place in which to start a CIE or a galaxy of them because it already has editors, indexers, surrogators (abstractors), and other personnel. It also has subscribers.

The success of one CIE should be used to promote CIEs in other subfields of a galaxy and in other galaxies. Anticipated benefits should attract support for CIEs from editors, employers, indexers, librarians, translators, professional searchers, professional societies, publishers, scientific societies, surrogators, translators, and users. Scientific and professional societies, through financial assistance from employers, could be sources of capital, initiative, management, and supervision of CIEs.

CONCLUDING REMARKS

Implementation, even if experimental, faces formidible obstacles including publishers; the NIH (Not Invented Here) syndrome that seems based on fear of disclosure of inadequate creativity, ingenuity, knowledge, and wisdom; risk timidity; and the comfortable paralysis of custom. However, anticipated benefits may prevail.

It now seems possible to improve existing information systems by creating CIEs and their galaxies—or reasonable facsimilies thereof.

This paper elaborates a prediction of the advent of CIEs.[8]

REFERENCES AND NOTES

(1) Bernier, C. L. "Ethics of Knowing". *J. Am. Soc. Inf. Sci.* **1985**, *36*, 211-212.
(2) Bernier, C. L.; Yerkey, A. N. *Cogent Communication. Overcoming Reading Overload*; Greenwood: Westport, CT, 1979; pp 31-38.
(3) Bernier, C. L.; Gill, W. N.; Hunt, R. G. "Measures of Excellence of Engineering and Science Departments: A Chemical Engineering Example". *Chem. Eng. Educ.* **1975**, *9*, 194-197.
(4) Bernier, C. L. Book review of P. Wilson's *Second Hand Knowledge: An Inquiry into Cognitive Authority*; Greenwood: Westport, CT, 1983. *J. Am. Soc. Inf. Sci.* **1984**, *35*, 255-256.
(5) Bernier, C. L. "Terse Literatures". *Encycl. Libr. Inf. Sci.* **1980**, *30*, 312-330.
(6) Myatt, D. O.; Upham, T. E. "A Quantitative Technique for Designing the Technical Information Center". *J. Chem. Doc.* **1961**, *1*(3), 18-29.
(7) Bernier, C. L. "Measurement of How Well Professional People Keep Up with Their Technical Literatures". *J. Am. Soc. Inf. Sci.* **1971**, *22*, 292-293.
(8) Bernier, C. L. "Development of Indexing and Indexes". *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 164-170.

Supply of Information on Chemical Reactions. An Advanced, Topology-Based Method

R. FUGMANN,* G. PLOSS, and J. H. WINTER
Hoechst A.G., Postfach 800320, 6230 Frankfurt am Main, Federal Republic of Germany

Received October 31, 1986

Several pitfalls and problems that the authors have encountered and—in part—overcome are described. The idea of the atom identification (AI) number, which has been in use in IDC (International Documentation in Chemistry) firms since 1960, has proved an efficient device for optimizing the precision and recall ratios for reaction searches and thus endows reaction information systems with high survival power. In particular, it makes it possible to find reactions in which the required (sub)structures of educts and products are separated from one another by unforeseeable intermediates. It also promises to overleap even the document boundaries through which these structures may be separated in the literature.

INTRODUCTION

In a large majority of the queries concerning chemical reactions, the inquirer has a specific substructure in mind, and he is seeking pathways that are described in the literature as leading to this objective from a different substructure, which should be contained in an educt. An efficient documentation system for chemical reactions must render it possible to formulate such queries without having to place restrictions on the type and number of possible intermediate steps. Nor should it be necessary for the inquirer to restrict himself to the retrieval of those publications in which only the *direct* conversion of the educt substructure into the product substructure is described. No inquirer can really predict which ingenious and useful indirect routes, leading from the educt structure in question to the product structure being sought, have already been described in the literature.

It is only with the help of a search process that disregards intermediate steps that all relevant documents can be retrieved with sufficient accuracy from a reaction file in which not only direct conversions from structure I to structure II are contained, as shown in Figure 1, but also indirect conversion processes are contained, which are also depicted in Figure 1.

Frequently the inquirer feels certain that the reaction being sought proceeds only in one step and can have been described in the literature only in this way, as, for example, the conversion of structure I to structure II in Figure 2. If, however, the capability is available of conducting searches that overleap intermediate steps, it is always surprising how many multistep pathways for the conversion process are encountered in the literature. An example of this is also shown in Figure 2.

An obvious way to conduct searches that leap the intermediate steps would appear to be that of requiring the co-occurrence of the educt and product structures in the same document. However, such generalized search parameters will lead to many responses that do not satisfy the search objective. For example, anyone requiring the mere co-occurrence of structure I from Figure 1 as the educt and structure II as the product will receive the three irrelevant documents with the contents presented in Figure 3.

