Polymer Information: Storage for Retrieval, or Hide and Seek? Introduction†

STUART M. KABACK
Exxon Research and Engineering Company, P.O. Box 121, Linden, New Jersey 07036

Received July 7, 1991

The scientific and patent literature dealing with polymers continues to grow rapidly, but searching this literature can be particularly difficult. This paper and those following it deal with the current state of the art of searching the patent literature for information on polymeric materials. This is an imperfect process, and some of the problems will become clear from these papers.

When the ACS met in Philadelphia in the fall of 1984, there was more happening on the program of the Division of Chemical Information than the Open Meeting of the CAS Committee. One session was taken up by the Skolnik Award Symposium in honor of Monty Hyams, and I was privileged to be one of the speakers.

My title at that time was Polymer Patent Information Systems Could Be Even Better. If I had wanted to be lazy about it, I could have scratched the word *patent* and reused the title for this symposium, because the underlying theme is the same. There is a lot out there, and a lot of it is quite good and quite helpful—but there is a lot of room for improvement. The problem by and large is not that polymer information is not documented; I think it is fair to say that there is considerable documentation of both patent and nonpatent information on polymers. The problem is more one of locating the information you need once it has been secreted away by the various documentation organizations—and locating it accompanied by a minimum of extraneous, undesired information. In other words, the well-known target of combining high recall with high relevance.

High relevance: that is the key to the whole situation. You can always get high recall from a file that contains relevant information by dumping the entire file, but that is probably not wise with the polymer literature. The volume of this literature is huge, and it continues to grow apace. I can recall that when I first started paging through Derwent's Plasdoc each week, in the late 1960s, the yield was on the order of 400 basic polymer patents per week. By the time the Chemicals Patent Index (CPI) started in 1970, the weekly average was about 515, and 10 years later, in 1980, there were still a bit under 650 new items each week. But by 1990 the weekly count had zoomed to nearly 1100. The issue for week 9050 was, I believe, a new all-time record, and during March 1991 the grand total of items in ORBIT's version of the Derwent file—which of course includes some APIPAT-derived items—pushed over the 900 000 mark.

Don't try to memorize the numbers; my only point in bringing them up was to give an indication of the size of the mountain we keep trying to climb, and how rapidly it is growing. And that is just the patent subset. Polymer scientists and technologists are learning rapidly how to do more and more sophisticated things with polymers. The challenge is to find ways of documenting their accomplishments with effective and discriminating systems that can be applied consistently by indexers on the input side, and can be understood and utilized by users on the output side. Not only must these systems work effectively; they must have costs low enough so that they will be used, but high enough to provide the database producers and online hosts with sufficient income to produce, maintain, and further improve them.

Sounds like a snap! At this point, let us see what kind of progress we are making toward this particular utopia of mine.

† Introduction to the Symposium of this name, presented at the 201st National Meeting of the American Chemical Society, Atlanta, GA, April 16, 1991.

There's More to a Polymer Than Just Its Build†

STUART M. KABACK
Exxon Research and Engineering Company, P.O. Box 121, Linden, New Jersey 07036

Received July 7, 1991

Much of the effort being devoted to the development of polymer information systems is aimed at being able to identify and distinguish complex structures, and at times it seems that insufficient attention is devoted to polymer systems based on relatively simple structures, such as the lower olefin homo- and copolymers. The volume of journal and patent literature on such polymer systems is very large and is growing rapidly, making it essential for information systems to develop ways of discriminating among references with subtle differences.

What am I looking for when I search the polymer literature? Usually it is not an unusual polymer per se. My polymers are the polyethylenes and the polypropylenes, the thermoplastic olefin copolymers such as linear low-density polyethylene, the elastomeric olefin copolymers such as EPM and EPDM, and butyl rubber. Then there are the more mysterious and somewhat complicated petroleum resins, so mysterious that Derwent still treats them as natural polymers even though they

† Presented at the Symposium Polymer Information: Storage for Retrieval, or Hide and Seek, at the 201st National Meeting of the American Chemical Society, Atlanta, GA, April 16, 1991.

