(18) Hunter, D. G. N.; McKenzie, H. R.; "Experiments with Relaxation Algorithms for Breaking Simple Substitution Ciphers". *Comput. J.* **1983**, *26* (1).
(19) Schubert, W.; Ugi, I. "Constitutional Chemistry and Unique Descriptors of Molecules". *J. Am. Chem. Soc.* **1978**, *100*, 37-41.
(20) Morgan, H. L. "The Generation of a Unique Machine Description for Chemical Structures—A Technique Developed at Chemical Abstracts Service". *J. Chem. Doc.* **1965**, 107-113.
(21) Lynch, M. F.; Willett, P. "The Automatic Detection of Chemical Reaction Sites". *J. Chem. Inf. Comput. Sci.* **1978**, *18*, 154-159.
(22) Freeland, R. G.; Funk, S. A.; O'Korn, L. J. Wilson, G. A. "The Chemical Abstracts Service Chemical Registry System. 2. Augmented Connectivity Molecular Formulae". *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 94-98.
(23) Wipke, W. T.; Dyott, T. M. "Stereochemically Unique Naming Algorithm". *J. Am. Chem. Soc.* **1974**, *96*, 4834-4842.
(24) Kitchen, L.; Krishnamurthy, E. V. "Fast, Parallel Relaxation Screening for Chemical Patent Data-Base Search". *J. Chem. Inf. Comput. Sci.* **1982**, *22*, 44-48.
(25) Kitchen, L.; Rosenfeld, A. "Discrete Relaxation for Matching Relational Structures". *IEEE Trans. Syst. Man. Cybern.* **1979**, *SMC-9*, 869-874.
(26) Figueras, J. "Substructure Search by Set reduction". *J. Chem. Doc.* **1972**, *12* (4), 237-244.
(27) Sussenguth, E. H., Jr. "A Graph-Theoretical Algorithm for Matching Chemical Structures". *J. Chem. Doc.* **1965**, *5*, 36-43.
(28) Welford, S. M.; Lynch, M. E.; Barnard, J. M. "Computer Storage and Retrieval of Generic Chemical Structures in Patents. 3. Chemical Grammars and their Role in the Manipulation of Chemical Structures". *J. Chem. Inf. Comput. Sci.* **1982**, *21*, 161-168.
(29) Cheng, J. K.; Huang, T. S. "A Subgraph Isomorphism Algorithm Using Resolution". *Pattern Recog.* **1981**, *13* (5), 371-379.

DARC-SYNOPSYS. Designing Specific Reaction Data Banks: Application to KETO-REACT

R. PICCHIOTTINO,* G. GEORGOULIS, G. SICOURI, A. PANAYE, and J. E. DUBOIS*

Association pour la Recherche et le Developpement en Informatique Chimique, 25 rue Jussieu, 75005 Paris, France, and Institut de Topologie et de Dynamique des Systemes, associé au CNRS, Université Paris 7, 75005 Paris, France

Received February 3, 1984

On the basis of the Entity/Relationship approach, specific reaction data banks have been designed by modeling data in a logical scheme and by proposing a compatible physical scheme that takes access optimizations into account. This architecture determines the general organization of the IGRES-RECRE acquisition and retrieval software incorporating original computer validation procedures that improve the quality, exactitude, and coherence of reaction data and thereby ensure bank reliability. The RECRE software interactivity assists retrieval by letting the user break down a question into formalized elementary requests (DARC structure and substructure searching, nonstructural searching, logical operations, output) in a quasi-natural language. This methodology is illustrated on the KETO-REACT data bank in the framework of the DARC-SYNOPSYS expert system.

INTRODUCTION

Tools for computer-aided design in chemistry, developed within the last 10 years,[1,2] contribute to the progress of expert systems in artifical intelligence.[3-6] Many systems for the computer-aided organic synthesis design of a given molecule have been built over this period.[7]

According to Gund,[8] the design of a synthesis involves an overall approach, which is the planning stage, and a local approach by which the experimental description of a reaction becomes accessible. Today's computer-aided synthesis systems[7] use both approaches simultaneously, at the expense of the retrieval of detailed reaction data. Actually, though such systems are built from an in-depth literature analysis,[9] there is currently a real need[8] for tried and proven procedures to (*i*) *constitute specific reaction data banks and* (*ii*) *access their data*. The specific reaction data banks implemented in the DARC-SYNOPSYS expert system[6,10-14] have been designed to meet this need.