Intermediates-overleaping searches have been performed at IDC[1] for many years with the help of the GREMAS system. Here in the process of indexing each reaction, certain criteria are used to determine which atoms and bonds should be considered to be involved in the reaction. These reaction sites are then especially characterized, and at the same time it is determined which carbon atoms in the educt and product "correspond" to each other in the sense that they only represent different reaction stages of one and the same atom. This correspondence is indicated by assigning to them the same, arbitrarily chosen "atom identifying (AI) number". A carbon atom retains this number through all the stages through which it passes in the document concerned. In Figure 1 the numbers entered beside the "reacting" carbon atoms are these AI numbers.

0095-2338/88/1628-0047$01.50/0

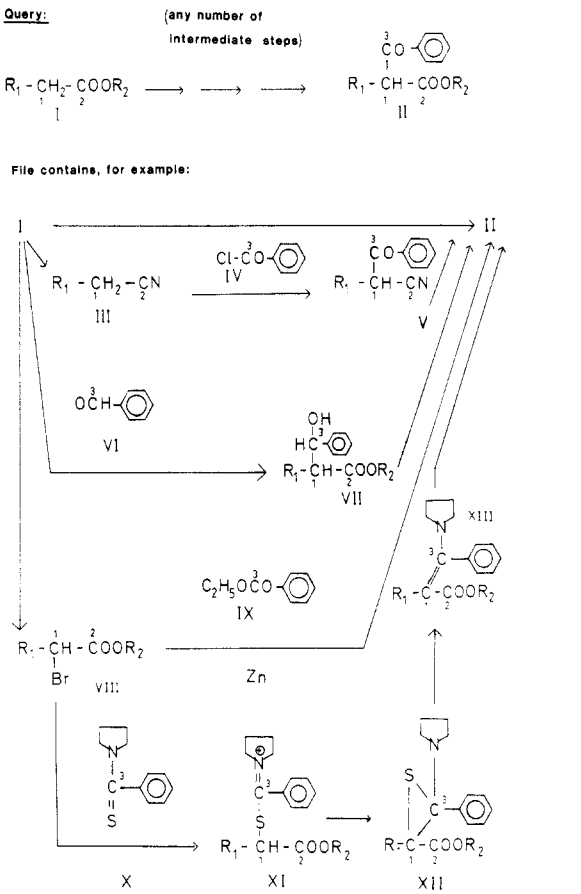


Figure 1. Reactions for the acylation of an acetic ester described in unpredictable steps.

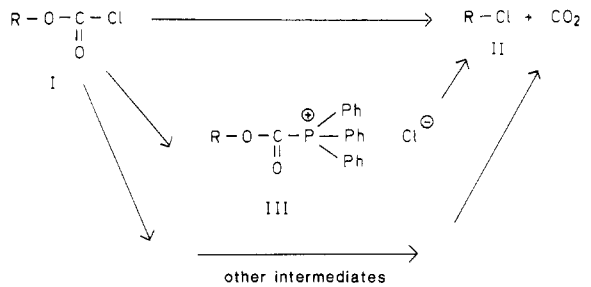


Figure 2. Reaction assumed to occur in one step but described in several steps.

When indexing has been performed in this manner, the query can be formulated with corresponding precision. For example, it can be required that the CH group on the radical R_1 of product II in Figure 1 be derived from the CH_2 group of the acetic acid ester I, and this condition is not satisfied by any of the noise cases in Figure 3. The further advantages of this type of indexing are described in ref 3.

In formulating queries for intermediate-overleaping searching it is easy to make a mistake that will lead to the loss of relevant information. In Figure 1 it would appear obvious to require for educt I a reaction at the CH_2 group (with AI number 1). Likewise, one might with equal confidence require for product II that the carbonyl group (with AI number 3) should have reacted.

These conditions, however, may not be fulfilled in a perfectly relevant multistep reaction. An example of this is the reaction sequence also shown in Figure 1 that proceeds in three steps via III and V. Here the atom with the AI number 1 does not react in educt I, and, again, when product II was formed, the reaction did not take place on the carbonyl group (AI number 3). If these conditions had been specified, this reaction would