0095-2338/91/1631-0439$02.50/0

Table I. Chemname Postings for Lower Olefin Homopolymers

		CA postings	prepns
polyethylene	9002-88-4	71954	6592
polypropylene	9003-07-0	34598	3735
isotactic PP	25085-53-4	4909	914
syndiotactic PP	26063-22-9	118	27

Table II. Chemname Postings for Lower Olefin Copolymers

		CA postings	prepns
ethylene–propylene	9010-79-1	7019	1370
EP-isotactic	56453-76-0	31	15
EP-syndiotactic	29160-11-0	216	61
EP-block	106565-43-9	640	90
EP-isotactic block	115404-65-4	13	5
EP-alternating	106974-59-8	2	0
ethylene–butene	25087-34-7	2805	1017
EB-isotactic	54570-68-2	10	0
ethylene–hexene	25213-02-9	765	325
ethylene–4-methylpentene	25213-96-1	478	140
ethylene–octene	26221-73-8	663	188

are actually synthetics prepared by polymerizing a variety of mixed unsaturated hydrocarbon feeds by cationic, thermal, or even free radical routes. There are even some non-hydrocarbon species, things like ethylene–vinyl acetate, ethylene–methyl acrylate, or PVC. Whatever they are, they're almost always on the straightforward side, at least in their gross components.

From this you might draw the reasonable conclusion that my job in searching this literature is a simple one. WRONG! Go back to your drawing board. There are times when the complications and frustrations of dealing with this body of literature almost—almost but not quite—almost make me yearn to emigrate to the land of Markush. The first point is that the things I deal with are found all over the place. If there is such a thing as more than ubiquitous, that is what they are. Take the nice simple lower polyolefins. If we look at Registry Number postings in DIALOG's CASearch File as of mid-April 1991, shown in Table I, we find almost 72000 for polyethylene, of which nearly 6600 were assigned the preparation role by DIALOG. That's bigger than a lot of respectable databases. It is nearly 30% the size of Derwent's Farmdoc segment, which represents almost 30 years of the world's pharmaceutical patents. For polypropylene there are about half as many postings, along with a sizable chunk specifying isotactic polypropylene and a smattering on syndiotactic polypropylene. If this is not your usual area of interest, let me point out to you that the bulk of items on ordinary polypropylene really refer to the isotactic polymer. One really must look at all of them when searching.

How about the simplest of copolymers? Some numbers are given in Table II, and once again they are sizable. By now we have at least six valid Registry Numbers for binary copolymers of ethylene and propylene, but they do not cover it all. There are a great many postings under ethylene–propylene rubber, without any Registry Number. Of course there are lots of variants of ethylene–propylene rubber including additional comonomers, especially a nonconjugated diene—there are several dienes that are popular. This is starting to get complicated, and I am only talking about the simplest of polyolefins.

And if you are looking at the copolymers with higher olefins on the bottom part of this table, a lot of them are things that are called linear low-density polyethylene, or its two subspecies very low-density polyethylene and ultra low-density polyethylene. Linear low-density polyethylene is not really polyethylene, but rather a copolymer. Actually there are some terpolymers, too, but I've omitted them here. Many of the references to LLDPE, VLDPE, or ULDPE are not included in these data, because the nature of the specific higher olefin used is not identified. Lots of the references to LLDPE, VLDPE, or ULDPE cannot be pinpointed, because they are not indexed or designated with any consistency.

My problem is that I often have to be able to locate references on LLDPE, VLDPE, or ULDPE, without making it a lifetime project. I have to be able to distinguish references to ethylene–propylene copolymers which are elastomeric from those which are rigid thermoplastics or those which are somewhere in between; to distinguish those with 1 or 2% ethylene from those with 1 or 2% propylene; or how about trying to spot in that vast pile of references a ternary blend that includes a propylene–ethylene block copolymer containing relatively small amounts of ethylene in an impact-modifying block, as well as two elastomeric ethylene–propylene copolymers with differing comonomer ratios?

Depending on the information system I am using, those three components of the blend may very well be described by the same indexing terms. There certainly is not any system that can come close to spotting such a blend neatly and separating it from all sorts of other blends. Think for a while about how that problem might be solved; while you're doing so, I'll be worrying about how to separate narrow molecular weight distributions from broad ones, bimodal from polymodal ones; how to find molecular weights that are very high, or relatively low; how to find regular or irregular comonomer distributions; and how to distinguish polymers with differing patterns of residual unsaturation.

Why should I worry about these things? Because they make all the difference in the world in the usefulness of polymers. And that is where polymers are so different from most other chemicals. When you make 2-propanol, toluene, or phthalic anhydride, your product is 2-propanol, toluene, or phthalic anhydride, respectively. True, there may be differences in purity from one synthesis to the next. True, with some substances there is room for various types of isomerism. But when you are dealing with polymers, it is very, very difficult for two syntheses of what is nominally the same thing to actually give you the same thing.