Unlike chemical compounds,[15] which are generally described by "hard data",[16] reactions are mostly described by "soft data",[16] so they are more difficult to define and organize. Both types of data serve to define the field of a reaction that can be organized conveniently on the basis of its structural data.[14]

Our approach to reaction modeling consists in identifying this set of data and in conducting a logical analysis of these data and of the relationships between them in order to attain a clear description of the various functions (input, validation, search, output) of a specific reaction data bank. This results in the proposal of a data model irrespective of its subsequent use, e.g., documentary search or computer-aided synthesis. The logical and physical schemes resulting from this data model express the architecture of the specific reaction data banks that is intended for straightforward integration into a chemical Data Base Management System (DBMS). This methodology is illustrated hereafter on the KETO-REACT data bank in the framework of the DARC-SYNOPSYS expert system.[6]

ORIGIN AND NATURE OF DATA

Bank Coverage. The coverage of a reaction data bank is expressed by its *scope* and *exhaustivity*, which should be defined precisely. The DARC-SYNOPSYS data banks are compiled from existing organic chemistry periodicals or compendiums,[17,18] whereas other systems rely mostly on compendiums. Although the exhaustivity of a compilation of reactions of general interest has been recognized as an illusory goal,[19] this remains conceivable in a limited and strictly defined area.

The methodology we proposed in 1969 for selecting reaction data for the preparation of families of compounds deals with the two above-mentioned factors.[20] In applying this methodology to the synthesis of ketones, a bibliography made it possible in a first stage to select from the literature those articles mentioning at least one method for preparing aliphatic and acyclic ketones having a first or immediate environment limited to B, designated hereafter as *first EB ketones* (Figure 1). Such an approach provides an exhaustivity criterion that

0095-2338/84/1624-0241$01.50/0

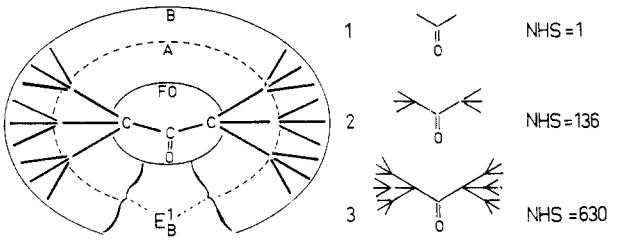


Figure 1. First EB ketones. Each compound in this family is characterized by a focus (FO) and a saturated carbon-bearing environment limited to no more than two layers (A and B) around the focus, i.e., the first Environment is limited to B (first EB). For example, for compound **1**, layers A and B are empty; for compound **2**, layer A is full of carbon atoms; for compound **3**, layers A and B are full of carbon atoms. The population with the thus-defined 630 ketones is organized into a HyperStructure[45] in which each compound is unambiguously identified by a number called NHS between 1 and 630.

materializes as a tally of the specific reactions yielding first EB ketones,[20] and lets the principal *methods* (set of predefined structural changes) for preparing this type of product be identified (Table I).

The bibliographic references constituting the *documentary fund* for the KETO-REACT bank are obtained in two different ways, depending on the year of publication. For the period prior to and including 1968, the 88 references cited in our methodology proposal paper[20] are used; for the period from 1969 up to and including 1975, bibliographical references are gleaned manually from *Chemical Abstracts*. First EB ketones are initially sought in the *Formula Index* (search for $C_nH_{2n}O$ molecular formulas with n between 3 and 27). For data obtained from the *Formula Index*, references are either retained or discarded after the articles have been analyzed.

The KETO-REACT documentary fund contains 136 references corresponding to 2788 specific reactions[21] extracted and classified according to principles described hereafter. An update of this fund with the help of EURECAS[15] via a DARC substructure search and a textual search of CAS files is currently under way.

Classification of Reaction Data. The wide variety of reaction data and their presentation in the literature complicate the description of a reaction. The different criteria until now proposed as a basis for a reaction definition are (i) bond cleavage and/or formation,[22-25] (ii) valency electron redistribution,[26] (iii) electronic process sequences,[27,28] and (iv) energy variations.[29,30] Given this lack of consensus on a definition and on a standard representation of a reaction, we have chosen to divide reaction data into four groups: main experimental data concerning reaction equations (structural diagram, curtailed experimental conditions, and bibliographical references) leading to the preparation of products identified by an experimental result, e.g., yield or physicochemical data; experiment interpretations, e.g., hypothetical reactions and reaction mechanisms, possible extensions of the reaction scope, reaction analogies, or data clustering; experimental data with no direct bearing on a reaction—product spectra, kinetic information, etc.; details regarding experimental conditions—detailed experimental protocol, product brand names, etc.