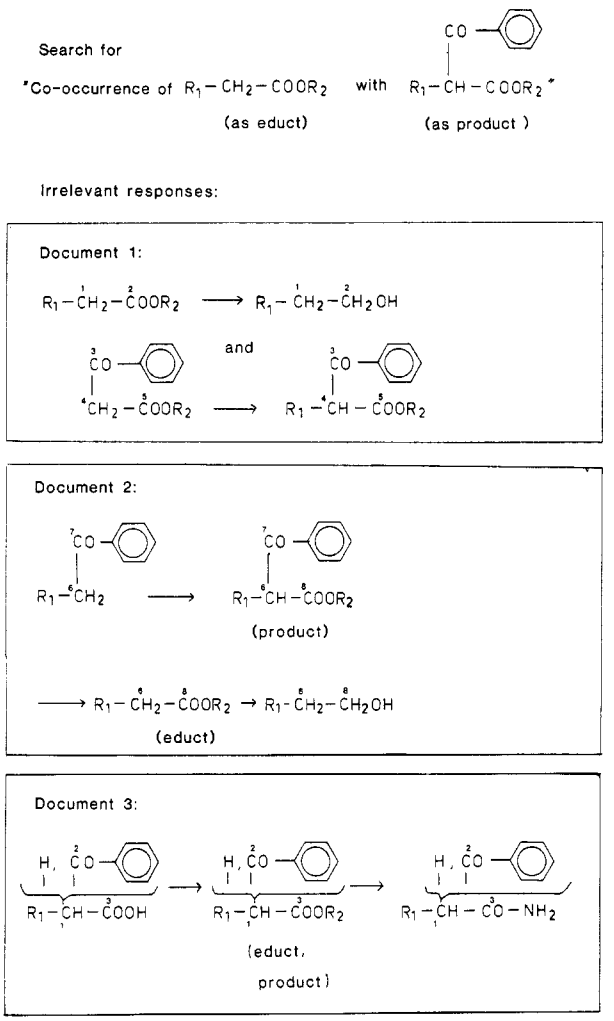


Figure 3. Irrelevant responses to a query merely requiring co-occurrence of educt and product.

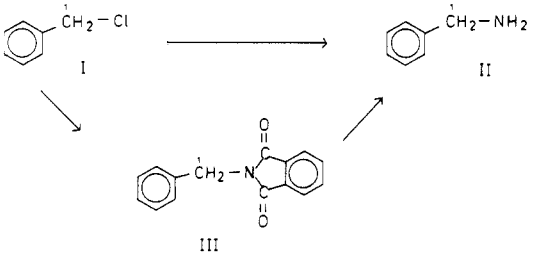


Figure 4. Danger of information loss when reaction is described in multiple steps.

not have turned up during retrieval.

Figure 4 presents a reaction in which only one of the two conditions is satisfied that one might be inclined to stipulate in a query without more careful consideration. Here the replacement of a chlorine atom by an amino group is required. While the carbon–chlorine bond is broken in "starting" educt I, the carbon–nitrogen bond is formed in intermediate step III and by no means in final product II. Again this reaction would not have been found if the bond changes had been required in both starting educt I and final product II.

Thus, the requirement that a reaction site in a reaction step be active (i.e., undergo a change in one of its bond relationships) *presupposes knowledge of the product that directly succeeds it.* If this product is not known or if any commitment with respect to the nature of this product is to be avoided, then one must not be forced to specify a particular atom in the educt as a reaction site. Otherwise, one would, though only in a concealed way, also undesirably predefine the kind of reaction

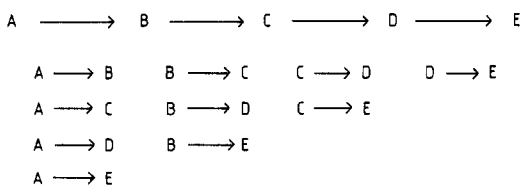


Figure 5. Attempt to represent a multistep reaction by several one-step reactions.

that the educt under consideration is to undergo. The same applies analogously to the product. If the reaction site in the product is specified, a presupposition is then also imposed on the nature of the immediately preceding educt. It is precisely such presuppositions that must be avoided in intermediates-overleaping searches.

Hence, intermediates-overleaping searches require a kind of indexing, by means of which a reaction site can be found at *any reaction stage* between a starting educt and an end product. This should be possible regardless of whether this reaction site was active by undergoing a bond change in one of these structures and also regardless of the molecular environment of the reaction site in one of the intermediates.

The query of Figure 1 should accordingly be formulated as follows: (1) Educt I and product II must co-occur in the same document. (This question will be discussed in more detail later.) (2) The CH_2 group in educt I must have the same AI number as the CH group in formula II (in this case the AI number 1). (3) Any document retrieved in response to the query must describe a pathway *from* educt I *to* product II (and not, for example, in the reverse direction). (4) At *some* stage on the way from I to II (in Figure 1) the atoms with AI numbers 1 and 3 in Figure 1 must have changed their bonds. Only this condition can exclude, for example, document 3 in Figure 3.

Note that no particular molecule can be specified as the one in which the carbon atom with AI number 3 becomes linked on the way from educt I to product II. Nor is it permissible to specify that this carbon atom must appear as a keto group directly after being linked on. Otherwise, reaction pathways would be excluded in which instead of the acid chloride an aldehyde is used that intermediately yields a carbinol (cf. Figure 1, VII) that then yields the keto group only in a later step.

It would hardly be satisfactory to solve the problem by merely breaking down a multistep reaction into all the individual steps with which one would like to be able to retrieve the reaction sequence at a later time, as for instance according to the scheme shown in Figure 5. Such an approach has occasionally been pursued. The storage capacity and machine time required for searches involving long and highly ramified synthesis pathways would probably be prohibitive.

We have also investigated the question of how an inquirer can retrieve a synthesis pathway that has been described in multiple steps and in separate publications. Below we describe a method that promises to solve this problem.

