gerber, D. W., "Handling Commercial Product Names at Chemical Abstracts Service," *J. Chem. Doc.*, **14**, 92–5 (1974).
(2) See CA Volume 76 Index Guide (1972), Section II, paragraphs 9–10, pp 5I–9I, and Section III, pp 13I–19I for a discussion of the CA General Subject Index.
(3) Data element statistics for Volume 71 are presented by Zipperer, W. C., Stearns, R. E., Jr., and Park, M. L., "The Integrated Subject File. I. Data Base Characteristics," *J. Chem. Doc.*, **13**, 92–8 (1973).
(4) Lancaster, F. W., "Vocabulary Control for Information Retrieval," Information Resources Press, Washington, D. C., 1972, Chapter 18, p 167.
(5) The Control File is a component of the Chemical Abstracts Service's User Aid package, which provides assistance to the chemist and information scientist in the form of 13 Search Aids to be used in accessing the various CAS products and files. These Search Aids are described in "Chemical Abstracts Service Search Aids for the 9th Collective Index Period (1972–1976)" (International Standard Book Number: 8412-0198-6, Library of Congress Number: 74-80986), June 1974, which is available by contacting the Marketing Department of Chemical Abstracts Service.
(6) See CA Volume 76 Index Guide (1972), Section II, paragraph 10, p 8I.
(7) Reference 6, Section I, paragraph 6, p 2I.
(8) Other programs for detecting and correcting spelling errors are described by (a) Alberga, C. N., "String Similarity and Misspellings," *Commun. ACM*, **10**, 302–13 (1967); (b) Blair, C. R., "A Program for Correcting Spelling Errors," *Inf. Control*, **3**, 60–7 (1960); (c) Damerau, F. J., "A Technique for Computer Detection and Correction of Spelling Errors," *Commun. ACM*, **7**, 171–6 (1964); (d) Davidson, L., "Retrieval of Misspelled Names in an Airline Passenger Record System," *Commun. ACM*, **5**, 169–71 (1962).

Polymer Nomenclature, Classification, and Retrieval in the Du Pont Central Report Index†

JOHN L. SCHULTZ

Central Report Index, Information Systems Department, E. I. du Pont de Nemours and Company, Inc., Wilmington, Delaware 19898

Received March 14, 1975

A comprehensive nomenclature system for polymers has been devised in which carbon–carbon addition polymers and polymers of unknown structure are named from their starting monomers; polyamides, polyesters, and polyurethanes from prescribed monomers; and all other polymers of known structure from their structural repeating units (SRU's). Rules are given for aftertreated, graft, and chain-extended polymers. Polymers are classified, for generic retrieval, according to features obvious from their names: each polymer is posted in the Descriptor File under registry-number descriptors corresponding to the registry numbers of its monomers (including artificial monomers derived from SRU's), and a small number of class descriptors are used for overall features of the polymer. Correct registration of polymers is aided by lookup and matching of names and molecular formulas. The molecular formula of a polymer is generated by computer as the sum of the molecular formulas of its monomers. Name and molecular formula lookup are verified by a computer check on the uniqueness of descriptor combinations. Families of polymers are retrieved by searching the Descriptor File. The distinctive feature of the search system is the retrieval of polymer structural details through monomer structural details. A family of monomers can be retrieved by searching the Descriptor and/or Chemical Topology Files for any specified features. The registry numbers so retrieved are translated into the corresponding registry-number-descriptors and automatically resubmitted to the Descriptor File to retrieve the polymers posted under them.

INTRODUCTION

The Du Pont Central Report Index and its handling of chemical information have been described in earlier papers.[3,4] Briefly, the Central Report Index indexes and retrieves the information contained in Du Pont proprietary documents. These functions are carried on by information chemists. Documents are indexed under chemical terms and general terms. Chemical terms denote individual chemical compounds and related concepts, incuding polymers; general terms denote all other concepts. Each general term is the name of a concept, e.g., OXIDATION; each chemical term is a seven-digit alphanumeric registry number assigned sequentially to a compound when it first enters the system. The registry number ties together a series of files, used for identification, document referencing, and classification of compounds. This is shown schematically in Figure 1; for details see an earlier paper.[4] All these files, except the Molecular Formula Card File, are computerized on the IBM 360 and are therefore interlinked to perform a variety of storage and retrieval functions. Currently the files cover 73,000 documents with 116,000 chemical terms: 94,000 for nonpolymers and 22,000 for polymers. These 22,000 polymers have been encountered in the indexing of such major technologies as elastomers, films, plastics, and synthetic fibers over the last 25 years.

DISTINGUISHING AMONG "DIFFERENT" POLYMERS

A polymer is not an individual chemical compound but rather a collection of compounds differing in such chemical properties as molecular weight, linearity, and sequence of structural repeating units (SRU's). This raises the problem of how and in what detail to distinguish between one poly-

† Presented before the Division of Chemical Literature, 169th National Meeting of the American Chemical Society, Philadelphia, Pa., April 8, 1975.