In the version of KETO-REACT described herein, only the first group of data, called reaction-related information, is considered. All the reactions mentioned in an article and consistent with this definition will be retained. This extends the coverage of the bank, initially limited to the preparation of first EB ketones, to the preparation of intermediate products involved in the synthesis of these first EB ketones. For handling purposes, reaction-related information is divided into the following types: bibliographical references, structural diagram, experimental conditions, method for preparing each ketone end product (Table I), and yield.

Table I. List of Methods[a] for Preparing First EB Ketones: Predefined Structural Changes Included in KETO-REACT

- (A) reactions with unchanged carbon skeleton
  - (1) catalytic hydrogenation of unsaturated ketones
  - (2) secondary alcohol oxidation
  - (3) secondary alcohol dehydrogenation
  - (4) starting from 1,2-glycol monoethers
  - (5) acetylenic compounds and analogous hydration
  - (6) starting from *gem*-dihalides
  - (7) starting from *vic*-dihalides
  - (8) starting from olefins
- (B) alkylation reactions
  - (1) ketone alkylation
  - (2) 1,4-addition on ethylenic ketones
  - (3) bromoketone alkylation
  - (4) starting from sulfur derivatives
  - (5) starting from ketenimines
- (C) rearrangements
  - (1) 1,2-glycol rearrangment
  - (2) aldehyde and ketone rearrangment
  - (3) allylic alcohol rearrangement
  - (4) epoxide rearrangement
  - (5) 1,3-glycol rearrangement
  - (6) alkene oxidation
- (D) reactions between acid derivatives and organometallics
  - (i) organometallic compound type
    - (1) organozinc compound
    - (2) organocadmium compound
    - (3) organomagnesium compound
    - (4) organolithium compound
    - (5) other organometallic compounds or mixtures
    - (6) organocopper and complex compounds
  - (ii) acid derivative types
    - (1) acid chloride
    - (2) anhydride
    - (3) ester
    - (4) nitrile or amide
    - (5) acid salt
    - (6) ketone
    - (7) carbon dioxide
    - (8) ethyl carbonate or carbon monoxide
    - (9) oxazoline
    - (10) oxazine
- (E) miscellaneous methods
  - (1) with diazomethane or ethane
  - (2) thermic decarboxylation of acid (salt or anhydride)
  - (3) oxo process
  - (4) other methods
  - (5) starting from two acid derivatives
  - (6) reaction with boranes
- (F) degradation methods
  - (1) oxidation or ozonolysis of ethylenic compounds
  - (2) cleavage of β-ketoeter
  - (3) cleavage of β-diketone
  - (4) malonic acid decarboxylations
  - (5) Darzens method
  - (6) miscellaneous degradations

[a] See footnote 20.

BANK ARCHITECTURE

The architecture of a specific reaction data bank is depicted by a model, the schematic description of which must be conducted at three levels:[31] logical, external, and physical. The *logical scheme* describes the meaning of the data and the logical ties uniting them. It expresses abstractly the choices in representation made in designing the bank. An *external scheme* (not discussed here because it can be deduced from the logical scheme) describes the user's view of the bank. The *physical scheme*, in keeping with the logical scheme, describes the concrete realization of the bank and reveals the implementation choices.

Logical Scheme. The *Entity/Relationship (E/R) approach*, the principles and vocabulary of which are recalled hereafter, serves to describe the logical scheme of specific reaction data

