

Chemical Structure Searching in Derwent's World Patents Index†

STUART M. KABACK

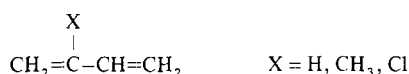
Information Research and Analysis Unit, Analytical and Information Division, Exxon Research and Engineering Company, Linden, New Jersey 07036

Received June 7, 1979

Derwent's World Patents Index (WPI) contains information on general chemicals since 1970 and on pharmaceutical, agricultural, and polymer-related chemicals since different points in the 1960s. Chemical products and significant components of chemical formulations are indexed for structural retrieval by a variety of codes and terms. Foremost among them are the Farmdoc-Agdoc-Chemdoc multipunch codes, but the user must keep in mind other retrieval parameters which are available to enhance and refine retrieval. This paper surveys the currently available techniques for retrieving information on chemicals from WPI, and highlights some developments which could significantly improve chemical structure retrieval from WPI in the future.

In the first half of the nineteenth century, there existed a popular "theory of substitutions" which said, in essence, that atoms of a given element in a compound could be replaced by atoms of almost any other element, without significant changes in properties. In 1840, the famous chemist Wöhler—functioning through his alter ego SCHWindler—effectively burst the substitution balloon by reporting his success in sequentially replacing all of the atoms of manganous acetate with chlorine atoms.¹ The end product was a multichlorine substance that was said to show promise as a synthetic fiber.

Despite Wöhler's efforts, though, a vestige of the substitution theory lives on today as the Markush claim or disclosure. In a Markush claim, one or more structural features of a compound is subject to variation. The simple Markush shown here



represents butadiene, isoprene, or chloroprene. Figure 1 illustrates a far juicier Markush, taken from a real patent specification. It has 12 R substituents; an X and a Y, each of these an amine derivative with substantial variability; certain pairs of substituents which may form rings (again with a degree of variability); and the anion which can be anything. Since many of the substituent possibilities here are open-ended, a full permutation of all possible compounds would literally approach infinity, one patent representing more potential compounds than the entire CAS Registry.

For better or worse, Markush claims are allowed in patents. A patent may include examples for only a handful of compounds, but its claims may encompass hundreds, thousands, and on to infinity. A chemical structure retrieval system for patents must be capable of handling all of those compounds, not just the ones that are really there but every one encompassed by the claims. Failure to do so may lead to the inadvertent infringement of someone else's patent. The claims of examined patents, such as United States patents, tend to be limited to reasonable bounds. Markush structures are allowed, to be sure, but they tend not to be the sort of open-ended thing shown in Figure 1. Unfortunately the vast bulk of the patent literature appearing today consists not of examined patents, but rather of unexamined published applications, which spew out from various patent offices by the hundreds of thousands each year.^{2,3}

In designing a structural retrieval system for an environment that allows for such multiplicity and variability, it becomes necessary to make certain compromises. It may be reasonable in some files to portray every minute structural facet of a molecule. In a large file it becomes extremely difficult and costly to do so. It is quite clear that full substructure search is a practical impossibility in a Markush-ridden patent file; at least, it is, given today's storage and retrieval capabilities.

Derwent's chemical information files have thus settled on a series of fragmentation codes. Separate codes exist to cover natural products, steroids, dyestuffs, and general chemical structures. This paper will not consider the first three of these; rather, it will focus on the retrieval of general chemical structures from the Farmdoc file of pharmaceuticals, the Agdoc file of agriculturals, and the Chemdoc file of general chemical compounds, with emphasis on Derwent's multipunch code.

The Farmdoc-Agdoc-Chemdoc multipunch code is a fragmentation code. To be sure it is one which is capable of showing some fine structural details, but by and large it deals with the gross chemical features of a compound: rings and chains, functional groups, and the inorganic ingredients.

Let us set the stage with a quick bit of history. The coding system was first developed to handle pharmaceutical chemicals, in the Farmdoc service that started in 1963. Certain modifications were made to deal with pesticides and fertilizers when Agdoc began in 1965. In 1970 Derwent began to deal with general chemicals in Chemdoc, and additional modifications were made in the retrieval system. Further changes were made at various points along the way.

