J., and determine the article numbers associated with his name. He would then look under "Benzodiazepines" in the subject index. A comparison of article numbers would lead him to the article in question.

If a reader desires even more information about a particular article, an OATS postcard, available from ISI upon request, with the ISI accession number, can be sent to the Institute for Scientific Information, and a copy of the original article will be furnished to the reader by return mail.

INDEX CHEMICUS REGISTRY SYSTEM

As mentioned above, structural diagrams are not contained on tab cards or magnetic tape. Because demand is growing for substructure searching by computer, it has become imperative to adopt a computer-usable language to depict structural diagrams. The Wiswesser Line Notation was chosen by ISI, from among several such languages proposed for the *Index Chemicus Registry System*, because of the relative simplicity of this notation, its adaptability for computer searching, and the ease by which these notations can be retranslated into molecular formulas and even structural diagrams.

The more than 150,000 new compounds indexed annually in *Index Chemicus* are available in the line notation form on magnetic tape and computer printouts from the beginning of 1968. We plan to convert the *Index Chemicus* back file of 800,000 compounds into this linear notation in the near future.

A major new feature has been added to *IC* for 1969. Starting with the first issue of the year, all articles containing reports of new chemical reactions are being indexed, even though these reactions are for the synthesis of old rather than new compounds. Wherever possible, the synthetic flow diagram will be shown, and the end products of a reaction which are old compounds will be clearly indicated in the *IC* indexes.

The new reaction code will also appear in the magnetic tapes for the *Index Chemicus,* along with the codes for instrumental and analytical methods.

LITERATURE CITED

(1) Elias, A. W., G. S. Revesz, and G. H. Foeman, "Effects of Mechanization on a Chemical Information Service," J. CHEM. DOC. 8, 2 (1968).

(2) Garfield, E., and G. H. Foeman, "Statistical Analyses of International Chemical Research by Individual Chemists, Languages and Countries," Presented at the Div. Chem. Lit., ACS, Chicago, Ill., September, 1964.

(3) Sher, I., J. O'Connor, and E. Garfield, "Rotaform—A New Index for Chemical Searching of Chemical Compounds," J. CHEM. DOC. 4, 49 (1964).

(4) Sher, I., A. W. Elias, and E. Garfield, "Control and Elimination of Error in ISI Services," J. CHEM. DOC. 6, 132 (1966).

(5) Sher, I., G. H. Foeman, E. H. Baus, "A Slide Rule for Calculation of the Number of Double Bonds and Hydrogen Atoms," Presented at the 145th National Meeting, ACS, September 1963.

The IDC System for Chemical Documentation

ERNST MEYER

Badische Anilin- & Soda-Fabrik AG, 67 Ludwigshafen am Rhein, Germany

Received November 12, 1968

After a decade of systems development, firms of the European chemical industry founded a corporation (IDC) to make the chemical journal and patent literature accessible by efficient computer methods. A comprehensive and critical review of these methods accomodated to the four most important types of chemical data is given.

Chemistry probably more than any other science builds on results which are experimentally reproducible and important even after decades. Compounds described many years ago may be today's starting material and have to be synthesized again; their properties, use, and preparation are continually of interest.

The chemical industry, therefore, requires especially good documentation systems and has been willing to invest monies to avoid unnecessary duplication of experimental effort and to have easy access to the information needed. Numerous documentation services are offered and bought; yet they cannot satisfy all needs and have to be expanded and improved constantly.

The chemical industry not only purchases commercially available services, but widely indexes journal and patent literature in addition to their internal reports. The individual companies, however, in being confronted with the task of coping with the rapidly growing amount of literature, have been considering far-reaching improvements.

This was accomplished by developing new methods of documentation, making full use of modern computer techniques, and by consolidating firms of similar interests—i.e., by founding a joint corporation for centralized documentation to avoid duplication of input effort. After having developed and tested methods covering important areas of chemistry, the International Documentation in Chemistry (IDC) was founded in the spring of 1967[1] to index important areas of the journal and patent literature and to supply the corresponding magnetic tapes for searches. This paper reviews the documentation approach since, up to now, only excerpts have been published.