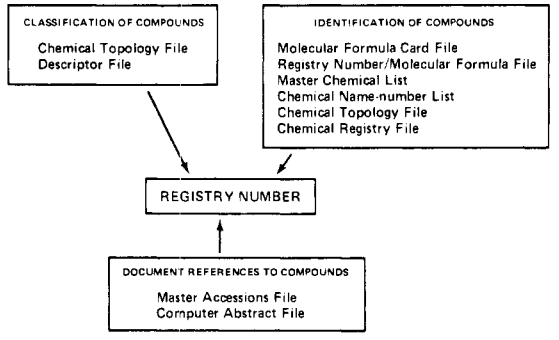


Figure 1. Files for handling chemical information.

mer and another, or to decide when one polymer is "different" from another for purposes of indexing and retrieval.

At the Central Report Index we regard two polymers as different (and therefore assign different registry numbers to them) when they can consistently be given different names according to the nomenclature system described below. In general, the same name and registry number are assigned to all polymers that have the same, or ideally the same, combination of smallest possible SRU's, no matter how or from what reactants they were made. Any distinctions among polymers differing in features such as the following are made by indexing, in combination with the registry number, appropriate general terms rather than by assigning different names and registry numbers: high vs. low molecular weight; comonomer ratios (proportions of the various SRU's in copolymers); random vs. block vs. alternating copolymers; linear vs. branched vs. crosslinked; head-to-tail vs. head-to-head; atactic vs. tactic.

GOALS IN INDEXING AND RETRIEVING POLYMERS

The system described here was set up to attain the goals listed below in the areas of nomenclature, registration, classification, and retrieval (details of implementation are shown in later sections).

Goals in Nomenclature

- To name each broad type of polymer consistently for repetitive identification but in a way that conforms as closely as possible to the way the chemist usually thinks of it. Accordingly, carbon–carbon addition polymers are named from their starting monomers; the large condensation classes polyamide, polyester, and polyurethane are named from prescribed (formalized) monomers; and all polymers of known structure are named from their SRU's.
- To provide a firm basis for consistent identification and registration. The same name should always be associated with the same registry number, so that a polymer is registered as new only if its name differs from all other names in the system.
- To provide a clear-cut basis for classification. Each polymer is classified, directly or indirectly, according to features that are deducible from its name. (As shown under Classification, some of this classification is done indirectly through the structural records associated with the monomers whose names appear in the polymer name.)

Goals in Identification and Registration

- To attain consistent registration as discussed in an earlier paper[4]: to assign the same registry number to the same polymer every time, to provide tools for easily finding the existing registry numbers for old polymers, and to enter as new only those polymers that are truly new to the system.
- To provide computer-maintained tools for registration by lookup of both name and molecular formula.
- To provide a computer check on the accuracy of manual lookup and registration using the above tools.

Table I. Analysis of the 493 Central Report Index Polymer Search Inquiries for the Year 1973

Type of inquiry	% of total[a]
For specific monomer or monomer/class combinations (e.g., 6–6 nylon)	79
For families of polymers	30
Polymers of a given class (e.g., polyesters)	12
Polymers from specific monomers (e.g., polymers of ethylene)	10
Polymers from families of monomers (e.g., polymers of aliphatic diols)	8

[a] Percentages of subtypes of polymer inquiries add up to more than 100 because some inquiries called for both specific polymers and families of polymers.

Goals in Classification and Retrieval

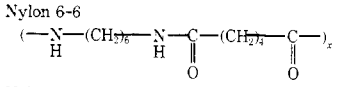
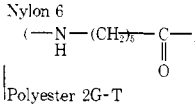
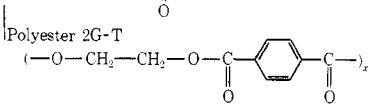
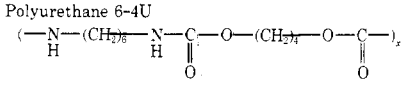
- To group polymers into classes on the basis of structural features so that information on families of polymers can be retrieved.
- To provide a clear-cut, reliable classification procedure with a minimum of difficult judgments and borderline cases. As mentioned above, polymers are classified according to features that are deducible from their names.
- To afford direct access to the polymer features most frequently requested in search inquiries.
- To use the already existing class retrieval system for nonpolymers for access to detailed structural features of polymers.

DETERMINATION OF RETRIEVAL AND CLASSIFICATION REQUIREMENTS

To help define and attain the goals listed above, the 493 polymer search inquiries submitted to the Central Report Index during the year 1973 were analyzed. The results are shown in Table I. The following conclusions were drawn from this analysis.