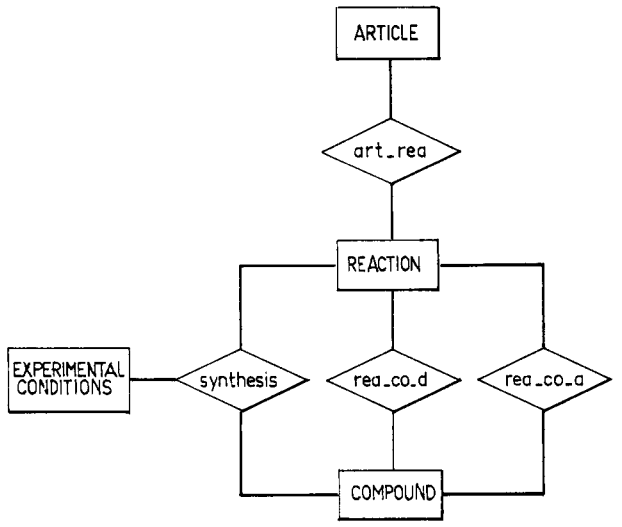


Figure 2. Logical scheme of a specific reaction data bank. A key is associated with each entity. For the entities (rectangles) article, reaction, compound, and experimental conditions, the associated keys are NUBIB#, NREAC#, NCO#, and NUMO#, respectively. Each entity is described by a key and the list of its attributes (Figures 4–7): article (NUBIB#, origin, author list, title, coden, volume, page, year); reaction (NREAC#, comments); compound (NCO#, topological description); experimental conditions (NUMO#, solvents, catalysts, reactants, temperature range, keywords, comments). In the "article" entity, the author list attribute is considered as an item, so as not to complicate a minor aspect of the logical scheme. The relationships (diamonds) are art_rea, i.e., the article (NUBIB#) contains the reaction (NREAC#), rea_co_d, i.e., the reaction (NREAC#) has the compound (NCO#) as starting compound, rea_co_a, i.e., the reaction (NREAC#) yields the compound (NCO#) (the method is the attribute of this relationship), and synthesis, i.e., the reaction (NREAC#) yields the compound (NCO#) via experimental conditions (NUMO#) (the yield is the attribute of this relationship).

banks. As initially proposed by Chen,[32] an E/R model[33,34] includes a set of entities characterized by a set of attributes as well as by a set of relationships between these entities. An *attribute*[34] is the smallest amount of information representing a characteristic of an entity or a relationship. *Entity* means the model of a concrete or abstract object that cannot be more finely defined as an articulation of entities and relationships. An entity is named and identified by a primary key. A *relationship* is the model of an association between objects.

A top-down design inspired by Batini's methodology[35] applied to reaction-related information yields its logical model (Figure 2). This approach provides a logical scheme that is not unique but that constitutes an efficient representation of the modeling. With this approach, a data model becomes easily understandable, and its concrete use by computerized procedures can be described. With this model, a physical scheme of the bank can be proposed.

Physical Scheme. The physical scheme of a data bank provides the general structure of the files containing the acquired information. Two types of files can be distinguished in this structure: *primary files* (resulting directly from acquisition) and *secondary files* (created from primary files, they can serve to optimize certain accesses).

The primary files used come under four headings: bibliography, compounds, reactions, and experimental conditions (Figure 3). Each group contains the description of one of the entities of the logical scheme. The reaction file contains all the relationships of the logical scheme. The compound file contains a topological and one-to-one description of the structures.[15] The bibliography and experimental condition files contain the attributes of the corresponding entities. With this primary file structuring, other information can be accessed efficiently from reactions.

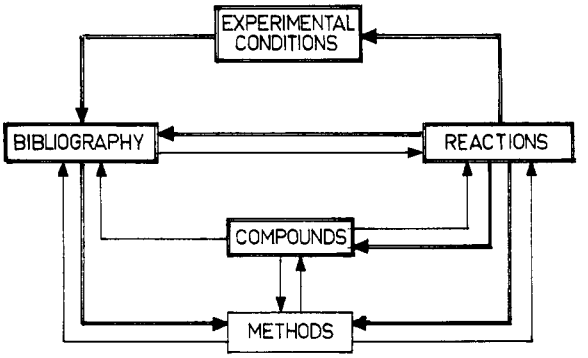


Figure 3. Physical scheme of a specific reaction data bank. The bibliography, reaction, compound, and experimental condition files are accessible via the keys of the corresponding entities. These primary files contain the description of the attributes of each entity and ensure the connections indicated by the double-shafted arrows. The "method" index makes it possible to optimize the accesses to predefined families of reactions. The single-shafted arrows indicate the existence of secondary (inverted) files intended for the optimization of corresponding accesses.