Our paper presents experience compiled during more than 25 years of reaction indexing and is intended as a suggestion to the design of more effective and precise documentation methods in this field.

INDEXING WITH AI NUMBERS

A recommendable variant of working with AI numbers consists of expanding the topological connection table for structural formulas by one column, in which for each atom of a molecule the AI number arbitrarily assigned to this atom is also recorded. The AI numbers are thus assigned to the atoms in a manner similar to that carried out for charges or abnormal mass numbers, for example.

Consider the multistep reaction sequence of Figure 6. Here, for each atom that undergoes a bond change a "reaction vector" could be used to indicate the type of bond change and in which structural formula (expressed by the document internal formula number) this bond change occurs (cf. Figure 6). The information in the reaction vectors would be sufficient to generate from all educts the connection tables for all subsequent products. In this process the AI numbers of the educts are simply retained. For example, in the reaction sequence of Figure 6, in order to derive the connection table for product III, which is formed from starting educts I and II, it would be sufficient to add together the connection tables of both structural formulas and, further, to express the transition of both single bonds of atom 15 to the double bond between the atoms with AI numbers 1 and 15. In this way a connection table would be generated for structural formula III with all AI numbers originating from the educts. No new AI numbers would have to be assigned to product III.

The Educt–Product Syntax. The reaction vectors of Figure 6 also contain comprehensive information on how the various educts, intermediates, and products of Figure 6 are syntactically linked with each other. A lucid and self-explanatory graphical representation of these linkages is given in Figure 7. It is, for example, obvious that compound VII is accessible from educts VI and III but not from educt VIII.

Joining Separately Published Reaction Steps. With conventional documentation methods it is difficult to find multistep pathways that have been torn apart by having been published in separate parts in different articles. No inquirer can predict at which point the multistep synthesis pathway in Figure 6, for example, might be broken and divided between different documents. Consequently, one can never be certain which educts I, III, or even VII can be required to co-occur with product IX. If the reaction sequence being sought is described only in separate steps in the literature and if the separation in two different publications has occurred at the isocyanate intermediate III, then the requirement for the co-occurrence of amine (I in document A, Figure 8) and chlorinated urea (V in document B, Figure 8) in the same document cannot be met.

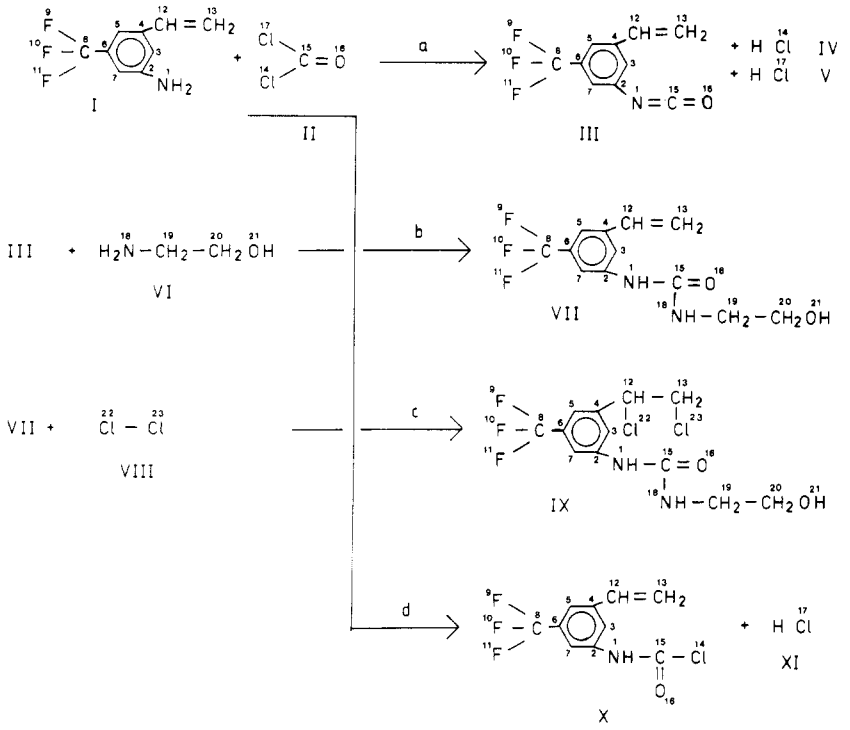
The concept of the AI number opens up the possibility of accurately locating reaction sequences that cross document boundaries. The inquirer would not be forced to make assumptions about where and how often a synthesis pathway might have been interrupted and, hence, which particular intermediates can be required to co-occur in a document. He would also not be forced to make the assumptions mentioned above concerning the type of reaction that a precursor should undergo in its first reaction step. The procedure as described in the following promises to achieve that goal.

We consider Figure 8 and set ourselves the task of finding a reaction sequence that starts from aromatic amine I in document A and leads to chlorinated urea V of document B.