- Polymers are better indexed as integral, bound monomer combinations or monomer/class combinations than as their separate monomers and classes (79% of the polymer inquiries were for specific monomer or monomer/class combinations). Thus 6-6 nylon, e.g., is indexed as Polyamide-adipic/1,6-hexanediamine rather than as adipic acid, 1,6-hexanediamine, and polyamides. This objective has been accomplished by naming and assigning a separate registry number to each individual monomer combination or monomer/class combination. (Note the use of separate general terms rather than different registry numbers for distinguishing among various combinations of the same monomer—as shown above under Distinguishing Among "Different" Polymers—and for handling end groups and indefinite aftertreatments—as shown below under Nomenclature.)
- Polymers of important classes should be directly retrievable by their classes if they cannot easily be retrieved through the structures of their monomers (12% of the inquiries were for polymers of a given class, mostly such large classes as polyamides and polyesters). This has been accomplished through the class descriptors described below under Classification and Retrieval.
- Polymers should be directly retrievable by their monomers (10% of the inquiries were for polymers from specific monomers). This and the next item have been accomplished through the registry-number-descriptors described below under Classification and Retrieval.
- Detailed substructural access to polymers should be through the structural details of their monomers and thus through the already existing structure classification/retrieval system for nonpolymers.[4] (Past experience had

Table II. Central Report Index Names of Polyamides, Polyesters, and Polyurethanes

Polymer	Central Report Index name
Nylon 6-6 	Polyamide-adipic/1,6-hexanediamine
Nylon 6 	Polyamide-hexanoic, 6-amino-
Polyester 2G-T 	Polyester-ethylene glycol/terephthalic
Polyurethane 6-4U 	Polyurethane-1,4-butanediol/1,6-hexanediamine
Polyamide-ester 6-6/2G-T or 6-T/2G-6	Polyamide-ester-adipic/ethylene glycol/1,6-hexanediamine/terephthalic

shown that a topological system for retrieving detailed structural features of complete polymer SRU's was inefficient and not geared to answer the majority of inquiries for families of polymers—those specifying polymer class or specific monomers.) This has been accomplished as shown under Retrieval.

Implementation of the current system to meet the above goals and requirements is detailed in succeeding sections.

NOMENCLATURE

Carbon–Carbon Addition Polymers—Names Based on Actual Monomers. These are named by prefixing "Poly-" to the *Chemical Abstracts* name(s) of the starting monomer(s). Multiple monomer names are placed in alphabetical order. Examples: Poly-propene; Poly-maleic anhydride/styrene; Poly-acetylene.

Large Condensation Classes: Polyamides, Polyesters, and Polyurethanes—Names Based on Prescribed Monomers. These are named by prefixing a class name to the *Chemical Abstracts* name(s) of *prescribed* (formalized) monomers in alphabetical order. Each pure and mixed class has its own prefix: Polyamide, Polyamide-urethane, Polyester, Polyurethane, etc. Each class of polymer is arbitrarily assumed, for purposes of registration and nomenclature, to be made by the reaction of a complementary pair of prescribed functional groups, present in prescribed monomers. These functional groups are as follows:

Polymer class	Prescribed functional group pair	
Polyamide	Acid	Amine
Polyester	Acid	Alcohol
Polyurethane*	Alcohol*	Amine*

Acid, alcohol, and amine denote any functional group that can act like acid, alcohol, or amine groups in forming polymers. Thus, acid includes not only carboxylic but also sulfonic, etc.; alcohol includes also phenol, thiol, etc. The word "acid" is omitted from the name for brevity. Examples of polymer names are shown in Table II. Prescribed monomers need not be the same as the actual starting monomers. The Central Report Index uses them in nomenclature to avoid assigning several different names to what is ideally the same polymer made from several different combinations of reactants. The commercial polymer 2G-T, e.g., has been made from well over a dozen combinations of common reactants.

All Other Polymers of Known Structure—Names Based on SRU's. Polymers of known structure other than carbon–carbon addition polymers and the large condensation classes shown above are named from their final SRU's because there are no obvious monomers for them. Naming them requires two steps: (1) orient the SRU from left to right, starting with one of the backbone atoms as the head (or senior) atom; (2) name the bivalent radicals as they appear in the SRU and prefix "Poly-" to the result.

This system is similar in principle to that used in *Chemical Abstracts*[2] except in the details of the rules used for orienting the SRU. In the Central Report Index system the head atom is the atom of highest atomic number, chosen by an adaptation of the Cahn–Ingold priority conventions,[1] that has at least one non-ring bond to another backbone atom. If there are two or more atoms of the same highest atomic number, priority among them is decided by priority of atoms attached to them. If the head atom is in a ring, the rest of the ring lies to its right in the oriented SRU. If not, the priority conventions are then used to orient the SRU by choosing which backbone neighbor of the head atom lies to its right. Special conventions are used to avoid splitting functional groups when orienting the SRU. (By contrast, the Chemical Abstract system orients the SRU by selecting a senior *radical*.)

Examples of polymer names are shown in Table III.