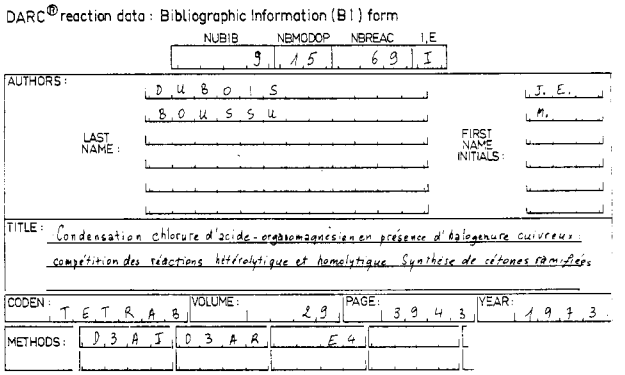


Figure 4. Bibliography Information (BI) form. The BI form is indexed by a number (NUBIB) and contains, for each article, the following information: number of experimental conditions (NBMODOP); number of reactions (NBREAC); article origin (Internal or External); last name and first name initials of each author; title in the language in which the article was published; coden of periodical of origin; volume, page, and year; list of methods (Table I) yielding compounds containing the synthon C–CO–C.

Constructing secondary files (inverted files) made possible efficient access to reactions starting from other entities (an obvious need for such banks), all entities via the "method" attribute, thereby making possible the retrieval of predefined families of reactions, and all entities starting from the compounds, the complete structural description[15] of which is the basis of a great many retrievals. No secondary file was created for experimental conditions because, in the software version discussed hereafter, these data contribute essentially to improve reaction output.

ACQUISITION OF REACTION DATA

Various studies deal with acquisition via a specific language[9,19,36] or via transcoding.[37] However, these approaches cannot be used to acquire structural data in graphic form, even though complex chemical constraints can be expressed in natural language without any check on their relevance.[8] The language we have designed for acquiring reaction-related information, called IGRES,[38] is based on four well-specified information forms. Once these forms have been filled in with data indexed from literature, compendiums, or private sources, structural data are input graphically, whereas all other data are input alphanumerically. During or after the input stage, data are processed in validation procedures that produce diagnoses used to correct the errors.

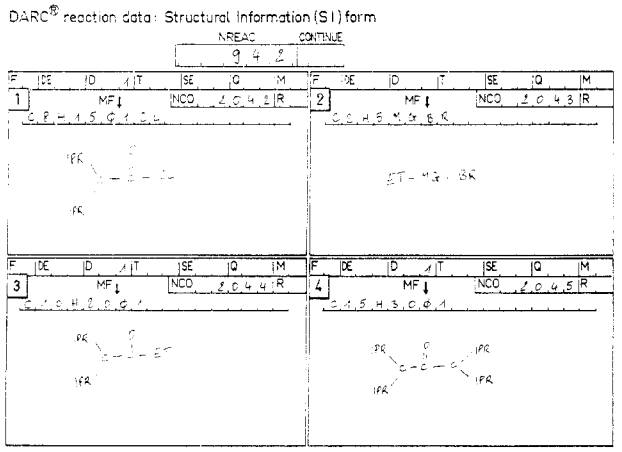


Figure 5. Structure Information (SI) form. An SI form is indexed by a number (NREAC) and contains, for each reaction, the description of compounds in the structural diagram. If it contains more than four compounds, there is a box (continue) for continuation which then indicates the presence of another information form indexed by the same NREAC and containing the other elements of the diagram. The maximum number of compounds per diagram is limited to eight (Figure 7). Each compound is described in a space divided into two zones: a label zone (control information) and an image zone. The label zone contains successively the description of secondary elements [series of numbers indicating the presence of ring closures (F), delocalized bonds (DE), double bonds (D), triple bonds (T), valency (SE), charges (Q), and isotopic mass (M)], the compound number (NCO) and an indication of redundancy of the compound, and the molecular formula. The image zone contains the structural diagram of a molecule drawn according to DARC conventions.[39]

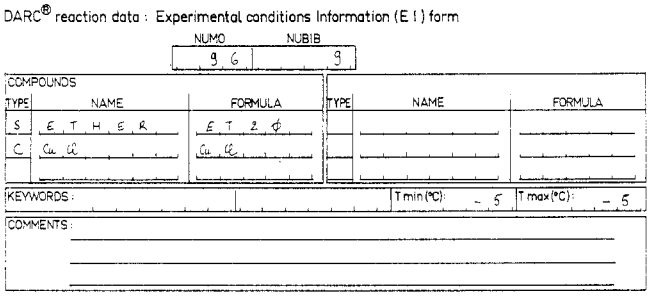


Figure 6. Experimental condition Information (EI) form. An EI form is indexed by a number (NUMO) and contains, for each experimental condition, the article number (NUBIB), the list of compounds used, with each described by a type (i.e., R for reactant, S for solvent, C for catalyst, and X for undetermined), a name and a formula, a set of two key words bearing on the nature of the experimental conditions or on the indication of particular conditions (e.g., nitrogen inflow, pressure, radiations), temperature range, and comments on the experimental conditions.