The first step in this procedure would be to make certain that the aromatic amine I appears as an educt in the literature file and also that the chlorinated urea V is stored as a product but not necessarily in the same document. Only under these conditions is it reasonable to undertake a search for the reaction sequence in question.

Thus, V may have been found in document B, and the educt–product syntax (cf. Figure 9) would show two educts for product V, namely, elemental chlorine (IV) and olefinic urea III.

For the machine program it would be obvious that the path back to the starting educt (amine I) cannot lead via elemental chlorine as an intermediate, for there is not a single chlorine atom contained in this starting educt. Thus, educt IV can be disregarded in the search for the desired reaction sequence.



Reaction number	AI-number of the atom where bonds change	contained in educt number	Breaks ... bond(s)	to atom with AI-number	Makes ... bond(s)	to atom with AI-number	of educt number	With formation of product number
a	1	I	two single	—	one double	15	II	III
	15	II	two single	14,17	one double	1	I	III
	14	II	one single	15	one single	—	I	IV
	17	II	one single	15	one single	—	I	V
b	1	III	one single	15	one single	—	VI	VII
	15	III	one single	1	one single	18	VI	VII
	18	VI	one single	—	one single	15	III	VII
c	22	VIII	one single	23	one single	12	VII	IX
	23	VIII	one single	22	one single	13	VII	IX
c	12	VII	one single	13	one single	22	VIII	IX
	13	VII	one single	12	one single	23	VIII	IX
d	1	I	one single	—	one single	15	II	X
	17	II	one single	15	one single	—	I	XI
d	15	II	one single	17	one single	1	I	X

Figure 6. (Top) Multistep branched reaction sequence with AI numbers at the reaction sites. (Bottom) Multistep branched reaction sequence with reaction vectors.

This is not true, however, for isocyanate III in document B, which will be discussed in more detail below.

Another criterion for deciding whether the reaction pathway being sought could proceed via a particular educt could consist of the requirement that this educt must introduce at least one carbon atom into the product. In this way, for example, one could exclude benzoyl peracid as an educt for an epoxide.

Educt III has in turn two educts in document B, namely, I and II, without either one of them being identical with starting educt I in Figure 6A that is being sought (or with I in document A of Figure 8). Consequently, the reaction pathway being sought has not yet been found. Now all educts that have not been excluded for the reasons given above must be collected in an "educt collective file". For document B this would hold for educts I and II.

An analogous search procedure would be initiated from starting educt I in document A (and also in all other documents in which A has been described as an educt). I in document A of Figure 8 has successors, as can also be seen from the educt–product syntax of Figure 9, namely, the products III, IV, and V. Again, the two successors IV and V could be disregarded as intermediates on the way to V in document B,

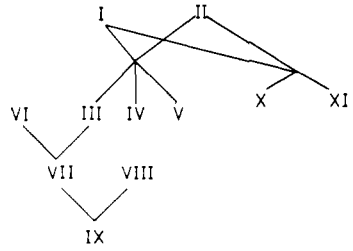


Figure 7. Educt–product syntax of reaction sequence in Figure 6.

for they do not contain a single carbon atom. No other direct or indirect successor of I is described in document A. Since the successor III fails to be identical with the target product V, again III (and generally all successors of an educt that are not identical with the target product in question) must be collected in another file (which we call the "product collective file").

The machine program must now begin a search for structural formulas that appear in both collective files. If such a common subset of structural formulas is found, such as formula III in document A and formula I in document B of Figure 8, then it could represent a link between the two publications in which a reaction sequence being sought is described in separate parts.

Final certainty concerning this question can only be provided by a comparison of the AI numbers in starting educt I of document A with those of V in document B. In particular, the nitrogen atom with AI number 1 in document A must correspond to the nitrogen atom with AI number 30 in document B in the sense described above. It would, however, be purely coincidental if in both documents, which had certainly been indexed entirely independently from one another, the corresponding atoms had been assigned the same AI numbers, since in each document these numbers were assigned in a completely arbitrary way.

The task is now to determine whether the nitrogen atom with AI number 1 in document A really does correspond to the nitrogen atom with AI number 30 in document B. If this can be confirmed by the machine program, then the link being sought to connect the two documents A and B has been found.

At this point let us recall to mind the Morgan numeration of atoms in a structural formula and an essential characteristic of these Morgan numbers: each and every atom in a structural formula is always assigned one particular Morgan number in a unique and unambiguous manner regardless of the context in which such a structural formula might appear in another place.[2] Thus, the nitrogen atom with AI number 1 in document A, formula III, is always assigned the Morgan number (M) 11, *and exactly the same Morgan number* will also be assigned to this nitrogen atom in formula I of document B, where it might have been assigned AI number 30. In order to build a bridge for the AI number comparison of documents A and B, one would determine for educt I in document B which Morgan number there applies for AI number 30 for the nitrogen atom. The applicable Morgan number in this case, M 11, would be looked up in the connection table of III in document A, from which the "new" AI number 1 for this atom could be read. It is thereby determined *that AI number 30 in document B corresponds to AI number 1 in document A.* Thus, for the remaining search with AI numbers in document A, AI number 30 must be replaced by AI number 1.