The history of the Derwent products colors strongly the capabilities of the chemical retrieval system. The system originally had to deal with the relatively complicated structures that are generally found in the pharmaceutical field. It was not well adapted to coping with information on relatively simple compounds, such as petrochemicals. Further, since the pharmaceutical industry is oriented to end-products, no attempt was made to provide retrieval for starting materials or intermediates, except where they themselves might be novel.

Retrieval at the start was based on the standard IBM punch card, used with a card sorter. Each structural feature was expressed by a single punch position. Given the limited number of coding positions on a card, it was necessary to be economical in the use of those positions. Thus, while many structural features were coded in some detail, others had to be ignored, or at least thrown into catch-all categories.

The early file was relatively small, and detailed abstracts

† Presented at the 13th Middle Atlantic Regional Meeting of the American Chemical Society, West Long Branch, N.J., March 21, 1979.

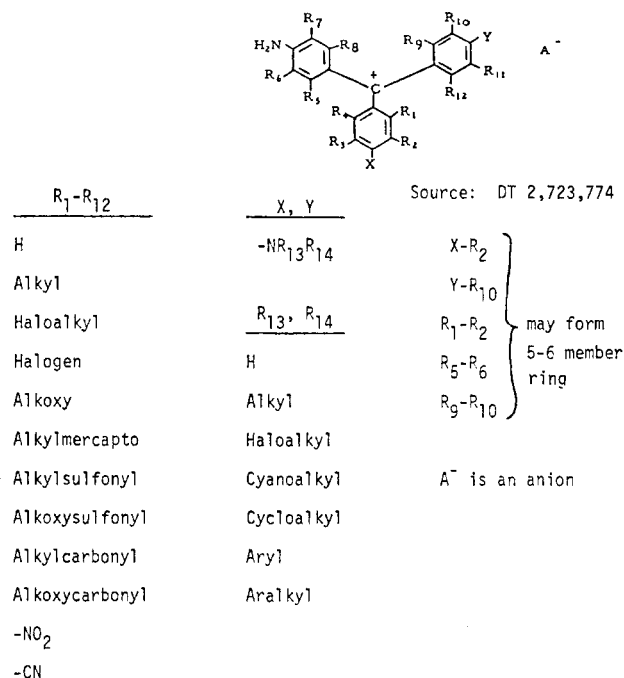


Figure 1. A complicated Markush structure.

were printed on the IBM cards. Given output in the form of concise descriptions of the invention, it was not too essential to gear the system for high precision. Rather, the aim was for high recall.

In keeping with the desire to minimize the bulk of punch cards to be handled, multiple compounds from a given patent were generally coded on the same card. This led to scrambling of their structural features and further imprecision in retrieval.

Finally, the overall body of chemically related information that comprises Derwent's Central Patents Index is of necessity segmented in 12 sections, marketed independently.³ Few subscribers are interested in all 12 sections or can afford the purchase of the entire data base. Compounds are indexed for retrieval from only three of those sections (Farmdoc, Agdoc, and Chemdoc) although there are certain structural retrieval capabilities for chemicals in Plasdoc, the polymer section of the Central Patents Index. Many patents are classified in more than one CPI section: for example, a patent on an inorganic substance may be found in both Chemdoc and in Section L, dealing with refractories, glass, ceramics, and electrochemistry, and may thus be retrievable by structure search. On the other hand, it may be found only in Section L, and thus the structure may not be retrievable.

The overall file is no longer a small one. Farmdoc is more than 15 years old, and even the newest component, Chemdoc, is approaching its 10th birthday. Together Farmdoc, Agdoc, and Chemdoc add nearly 30 000 new patents to the file each year. The file is no longer punch card limited, or even limited to batch search on magnetic tape, but now exists as part of the vast on-line World Patents Index (WPI) file that encompasses both chemical and nonchemical patents, and that includes a variety of retrieval parameters in addition to the multipunch codes.⁴ We shall examine some of the implications of these factors in due time, but at this point let us look at the multipunch code itself and see how it works.