Polymers of Unknown Structure. These are named by prefixing "Poly-" to the name(s) of starting monomer(s) in alphabetical order.

Chain-Extended Polymers (Segmented Elastomers). These are named by treating the soft segment as a prescribed monomer, placing its name alphabetically with the names of other prescribed monomers, and prefixing the appropriate class name as shown above for polyamides, polyesters, and polyurethanes. Example: a polyurea–urethane made from hydroxyl-ended polyoxytetramethylene capped with methylenedi-*p*-phenylene isocyanate and chain-extended with hydrazine is named as: Polyamide-urethane-aniline, 4,4′-methylenedi-/carbonic/hydrazine/poly-oxytetramethylene.

Graft Polymers. These are named by naming the base polymer, followed by the word "graft", followed by the name of the grafting unit without the prefix "poly". Example: a polymer made by grafting vinyl acetate onto polystyrene is named as: Poly-styrene, graft acetic acid, vinyl ester.

End Groups. A polymer with specified end groups is indexed as the base polymer plus appropriate general terms denoting the structural features of the end groups. (A former system of separate names and registry numbers for each base polymer/end group combination caused scattering of information on a given base polymer.) These general terms are selected from a complete, closed set according to precise rules. Example: α,ω-dihydroxy-poly-oxytetrameth-

* The carbonyl group in –N(R)C(=O)O– is ignored in deriving the prescribed functional group pair.

Table III. Central Report Index Names of Polymers Other Than C–C Addition Polymers
Polyamides, Polyesters, and Polyurethanes

Oriented SRU	Central Report Index name
—O—CH_2—CH_2—	Poly-oxyethylene
—O—CF_2—CF_2—	Poly-oxy(tetrafluoroethylene)
—S—CH_2—N⟨ ⟩N—CH_2—O—$(CH_2)_4$—	Poly-thiomethylene-1,4-piperazinediylmethylene-oxytetramethylene
—O—CH_2—CH_2— and —O—$(CH_2)_4$—	Poly-oxyethylene/oxytetramethylene

ylene is indexed as Poly-oxytetramethylene (the base polymer) and the general term END GROUPS, HYDROXY.

Aftertreated (Post-Reacted Polymers). These are handled in either of two ways.

(1) If the aftertreatment causes a definite result at a definite site, the resulting polymer is named as the original base polymer, followed by the phrase "aftertreated to", followed by the name of the new monomer or SRU produced by the aftertreatment. No distinction is made among various degrees of completeness (even 100%) of aftertreatment. Example: an ethylene/methacrylic acid copolymer converted to sodium salt is named as: Poly-ethylene/methacrylic acid, aftertreated to methacrylic acid, sodium salt.

(2) If the aftertreatment is indefinite or occurs at an indefinite site or number of sites within an SRU, the resulting polymer is indexed as (a) a name consisting of the base polymer name followed by the word "aftertreated" plus (b) appropriate general terms denoting the structural features introduced by the aftertreatment. These general terms are selected from a complete, closed set according to precise rules. Example: a chlorosulfonated ethylene/propylene copolymer is indexed as: "Poly-ethylene/propene, aftertreated" plus the general term POLYAFT-SULFONYL CHLORIDE.

Special rules are used for other, rarely encountered nomenclature situations.

CLASSIFICATION

Each polymer is classified when it first enters the system by inspecting its name and posting its registry number under descriptors in the Descriptor File. As explained in an earlier paper,[4] each descriptor denotes some structural or other feature and is addressed by a descriptor name up to 23 characters long. Two types of descriptors, explained in detail later in this section, are posted to polymers.

(1) Registry-number-descriptors, one for each monomer that appears in the polymer name. The monomer term thus becomes an attribute of the polymer term. This is the most distinctive feature of the Central Report Index classification system. As shown below under Retrieval, it ties the polymer class retrieval system into the nonpolymer class retrieval system. Registry-number-descriptors afford access to the structural details of polymers through the structural details of their monomers and thus obviate any separate structural classification/retrieval system for polymers. Through registry-number-descriptors, therefore, a polymer is classified not only directly according to its own class and monomers, but also indirectly according to the structural features of its monomers.

(2) Class descriptors for certain other features that are deducible from the polymer name but are not obvious from its monomers.

Registry-Number-Descriptors. A polymer is posted under a registry-number-descriptor for each monomer that appears anywhere in its name. The descriptor name is a seven-digit name identical in spelling with the registry number for the monomer. Example: the monomers for "Polyester-ethylene glycol/terephthalic, graft styrene" are ethylene glycol, terephthalic acid, and styrene, with registry numbers 000309E, 000920F, and 000609B respectively. The polymer is therefore posted under registry-number-descriptors 000309E, 000920F, and 000609B.