When somebody synthesizes 2-propanol, it probably does not matter very much what kind of reaction vessel was used. Not so with a polymer, though. It makes all the difference in the world whether the reactor was a stirred autoclave, a tubular once-through flow system, or a fluidized bed operating in the gas phase. Each of these will produce a polymer with very different physical characteristics—molecular weight, molecular weight distribution, comonomer distribution. Yet it remains very difficult to identify consistently which polymer syntheses use which types of equipment. Derwent will code the equipment if it is claimed explicitly, but that is not enough. Engineering details have never been a strong point with Chemical Abstracts. Just as a single example, let me refer to a study I did awhile ago that was concerned with gas-phase polymerization processes of a particular type. It presents something of a microcosm of the problems we face.

At the time the study was carried out, I judged that there were 38 U.S. patents that were relevant. Every one of them had claims that read on a gas-phase polymerization process, so when I talk about retrieval failures I am speaking of significant ones, not just the failure to pick up some incidental disclosure.

A summary of the results is shown in Table III. Fewer than half of the patents were given the international patent class (IPC) on gas-phase polymerization by any patent office; the U.S. Office assigned that international class to only seven of them, and even its own class—(US-CL) a cross-reference collection rather than a primary subclass—was given only to some 30%. IFI Plenum (UNITERM) did somewhat better. Five of its 11 misses were early references, issued before the

Table III. Retrieval Problems, Gas-Phase Polymerization Study

	yes	no
IPC C08f-002/34	16 (7)[a]	22 (31)[a]
US-CL 526/901	11	27
UNITERM 07942	27	11[b]
WPI 358 or 59&	32	6
Chemical Abstracts	19 (30)[c]	19 (8)

[a] IPC only on 7 U.S. patents, but also on 9 foreign equivalents. [b] Five patents indexed before index term established; 3 not yet indexed when search was run—1 OK, 2 failures. [c] 9–11 patents retrievable if abstracts searchable.

vapor-phase indexing term was added to their system. On the other hand IFI tends to lag somewhat in its indexing. For one thing, it does not have the advantage of being able to index a quick-publishing foreign equivalent before the U.S. patent issues. In this instance three patents had not yet been indexed at the time the search was carried out. Only one of those three was subsequently given the indexing term for vapor phase.

Derwent with their WPI had the best overall performance, with just six misses. Most of the misses were patents in which an unexamined Japanese application, or Kokai, was the basic patent in the Derwent system. The abstracts of Kokai that Derwent obtains from its Japanese agents are often rudimentary, and in these instances did not convey to Derwent's coders the need to apply the term for vapor-phase polymerization. There are instances in which Derwent recodes documents that appeared first as Kokai when equivalents appear in other countries. Too bad that policy is not applied more broadly; too bad it was not applied here.

Searching the indexing of Chemical Abstracts retrieves half of the relevant patents. If one is able to search the CA abstracts as well, another 9 or 11 references can be located, depending on the search strategy used. In doing a search for U.S. patents on CA, one definitely wants to be able to search on a host that offers the Derwent system, so that U.S. equivalents to foreign patents can be located at minimal cost. One also needs to be able to search on a host that enables one to find the relevant references that are contained in CA. Only STN, of course, provides the abstracts—but STN has no Derwent file. This searcher, at least, is thoroughly fed up with the fact that it is still impossible to utilize CA information most effectively, because of policies and political decisions.

You would think that with all the databases I used I would have retrieved all 38 of the relevant patents. I did eventually find all of them, of course, but only 36 of the 38 by the primary search strategy. One was picked up because it was a divisional of a patent that *was* correctly retrieved, and Derwent's family capability led to it. Another was found only through a bit of serendipity. As I have been saying for years to anyone who would listen, you had better use multiple databases if you need to do a complete job. It is scary to be reminded over and over that, even with multiple files, you are still sometimes on shaky ground.

This example is by no means an isolated instance of the importance of unindexed information in CA's abstracts. I have been grumbling for years about CA's failure to index catalyst supports in many patents, on the dubious basis that the support involved is a common one—something like silica or alumina. When I started complaining about things like this to CAS management in the early 1980s, the centerpiece of my argument was a patent that started off by saying that supported polymerization catalysts were of course well-known, but were beset by problems. The patentee indicated that he or she had managed to solve those problems and produce a new and improved supported catalyst. I just could not see how any rational policy decision could cause the support in such a patent to remain unindexed.