A really efficient documentation system has to meet many requirements. We shall discuss only the most important ones. The IDC system particularly emphasizes the following objectives:

Adaptation to Variable Needs of Users

The services have to satisfy many and heterogeneous users asking quite different questions. Therefore, the retrieval facilities in particular had to be designed for flexibility. The query in system language has to reflect the real problem of the user. By pressing it into a rigid documentation scheme, the formally correct recall will contain a lot of nonapplicable answers (false drops). To make exceptionally multipurpose searches possible, no information of value for retrieval must be lost in indexing a document. This important criterion for the future value of the system was emphasized by developing it as well as by applying it.

Utilizing Computers Effectively

Transition of proved punched card systems to the computer just is not enough. Electronic data processing provides new and very versatile possibilities for more relevant and flexible searching, such as logical operators—e.g., intersection or negation of individual search terms, automatic inclusion of narrower terms, syntactic relations, etc. It was essential to use all of these facilities to minimize the amount of false drops and the search effort without unduly increasing the input cost.

Economy

Even the best documentation of a subject as extensive as organic chemistry with its abundance of data will fail if the cost of encoding and retrieval is prohibitive. Thus, optimum operating costs were an important objective in developing the IDC system. A multitude of problems were thereby encountered, a few examples of which may suffice:

The option of specifying generic structural formulas with alternate groups (Markush formulas) is essential for indexing patents. The file would become by far too voluminous if these structures had to be specified by each of the possible individual compounds.

Searching is particularly efficient when a multistep code system is used, in which one step is machine-generated from the other one. This holds true in particular if the scanning of one step is less selective but far less expensive, and may thus be used for preselection. Only a small part of the file then has to be searched by the other comprehensive procedure.

The cost of input was substantially reduced as well. To this end a special device was developed[5] which scans structures drawn on grid sheets and stores them on punched tape for topological recording. A great part of the coding, originally done by chemists, could thus be transferred to clerks. This machine is not required for coding queries.

THE IDC DOCUMENTATION SYSTEM (FIGURE 1)

In indexing low molecular organic chemistry—emphasized by IDC up to now— there are four essential types of important specifications: structural formula, reaction, numerical data, and nonstructural information to be described by words only (called "keyword information"). For each of these specifications a separate coding and search program was developed.

The file consists essentially of the three types of magnetic tapes: the "master tapes" containing mainly information about the structural formula, a short reaction code, and code symbols of the most important, most often asked-for keyword information. The "supplementary tape" ("E-tape") contains a detailed reaction code, keywords, and numerical data, and some structural code of special classes of compounds like steroids and peptides. Additionally, there is a tape with topological structure descriptions—as far as compounds have been recorded topologically.[3] Usually it is sufficient to search the master tape. However, the common "file unit number" allows the retrieval of all specifications on all tapes assigned to one compound or Markush formula. In addition, it is possible to intersect queries recalling only those documents which simultaneously answer different single queries—e.g., publications describing all the chemical compounds asked for. This type of query, which is very useful, is facilitated by the "serial file principle" used by IDC—i.e., information is posted to documents, which are stored chronologically. At the same time, this principle reduces machine input cost since—contrary to the "inverted file principle"—the existing file is not changed.

Now, the documentation of the four types of information will be dealt with separately. There is, however, only one search program for each type of tape allowing for all the information in this tape.

Structural Formulas. For about a decade, structural formulas have been encoded and searched by the GREMAS system[2] developed by R. Fugmann of Farbwerke Hoechst; at first, in Hoechst only, since 1962 also at Farbenfabriken Bayer and Badische Anilin- & Soda-Fabrik (BASF), and, since 1967, at IDC.