Expressed in another way, the amine nitrogen atom in I of document A must bear AI number 1 if it is to correspond to the amide nitrogen atom of V with AI number 30 in document B. Only if this condition is fulfilled can it be established that I in document A is a genuine educt of V in document B in accordance with the reaction sequence being sought.

In principle, the number of documents among which the individual steps of a multistep synthesis pathway are divided would play no part in the search for a synthesis pathway. A structural formula that appears both in the collective file of products in synthetic progression and in the collective file of educts in synthetic retrogression always marks the synthesis pathway from the starting educt to the end product. Determining the syntactic position of such a connective link in the product–educt syntax reveals the path that leads via this link from the starting educt of the final product.

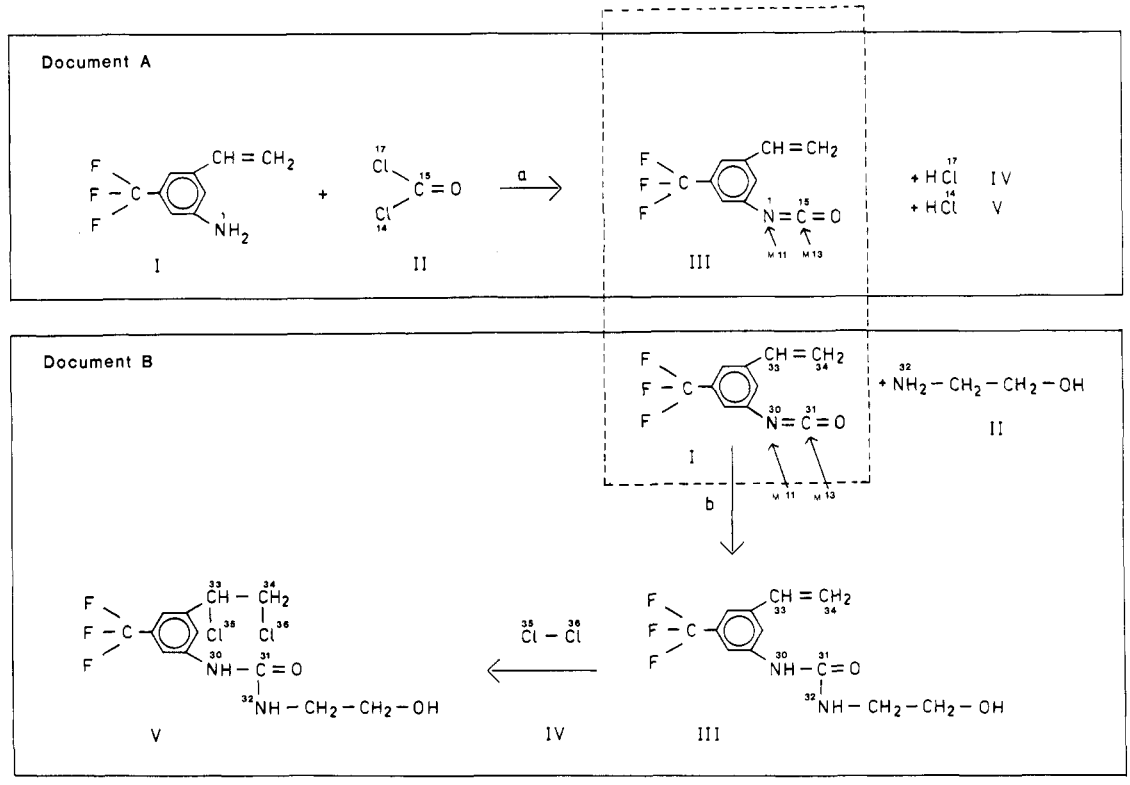


Figure 8. Interruption of a reaction sequence by its being continued in another document.

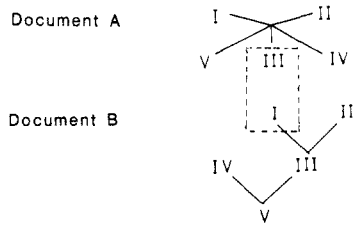


Figure 9. Educt–product syntax of the reaction sequence in Figure 8.

DISCUSSION

Work with AI numbers has been carried out on a large scale with the GREMAS system for reaction documentation for 26 years.[3] It has hitherto been restricted, however, to carbon atoms that participate in the reaction. Experience with tens of thousands of searches has shown that for large files that are very actively searched it is essential to be able to recognize corresponding atoms and not to be forced to specify assumed intermediate steps while searching or even to exclude the possibility of any intermediate steps between the educt and product structures in question. This may well occur, if only in a hidden manner, if individual reaction centers in the educt and/or product structure are specified in the query or if a one-step reaction from the educt to the product is demanded.

