation replace some of the long-time rivalry between CAS and Derwent.

Patent databases reflect a lot of intellectual effort invested in the design and construction.

CAS recently gathered a dozen of the more notoriously outspoken members of the patent information community in Columbus, Ohio, to participate in a focus group, with CAS staff discussing issues and needs arising from mounting the Derwent files. This turned out to be a most stimulating brainstorming session, featuring lots of juicy ideas, both realistic and blue-sky, flying in all directions. I took advantage of the opportunity to air an idea dear to me for a while now, one that all the major online hosts should consider — in my opinion.

Almost ten years ago, Stuart Kaback mused that it would be great if we didn't have to do all that cross-file and multifile searching — if the indexing from the different patent databases, with all their varied strengths, were merged into one mega-patent database.[2] A small part of that dream has come true: WPI and APIPAT have indeed merged on ORBIT. This was a true physical merger: Records common to both databases were identified, and API indexing was added to corresponding Derwent records.

Unique APIPAT records were then added separately to the WPI database. The result is a synergistic whole far greater than the sum of its parts, because patents originally in both files can now be searched with a mix of API and Derwent indexing. Searchers can frequently retrieve patents unobtainable by either API or Derwent indexing alone.

The problem? The merger proved far more difficult, time-consuming, and expensive than anyone — Derwent, API, or ORBIT — had anticipated. ORBIT may not try another such merger; nor, for that matter, may other online hosts. Furthermore, API and Derwent have a history of joint ventures; for instance, API does its indexing from Derwent documentation abstracts. Other patent database producers may not want to cooperate to that extent with firms they see as rivals.

So what's the solution? Are we forever fated to juggle lots of patent databases in order to do comprehensive searches? Maybe not. The idea that I brought up at the CAS focus group I've dubbed "virtual file merging." Some of the building blocks for it are already in place. Host computers let us search many databases at the same time, and they can recognize the same record in two or more of these databases. The next logical step is for host computers to recognize and combine the indexing for that record from all the databases where it appears, and to let us retrieve that record with a combination of indexing from different databases.

Let's take a rather simple example. Say that I'm searching IFI and WPI at the same time for a certain composition in a certain application. I've created a set of search terms for the composition, combining IFI terms (chemical compound and fragment terms) and WPI terms (chemical coding). I've created another set of search terms for the application, again combining IFI terms (general term indexing) and WPI terms (Derwent manual codes). In a normal OneSearch situation, the computer would search each database with the indexing available in that database, string the separate sets of results

together end-to-end, and get rid of duplicates upon request.

In a virtual file merging scenario, the computer would recognize the same patent in different databases and pull that patent when for instance, it was indexed for composition only with WPI chemical coding and for application only with IFI general terms; or, alternatively, indexed for composition only with IFI fragmentation terms and for application only with Derwent manual codes. In other words, it would retrieve records indexed for both my composition and my application, but each part from a different database.

Records would not be retrieved by searches of the separate databases. In a worst-case scenario, imperfect duplicate detection algorithms (or typos in records) could cause the computer to fail to recognize the same record in different databases. In those cases, it would simply revert to the OneSearch situation now in existence, retrieving a record only by separate database search parameters. I would lose nothing; I could only gain unique records.

Keep in mind that this capability would by no means be limited to patent databases. As duplicate detection algorithms for literature databases become more sophisticated and reliable, it could work anywhere.

Is this scenario technically feasible? I discussed computer needs and possible problems with Sophie Hudnut of Dialog and Lindley McGrew of ORBIT. The challenges are considerable.

First, the host computer must be able to link records between databases. This might be an offshoot of duplicate detection, or it might be a completely different algorithm.

Next, let's suppose I search two or more databases with several Boolean sets, each containing a mix of indexing terms from different databases. For each search statement produced by a group of terms joined by a Boolean OR, the host computer must be able to look at the records that the search logic pulled from each database, and then pull and add to the set corresponding records from the other databases, even if not they are retrieved directly with the indexing in those databases. A record's indexing from all its databases must be available for further search refining.

Finally, the host computer must recognize the same records from different databases and then do Boolean operations on them. The problem with this last demand is that, when we search, the computer looks in inverted indexes of the terms we're searching and pulls some record identifier — usually an accession number — for records indexed with those terms. When sets are combined with a Boolean AND, the computer matches accession numbers. And, of course, one record in three databases would have three different accession numbers. So either preassigned cross-reference numbers would have to proliferate greatly, or the host computer would have to identify duplicates early in a search and assign on-the-fly dummy numbers that work across databases.

Another little complication stems from the patent databases' different definitions of what constitutes a patent family. If the hosts define duplicates by priority information, then duplicates will occur *within* databases, most notably in WPI. (well, no; *most* notable in INPADOC as mounted on ORBIT).

Is this scenario politically feasible? As long as we choose to print or display whole records from one (or more) databases in which they appear, per-record charges should be fairly straightforward. Search term and time charges could get a little more complicated. Serious technical problems could arise in developing computer algorithms to track data from the same record in different databases and assign each database a fair share of the revenue in each search. Search pricing should certainly be set to ensure that the database producers don't lose income from the virtual mergers.

If what I've proposed is technically feasible, and if we searchers want it badly enough and rattle enough cages, the political problems should be solvable. Online hosts: Get busy! The first one of you to perfect this capability will win the undivided loyalty of many searchers.

Notes

1. Patricia A. Dorler, E.I. DuPont de Nemours, CR&D, BMP 14–1100, Wilmington, DE 19898 (302) 992-2652; fax 302-892-1997.

2. Kaback, Stuart M., "Online patent searching: the realities." *Online*, vol. 7, no. 4, July 1983, p. 22–31.

STATEMENT OF OWNERSHIP AND MANAGEMENT

DATABASE SEARCHER is published monthly, with July/August, November/December and February/March combined by Meckler Corporation, 11 Ferry Lane West, Westport, CT 06880, Alan M. Meckler, publisher; Norman Desmarais, editor. Meckler Corporation, owner. Shareholders of more than one percent of stock: Alan M. Meckler, 11 Ferry Lane West, Westport, CT 06880; Meckler Corporation, 11 Ferry Lane West, Westport, CT 06880. Second Class postage paid at Westport, CT and at additional mailing offices.

Extent and Nature of Circulation

("Average" figures denote the number of copies printed of each issue during the preceding twelve months; "Actual" figures denote the number of copies of the issue published nearest to filing dated October, 1992.)

Total Number of Copies Printed: Average 1,716; Actual 1,960.
Paid Circulation: Average 0; Actual 0.
Mail Subscriptions: Average 1,016; Actual 1,100.
Total Paid Circulation: Average 1,016; Actual 1,100.
Free Distribution: Average 250; Actual 400.
Total Distribution: Average 1,266; Actual 1,460.
Office Use, Leftover, Unaccounted, Spoiled after Printing: Average 450; Actual 460.
Total (sum of previous two entries): Average 1,716; Actual 1,960.