Essentially, the GREMAS code is a fragment code—i.e., each "term" describes an individual fragment of the formula. Nevertheless, there are important differences in the usual fragment codes. Some of the most important features are the following ones:

> Most of the individual terms are based on a faceted hierarchical principle and can be searched with regard to the presence or absence of individual code symbols in each term. A query for different broader concepts (compound classes) will therefore automatically recall the pertinent narrower terms. For instance, the term NGD characterizes aromatic carboxylic amides; the query terms NG* ("all carboxylic amides"), N*D ("aromatic carboxylic acid derivatives") and N** ("all carboxylic acid derivatives") serve as well. These generalizations may be even restricted in certain ways.
>
> In Markush formulas one differentiates between "constant" and "alternative" terms—i.e., between terms applying to all of the compounds described by the formula, and those which apply only to one or some of them. Nevertheless, negative search conditions (restrictions) can be used, because the alternative terms are in a part of the file unit not answering restrictions. Asked-for compounds are recalled, even if the Markush formula comprises compounds containing forbidden fragments.
>
> In addition to the *ca.* 3500 fragment terms consisting of three characters, there is a theoretically unlimited amount of syntactic terms, describing the simultaneous occurrence of different chemical functions in the same molecule—i.e., at a carbon chain or a ring. These terms are derived easily from the formula and need not be looked up in a code table. In combination with the fragment terms, these very syntactic terms make it possible to reconstruct the formula from the code more exactly than presumably from any other fragment code. And the possibility of reconstruction of the formula is a reliable criterion for selectivity in searching.

To be sure, not all formulas can be reconstructed unambiguously; thus, there will be still some false drops which might be quite annoying at times, especially if the file is growing rapidly. For such instances a subsequent search using a more accurate system—e.g., a topological one—would be desirable. That is one of the reasons the IDC is about to improve its input by topological methods. This transition is being tried currently on a larger scale.

Topological methods allow an unambiguous reconstruction of the formula from its coding; therefore, no false drops will be encountered in searches for defined structures or partial structures. (The GREMAS code allows furthermore the storage of generic structural concepts as "lower alkyl," "saturated aliphatic oxo-carboxylic acid derivatives," etc.) Those partial structures may have any shape and size. This system thereby differs from linear notation systems, in which an unambiguous reconstruction formula is possible, too, but in which searches are limited to a certain selection of the provided concepts, except by first transferring the notation into a topological form, as is theoretically possible for some systems. Unfortunately, the known notation systems do not cover important classes of compounds and could do this only with a large number of rules and conventions.

Fragment codes like GREMAS are unrivalled, when supplemented by superimposed screens,[4] in search speed—especially in searches for partial structures. Search of topological file structures is, on the other hand, extremely expensive when not combined with preselection methods geared to both the computer and question problems.

Considering this, IDC took a three-step approach for the documentation of structural formulas:

> Input, without any loss of information, by the topological system.
>
> Computer-generated GREMAS coding using a very sophisticated program. The GREMAS coding is stored on the master tape, and is flexible and selective enough for most of the requests.
>
> Finally, from the GREMAS coding, a superimposed bit code[4] is automatically generated and stored in a special part of each master tape file unit.

The bit code of the third step consists of five machine words of 32 bits each. In searches, the first word checks negations of "Genus symbols,"—i.e., the first letter of the GREMAS term. In each of the other four words, certain bit combinations—for two-character combinations from GREMAS terms—are set to "1." These combinations for all terms of one formula are superimposed in a 128-bit-field. Some "false drops by superimposition" in preselection are the result, but scanning for the presence of a large number of certain combinations is done simultaneously and extremely fast.

Search is a three step procedure as well: First, the computer generates the superimposed screen from the GREMAS encoded query and compares it with the superimposed bit code of the master tape. On the average this screen eliminates 99.5% of all file units extremely fast, because they definitely cannot give a hit. Only the remaining 0.5% are simultaneously searched with the GREMAS system retrieving, on the average, only about 10 to 20% irrelevant material—with regard to the original inquiry. Should the number of false drops at times be quite substantial and the document screening by a chemist too time-consuming, a search of the topological tape may follow. As the compounds to be checked this way concern a small part of the file only, the cost of the inherently expensive iterative search is negligible. And searches of the GREMAS master tape are inexpensive.

Scanning of a file of 500,000 file units (one-third of which represent Markush formulas, the remainder well-defined compounds) takes an average of about 20 to 30 seconds per query in multiprogramming mode (OS/MVT) on an IBM 360/65. Simultaneous processing of about six queries, as customarily done, has no significant effect on the time required.