Derwent instruction manuals explain that all new chemicals are coded, as well as uses or activities of known compounds that are essential to the invention. Note that the present system does not provide retrieval for specific compounds by a single indexing term; rather, a full or partial structure must be described in terms of the appropriate multipunch codes. The structural features of the multipunch code for chemicals

Table I. Structural Elements of Farmdoc/Agdoc/Chemdoc Chemical Code

essential groups	1970 ^a
bridge structures	1970
carbon chains	1972
inorganics, organometallics	supplemented 1970
rings: fused heterocyclics	
monocyclic heterocyclics	
aromatics	
alicyclics	
functional groups: common	
less common	
ring systems present	1970
basic group	

^a Dates are given for major portions of the code added after 1963.

are outlined in Table I. The code also includes terms which describe chemical reactions taking place, properties and uses of the compounds, and other nonstructural information. While this type of information is extremely important in some instances, we will examine here primarily the structural features.

The first feature, the section of the code dealing with essential groups, is important because of Markush claims. Consider the simple Markush given earlier, encompassing butadiene, isoprene, or chloroprene. If one were interested in patents on hydrocarbon dienes such as butadiene or isoprene, applied the appropriate punches for the diene system, and then negated all patents dealing with halogenated substances, one would eliminate patents with a Markush claim such as this, patents which encompass the compounds one wanted along with those one did not want.

Derwent solves this problem with the concept of groups which *must* be present. Structural features that have to be present are coded in this section of the code, while Markush features that may or may not be present are not coded here. Thus, a patent specific to chloroprene is coded for halogen essentially present; a Markush containing the chlorine only as an option is not given this code. In searching for hydrocarbon dienes one negates patents that must have halogen present, and is able to retain those in which the halogen is just an option. Actually in this case the problem could have been solved by using the code for hydrocarbon, but in more complex molecules this would not be possible, and the essential group concept becomes important in limiting false drops.

The section on bridge structures describes linkages between rings. Thus, one can distinguish among biphenyl, diphenylmethane, stilbene, and diphenyl ether, or among diphenyl ether, naphthyl phenyl ether, and phenyl pyridyl ether, using terms that describe the type of each ring and the nature of the linking group.

The section on carbon chains was added to the system rather late. This is a consequence of the system's origins in the realm of complicated compounds, and general lack of concern with anything so mundane as the nature of alkyl and alkylene groups. The present system can show the existence or absence of monovalent carbon chains, as well as of carbon chains bonded to two or more nonchain groups. It can show whether the chains are straight or branched; whether they are attached to rings, heteroatoms, or certain types of functional group; and whether there are one, two, or more of such chains. There is also a limited capability for showing the length of chains.

Inorganic compounds and the inorganic aspects of organic molecules were codable to some extent in the original Farmdoc and Agdoc code. This included a fairly detailed system for dealing with compounds containing boron, silicon, phosphorus, arsenic, selenium, and tellurium. Among the metals some, such as magnesium, aluminum, tin (eight metals in all), had their own punch position. Others, however, were not so fortunate. Thus, for example, the punch position 131 encompassed 10 different metals, including molybdenum, rhodium,

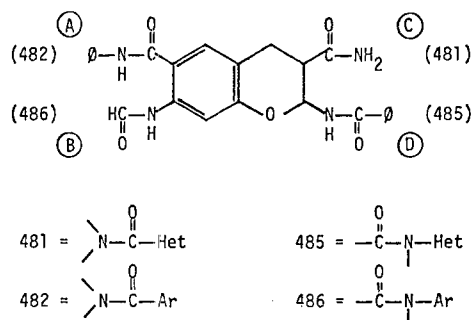


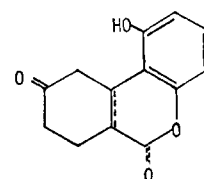
Figure 2. Coding of environment—common groups.

palladium, silver, and cadmium—not a very discriminating indexing term.

At the start of Chemdoc in 1970 additional punch positions were allotted to some of the individual metals, and further positions were set aside to improve the specificity of inorganic indexing with respect to the common elements hydrogen, oxygen, sulfur, nitrogen, halogens, and carbon. The system is able to distinguish such concepts as a fully inorganic compound; an organic anion bonded to an inorganic cation, or vice versa; and whether a metal is bonded to an organic anion at carbon or at a heteroatom.