With the procedure described here we have set ourselves the objective of being able to trace and relate to one another all transformations that a particular reaction site undergoes when passing through a sequence of reactions. As we have seen, this requires that each atom that actively participates in a reaction be reliably recognizable in *all* educts, intermediates, and products. This is accomplished by identifying all atoms with AI numbers.

No contemporary operational reaction documentation system is encountered in which unambiguous document-internal atom identification is carried out. One is often contented with merely recording for each individual step which bonds are broken and made anew, for example, the breaking of a carbon–hydrogen bond in the methylene group of ester I in Figure 1 and the making of a carbon–carbon bond in the direct transition from I to II. However, if then in a later query a reaction on such a methylene group is required, and in the final analysis this methylene group is to be linked to a benzoyl radical, such a query can only be formulated for *any* methylene group in an ester of type I in general, i.e., only in an undesirably generalized form. A molecule in question may have several of these methylene groups, especially such groups as do not correspond at all to the methylene group in the original acetic acid ester.

It does not mean that a solution has been found if, to better identify this methylene group, its exact position in the overall molecule is made one of the search parameters. This is because this embedment may have changed in an unpredictable way before the establishment of the bond with the benzoyl radical. Nor can one require of the added carbon atom that it should already exhibit the acyl structure immediately after this reaction, as can also be seen in examples VII and XI of Figure 1. The noise that arises through such inexactness in the formulation of the query, the latter being imposed by the inexactness of indexing, can easily become prohibitively voluminous as the file continues to grow.

In regard to the correspondence relationships that prevail between the atoms in a sequence of reaction steps, they are occasionally difficult to recognize. Often a chemist must enlist every bit of his/her knowledge of synthetic and theoretical chemistry for this task, and now and then a solution can be found only through an added study of the literature. This is especially true of many rearrangement reactions.

We have not yet encountered a procedure for which we can verify that it can accomplish this task satisfactorily by following a purely algorithmic, programmed route. In particular, the reliable recognition of reaction sites in educts and products is still an unsolved problem if the entire task is to be turned over to a computer program.[4,5]

On occasion a reaction even fully resists all attempts to analyze the details of its course. As an example, in ester hydrolysis it is not always clear whether the oxygen atom of water is subsequently located in the carboxylic acid or in the alcohol. In such cases the remaining uncertainty could be expressed and rendered harmless to the search by assigning "*" as the AI number to any atom whose origin is unclear. As a precaution the program would equate the asterisk with each AI number that has already been used in the document concerned. At the most, noise could arise, but a loss of information would be avoided.

The idea of recording the correspondence relationships of *all* atoms in the compounds taking part in a reaction regardless of whether these atoms participate (however this is defined) in some way was discussed early within the circle of IDC companies.[6] It was not pursued further at that time, however, for extrinsic reasons, but now technological progress has removed the obstacles existing then.

Working with AI numbers and the storage of a product–educt syntax is rather expensive, as experience[3] has shown. On the other hand, the method described here also brings savings in comparison with conventional techniques because all products arising out of the educts could be algorithmically generated with the help of the reaction vectors.

If, however, indexing is not carried out with the high precision proposed here, searches in the file are more expensive, and this expense increases steadily, especially once the use of the reaction file is in full swing. If several of the measures recommended here are omitted at the same time in an information system, the survival of such a project is jeopardized because the constantly increasing volumes of noise will soon no longer be manageable within an acceptable length of time. Then it becomes obvious that the savings made in a more superficial indexing and storage have actually had very costly consequences, namely, the gradually developing worthlessness of all the work hitherto devoted to the project.

Special caution should be exercised in regard to those saving measures that can lead to a loss of information. Such a system may well perform a useful service for information seekers who do not have to insist on complete retrieval of stored information, but for a file of a company's internal reports or of patented reactions, great value must be placed on the completeness of retrieval. Gaps of this kind can lead an information system for chemical reactions to the verge of total worthlessness.

This must be borne in mind when, for example, one omits the consistent and precise storage and labeling of all educts, intermediates, and products, and, in particular, the consistent labeling of all atoms that are involved in a reaction.

Where less exacting requirements in respect to the accuracy and completeness of retrieval are imposed, simplified variants can be justified. Our objective in this paper has been to depict a technique that satisfies stringent requirements, to suggest the materialization of an approach like this, and at the same time to show the specific consequences of savings effected on the indexing and storage side. Whether these consequences can be tolerated in the specific situation in which an information system is now and will find itself in the foreseeable future must be decided separately for each individual case.

ACKNOWLEDGMENT

Substantial contributions to the development of this to-

pology-based approach were made by Drs. U. Doelling and I. Wilhelm and M. Jaeger.