These results could be achieved only by combining several systems and exploiting the advantages while suppressing the disadvantages of each one. The topological search is admittedly expensive, but ensures utmost flexibility of queries and highest selectivity in searching. The GREMAS system, born out of chemical search practice, is intended for daily use and makes full utilization of the possibilities of data processing; it is fully geared to the kind of queries a chemist usually asks (to the extent that they cannot by looked up any faster in indexes or handbooks)—contrary to linear notations which were developed originally not to search compound classes but to file well-defined compounds. This fragmentation system, while thus applicable to actual problems, is not yet machine-ready from the economic point of view, however. It is therefore supplemented by an especially economical screening procedure, which naturally is not as selective as the other two approaches.

Using several systems simultaneously is reasonable and economical provided that only one system—obviously the one with the least loss of information—is used for input, and the encoding according to the other codes is done by computer. Understandably, it took years of programming to develop a really efficient integrated system; and there are certainly not many organizations willing to invest the required capital. Furthermore, a team of experienced and specialized programmers is needed, because the extensive programs have to be adjusted from time to time to new types of machines to keep up with the rapid progress in data processing. It is therefore easy to see why it took so many years before the IDC system was considered ready. On the other hand, its central part (GREMAS) was brought to maturity during this time by using it extensively for years. This applies especially to its use in industry, the specific requirements of which for indexing patents were taken into account right from the beginning, although the development was complicated and delayed considerably by this condition.

To be sure, not only the retrieval but also the storage should be as economical as possible. Considerable progress was achieved here, too: clerical staff or machines can now handle a good part of the encoding which previously had to be done by chemists.

Reactions. There are two types of GREMAS reaction descriptions. For all intermediates and end products "short reaction terms" are stored on the GREMAS master tape. They are formed by specifying the atoms taking part in the final step of the production of the compound in question with special three character terms describing their chemical functionality. Searching the master tape is therefore sufficient for most of the reaction queries.

The nontrivial reactions are additionally described in more detail on the E-tape. Essentially, this description consists in recording the GREMAS terms of each reactant

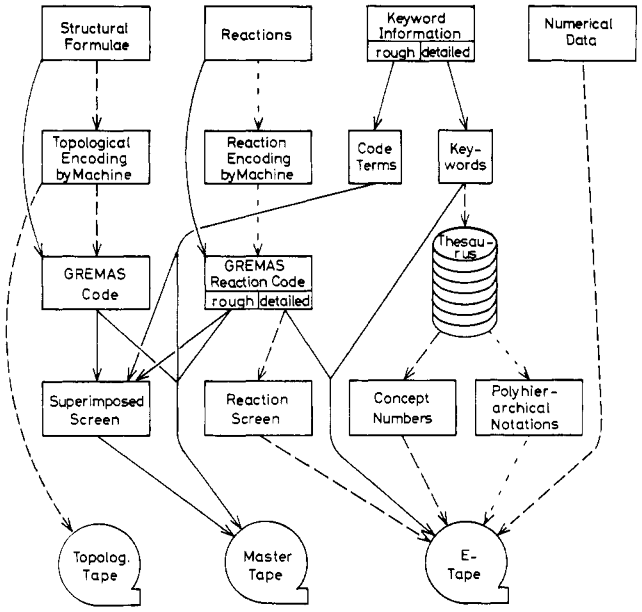

Figure 1. The documentation system of IDC

——— Operative
– – – Operational
--------- Planning Stage

atom, and specifying it before and after the reaction. Moreover, all atoms taking part in a multistep reaction are identified by a number. This "linking" allows searching, without false drops, for the initial and final state of the same atom with no regard to intermediaries possibly described. Description of reaction conditions, inorganic auxiliary agents, etc., can be filed, too, and linked to the appropriate reaction. Widely different questions can be answered by the reaction description file, and many generalizations are permitted in searches for types of reactions. A superimposed screen similar to the one used for structural formula retrieval supplements the reaction documentation, making it a very economical two-step method.

Up to now, no system is known to us which approaches a similar diversity in searches for types of reactions. Nevertheless, there are still some particulars of reactions not searchable; for instance, the influence of substituents at some distance from the reaction site, or the presence of functional groups not changed. These items can be handled by a topological reaction description which is in preparation.