I have not attempted a sophisticated study of the situation here, but what I did do was a fairly simple-minded search that combined the concepts of polymerization and catalyst, with the word silica or SiO_2 in the abstract, but without silica or its Registry Number in the basic index. Table IV shows numbers for the past decade. As you can see there were from 30 to nearly 60 references each year that met the criteria of the search. A spot check for relevance suggested that perhaps 40–70% of these references might be relevant to my interests. There is a lot of meat tucked into those abstracts that does not get indexed. I believe it must be made more broadly available.

Table IV. Catalyst Support Information in CA Abstracts but Not Indexed[a]

'80	'81	'82	'83	'84	'85	'86	'87	'88	'89	'90
38	30	33	57	37	30	47	44	47	53	50

[a] Silica or SiO_2 in abstract. Silica or 7631-86-9 *not* in basic index.

How is the decision made whether or not to index a given bit of information? Clearly it ends up being a subjective decision, governed by guidelines established by each given database producer. I would like to shake up some of that thinking right now, especially, but not only, in the area of polymer chemistry.

I referred earlier to the substantial differences in characteristics of what are nominally the same polymer, made of the same ingredients, depending on the manner in which it was produced: molecular weight, molecular weight distribution, gross composition, composition distribution, and so forth, properties which define whether the product is an elastomer or a rigid thermoplastic, whether it is heat sealable, or useful for the preparation of blown films, or melt-blown fibers, or injection-molded articles, or whatever. Polymer chemists and engineers are learning more and more effectively how to make their charges jump through hoops. Karl Ziegler and Giulio Natta might be astounded at the degree of control available today in their eponymous polymerization process. By changing catalyst components in subtle ways, by using hybrid systems with more than one type of catalyst, or by changes in conditions, they can do far more than just get olefins to link together in relatively linear fashion under relatively mild conditions. Polymers with desired microstructure and the various properties to which I have referred can be produced virtually at will. And polyethylene A may be very different from polyethylene B, itself very different from polyethylene C. When I read in an abstract that polyethylene is used in a given application, or when I read that a propylene–ethylene copolymer is subjected to maleinization, I would very much like to know how that polyolefin was produced. If the patent has a substantial disclosure of the production method, of the catalyst used, *that information may be invaluable to me, even though it is not the subject of the invention itself.* Polymers are both chemicals and materials. In viewing a polyolefin as a chemical reactant, it is critical for me to know if it was produced by a catalyst system that gives a product in which residual unsaturation is overwhelmingly at the chain end, rather than one or more units into the chain. In viewing it as a thermoplastic, it is critical for me to know whether the molecular weight distribution is broad, implying good flow properties, or narrow, suggesting possible melt processing difficulties.

Therefore, database producers, please remember that the invention itself is not the only important information in a patent. What happens at the end of the road can be very different, depending on the route you used to get there. Thoughts of this sort are especially meaningful to me these days because I have spent a lot of time putting together an absolutely complete—so far as I can make it complete—collection on the production of a type of polymer by a specific

type of catalyst system, including uses for the polymers so produced. Some of the polymers involved are fairly low in molecular weight, used as reactive intermediates, and the location of residual unsaturation is especially important for them, as I indicated earlier. It was very embarrassing recently when one of our attorneys came across a patent that belonged in my collection, but was not there. It turned out that the key information on relevance was indeed in the Derwent documentation abstract, but not in the computer-searchable portion. (I should add that Derwent's coding system at present is nearly useless for most polymerization catalyst information. The situation I described in 1984[1] has not gotten any better. Hopefully the revised system that was discussed here this morning[2] will help.) All of this sent me with my fine-tooth comb to check the patents of one particular company. By the time I finished I had located several additional items in which the catalyst of interest had been used, but was not itself a part of the invention. Some items were locatable in Derwent but not CA, and some in CA but not Derwent. There are still a couple that I suspect are relevant, but the patents are in Japanese and I did not obtain full translations.

The point I am hammering at is that the database producers can do a great deal to help information users by construing more broadly the significant information in the documents they cover, and then providing us with access to that information—if not via indexing, at least by including the information in their abstracts, and making the abstracts more fully searchable. And if I have pointed my finger at CAS regarding the need to have searchable abstracts available in all versions of the CA File, let me not fail to remind Derwent once again of the great need to make their documentation abstracts searchable, not just the alerting abstracts. I understand full well that the graphics create problems, but that should not prevent us from being able to search the rest of the text. If necessary, markers could be provided so that graphics could be tied in at a later date; in the meantime, we need the full text of these abstracts without any further delay.