Rings are the bread and butter of the coding system. Twenty-four columns of the punch card (288 coding terms in all) are devoted to the coding of rings. Half of those terms deal with fused heterocyclic systems, one-fourth with mononuclear heterocyclics, and the remainder with aromatics and alicyclics. Many ring systems have a special coding term all to themselves, and with very common systems there are even separate terms to show different degrees of hydrogenation. Thus, there are three separate codes for pyrroles, pyrrolines, and pyrrolidines. With more complicated ring systems this type of specificity is not available within the basic system, but in 1972 Derwent introduced an important extra coding feature, the use of Patterson's Ring Index numbers to index any ring system that does not have its own unique position in the multipunch system. Thus, dibenzofuran is coded 175, and symmetrical dibenzopyran 176. Any three- or higher ring system with a single oxygen as heteroatom is coded 177, but is also coded with the Ring Index number. Where a ring system does not appear in the Ring Index or its supplements, Derwent creates its own number for the system. This move to the use of Ring Index numbers was Derwent's first sharp break with the limitations of the IBM card and has important implications for the future of the retrieval system in WPI.

Common functional groups include amines, carboxylic acids, esters and amides, hydroxy, and so forth. Since these groups appear so frequently, the system attempts to show more of their immediate environments than it does with the less common groups. Thus, it will indicate whether they are attached to a heterocyclic, aromatic, alicyclic, or aliphatic carbon atom. Where the linkage to carbon may be at two dissimilar points in the functional group, as with an ester or an amide, some information but not all is conveyed, though a hierarchical system of attachments. The order just used—heterocyclic, aromatic, alicyclic, aliphatic—constitutes the descending hierarchy. A secondary hierarchy is established, in which the attachment to the carbon of the functional group outranks the attachment to oxygen in the ester, or to nitrogen in the amide. Only the highest hierarchical code is used for any group. Figure 2 shows some examples of this coding. Amide group A is attached on both ends to an aromatic carbon and is coded 482 (amide attached at the carbonyl carbon to aromatic carbon). Group B is coded 486 (amide attached at N to aromatic carbon). Group C is coded 481 (amide attached at the carbonyl carbon to heterocyclic carbon). And Group D



177/B2	fused hetero, ≥ 3 rings, 1 oxygen
49-/B2	1 aromatic -OH
50-/B2	1 alicyclic =O
[50&/B2	1 heterocyclic =O]
03580/RR	dibenzo[b, d]pyran system

Figure 3. Retrieval of dibenzopyranone derivatives.

is coded 485 (amide attached at N to heterocyclic carbon). Note in this last instance that the hierarchy of heterocyclic outranks the carbonyl-nitrogen hierarchy. The hierarchy precludes showing that the attachment to carbonyl is aromatic, as opposed to alicyclic or aliphatic. Note finally that one can distinguish between attachments to an aromatic and a heterocyclic ring in fused hybrid systems.

With the less common functional groups such as sulfonate, carbamate, acid halide, and so forth, there is no attempt to show the environment of the group, merely that the group is or may be present. Many complicated functional groups cannot be represented by a single code and must be represented by the codes for their largest fragments. Thus, semicarbazones are coded as urea, hydrazine, and imine.

The section of the code dealing with ring systems present can be very useful in giving a count of the four types of ring system within an overall molecule. The section for basic group categorizes the molecule in general, again using a hierarchical system in which inorganics or organometallics outrank fused heterocyclics, mononuclear heterocyclics, aromatics, and aliphatics, in that order. This section probably had its greatest importance in the days of card sorters, when it could be used to limit the overall file to subsets for more detailed sorting, but it still can serve a useful purpose. For example, identification of a molecule here as an aliphatic means that it is not necessary to use a series of coding terms in the section on ring systems present showing the absence of each ring type.