REFERENCES AND NOTES

(1) International Documentation in Chemistry, Otto Volger Strasse 19, 6231 Sulzbach/Taunus, Federal Republic of Germany.
(2) Morgan, H. L. "The Generation of a Unique Machine Description for Chemical Structures—a Technique Developed at Chemical Abstracts Service". *J. Chem. Doc.* **1965**, *5*, 107-113.
(3) Fugmann, R.; Kusemann, G.; Winter, J. H. "The Supply of Information on Chemical Reactions in the IDC System". *Inf. Process. Manage.* **1979**, *15*, 303-323. Fugmann, R. "The IDC System". In *Chemical Information Systems*; Ash, J. E., Hyde, E., Eds.; Ellis Horwood: Chichester, U.K., 1975; p 195.
(4) See also: Valls, J.; Schier, O. "Chemical Reaction Indexing". In *Chemical Information Systems*; Ash, J. E., Hyde, E., Eds.; Ellis Horwood: Chichester, U.K., 1975; p 255. Valls, J. "Reaction Documentation". In *Computer Representation and Manipulation of Chemical Information*; Wipke, W. T., Heller, St. R., Hyde, E., Eds.; Wiley: New York, 1973; p 92.
(5) See also: McGregor, J. J.; Willett, P. "Use of a Maximal Common Subgraph Algorithm in the Automatic Identification of the Ostensible Bond Changes Occurring in Chemical Reactions". *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 137-140. (Here, the algorithmic identification of reaction sites is considered "sufficiently precise".)
(6) Vogt, F. BASF Report, 1968, unpublished.

Data Access Subroutine Package for Spectrometric Data Bases

CHENG QIAN[‡] and CHARLES L. WILKINS*

Department of Chemistry, University of California, Riverside, California 92521

Received April 14, 1987

A data access subroutine package for spectrometric data bases has been developed. The subroutines are high-level language callable. They can transform logical records of a data file to virtual records consisting of selected data fields with user-specified data format for further processing or, conversely, fill a record of a data file with data transformed from a virtual record. By careful balance of its functions and the overheads, the package has been made very compact, resulting in low development cost and reasonable execution efficiency. Application examples are given to show its versatility and usefulness.

INTRODUCTION

During the past decade, a number of multispectrometric data-base systems have been implemented. Some examples are CIS of NIH/EPA,[1,2] Pluridata in the DARC system,[3] and a retrieval system developed by Koptyung and co-workers.[4] These systems integrate large spectrometric data bases of tens of thousands of spectra as well as related structural data bases. They can retrieve the data in an interactive mode by using their built-in search and retrieval software. The software is very user-friendly. However, this software works in a "closed" environment where users can search data using only procedures dictated by the software and retrieve the data only in a predetermined format. Molecular Design Ltd. has developed a more flexible system, called MACCS, that can be used for structural and associated textual databases[5] but is still end user oriented, although it, too, is a closed environment because of its proprietary nature.

In some research laboratories, users work in an "open" environment, where they may actually be developers of new library search algorithms or other application programs in which data retrieval is only the first step for further program operations. Therefore, more flexible data-base software is needed. Such software should function as an interface between user application programs and data bases, keeping both independent from each other. This is, in fact, a part of the goals of various data base management systems (DBMS).[6] Using a relational DMBS, Morffew and others have handled molecular geometry data in a molecular graphics system.[7] However, there is still a lack of technically and economically suitable commercial DBMS for large multispectrometric and structural data-base management. Most commercial DBMS are based upon algorithms that combine a set of individual simple keyword searches by Boolean operations. They can hardly support spectrometric and structural data bases, owing to the special characteristics of queries required for these data bases. Table I lists various types of queries and typical search methods, showing that present DBMS support the queries listed in columns 1 and 3. In contrast, the queries raised frequently with respect to spectrometric and structural data bases are those in columns 4 and 5, sometimes simplified to those listed in column 2 for presearch purposes. Thus, the DBMS are of limited utility for spectral library searches. Moreover, because they usually require much CPU time, they can hardly meet the time-response criteria necessary to make possible searching of large spectral and structural libraries on supermini computers that are popular in chemistry laboratories. In other words, the use of commercially available DBMS for spectrometric and structural data bases is generally expensive and ineffective, if not impossible. Therefore, we have developed a flexible software tool that provides data access facilities to user programs and includes the potential for future expansion of searching features. This is the idea behind the data access subroutine package (DASP) described here.

DASP's approach is to provide a set of compact and efficient application program callable subroutines for convenient data access while maintaining data-program independence and allowing search algorithms of all varieties to be realized in the user's programs. Because variable-length fields of binary data are used extensively in spectrometric and structural data-base organization, equal attention has been paid to data processing of fixed-length as well as variable-length fields. Furthermore, DASP is carefully designed to fit the needs of integrating and reorganizing multisource data bases, such as those that are required to support integrated gas chromatography–infrared mass spectrometry systems.[8]

PRINCIPAL FEATURES OF DASP

Data Model. Spectrometric or structural data often aggregate as units of "spectra" or "structures". Data from different spectra or structures are seldom related, in terms of either data organization or application. On the other hand,

[‡]Present address: Chemical Abstracts Service, P.O. Box 3012, Columbus, Ohio 43210.

0095-2338/88/1628-0053$01.50/0