Numerical Data of Measurements. The machine program of the IDC system allows storage and retrieval of data and ranges of measurements—e.g., ranges of pressure or temperature. A posting is relevant if the asked-for range at least partially overlaps the filed one. Programming such a comparison is no problem, but it is rather difficult for the analysts to decide which numerical data are worth indexing. Up to now, not much use was made of the option to store such data, because of the extra effort involved, and because false retrieval is, as a rule, not substantial by searching without those data. They might, however, be of importance for processes, especially in patents. By searching for structural formulas of main industrial products in combination with rather broad keywords, many documents may be recalled which could be restricted by additionally searching for numerical property or process data.

Nonstructural Information. The storage of structural formulas and reactions is not sufficient for frequently encountered substances. Here, one should be able to retrieve them in combination with other facts describable, as a rule, by words only, as e.g., applications, biological activities, chemical engineering details, products of unknown structure, and many more. The documentation method for nonstructural concepts should be as future-oriented as that for structural concepts. This is achieved by encoding the essential—e.g., the possibly asked-for—details as exactly—and that they are reproducible as correctly—as possible. This is accomplished best if the analyst is permitted to choose those words for input which seem to characterize the fact best without restriction by a fixed vocabulary. Thus the concepts will be neither too broad nor too narrow.

This procedure gives in the first stage a nonstandardized file, though, from which retrieval is both difficult and expensive. But it is not impossible if a thesaurus of all keywords is available in which all the pertinent words for setting up a query can be found. But one has to search for all synonyms of the concept asked-for, and, in most cases, its narrower terms and their synonyms as well. This approach becomes expensive when the file reaches a certain size.

A standardization which combines at least the synonyms and better establishes hierarchic relations is therefore necessary. The preparation of such a standardized system is fortunately not too critical because the magnetic tapes with the nonstandardized input form are kept and may be used later for a new computer-generated standardization with an improved concept system, if the need arises. It is therefore not essential to consider too carefully whether two concepts are really synonyms, or whether a classification will meet all later requirements. Wrong decisions do not devalue the file; they can be corrected by computer. The only provision is that not too much information gets lost during the original input. These are the principles of the IDC operation.

Actually, the IDC system is not quite as "open" with regard to nonstructural concepts as it seems to be. For economic reasons, the most important facts are stored together with the structural formula on the master tape as definite three character terms. These terms specify roughly whether the compound is a starting material, its utility, its biological activities, and similar things. Whenever a query for a substance or compound class is selective by itself, these specifications usually suffice, and a search of the E-tape is unnecessary. Furthermore, concepts of some categories which are asked for infrequently and only generically—although often fully described in publications (e.g., some physical properties and model considerations)—are not specified in full detail on the E-tape. Here, the analyst uses a vocabulary of rather broad concepts; a list of relevant terms serves for orientation.

Three stages are scheduled for the "detailed keyword documentation" (on the E-tape) of the IDC system. First, the concept terms are stored as alphanumeric descriptions on the E-tape, as given by the analyst. They are already retrievable by a kind of "Codeless Scanning"[6] and the use of a thesaurus in setting up the query. Second, the free terms are substituted by concept numbers, the same one for synonyms and quasisynonyms. Hierarchical relations are not taken into account in this phase; a concept still has to be searched in addition to its narrower concepts. In a third and final stage, the individual concepts are to be represented by a polyhierarchical notation. Then, by searching for a broad concept the narrower ones may be automatically recalled.

Concept numbers are assigned to the individual terms (second phase) in the following manner:

After each input run, the terms just filed on the E-tape are sorted alphabetically and machine matched against the alphabetical thesaurus tape. Concept numbers are obtained for all terms already included in the thesaurus; these numbers are now in the E-tape. A printout of the remaining terms is submitted to a thesaurus editor, who checks each term to see whether a synonym has already been used by consulting a systematic thesaurus printout for comparison. If the thesaurus contains the concept (although not as the term in question) the new term is inserted into the thesaurus disk file. Now, the new term has its valid concept number for the next updating.