Having looked briefly at the chief elements of the multipunch code, let us consider a few examples of the code in action, using the on-line WPI file available via the System Development Corporation. Examine the dibenzopyranone substructure shown in Figure 3. It is a pharmaceutical compound, and so must be searched for using the B2 qualifier that identifies the Farmdoc chemical code. Note that if someone had obtained a patent on this structure as a general chemical without indicating pharmaceutical use, it would be coded in the general chemical code E3. For absolutely complete searching on any structure it may be necessary to search in each of the codes.

If one is content to search only from 1972 to date, one can use the Ring Index term and not worry about the 177 punch position, which is somewhat less specific. The Ring Index term in this case has the effect of cutting retrieval approximately in half: from 19 to 9 if one is looking for the structure containing both oxo groups, from 49 to 29 if one just requires the single oxo group in the alicyclic ring. Closer examination of the hits in the search for just the one oxo group shows that several of them involve the use of the compounds with both oxo groups as intermediates, but these dioxo intermediates were not indexed because they were not claimed as novel compounds. Remember, the system is geared to end products. Remember too that the 177 punch must be used to retrieve items from 1971 and earlier. It is often necessary in searching

Title Words (TT)

citric
citrate


Manual Code (MC)

E10-C02A citric, isocitric acid

International Patent Class (IC)

C07c-059/16 citric acid
C12d-001/04 citric acid by fermentation

Multipunch (MP)

050 C3-4 multivalent chain
053 unbranched chain
055 one such chain
059 no monovalent chain
061 tetra- or higher-substituted chain


063 chain attached to -O or -C
475 >1 aliphatic -CO₂H
491 1 aliphatic -OH
53& saturated aliphatic compound
59& no fused heterocycle
60- no monocyclic heterocycle
62- no aromatic ring
620 no alicyclic

Negation: all but 011 (essential -CO₂H) and 013 (essential -OH)**Figure 4.** Citric/citrate retrieval strategy.

Derwent files to use multiple search strategies to cope with changes in the indexing system. In any event some of the patents retrieved seemed strange, probably irrelevant, but a number of them were clearly on the mark. The Ring Index number is a highly specific and discriminating indexing term (there were only 100 postings for this term in the entire file when this search was run), and so there are relatively few false drops. A final comment on Ring Index numbers: where tautomerism causes the Ring Index to assign multiple numbers (the dibenzopyran system is such an example), Derwent ignores the tautomerism and uses just the lowest Ring Index number. Thus, the real Ring Index number for the system shown was 3581, rather than 3580. In practice this procedure causes no problems, so long as the user is aware of the indexing rules.

For a different type of retrieval example, consider patents on citric acid or its salts. There are no rings here, but rather a common aliphatic molecule, so let us anticipate something. The on-line WPI file provides numerous retrieval parameters other than multipunch coding.⁴ These include title words (which since 1978 have been supplemented by added keywords), Derwent manual codes, and international patent classes. Derwent's titles are not the original patent titles, which often have a meager information content; rather, they are augmented titles (essentially miniabstracts) which contain quite a bit of information. Let us use all of these retrieval parameters to get at citric. The strategy is shown in Figure 4. The multipunch terms were used in each of the codes: Farmdoc, code B2; Agdoc, code C2; and Chemdoc, code E3. Retrieval results for a period covering about two-thirds of 1978 are shown in Table II.

The combined strategy produced a total of 185 references. Examination of their abstracts and, in some instances, the full

Table II. Retrieval of Citric/Citrate Patents

	TT	MC	IC	MP B2	MP C2	MP E3	total
total retrieved	78	47	6	31	17	100	185
relevant	73	45	6	27	15	87	163
rejected	5	2	0	4	2	13	22
unique relevant items	35	3	1	14	8	40	101

Table III. Retrieval, Citric/Citrate Preparation

	TT	MC	IC	MP B2	MP E3	total
total retrieved	9	10	5	1	10	12
relevant	8	9	5	0	9	11
rejected	1	1	0	1	1	1
unique relevant items	0	0	1	0	1	2

- Honey jelly prodn. by adding honey to citric acid soln. ...
- Treating surface of austenitic stainless steel by finishing to a specific roughness and dipping in aq. soln. contg. sodium hydroxide, sodium citrate, and sodium chloride.
- Photochromic glass prodn. using tetrasodium ceric dicitrate complex...
- Composn. for treatment of spilt liq. caustic base by neutralisation and absorption contains citric acid, ...
- Biomass pretreatment for measuring lipid(s) content comprises treatment with citric acid, ...