Let us take stock so far. We need things that allow us to distinguish among polymer systems that appear to be relatively simple in their gross compositions, but that differ in many and sometimes very subtle ways, differences which to an increasing extent are introduced intentionally as the result of the growing skills of polymer scientists and technologists. The chemical and engineering details that caused these differences are important information that must be conveyed to information users, not only when it is the focus of the invention, but also when it is provided as incidental information. The systems available for describing polymerization catalysts must be improved, and improved very soon. It does us very little good that Derwent is capable of describing polymerization catalysts with the Markush DARC system if most polymerization catalyst patents are classified only in the Plasdoc section, and thus not subjected to DARC coding. Nor does it warm the cockles of my heart to hear the Derwent staff explain that catalysts will be structurally coded when they are NEW but not afterwards. That means that once a catalyst becomes significant by being used again and again, it no longer becomes newsworthy. On the contrary, the more often it is used, the more newsworthy it becomes!

The same sort of consideration holds for other chemical compounds used in polymer compounding. Plasticizers, for instance, because they happen to be products my company makes. We use patent information for more than ascertaining whether or not our inventions are novel, or whether our projects might infringe, or whether adversely held patents are valid. We use patent information for more than determining the state of the art in an area of potential R&D. We use it too, very importantly, for technico-economic information. We want to see whether a manufacturer is compounding its resins with dioctyl phthalate or with diisononyl phthalate, perhaps with some other phthalate, perhaps with some entirely different type of plasticizer. What is the difference between dioctyl and diisononyl, you ask? Is it not enough to know that a phthalate is being used? The answer to that is emphatically, *NO*! There are performance differences between these species. We have to know where and why our plasticizers, our solvents, are being used, so that we can better understand our markets and their potential. You may have the same sort of situation with your antioxidants, with your flame retardants, and so forth. It is fairly easy when the substance involved is novel, described as novel. All abstracting and indexing operations know that they have to do a good job indexing things that are supposed to be brand new. But when something seems mundane they appear to relax and say "That's ordinary; I won't bother with it". It is just those mundane things that appear in the literature over and over that have important economic significance. They must not be ignored; they are too important to us.

Let me offer to you a word, one which has assumed very negative connotations during the past half-century: *discrimination*. We must not discriminate against people because of color, gender, religion, or country of origin. Discrimination has gotten a very bad name! But when we are talking about information services, discrimination is everything. It is the ability to discriminate that makes the difference between information retrieval and hide and seek. And that is especially true with the kinds of substances I have been concentrating on, the commodity products, the ones that are the "most ubiquitousest".

And so we need to have IFI/Plenum change its system of not linking the components of copolymer systems, because patented systems increasingly include several different—and often interrelated—polymers. When each claimed component has several options, the number of combinations and permutations skyrockets, and you end up retrieving some references no matter what you were looking for. Similarly, the overcoding of alternatives by Derwent produces all sorts of false coordinations. Derwent has come a very long way from its early days when almost everything got overcoded onto one punch card record and false coordinations abounded. Those false coordinations could be tolerated in a file of 25 000 references, perhaps in a file of 100 000 references; they become intolerable in a file that approaches its first million.

Another way I would like to be able to discriminate is between claimed information, information that is exemplified, and information that is merely disclosed in patents but not substantiated. Once upon a time Monty Hyams of Derwent talked about providing that sort of discrimination, but the talk did not last long. Some of this was supposed to be part of CA's patent project, that became the short-lived Agpat and Pharmpat. They vanished for other reasons, but this proposed capability would have been a very valuable one. On some searches I really must know everything that was disclosed or everything that was really covered in the claims. On other searches I really only want to get the patents with real information. Oh, for the ability to distinguish among these categories.

Let me close by talking to you about one of the most deceptively simple products around today: the packaging laminate. Looks pretty straightforward to you, doesn't it? A nice, clear film. Perhaps it has a self-cling feature that enables it to be closed and reclosed. Perhaps it has an especially high tear strength that makes you despair of ever getting the darned thing open, but certainly protects the product before you buy it. That film lets you get a clear look at what is inside, while making sure that the oxygen in the air does not spoil the contents—or perhaps by making sure that a controlled amount

of oxygen and moisture does get in to ensure desirable product properties. That simple film may consist of two, three, five, or seven layers. There may be core structural layers; surface heat seal layers or cling layers; oxygen and/or moisture barrier layers; adhesive tie layers to hold the thing together. There may be tackifier additives, antioxidants or prodegradants, crystallization nucleating additives, plasticizers, or optical clarity promoters. One or more layers may be a blend rather than a single polymer. That simple film is probably not simple at all, but highly complex.