For polymers whose names are based on SRU's, an artificial monomer is created from each SRU by filling up each free valence with a dummy atom A. Such monomers are therefore called "A-atom-monomers". Example: for SRU-O-CH_2-CH_2-, the A-atom-monomer is A-O-CH_2-CH_2-A. A-atom-monomers are registered and classified in the system like all other nonpolymers, and their registry numbers are used as registry-number-descriptors for appropriate polymers. Example: the registry number of A-O-CH_2-CH_2-A is 025001C, whence Poly-oxyethylene is posted under the registry-number-descriptor 025001C. Besides the usual appropriate descriptors, each A-atom-monomer is posted under one of the following distinctive descriptors:

SRU-BACKBONE-C-NON-C: the path between A atoms contains both carbon and noncarbon atoms. Example: A-O-CH_2-CH_2-A.

SRU-BACKBONE-C-ONLY: the path between A atoms contains only carbon atoms. Example: A-CF_2CF_2-(*p*-$C_{L6}H_4$)-A.

SRU-BACKBONE-NON-C-ONLY: the path between A atoms contains only noncarbon atoms. Example: A-Si$(CH_3)_2$-O-A.

Use of these descriptors in searching is illustrated below under Retrieval.

Class Descriptors. These are posted by examining the polymer name for the presence of a small number of features. These class descriptors cover (a) the large condensation classes polyamide, polyester, and polyurethane, which are frequently requested in searches and are not conveniently retrievable though the structure of their monomers, and (b) the features "aftertreated", "graft", and "homopolymer", which would not be retrievable at all without the class descriptors. Class descriptors are assigned as follows:

If the polymer name begins with the prefix for one of the large condensation classes (pure or mixed polyamide, polyester, or polyurethane), one posts a descriptor whose name is identical with the prefix, e.g., POLYAMIDE or POLYESTER-URETHANE.

If the word "aftertreated" occurs in the polymer name, one posts the descriptor POLY-AFTERTREATED.

If the word "graft" occurs in the polymer name, one posts the descriptor POLY-GRAFT.

If neither "aftertreated" nor "graft" occurs in the polymer name, one posts the descriptor POLY-HOMO for (a) any polymer with only one monomer or SRU in its name, and (b) a pure or mixed polyamide, polyester, or polyurethane with only one pair of complementary monomers in its name, e.g., Polyamide-adipic/1,6-hexanediamine.

REGISTRATION AND IDENTIFICATION

Polymers are registered with the aid of two computer-maintained manual lookup tools (names and formalized

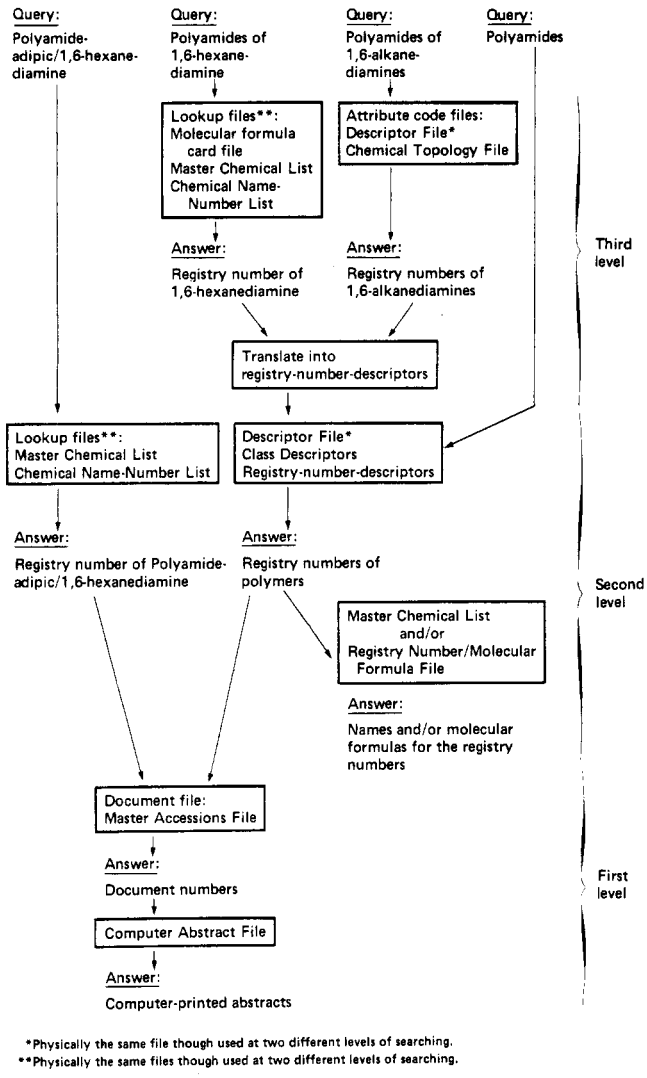


Figure 2. Retrieval of polymer information by retrospective searching.

molecular formulas) and a computer program for verifying uniqueness. If neither the name nor the molecular formula of a polymer is found in the lookup tools, the polymer is assumed to be new and entered into the system. As shown below, a computer program then checks for possible duplication of polymers already in the system by comparing combinations of descriptors for uniqueness.