Figure 5. Sample citric/citrate references not in chemically coded sections of data base.

patent specifications, showed that 163 out of the 185 items contained at least a disclosure relevant to citric acid or its salts. Among the items rejected several dealt with citrate esters. The bulk of the rejects were items retrieved by the multipunch code. In using a nonspecific fragmentation code one expects to retrieve a certain number of nonrelevant items. Only a harsh critic would object to the level of relevance in this search.

But consider the bottom line of Table II, showing unique retrievals for each of the parameters. Of the 163 relevant items, 101, or 62%, were retrieved by only one of the six retrieval parameters. It is not surprising that many items multipunch coded for citric acid did not have citric or citrate in the title; in many cases the patent dealt with a long list of food acidulants, including citric acid, or any of a series of buffers, including citrate. It is perhaps surprising that the system worked so well in retrieving these by multipunch. What is astounding, though, is the large number of items in which citric or citrate was found in the title, but was not multipunch coded. Closer inspection shows that 23 of the 35 patents in this group did not appear in Farmdoc, Agdoc, or Chemdoc. A sampling of these is shown in Figure 5. Remember that chemicals cannot be retrieved by structure unless they are in one of these three sections. In some instances, as shown here, the availability of other search parameters permits us to supplement the multipunch search. In using the Derwent system it is especially important to consider carefully each of the available search parameters, and use as many of them as possible, always provided, of course, that they do not produce excessive numbers of false drops.

It is interesting to examine a subset of the citric acid references: those dealing with preparation of citric acid or a citrate salt. Here the performance of the system is markedly different, as shown in Table III. The one reject dealt with isocitric, rather than citric acid. The overlap among retrieval parameters was far better, but note that, even here, no single retrieval method was able to find all of the hits. Note that it would have been possible to limit the multipunch search to items on citric acid manufacture, as opposed to end uses, by

131	Rh, Mo (OR Y, Zr, Nb, Tc, Ru, Pd, Ag, Cd)
081	Metal bonded to inorganic anion
085	Inorganic anion containing metal
089	Inorganic compound
104	No hydrogen
100	No halogen
101	No sulfur
102	No nitrogen
103	No carbon
109	No P, As, B, Si, Se, Te

Figure 6. Retrieval of rhodium molybdate, Rh_2MoO_6 .

adding to the search strategy the term 659, production of a known compound.

A footnote to the citric acid search. Buried in a remote section of the Chemdoc code is a term, added in 1972, which enables one to specify exactly three carboxylic groups. Our strategy had not been so specific; it relied on a term for two or more aliphatic carboxy groups. When the three-carboxy term was added to the search strategy, it led to the rejection of 11 items. Six of these were among the items that had been manually rejected, the group of false drops. Five, however, were good references. It would appear that these documents had been indexed for a hydroxylic multiacid, but that citric itself had not been indexed, and when the strategy was refined to zero in on citric, the references were rejected by the computer. There is an important inference to be drawn from this result. Where you can afford to, use as few key search terms as possible. The more restrictive your search strategy, the greater your chance of eliminating wanted references. This advice holds, of course, for most data bases, not just Derwent's, but it is especially important in multipunch searching, where one can easily develop a string of dozens of punch codes which could be used to describe the desired molecule.

Consider an example of inorganic retrieval. There seems to be just one patent dealing with rhodium molybdate, whose key multipunch terms are shown in Figure 6. Application of this strategy in Chemdoc produces over 900 hits. Clearly relevance has been hurt by the punch code that covers ten different metals. If we limit the retrieval to items with rhodium in the title, however, we get just 12, including the one relevant item. The low relevance in the multipunch search is perhaps an extreme case, but it is not an isolated instance. Large numbers of false drops occur frequently, especially when searching for inorganics, aliphatics, and relatively simple cyclic molecules. It is often possible to limit excessive false drops by intersecting a set retrieved by multipunch searching with other parameters: manual codes, title words, international patent classes. When completeness is essential, though, this runs the risk of eliminating good references.