Now if I am trying to find whether a novel kind of polymer, or novel kind of additive, has been used in a film, it probably will not be too difficult. But the questions my colleagues and I end up dealing with involve different combinations of those same old commodity substances, the good old simple polyolefins. The potential hits are many, the differences from the prior art subtle. It is questions such as this that make me pray fervently for the success of Derwent's plans that, if achieved, would let me discriminate in the details of a single polymer's structure, and its blends with a second polymer, and separate them from another polymer system (which could involve similar or identical polymers) present in the same overall structure. It is questions such as this that remind me over and over that discrimination can be one of the highest virtues and not just a dirty word.

So here's to discrimination in its most positive sense; may its presence in polymer databases grow. May the database producers develop better ways of implementing discrimination in their systems, and may their staffs be able to apply these improved systems in a consistent fashion that will be useful for us—because a system that breaks down repeatedly is a system that cannot be trusted, and will not be used. May the online hosts who deliver the databases develop improved methods of handling systems that can become quite complex, if they start involving multiple layers of linking as has been proposed by Derwent.

May I offer a final prayer. Maybe it could even be possible for some of the competing database producers to work out some sort of coordination of their efforts, so that work duplicated by two or more producers could be consolidated, and each could concentrate on the special features that are its particular strengths. To a significant degree Derwent and API have done this in the petroleum and petrochemical area, for almost 20 years and with considerable success, and without violating antitrust considerations. Is it totally ridiculous for me to propose a broader implementation of cooperation? Cooperation that might result in less duplication of effort and free resources to do things that might look uneconomical today? Nine years ago, at a meeting in this very same hotel in Atlanta, I made some suggestions about synergistic database combination that might benefit information users. I admitted at the time that it seemed utopian, but some of what I suggested then has actually come to pass, and more is possible in the future. Perhaps if some of the leadership of the information industry were to get its focus out of courtrooms and onto issues such as these, we might see some of this take place too—with obvious benefits for information users and great benefits for information providers as well.

REFERENCES AND NOTES

(1) Kaback, S. M. Polymer Patent Information Systems Could Be Even Better! *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 371-379.
(2) Briggs, J. A.; Ferns, E. A.; Shenton, K. E. Improvements in Derwent Plasdoc System. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 454-458.

Online Searching of Polymer Patents: Precision and Recall

NANCY LAMBERT
Chevron Research and Technology Company, P.O. Box 1627, Richmond, California 94802-0627

Received June 1, 1991

Patent information specialists searching online databases for patents in polymer subjects have faced a number of problems in precision (how clean the search is) and recall (how comprehensive the search is). Some databases have been designed for high precision, but recall in these databases can be both poor and inconsistent over time. Other databases have been designed for high recall, and the searcher faces looking through many irrelevant references. This paper discusses precision and recall problems in three major databases that cover polymer patents: Chemical Abstracts, the Derwent World Patents Index, and the IFI Comprehensive Index to U.S. Chemical Patents. It mentions some solutions that the database producers are proposing and discusses additional solutions to be considered.

INTRODUCTION

One of the oldest problems in computerized information retrieval is how to balance precision and recall, and optimize both. Precision is also known as the relevance of a search: how many of the references retrieved in the search were, in fact, pertinent. Recall is the comprehensiveness of the search: how many of the relevant references actually in the database the search retrieved. Numerous case studies have attempted quantitative measurements of precision and recall in great statistical detail in all possible subject areas under all sorts of circumstances.

The reader will be relieved to know that this paper will not add to their number. Instead, it will look qualitatively at three major databases that index patents in polymer chemistry, discuss how well their indexing policies and systems work in terms of precision and recall, look at specific precision and recall problems that occur in some or all of them, and describe some solutions that might help with these problems.

The databases are

1. Chemical Abstracts, whose online file contains polymer registrations and international patents on polymer chemistry since 1967.
2. The Derwent World Patents Index, which has covered international polymer patents since 1966.
3. The IFI Comprehensive Index to U.S. Chemical Patents, which has covered U.S. polymer patents with its current indexing system since 1972, some aspects since 1964.

Improvements for all three databases are either being planned or actually under development that will clear up some of the problems discussed in this paper. But, with a few

0095-2338/91/1631-0443$02.50/0