Names. The polymer names created as described above are displayed in the computer-maintained Master Chemical List.[4] This list is printed in name, registry number, and molecular formula order.

Molecular Formulas. To eliminate the necessity of depending entirely on nomenclature for lookup, a computer program calculates a formalized molecular formula for each polymer. These molecular formulas are stored in the Registry Number/Molecular Formula File[4] and thence stored and displayed in the Master Chemical List as a lookup tool.

The molecular formula of a polymer is defined in the Central Report Index system as the sum of the molecular formulas of the monomers (real, prescribed, or A-atom-monomers) that appear in its name. The computer program calculates this sum by noting the registry-number-descriptors under which the polymer is posted in the Descriptor File, translating them into the corresponding monomer registry numbers, picking up the molecular formulas for these registry numbers from the Registry Number/Molecular Formula File, and adding them up. Example: for the polymer, "Polyurethane-Aniline, 4,4′-methylenebis(2-chloro-/poly-oxytetramethylene/toluene-2,4-diamine," the monomers and their molecular formulas are: 4,4′-methylenebis(2-chloroaniline) ($C_{13}Cl_2N_2H_{12}$), A-O$(CH_2)_4$-A ($C_4A_2O_1H_8$), and toluene-2,4-diamine ($C_7N_2H_{10}$). The molecular formula of the polymer is therefore $C_{24}A_2Cl_2N_4O_1H_{30}$.

Computer Check for Uniqueness of Descriptor Combinations. As a final check to verify that each polymer entered into the system is truly unique, a computer program examines the polymer portion of the Descriptor File after every updating. If it finds two or more polymer registry numbers posted under the same combination of descriptors (both class descriptors and registry-number-descriptors), it prints them out for human inspection. If this inspection shows that there are duplicates by the criteria of the Central Report Index system, the files are corrected accordingly at the next updating.

RETRIEVAL

General Scheme. Retrieval of polymer information, shown schematically in Figure 2, is analogous to the scheme for retrieval of nonpolymers detailed in an earlier paper.[4] The following points apply.

The retrieval scheme is based on a series of computerized files that are interlinked through registry numbers (Figure 1).

The entire retrieval sequence can be programmed to run without interruption through all the steps shown in Figure 2, or it can be stopped at any point to allow human intervention.

Search answer options include not only polymer registry numbers but also names and/or molecular formulas of the polymers, the accession numbers of the documents containing information on the polymers, and abstracts of the documents.

Searching proceeds through one, two, or three logical levels as shown in Figure 2. The third level is a special feature of the system. It utilizes the nonpolymer class retrieval system[4] to retrieve polymers through the structural features of their monomers. Details are shown below.

Multi-level Searching. Searching proceeds through one, two, or three logical levels as shown in Figure 2. The key to most multi-level searching of polymers, and the distinctive feature of the polymer retrieval system, is a computer program that translates the registry numbers of nonpolymers (i.e., monomers) into the corresponding registry-number-descriptors (explained above under Classification), which are posted to polymers. This translation makes all the resources of the nonpolymer retrieval system available to the polymer retrieval system.

The mechanics of searching for families of polymers are similar to those for nonpolymers; consult an earlier paper[4] for details. As explained there, each statement is in the form of an equation. The left-hand member is an answer name up to 16 alphanumeric characters long; the right-hand member consists of the search terms related by the standard Boolean operators "*" (and), "+" (or), "¬" (and not), and "." (end of statement).

A statement may be a final answer statement, shown by a $ before the answer name, or a subanswer statement. The answer from a final answer statement is printed for the user. The answer from a subanswer statement merely enters into the logic of some other answer statement through its answer name's appearing as a search term in that answer statement. A final answer name may also appear as a search term in another final answer statement. Answer names are used as search terms to avoid recoding logic that is common to two or more answer statements.

Search terms can be of several types, the most common of which are the following.

- Descriptor. The term for a descriptor that does not take a count is the descriptor by itself.

• Descriptor relation count. One type of term for a descriptor that takes a count consists of the descriptor followed by an operator (the relation) and a number (the count). The relation shows the comparison to be made between the count stored in the Descriptor File and the count specified after the relation; it determines whether the count stored in the Descriptor File meets the requirement of the inquiry. Thus AMINE-1 > 2 means that a compound must have more than two primary amine groups to be an answer to the inquiry. The relations used are = (equal to), > (greater than), < (less than), > = (greater than or equal to), < = (less than or equal to), and ¬= (not equal to).

• Descriptor relation descriptor. The other type of term for descriptors that take a count consists of two descriptors separated by one of the relations described in the preceding paragraph. It specifies the numerical relationship that must exist between the counts for the two descriptors. Thus RING-NON-C=RING-N means that, for a compound to be an answer to the inquiry, the number of nitrogen atoms in rings must equal the total number of noncarbon atoms in rings, i.e., no nonnitrogen heterocycles are permitted.