Of course, the more concepts you can intersect, the better your chances of eliminating unwanted references. To show this, we selected a patent on the synthesis of various trimetallic oxides. This is German Offenlegungsschrift 2515874, which describes the preparation of substances including $\text{Cu}_3\text{Mn}_4\text{ThO}_{12}$ and $\text{CaCu}_3\text{Mn}_4\text{O}_{12}$. However, a multipunch search failed to retrieve the patent. It turned out that it did not appear in Chemdoc, only in the electrochemical section L, and thus was not multipunch coded.

The examples could go on and on, and could show further the strengths and weaknesses of the system.⁵ Let us try to sum up, though. Major system strengths are shown in Figure 7. The comprehensive subject coverage and broad country coverage are enormous advantages. The system is capable of dealing with Markush structures, no matter how complex. A

- Comprehensive subject coverage, broad subject coverage
- Ability to deal with Markush structures
- Retrieves claimed structures even if not actually made or used
- Retrieves full or partial structures
- Availability of supplemental retrieval parameters (TT, MC, IC)
- Bibliographic retrieval parameters can refine search

Figure 7. Strengths of structure retrieval in WPI.

- Lack of simple retrieval for specific compounds
- Lack of structure retrieval for patents in uncoded sections
- Low precision due to Markush and overcoding of multiple compounds on single card
- Starting materials not coded
- Catch-all codes, functional groups without specific codes
- Limited capability to show fine structural details

Figure 8. Weaknesses of structure retrieval in WPI.

- Registration of common chemicals list (~2000)
- Coding of starting materials
- Role indicators for starting materials, products, present
- Separate coding of Markush and definite structures
- Reduce or eliminate overcoding; each compound separately entered
- Coding of each individual element, ring substitution patterns, additional functional groups
- Coding of molecular formulas
- Coding of non-BCE compounds; more IPC's; more keywords
- Improved catalyst coding

Figure 9. Possible refinements in structure coding.

compound encompassed by the claims is retrievable, whether or not it was actually made or used. Compounds are retrievable based on their full structure, or on just a portion of that structure. The additional retrieval parameters (titles, added keywords, manual codes, and international patent classes) provide valuable supplemental retrieval capabilities. Derwent's excellent bibliographic retrieval parameters are extremely important. They do not directly affect subject retrieval, but even there they can often be useful in limiting subject-based searches.

Some of the limitations are shown in Figure 8. Multipunch coding can involve a long string of punch positions, and it would obviously be desirable to be able to carry out specific retrieval on specific compounds, especially simple common compounds, without obtaining a lot of false drops. The inability to retrieve from uncoded sections can cause valuable references to be completely lost. The coding of Markush structures and the overcoding of multiple compounds on a single card are major contributors to low precision which, especially in the case of simpler structures, can be quite a problem. Failure to code starting materials is a substantial drawback in feedstock-oriented fields such as petrochemicals. Codes which cover more than one structural feature, and structural groups not codable by a single code, create obvious problems. The limited capability for showing fine structure results in retrieval of a candidate group which must then be manually screened to select the desired structures. Such groups can on occasion be forbiddingly large.

Derwent appreciates that the WPI file has limitations and, in conjunction with subscriber advisory committees, is in the course of developing substantial refinements in the coding

system. Many decisions are yet to be made regarding these code modifications, but it is possible to list some that are definitely in the offing, and others which could greatly improve retrieval from WPI. These are summarized in Figure 9.

A select list of about 2000 common chemicals will be assigned registry numbers. They will be registered as starting materials, as well as when they are products or are otherwise used, and their function will be described by a role indicator. While it would be best to have all starting materials coded, the registration of common starting materials should take care of the most urgent needs. For a while Derwent talked of creating its own registry numbers, but now it appears that they will use CAS registry numbers, a wise decision in light of the growing use of these numbers in the literature.