Information Forms and Data Input. The four IGRES information forms are a *Bibliography Information* (BI) form, a *Structure Information* (SI) form, an *Experimental conditions Information* (EI) form, and an *Associated Information* (AI) form (Figure 4–7). These four forms constitute the basic framework for indexing a specific reaction. This framework can easily be expanded by using other information forms corresponding to a reaction-field enhancement. Aside from indexed data taken from an article, the four information forms contain redundant information used in hereafter described validations. Compounds in EI forms are those having no carbon atom or contributing nothing to the creation of a new bond in at least one end product.[19] Unlike SI compounds,[39] EI compounds are described by the assignment of a type, name, and formula (Figure 6).

Input accomplishes the transfer of information form contents onto a computer-recognized medium: card, tape, or disk file. BI and EI forms are entered on-line via a PDP-11/35. They are validated interactively during input, and the resulting files

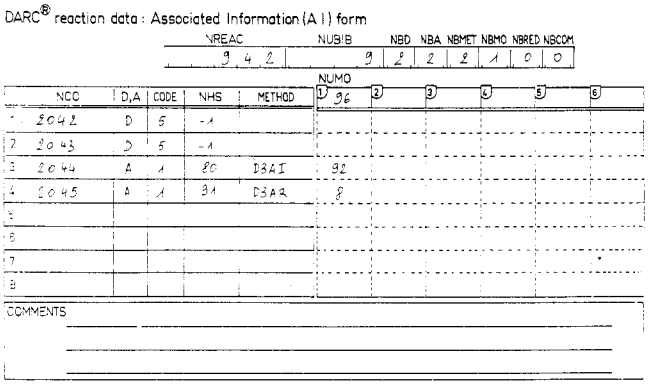


Figure 7. Associated Information (AI) form. An AI form is associated with an SI form and is indexed by the same number (NREAC). For each reaction, the AI form contains the following: a series of *control information* [article number (NUBIB), number of starting compounds (NBD) and of end products (NBA), number of methods (NBMET) and of experimental conditions (NBMO), number of redundant compounds in the structural diagram (NBRED), and number of lines in the comments (NBCOM)] experimental condition numbers (NUMO); (for each compound in the structural diagram) the compound number (NCO), its location in the diagram (Departure, Arrival), a code specifying the structural family of the compound and the nature of the yields associated with this compound, the compound NHS (Figure 1), the method (a code from Table I or −1 if the compound does not contain the C–CO–C synthon), and the yields for the indicated experimental conditions; comments on the reaction.

Table II. Intrinsic Check of AI Form: Error Diagnosis[a]

type of validation[b]	error diagnosis
item validation: checking the contents of an individual data item against some type specification or set of allowable values	NREAC, NUBIB, NBD, NBA, or NBMO equals zero (1) syntactic error (1)
interitem validation: verifying that a stated relationship among the data items or item attributes holds true	incompatibility between code and yield, code, and method (2)
batch validation: looking at a set of input transactions to verify a count, a control or a hash total, or an ordering	check of control information (2) redundancy of NCO NREAC and NUMO in the update (3)
data-base dependent validation: reference to files in the data base to determine the validity of the input transaction	redundancy of NCO NREAC and NUMO in the data base (3)

[a] The intrinsic check software diagnoses the indicated errors in three sequential steps (1–3). [b] See footnote 41.

Table III. Potential Intrinsic Checks and Cross-Checks[a]

	SI	AI	EI	BI
SI	1	2	4	3
AI		1	2	2
EI			1	3
BI				1

[a] The elements of the symmetrical four-dimensional (each entry corresponding to one type of information form) matrix indicate the state of the potential intrinsic checks and cross-checks. 1, intrinsic check; 2, direct cross-check; 3, indirect cross-check; 4, not implemented.

can be used directly by the retrieval software. Because of their sizeable volume, AI forms are entered off-line via a perforator. SI forms are input via a special device, the TOPOCODEUR,[40] or via standard graphic terminals in a less optimized fashion.

Data Validation. