Pure & Appl. Chem., Vol. 49, pp.1855 -1869. Pergamon Press, 1977. Printed in Great Britain.

IDC AND ITS METHODS OF OPERATION

Helmut Grünewald

Gesellschaft Deutscher Chemiker, Boschstrasse 12, D-6940 Weinheim, Germany.

Abstract - The paper summarises the activities of the Internationale Dokumentationsgesellschaft für Chemie mbH (IDC) and the systems used by IDC for storing chemical information and for its retrieval.

1. Introduction

It goes without saying that a modern documentation system is one of the bare necessities of chemical research and development in today's world. To be convinced of this, one need only recognise that the number of chemical publications which have appeared since the beginning of this century exceeds six million, and that another ten million are likely to be published before the year 2000. The question of whether a given problem has already been investigated is becoming more and more like a search for the proverbial needle in a haystack.

It is precisely at this point that the work of a documentation system comes into play. With the help of such a system, one should be able to extract from the vast number of documents available, that particular set which contains the answer to the question in hand. To make things more difficult, the selection should be as free as possible from irrelevant retrievals, and one also wants to be sure that no relevant publications have been overlooked. In actual fact, there is no documentation system in existence today which satisfies these desires in other than an approximate manner. Even if one is willing to be somewhat lenient, the goal is still quite demanding, and it is apparent that a great deal of groundwork will have to be laid before one can hope to reach it.

2. IDC

Working from the premise that it is less expensive to do the groundwork at one time and in one place, rather than several times in several places, some well-known chemical companies collaborated, in 1967, in founding the IDC Internationale Dokumentationsgesellschaft für Chemie mbH (International Documentation Society for Chemistry, Ltd.) They charged IDC with the task of processing the chemical literature in such a way that IDC member companies could search it with reliability and flexibility (1).

Today, IDC has eleven member firms from chemical industry in Austria, Germany, and Japan. These firms share the burden of IDC expenses proportionally, i.e. according to their size (based on the number of staff employed by each). In return, each is entitled to receive all files, computer software, products, and services provided by IDC. Membership in IDC is open to any chemical company in any part of the world.

However, IDC also performs searches on its files for non-IDC members. Therefore, by one means or another, every chemist can take advantage of the information capabilities described below.

3. IDC's Scope of Coverage

The literature of organic chemistry from 1959 forward has been encoded by IDC. In the beginning, information services from some of the member companies were used as source material, but in recent years IDC has shifted toward co-operation with other international services. Thus, the Central Patents Index of Derwent Publications, London, provides the basis for encoding the patent literature. In addition, there has been a partnership agreement between IDC and Chemical Abstracts Service since 1975, which gives IDC the right to unlimited use of all components of the CAS system, including files not yet publicly available, within the boundaries of the Federal Republic of Germany and within the confines of member companies outside West Germany.

IDC currently processes journals and patent literature in the area of low-molecular organic chemistry, as well as patent literature from inorganic and polymer chemistry. In addition, CAS material from the entire field of chemistry is processed.

In order to find the information one needs, some kind of label, so to speak, must be affixed to each document entering a documentation system, so as to permit one to determine the

content of each document through a rapid scan of the labels. Chemical publications can be distinguished formally, through their bibliographic data, or by content, using keywords, chemical structures, or reactions. All of these identifiers are used by IDC.

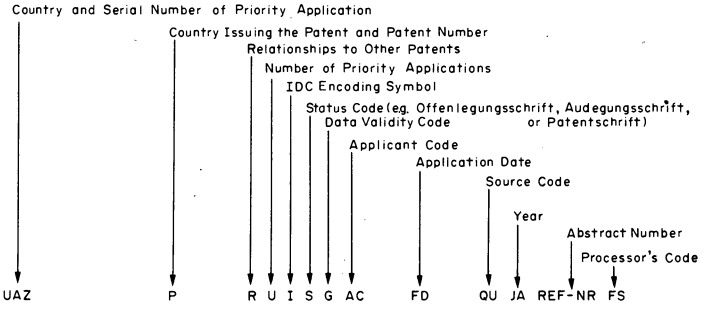
4. Bibliographic Data

The following types of bibliographic data are extracted by IDC from the journal literature:

- The title of the article
- The names of the authors
- The name of the journal
- The volume number of the journal
- The issue number of the journal
- The beginning and ending page numbers of the article
- The year the article appeared in the literature

Of these, only the title of the article, and perhaps the names of the authors, can provide information regarding the content of a publication. All the other data serve to identify the publication formally, but have no value in searches where the query pertains to the subject content.

The same is true of patent bibliographic data. But in this case, it may be important for legal reasons to know the patent situation of a particular enterprise, or to know all the patents which have been issued to protect a given invention in various countries, i.e. all the members of a "patent family". Such questions can be answered using the bibliographic data IDC brings together from the patent literature in its Patent Data File. This file contains the information items listed in Figure 1, and is made available on microfiche and in machine-readable form (magnetic tape). The data are sorted three different ways in the microfiche version; an excerpt from each is shown in Figure 1.



Sorted by country and Serial Number of the Priority Application

```
DT  2 339 865  BE     818 527    1 *      BAY  7. 8.73  OD  W 12 728 BA
      "        DT 74/  4 183     1        BAY  7. 8.73                I*
      "        DL     115 134    1        BAY  7. 8.73                BA
      "        DT  2 339 865     1   .    BAY  7. 8.73  OH 75/11 166 ID
      "        DT  2 339 865     1   .    BAY  7. 8.73  OT 75/ 1 883 BA
      "        DT  2 339 865     1   .    BAY  7. 8.73  CA 83/116715 CA
      "        FR  2 257 625     1        BAY  7. 8.73                I*
      "        JA 75/ 45 092     1        BAY  7. 8.73                BA
      "        LU     70 668     1        BAY  7. 8.73                I*
      "        NL 74/ 10 497     1        BAY  7. 8.73                I*
```

Sorted by country Issuing the Patent and by Patent Number

```
DT /2 339 863  DT  2 339 863     1   .    HOE  7. 8.73  OH 75/10 139 I*
DT /2 339 863        "           1        HOE  7. 8.73  CA 83/ 29195 CA
DT /2 339 865  DT  2 339 865     1   .    BAY  7. 8.73  OH 75/11 166 ID
DT /2 339 865        "           1   .    BAY  7. 8.73  OT 75/ 1 883 BA
DT /2 339 865        "           1   .    BAY  7. 8.73  CA 83/116715 CA
FR 72/ 28 596  DT  2 339 866     1        REN  8. 8.72  OD  V 15 842 DW
U7 /  279 800  DT  2 339 867     1        FLO 11. 8.72  OD  V 17 840 DW
```

Sorted by applicant Code and Date of Application

```
DT /2 339 865  BE     818 527    1 *      BAY  7. 8.73  OD  W 12 728 BA
DT /2 339 865  DK 74    4 183    1        "    7. 8.73                I*
DT /2 339 865  DL     115 134    1        "    7. 8.73                BA
DT /2 339 865  DT  2 339 865     1   .    "    7. 8.73  OH 75/11 166 ID
DT /2 339 865  DT  2 339 865     1   .    "    7. 8.73  OT 75/ 1 833 BA
DT /2 339 865  DT  2 339 865     1   .    "    7. 8.73  CA 83/116715 CA
DT /2 339 865  FR  2 257 625     1        "    7. 8.73                I*
DT /2 339 865  JA 75/ 45 092     1        "    7. 8.73                BA
DT /2 339 865  LU     70 668     1        "    7. 8.73                I*
DT /2 339 865  NL 74/ 10 497     1        "    7. 8.73                I*
```

Figure 1. The IDC Patent Data File.

The IDC Patent Data File is growing at a rate of about 500,000 entries per year, with each entry undergoing very careful scrutiny for accuracy. This file can also be made available to non-IDC members.

5. Subject Content Documentation - The IDC Thesaurus

Printed publications consist predominantly of words. Therefore, it must be possible to extract the most important words from the text and use them to characterize the content of the work. This sounds simple, but in reality, it is fraught with difficulties. In the first place, one must recognize which concepts would be meaningful in a search. Furthermore, there is no one-to-one relationship between words and concepts. There may be different words for the same concept (synonyms), or identical words for different concepts (homonyms). To make things worse, the same word can have different meanings in different languages ("false friends"). Further complications are introduced by relationships of the type illustrated in Figure 2, where one concept term implies other terms which one may have to consider in a search. And finally, there are syntactic relationships between the words in a text. These relationships carry substantive messages which are lost when the words are considered individually.

Hierarchical Relationships

1. Abstraction Relationships
   Broader Term (O)/Narrower Term (U)
   Example: Separation (O)
   Phase Separation (U)

2. Whole-Part Relationships
   Organic Term (S)/Component Term (T)
   Example: Motor Vehicle (S)
   Body (T)

3. Affiliation Relationships
   Reference Term (X)/Related Term (Z)
   Example: Coloring (X)
   Coloring Agent (Z)

Non-Hierarchical Relationships

4. Opposite Relationships
   Term (B)/Opposite Term (G)
   Example: Hydrogenation Catalyst (B)
   Hydrogenation Inhibitor (G)

5. Associative Relationships
   Term (B)/Associated Term (V)
   Example: Corrosion Protection (B)
   Surface Protection (V)

Figure 2. Relationships between terms denoting non-structural concepts.

Ultimately, then, the difficulty of subject content documentation lies in the fact that a given concept can be expressed in many ways, or simply implied by a sentence or phrase. If such diversity were carried over into a search file unfiltered, one could never be sure how to phrase a query in order to retrieve all pertinent documents pertaining to it. Instead, one must standardize the natural language expressions found in the original document. This is accomplished by translating these expressions into an indexing language.

To this end, IDC has developed a comprehensive thesaurus (2), i.e. a hierarchically arranged list of concepts and the words that denote them. Figure 3 shows one concept entry to illustrate how the thesaurus looks. The encoder selects from the thesaurus the particular word which is most appropriate for denoting a given concept. Sections of the thesaurus are used for the various branches of chemistry.

However, chemistry continues to develop. New terms come into being, and old ones may change their content and meaning. What happens in such cases? If the encoder cannot find a suitable term in the thesaurus to correspond to one found in a publication, he is then free to use

whatever word he chooses, and the term used will be processed for inclusion in the thesaurus. This means that the IDC Thesaurus is a "living" list which is constantly being supplemented and revised.

6. Syntactic Relations - The TOSAR System

Syntactic relations exist not only between the words of a text, but also between a compound described in the text and its transformations and properties. Let us assume, for example, that one is looking for all publications in which the reaction of Substance A and Substance B with Substance C is described - perhaps the copolymerization of butadiene and styrene in the presence of a catalyst. Then, any paper which describes the reaction of A with C or B

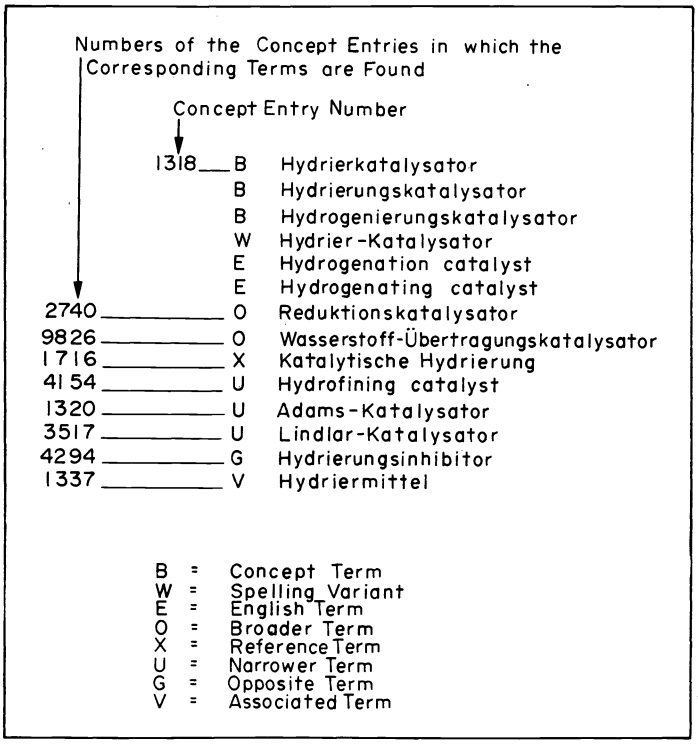


Figure 3. One concept entry from the IDC Thesaurus (abridged)

with C will be irrelevant. Therefore a process is needed which permits one to store the information regarding such relationships in a searchable form.

IDC uses the TOSAR System (3) for this purpose. TOSAR stands for "Topological Representation of Syntactical and Analytical Relations". The system can serve not only to describe the relationships between a compound and its transformations, but it can also be applied to other concepts and may even be used in disciplines other than chemistry.

In the TOSAR System, the interconnections between concepts are depicted through graphs of the type shown in Figure 4. A graph is a system of points and lines. In TOSAR, the points are assigned to concepts, while the lines represent the links between them. The starting points and end points of a given linkage are on different levels. The figures which emerge resemble the structural formulae of chemical compounds and can be handled in a similar manner when filing and searching.

7. Chemical Compounds and Reactions

Those in the field of chemistry have long been blessed with a great advantage, particularly with regard to documentation. This advantage lies in the availability of a highly formalized technique for describing the chemist's subject of interest. Being largely independent of natural language, the technique provides for unambiguous communication. We refer, of course, to the symbols for the chemical elements, and the formulae and reaction equations which one can construct using them and a few other symbols.

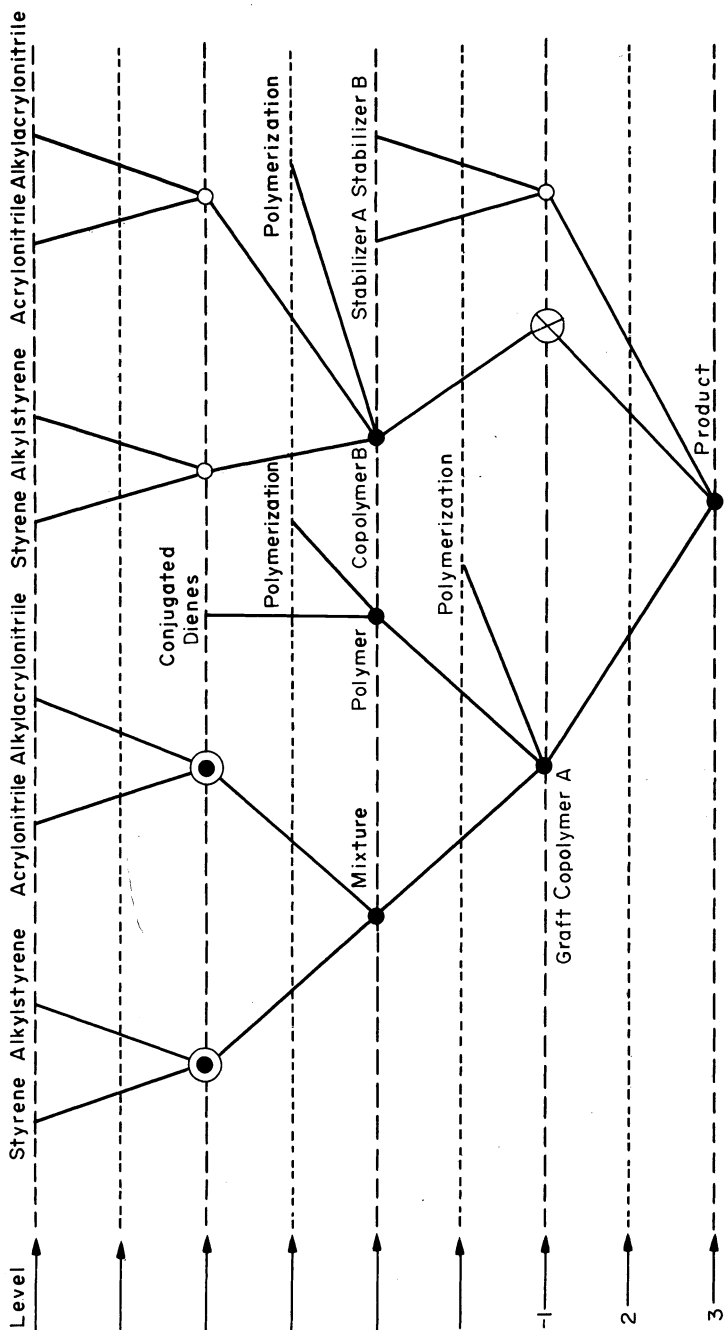


Figure 4. TOSAR graph representing a process that leads to a product consisting of a graft copolymer A, one of two stabilizers A or B, and possibly or a copolymer B.

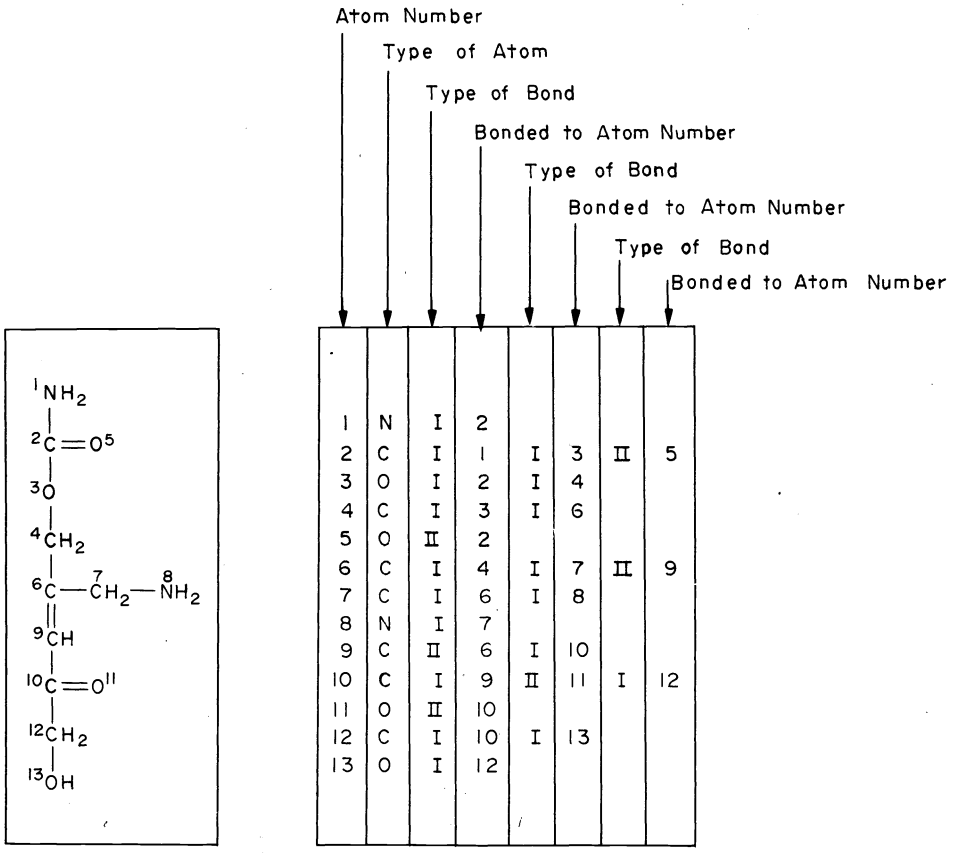


Figure 5. Connection table (right) for the compound shown on the left (simplified). Type of bond : I = single bond, II = double bond.

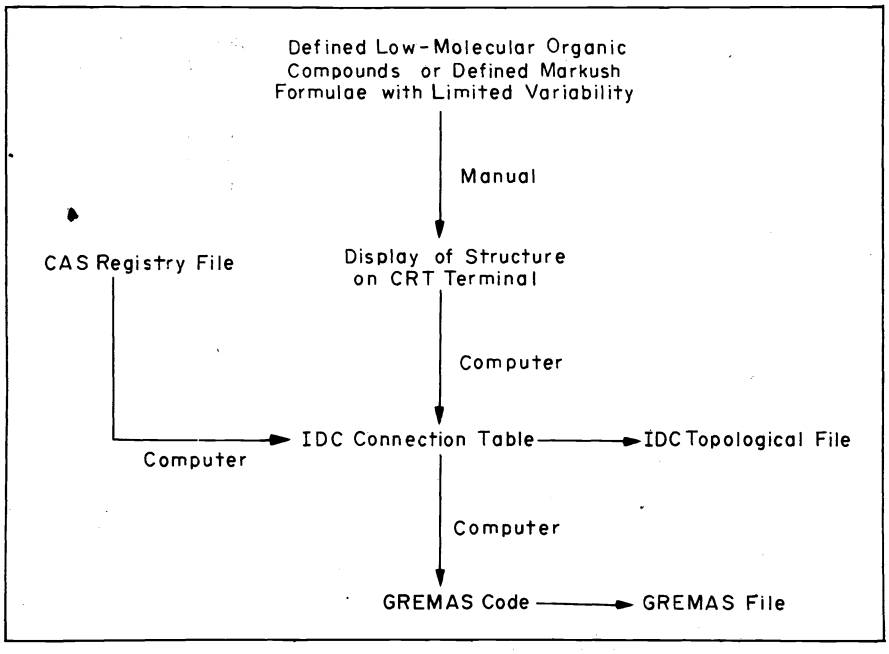


Figure 6. Input of low-molecular organic compounds in the IDC System.

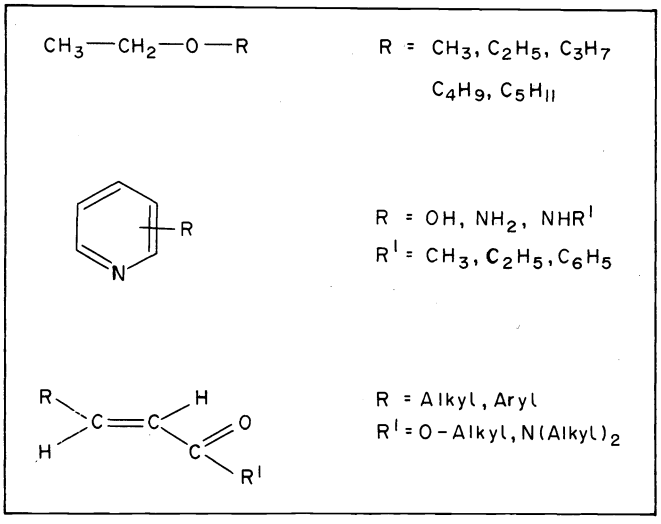


Figure 7. Structure diagrams which depict several distinct compounds or compound classes.

7.1. The Topological System

Every defined chemical compound can be described by means of a connection table like the one shown in Figure 5. One numbers all the positions of the compound (except those occupied by hydrogen) and then lists in a table what type of atom occupies each position, and what type of bond connects the atoms.

Connection tables, when supplemented with stereochemical detail, describe the structures they represent with no loss of information. Herein lies their virtue. Due to this feature, topology-based structure information systems (4) can be interconverted by computer, and this technique is used for linking the structure files of the Chemical Abstracts Service to the IDC system (cf. Figure 6). The drawback of connection tables is that they are relatively expensive to search. With each search, the connection table in question must be compared position for position with every connection table on the file, and one often discovers only at the end of this matching process that the structure wanted is not identical with the structure represented by a connection table.

In the chemical literature - particularly the patent literature - one often finds structure diagrams which are meant to depict several distinct compounds. They may even represent entire compound classes by the use of generalizing symbols or expressions. Some examples are shown in Figure 7.

In the IDC System, such formulae can still be handled topologically, provided that they contain no more than three centres of variable substitution, that the number of alternatives for a given centre is not greater than 9, and that all of these are defined, i.e. that there are no general terms of the type "alkyl" or "aromatic residue" at any position.

Generic substance descriptions such as "alkyl ester of an unsaturated carboxylic acid" or "halogenated alkyl aryl ether" cannot be handled with a system based on connection tables.

In the IDC System, however, such cases can be handled using the GREMAS Code. This code is also applied to fully defined compounds, because it permits much more rapid and less expensive searches in large files than one could hope to achieve using the topological approach. Of course, there is a price to pay for this advantage: precision and recall are generally less than 100%, but they are good enough to make the GREMAS Code a valuable tool for storage and retrieval of organic compounds.

7.2. Organic Compounds - The GREMAS Code

GREMAS stands for "Geneological Retrieval by Magnetic Tape Storage". The G, therefore, states that the code should permit one to search for compound classes (5,6)

To accomplish this, each compound to be stored is broken down into fragments, and a code is then assigned to each fragment. Additional codes indicate how the fragments are linked together.

7.2.1. Fully Defined Compounds

As is the case with chemical nomenclature, the GREMAS Code makes it possible to distinguish organic compounds according to their functional groups. To accomplish this, the code provides three-character expressions, which we shall call trigraphs, to describe the carbon atoms of a compound according to their "kind", i.e. according to the functional groups which are attached to a carbon atom, or of which a carbon atom is a part. The trigraphs begin with a letter and may have either a letter or a numeral in the second and third positions. The primary distinguishing feature for the classification of carbon atoms is the "hetero-orientation", i.e. the number of bonds the atom in question has with hetero-atoms (any atom other than H or C). The level of hetero-orientation determines the first letter, i.e. the GENUS of the trigraph. Carbon atoms, to which only one hetero-atom is attached by a single bond, will have a trigraph beginning with one of the letters from A to H; two single bonds from a carbon atom to two-hetero-atoms, or a double bond to a single hetero-atom, require that one assign the atom to one of the Genera I through M; triple hetero-orientation leads to Genus N or O; quadruple hetero-orientation to P or Q; while carbon atoms bound only to other carbon and/or hydrogen atoms go into Genus R.

One can see from Figure 8, that the meaning of a particular letter changes depending on the position it occupies in the trigraph. The letter B in the first position has a different meaning from a B in the second position; likewise, the letter A in the second position is different from A in the third position. All the details for assigning codes are displayed on a large chart, from which the appropriate trigraphs can be easily read.

Figure 9 shows three examples of encoded carbon atoms. For each compound there must be at least as many trigraphs as there are different functional groups. In actual fact, there are usually more. Namely, in those cases where several carbon atoms are attached to the hetero-atoms of a functional group, where a functional group is constituent in a chain and at the same time a substituent of a ring, and where there are further important characteristics of a structure which should be recorded, e.g. chain length or ring type. This can be seen in Figure 10.

The two structures in Figure 10 are isomers, so the code for each consists of the same trigraphs. This makes it necessary to introduce additional indicators, if one is to distinguish between the two. For this, one makes use of special "region descriptors", which define how the fragments represented by trigraphs are linked in the compound. A region descriptor always begins with the letter Y; then follows an indication of whether the term applies to a chain-like (R), alicyclic (S), aromatic (T), or heterocyclic (U) region within the structure; and, finally, come the initial letters of the trigraphs of all fragments which link together to make up this region. Figure 11 provides another example with region descriptors.

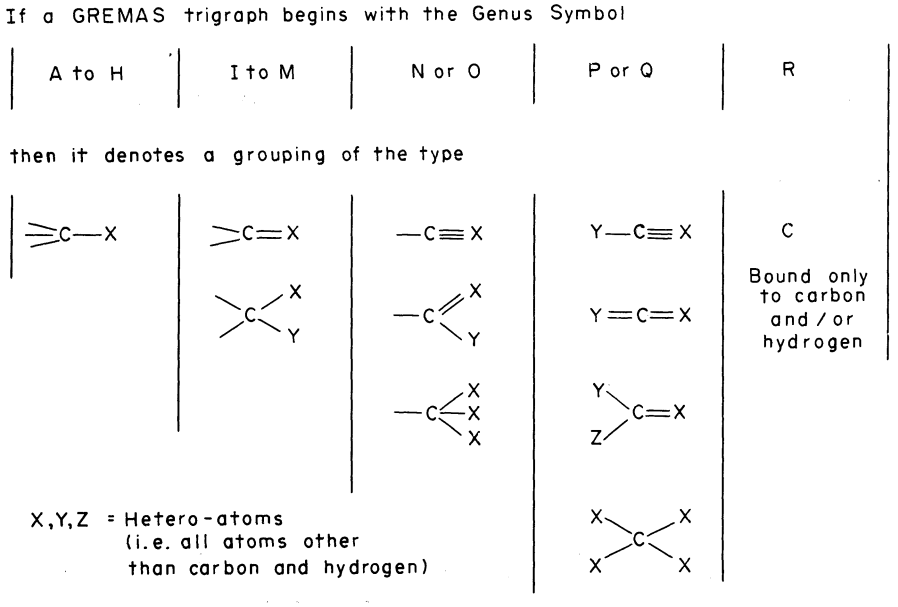


Figure 8. Construction of GREMAS trigraphs.

Examples		
Genus Symbol	Species Symbol	Subspecies Symbol
B Amines	A primary amine B secondary amine C tertiary amine D quaternary amine	A C atom is part of an aliphatic chain D C atom is a substituent on an aromatic ring
E Compounds where X = Oxygen	A alcohols B ethers D esters K hydroperoxides	E C atom is a substituent on a heterocyclic ring F C atom is part of an olefinic chain
H Halogen Compounds	A fluorine compounds B chlorine compounds C bromine compounds D iodine compounds	R C atom is part of an aromatic ring S C atom is part of a heterocyclic ring
N Carboxylic acids and their derivatives	B nitriles G amides N carboxylic acids O esters P anhydrides R halogenides	

The complete GREMAS Codes for all compounds whose structures can be described using topological connection tables can be generated nowadays by computer. The connection tables themselves can be generated using a special keyboard to draw the compounds on a CRT screen. Thus, encoding and storage of low-molecular organic compounds in the IDC system consists of the steps shown in Figure 6.

7.2.2. Generically Described Organic Structures

In GREMAS Code there are trigraphs which denote generalizations by use of the numeral "zero". Examples are:

SA0 = carbon ring
SB0 = aromatic ring
SE0 = nitrogen-containing ring
HOR = a halogen atom bound to an aromatic ring

This type of trigraph is used when the nature of a group is not known in detail. Figure 12 gives examples.

If, on the other hand, one is dealing with a structural formula in which alternatives are given for the substituents, this is expressed in the region descriptors: After the letter Y, with which each region descriptor begins, one first cites the unchanging portion of the structure, followed by a slash, then the variable portions. An example of this is given in Figure 13.

7.2.3. Reactions

Chemical reactions and reaction types can also be described using GREMAS Code (7). Here, of course, one must provide a special indicator to show that the code no longer deals with a single compound, but rather with the transformation of one substance into another. The indicator used is the character couplet "DR". One follows this digraph with those structure-oriented trigraphs which are different in the codes of reactants and products. This is demonstrated in Figure 14 using the example of the conversion of an aliphatic alcohol into an aliphatic chlorine compound.

7.3. Inorganic Compounds

In the IDC System, any material or substance which contains no carbon is considered to be "inorganic". Exceptions are elemental carbon, all carbides, carbonates, and pseudohalides, as well as CO, CO_2, and H_2CO_3 which are also considered to be inorganic.

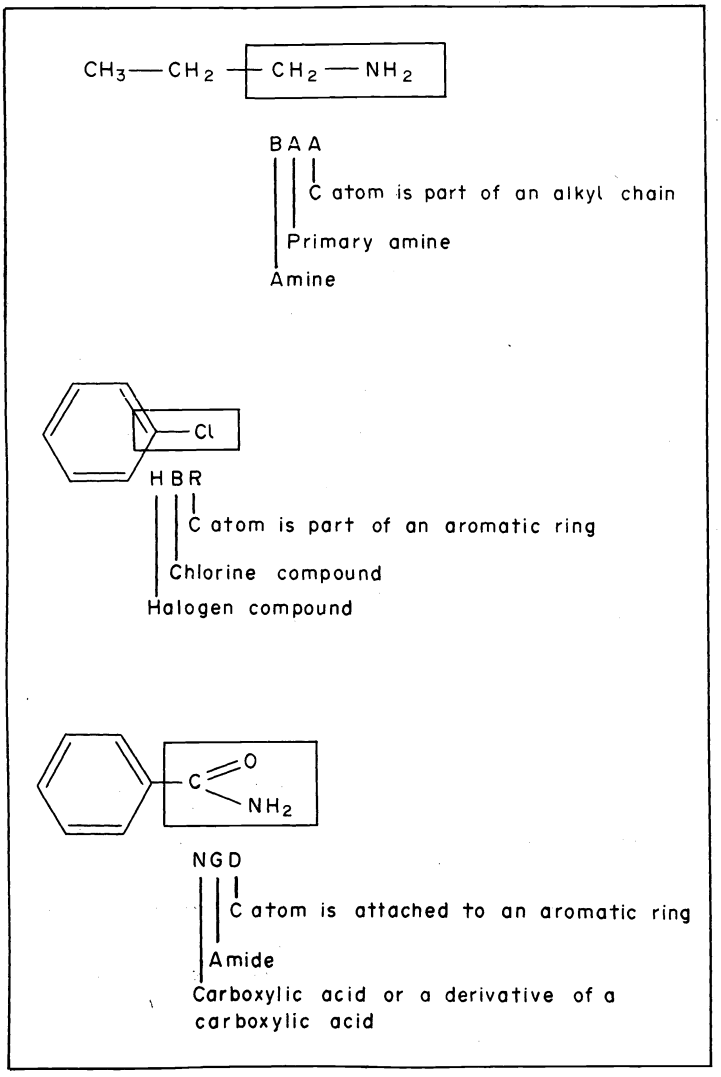


Figure 9. Examples of GREMAS trigraphs.

Special codes are not used for inorganic compounds. One simply uses formulae such as H_2SO_4, $NaNO_3$ or $(NH_4)_2Cr_2O_7$, which are familiar to the chemist. Since the computer can print only upper case letters, and no subscript characters, a numeral is placed after every element symbol. Otherwise, one would not be able to distinguish, for example, ØSCL2 as osmium (II) chloride from ØSCL2 as thionyl chloride. Therefore, $OsCl_2$ becomes ØS1CL2 and $OSCl_2$ becomes Ø1S1CL2.

Generalized designations also occur in inorganic chemistry, e.g. alkali-metal halides, or osmium compounds. In such cases, one is aided by "dummy symbols", for example, QA = alkali metal, QQ = halogen, QZ = optional element, and for an unknown number of atoms, one uses the numeral "zero" (0). With these conventions, "alkali-metal halides" becomes QA1QQ1, and "osmium compounds" becomes ØSOQZO.

If these aids are not sufficient, one has the possibility of adding explanatory information as free text. Thus, "oxides of precious metals" can be input as "MEOØO (EDELMETALLØXIDE)", since the dummy symbol ME means only metal.

In general inorganic compounds are substantially less difficult to encode than organic compounds. The reason is, of course, that structural differences between compounds with the same empirical formula occur less frequently in inorganic chemistry than in organic chemistry.

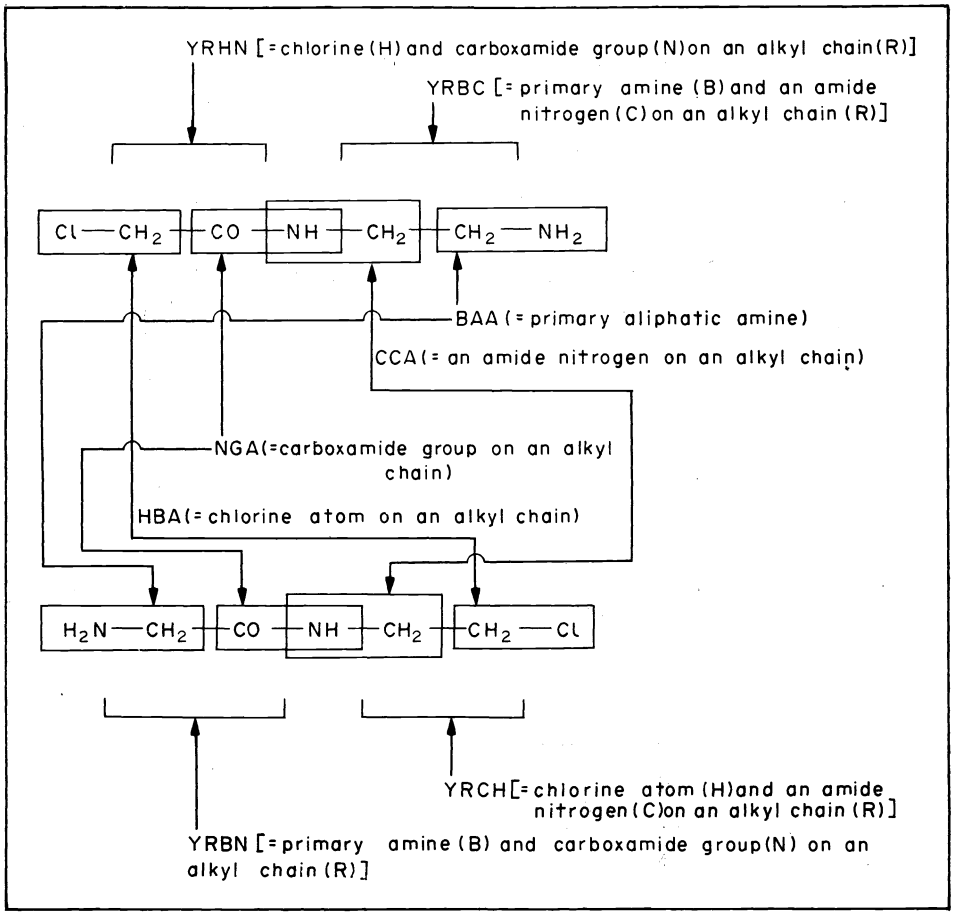


Figure 10. GREMAS Codes of two isomeric compounds.

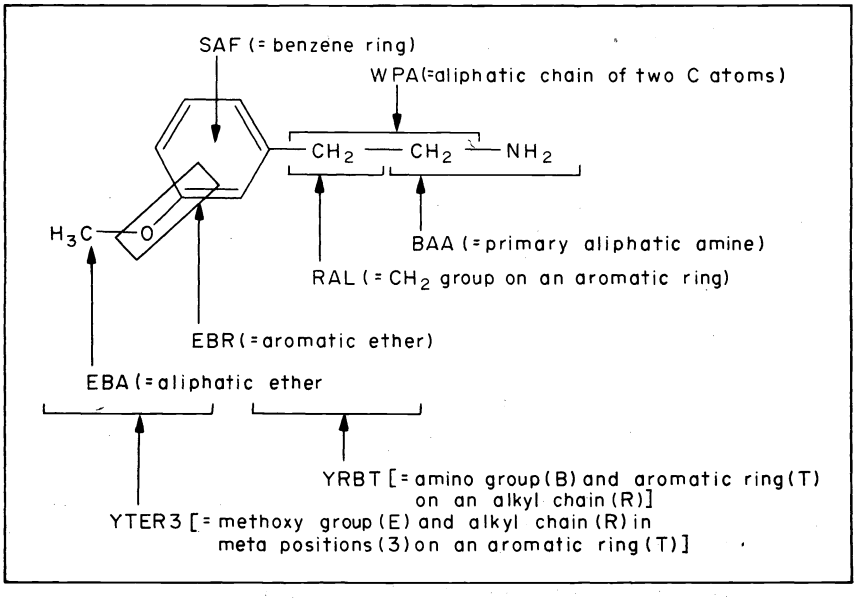


Figure 11. GREMAS trigraphs and region descriptors.

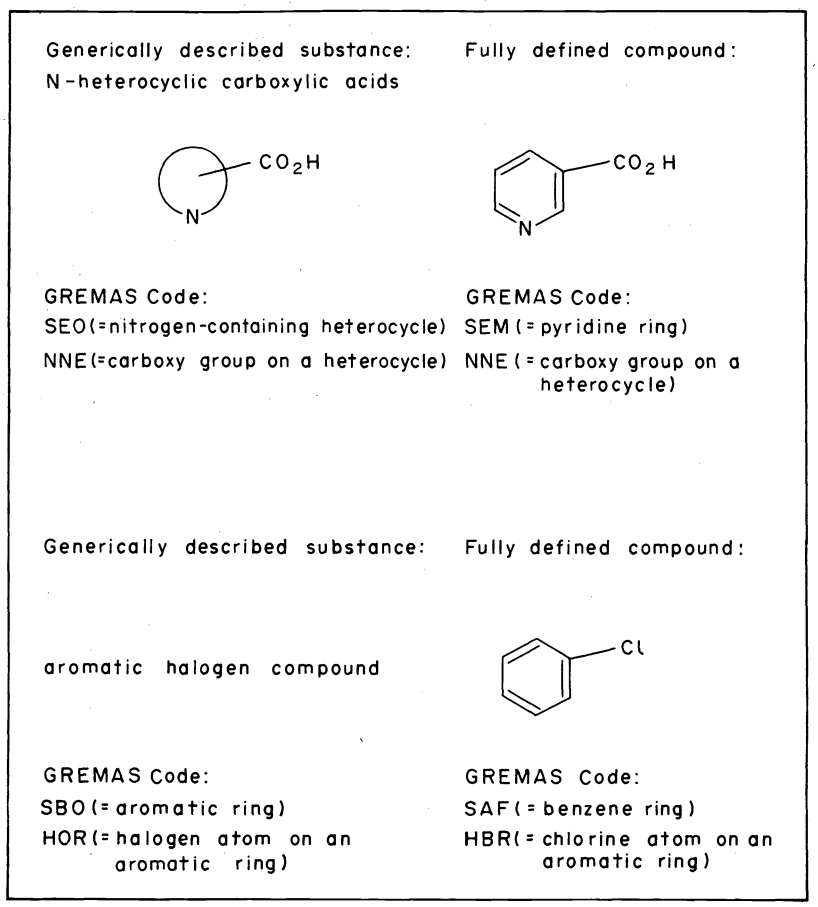


Figure 12. The use of GREMAS trigraphs which denote generalizations.

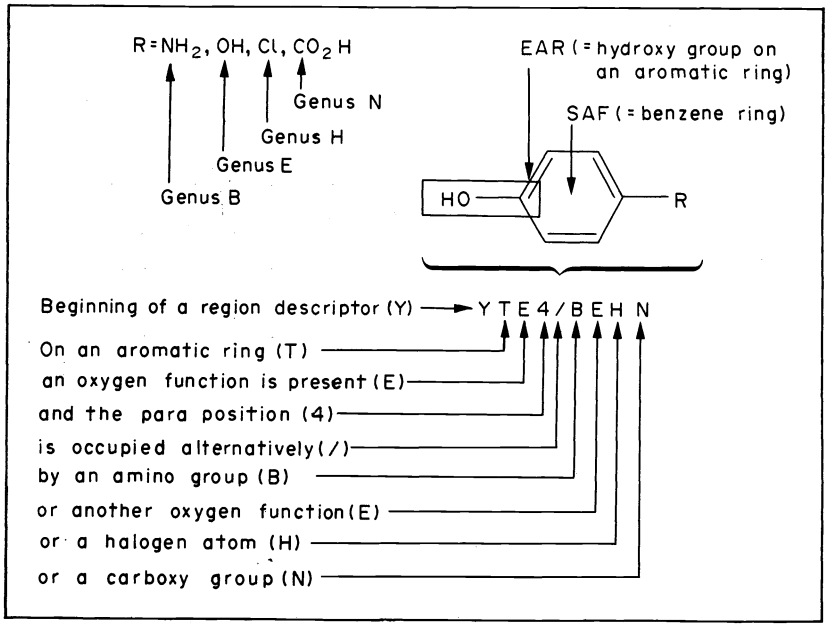


Figure 13. Encoding of a compound with variable substitution using the region descriptor.

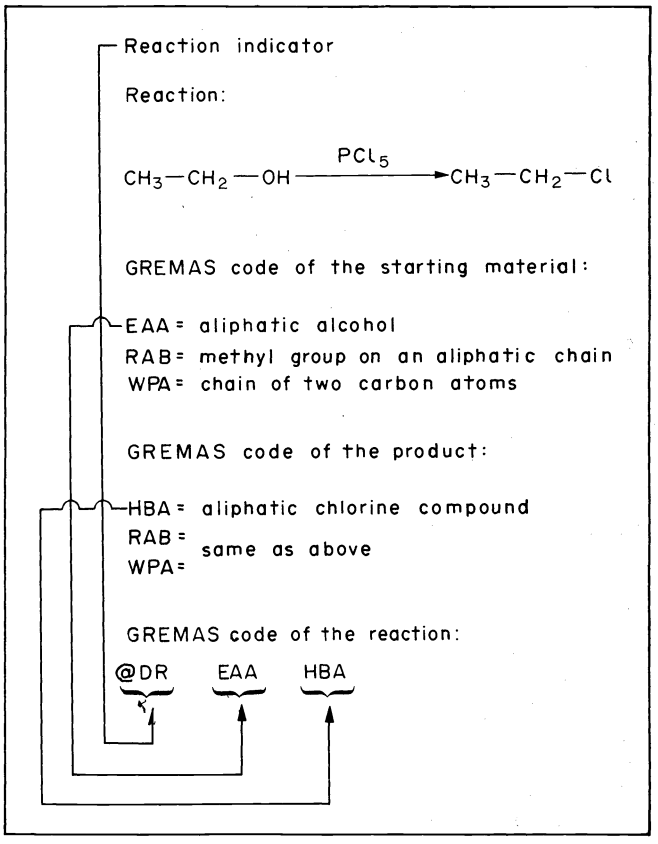


Figure 14. GREMAS Code of a chemical reaction.

8. The IDC Files

The components of the IDC System are brought together, in Figure 15, into a flow diagram.

The IDC Files are fed from three sources:

Chemischer Informationsdienst
Central Patents Index
Chemical Abstracts Service (CAS) Files

These sources are in double boxes in the diagram.

To the extent that IDC itself prepares information for input into its files, it does so from the secondary literature, i.e. from abstracts. In cases of ambiguity, however, the original literature is consulted. Every abstract processed by IDC is filmed, so that for every publication stored in the machine-readable IDC Files, there is a corresponding abstract available in the film file (microfilm or microfiche). All records in the various machine-readable files pertaining to a given publication are linked together by the abstract number. The abstract number also serves as the primary identifier of a publication in the search output.

One sees from Figure 15, that the different types of information (compounds, reactions, non-structural concepts, bibliographic data) taken from the chemical literature are stored on separate files. This has proved to be of benefit in searching.

The IDC Bibliographic Data File contains bibliographic data from publications which have appeared in the journal literature.

The IDC Patent Data File brings together bibliographic data from patents and patent applications.

In the IDC Thesaurus, one finds all the terms used to describe non-structural concepts arranged in hierarchical fashion.

The TOSAR File contains syntactic relations between compounds, reactions, and non-structural concepts, insofar as they are extracted from the patent literature of macromolecular

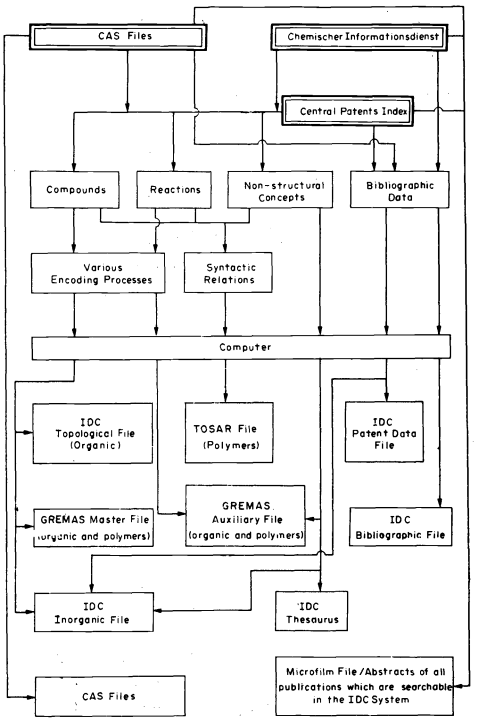


Figure 15. The IDC System.

chemistry. Low-molecular organic compounds are stored on IDC' Topology File in the form of connection tables.

The GREMAS Codes of low-molecular and macromolecular organic compounds form the content of the GREMAS Master File, while the reactions of these compounds as well as non-structural concept information related to them are stored on the GREMAS Auxiliary File.

All compounds and non-structural concepts from publications in the area of inorganic chemistry are stored together with the necessary bibliographic data on the IDC Inorganic File.

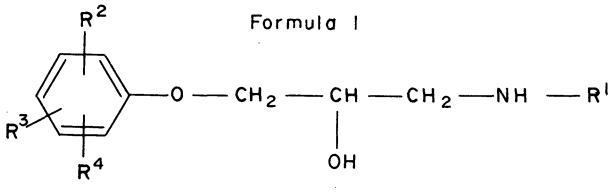
In addition, IDC searches some of the CAS computer-readable services in order to gain information in those areas of chemistry which are not covered by IDC's own sophisticated input process.

9. Searches

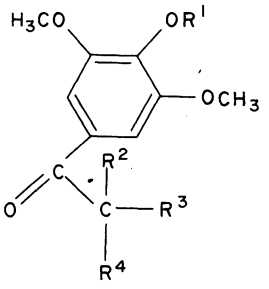
The wide variety of the IDC Files provide correspondingly wide flexibility in searching (8). To illustrate this point, some sample questions which can be answered using the IDC System are shown below:

Which 1-amino-3-aryloxy-2-propanols of the type shown in Formula 1 are known to be coronary therapeutic agents?

Formula I



Where are compounds corresponding to Formula 2 described in the literature?



Formula 2

R^1 = H or benzyl

$R^2 - R^4$ = Any single component other than heterocycles and condensed rings

How can chlorinated hydrocarbons be catalytically dehydrochlorinated in the gas phase?

Is there a process known for producing a scratch resistant transparent copolymer from propylene and methyl methacrylate?

What processes are known for generating hydrogen from water vapor by reaction with iron and chlorine?

In summary, one may ask for individual compounds, compound classes, partial structures, reactions, reaction types, technical processes, properties, and applications. Furthermore, these aspects can be combined in any manner desired. As the examples indicate, it is sufficient to formulate search queries in the normal language of the chemist. No knowledge of the encoding technique used by IDC is required.

IDC performs retrospective searches of publications from previous years, as well as current awareness searches for the most current literature.

In either case, one receives, as search output, a list of bibliographic data or abstracts from the secondary literature for all publications relevant to the search query. With this information, it is quite easy to find further details in the library without loss of valuable time. The cost of a search depends on the type and scope of the question asked, but this cost is always small compared to the time it would take the requester to achieve anywhere near the same result, if left to his own devices in the library.