• Subanswer names or final answer names. As shown above, an answer name may appear as a search term in another answer statement.

• If an answer from the Descriptor File is to be refined by a search of the Chemical Topology File, the answer name is followed by "(*)".

The translation of nonpolymer (i.e., monomer) registry numbers into registry-number-descriptors is coded into the above search logic as follows:

• For retrieving polymers of individual monomers, the appropriate registry-number-descriptors (which are identical in spelling with the registry numbers of the monomers) are coded as search terms in an answer statement to retrieve polymer registry numbers.

• For retrieving polymers from a family of monomers, an appropriate answer statement is written to retrieve the registry numbers of the desired nonpolymers (i.e., monomers). The answer name of the statement is then coded into a special search term in an answer statement to retrieve polymer registry numbers. This search term is of the form POLY(answername), where answername is the answer name of the statement that retrieved the registry numbers of the monomers. Any number of such POLY(answername) terms may appear in a statement.

EXAMPLES OF POLYMER RETRIEVAL

The examples given below show the answer statements coded to retrieve monomer and polymer registry numbers. They do not show the coding of any associated Chemical Topology File or Master Accessions File searches since these are the same for both polymer and nonpolymer searching. An example of such coding is given in an earlier paper.[4]

Example 1: "Polyamide-esters"
$AMEST = POLYAMIDE-ESTER.

Example 2: "Homopolyesters"
$HOMEST = POLYESTER * POLY-HOMO.

Example 3: "Any polymers of ethylene"
$ETPOL = 000303G.
(000303G is the registry number of ethylene)

Example 4: "Non-aftertreated 6-6 copolyamides"
$66CO = 000028J * 000337K * POLYAMIDE* POLY-AFTERTREATED¬ POLY-HOMO.
(000028J and 000337K are the registry numbers of adipic acid and 1,6-hexanediamine respectively)

Example 5: "Polymers from fluorinated diphenols"
FLUOROPHEN = F > 0 * HYDROXY-ARYL = 2.
$POLYFLUPH = POLY(FLUORPHEN).

Example 6: "Poly(fluorinated diphenol isophthalates)"
SUBA = F > 0 * HYDROXY-ARYL = 2.
$FINB = POLY(SUBA) * POLYESTER * 004730G.
(004730G is the registry number of isophthalic acid)

Example 7: "Polymers from vinyl monomers containing chlorine or fluorine"
ANSA = VINYL * (CL > 0 + F > 0).
$ANSB = POLY(ANSA).

Example 8: "Polymers from vinyl monomers containing both chlorine and fluorine"
CLF = VINYL * CL > 0 * F > 0.
$POLCLF = POLY(CLF).

Example 9: "Polymers of (a) vinyl monomers containing chlorine but not fluorine with (b) vinyl monomers containing fluorine but not chlorine"
SUBA = VINYL * CL > 0 F > 0.
SUBB = VINYL * F > 0 CL > 0.
$FINC = POLY(SUBA) * POLY(SUBB).

Example 10: "Polymers containing fluorine but not chlorine"
SUBF = F > 0.
SUBCL = CL > 0.
$FNONCL = POLY(SUBF) POLY(SUBCL).

Example 11: "Polyethers (polymers with recurring acyclic ether groups in the backbone)"
ETHA = ETHER-ACYCLIC > 0 * SRU-BACKBONE-C-NON-C.
$ETHB = POLY(ETHA).

Example 12: "Silicones"
SUBSIL (*) = A-SI-1-1 = 1 * A-O-1-1 = 1 * SRU-BACKBONE-NON-C-ONLY.
$POLSIL = POLY(SUBSIL).
(SUBSIL interfaces with a Chemical Topology File search for the substructure A-Si-O-A)

CONCLUSIONS

The system described here accomplishes the goals and requirements set forth above.

• Criteria have been established for indexing and registering polymers as individual concepts (bound terms) with provision for separate terms where these are useful for certain features.

• Firm rules have been established for deciding when one polymer is "different" from another for purposes of indexing and registration.

• Effective lookup tools have been provided for identifying and registering polymers: a consistent nomenclature system that conforms closely to the way polymers are visualized by the chemist; a simple, computer-calculated molecular formula file; and an automatic computer check on the accuracy of manual registration.

A classification system has been devised that requires a few simple judgments based on the polymer name and utilizes the existing nonpolymer classification system for structural details.

A retrieval system has been set up to answer the great majority of polymer inquiries directly, to make reasonable provision for the rest of the inquiries, and to utilize the structural retrieval system for nonpolymers to retrieve information on the structural details of polymers.

ACKNOWLEDGMENTS

The ideas brought together in the system described here grew, in one form or another, in various groups throughout Du Pont, including the Central Report Index. The Textile Fibers Department Report Index and the Central Patent Index devised systems of prescribed monomers. Dr. W. L. Alderson of the Central Research and Development Department applied the Cahn–Ingold rules to orienting

SRU's. The Central Patent Index and Dr. Alderson devised earlier systems for retrieving families of polymers through the features of their monomers or individual SRU's.

LITERATURE CITED

(1) Cahn, R. S., and Ingold, C. K., "Specification of Configuration about Quadricovalent Asymmetric Atoms," *J. Chem. Soc.*, 612 (1951).
(2) Committee on Nomenclature, American Chemical Society Division of Polymer Chemistry, "A Structure-Based Nomenclature for Linear Polymers," *Macromolecules*, **1**, 193 (1968).
(3) Montague, B. A., and Schirmer, R. F., "Du Pont Central Report Index: System Design, Operation, and Performance," *J. Chem. Doc.* **8**, 33 (1968).
(4) Schultz, J. L., "Handling Chemical Information in the Du Pont Central Report Index," *J. Chem. Doc.*, **14**, 171 (1974).

A Unique Chemical Fragmentation System for Indexing Patent Literature†

MARY Z. BALENT* and JANE M. EMBERGER

IFI/Plenum Data Company,** Wilmington, Delaware 19808

Received November 5, 1974

A new adaptation of a chemical fragmentation system provides a unique procedure for indexing and searching the specific chemicals, classes of compounds, and Markush structures found in patent literature. This computer-based system employs a POSSIBLE and MUST approach which allows generic structures to be searched with a minimum of false retrieval. The data base includes over 300,000 chemical and chemically related patents. Searches can be structured using the fragmentation system alone or in conjunction with general terms, compound terms, assignees, and U.S. Patent Office class codes.

INTRODUCTION

The IFI fragmentation system for organic chemicals is an adaptation of a scheme developed at Du Pont's Central Research Department in the late 1950's.[1] It was first used for patent literature by Du Pont's Central Patent Index in 1964.[2] In early 1972, the IFI/Plenum Data Division of Plenum Publishing Corporation purchased the rights to the Du Pont Company's system for machine retrieval of patent information, including the fragmentation system. During the remainder of that year, IFI successfully integrated the Du Pont system and data base into IFI's Comprehensive Index to U.S. Chemical Patents.[3]

This computer-based fragmentation system employs a POSSIBLE and MUST approach which, we believe, provides a unique method especially suited for indexing and searching the classes of compounds and Markush structures found in the patent literature. Also, since fragmentation is only a part of the total IFI index, the information scientist has the ability to pinpoint chemical information by utilizing fragments in conjunction with general terms, chemical terms, assignees, and/or United States Patent Office class codes.

FRAGMENTATION—DESCRIPTION

The IFI fragmentation system uses systematic rules and a controlled open-ended vocabulary. Chemical compounds are indexed in terms of substructural pieces that characterize them. The indexing terms or fragments are grouped into four categories:

1. atoms present terms
2. functional group terms
3. ring terms
4. configuration terms

Atoms present terms describe the number of carbon atoms and any specific halogen and metal atoms included in a compound. The 5-8 CARBONS and CHLORINE shown in Figure 1 are examples of atoms present terms.

Functional groups (FG's) are atoms or groups of atoms which characterize classes of compounds to which an indexed structure belongs. The system has a unique procedure for defining functional groups and allows for the introduction of new structures in a systematic and straightforward fashion. Examples of functional groups are the C to C DOUBLE BOND, HYDROXY, SULFONAMIDE, and HYDRAZIDE shown in Figure 1.

Functional groups are found in the fragment vocabulary list using atom counts. A name or linear structure is used to identify specific functional groups. For example, in Figure 1, the sulfonamide group is listed as NO2S SULFONAMIDE FG and the hydrazide as CN2O followed by a linear notation.

Ring systems are indexed by ring-type terms (ACYCLIC, CARBOCYCLIC (CARBO), HETEROCYCLIC (HETERO), FUSED OR BRIDGED), by degree of ring unsaturation terms (NO unsaturation, PARTIAL unsaturation, MAXIMUM unsaturation), by the number of ring units in a structure (ONE, TWO, THREE, FOUR or more) and by specific ring structure terms. The ring structure terms describe the skeletons of rings and follow the nomenclature established in "The Ring Index" by A. M. Patterson, L. T. Capell, and D. F. Walker. Their arrangement in the fragment vocabulary list is similar to that of "The Ring Index." A Roman numeral, used to denote the number of individual rings, is followed by a ring formula and a name. The benzothiophene ring in Figure 1 is described by HETEROCY-

† Presented before the Division of Chemical Literature, 168th National Meeting of the American Chemical Society, Atlantic City, N. J., Sept 10, 1974.

* Author to whom correspondence should be addressed.
** A division of Plenum Publishing Co., 227 W. 17th St., New York, N. Y. 10011.