At one point Derwent contemplated specifically coding each compound which was claimed or which appeared in an example, in addition to any Markush structures claimed. This would have been extremely valuable in improving precision, but apparently has been shelved. New overcoding rules which reduce the amount of overcoding on a given input record should help somewhat, but ideally each individual compound coded should be coded separately from all other compounds. The American Petroleum Institute's data bases have shown how each separate chemical entity can be coded, with its terms linked, so that they can be kept apart from other compounds in the document.⁶

Freedom from IBM card input allowed the introduction of Ring Index numbers, and an extension of this principle was planned to allow the specific coding of all elements, ring substitution patterns, and added functional groups. Hopefully Derwent will extend this principle in a general restructuring of the multipunch code format. Various proposals have been made in this area, but this point in particular is still very much up in the air.

Molecular formula can be a vital data element in solidifying a structural retrieval system. There is perhaps no simpler way of enhancing the retrieval from CPI (especially for inorganic compounds, but broadly for all compounds) than the coding of molecular formulas. A few multipunches plus a molecular formula would constitute a powerful retrieval tool. At present Derwent has made no commitment regarding the possible inclusion of molecular formulas in its indexing.

Other steps that would improve various aspects of subject retrieval would be the coding of chemicals that presently are not placed in Sections B, C, or E; more liberal use of added keywords; entering of all international patent classes assigned to a patent, even if they are related to previously assigned IPCs; and amplification of Derwent's catalyst codes, which presently lump together elements from different groups of the periodic table, a most unfortunate circumstance.

WPI is the world's most all-encompassing file of patent information. Even with its present retrieval limitations it is an invaluable information tool. Elimination of those limitations could provide us with an extraordinary patent information resource for the future.

REFERENCES AND NOTES

- (1) The story of Wöhler's letter, written in the name of S.C.H. Windler, is recounted in *CHEMTECH* 1978, 8, No. 12, 757.
- (2) Duffey, M. M. "Searching Foreign Patents", *J. Chem. Inf. Comput. Sci.* 1977, 17, 126-30.
- (3) Kaback, S. M. "A User's Experience with the Derwent Patent Files", *J. Chem. Inf. Comput. Sci.* 1977, 17, 143-8.
- (4) Kaback, S. M. "Retrieving Patent Information Online", *ONLINE* 1978, 2, No. 1, 16-25.
- (5) Several additional studies of retrieval from Derwent files may be found in the Proceedings, International Patents Conference, Stratford-upon-Avon, England, April 12-14, 1978; Derwent Publications Ltd., London.
- (6) Kaback, S. M.; Landsberg, K.; Girard, A. "APILIT and APIPAT: Petroleum Information Online", *Database* 1978, 1, No. 2, 46-67.

The Approach of the United States Patent and Trademark Office to Finding Prior Art[†]

ALFRED C. MARMOR

Patent and Trademark Office, Washington, D.C. 20231

Received May 14, 1979

The foundation for effective retrieval of technical information by the U.S. Patent and Trademark Office (PTO) continues to be the U.S. Patent Classification System. Utilizing this system, the PTO currently maintains over 22 million documents, including U.S. patents, foreign patents and applications, and a variety of nonpatent technical disclosures in a viable file array of approximately 100 000 distinct, defined subdivisions. Providing prior art searches requires the maintenance and updating of a very large search file and a complex classification system. Continually developing computer-based systems are permitting the PTO to take an increasingly sophisticated approach to search file management. These systems, as well as newly developing information systems technology, will provide the capability to find prior art more readily and more precisely from an ever-growing search file.

The United States Patent and Trademark Office (PTO) must discharge three primary functions. The first is to examine applications for patents and, after examination, to grant or refuse to grant a patent. The second is to examine applications

for trademarks and to register or refuse to register the trademark. The third is to collect, classify, and disseminate technological information disclosed in patents and related technical documents. This paper is concerned with the first and third functions.

The requirements for meeting the obligations under the first function, i.e., patent examination, make it necessary for the examiner to determine whether the invention meets the stat-

[†]Presented in the symposium on "International Aspects of Technical Information Retrieval", Division of Chemical Information, 177th National Meeting of the American Chemical Society, Honolulu, Hawaii, April 2, 1979.