If no synonym has yet been used, the editor sets up a new "concept unit" which has to contain additionally at least one broader thesaurus term, so that the computer can incorporate the term correctly into the systematic part of the thesaurus. A concept number of its own is then automatically assigned to the new term by the computer.

Homonyms—i.e., words with more than one meaning—are marked in the thesaurus and should be used as little as possible. If this happens, they are automatically submitted to a thesaurus editor who replaces them with unambiguous terms.

The programs for all these processes are completed and are being proved; they improve the keyword encoding and concept retrieval considerably. They are more extensive than one would assume. Flexible programs have to be available for the different phases to allow for constant corrections and improvements of both the thesaurus and the file. And there are thesaurus printouts of different types and arrangements to be prepared as aids for searches and classification.

Interpretation of keywords and allotment of concepts are handled in this approach by a few thesaurus experts and not by all indexers; this ensures consistency in the file. One of the most difficult problems is to decide whether a concept is a unit by itself or whether it is a combination of different concepts. Guidelines had to be set up which are still being checked and improved. How far complex concepts should be subdivided into uniterms depends as well on the syntactic aids available. Much research and development has yet to be done in this area, to which IDC hopes to contribute.

The third phase of the keyword documentation is not yet available because, even with computer assistance, the thesaurus editors are burdened considerably with the classification.

For the future, a polyhierarchical notation system is planned, allowing, under different aspects, correlation of several coexisting notations for one concept with the appropriate broader concepts. All concepts, which are assigned to a query term as narrower or equal ones under a certain aspect (of one facet), are then recallable automatically.

The second and even more the third stage of this documentation is already pretty much independent of language, although written words are input.

Expressions in different languages are for the time being treated as synonyms. When the meaning of a word in one language differs somewhat from its equivalent in another language, some of the notations, which will replace the word, may be varied correspondingly during the third phase. And even changes in meanings of words can be taken care of to a certain degree by using slightly different notations from a specified date on.

A considerable amount of work has still to be done to achieve all of these goals, but we are not pressed for time since all keywords are being stored and are retrievable, although with some extra effort.

OUTLOOK

Naturally, some aspects of the IDC system—as with any system—can be improved, and extensive research and development projects are therefore pursued at IDC. But, to our knowledge, its essential parts have been tested and proved more than those of any other chemical documentation system which is used to search large magnetic tape files. The basic concept seems to be sound, although some features may appear rather sophisticated. As a rule, only a complicated apparatus achieves the highest efficiency, and, as soon as someone is familiar with it, he will put up with the more difficult operation.

The IDC documentation system over-all has reached a point suggesting its world-wide distribution. Real computer processes are necessary to cope with the documentation of the literature explosion. The efficiency of all available systems should be compared and the best ones extended.

LITERATURE CITED

(1) Rüssmann, K. H., *BP-Kurier, Vierteljahresschr. BP Benzin Petrol. AG* **XX**, 12–15 (11/68).

(2) Fugmann, R., *Proc. IUPAC Congr.* p. 331–41, 1959; *Nachr. Dok.* **12**, 69–76 (1961); "Classification Research," Munksgaard København, pp. 341–67, 1965; Internat. Study Conference on Classification Research, FID/CR, working paper No. 15, Elsinore, 1964. Fugmann, R., W. Braun, and W. Vaupel, *Angew. Chem.* **73**, 745–51 (1961); *Nachr. Dok.* **14**, 179–90 (1963); Rössler, S., in preparation.

(3) Meyer, E., Proc. 26th ADI Meeting, pp. 131–2, Chicago, 1963; *Angew. Chem.* **77**, 340–45 (1965); *Angew. Chem. Intern. Ed.* **4**, 347–52 (1965). Meyer, E., and K. Wenke, *Nachr. Dok.* **13**, 13–19 (1962).

(4) Meyer, E., in "Mechanized Information Storage, Retrieval and Dissemination," Proceedings of the FID/IFIP Joint Conference, Rome, June 1967, pp. 280–8 North-Holland Publishing Co., Amsterdam, 1968.

(5) Meyer, E., *Nachr. Dok.* **13**, 144–6 (1962).

(6) Wegmüller, F., R. Becher, B. Hoffmann, and H. R. Schenk, *Experimentia* **16**, 383 (1960).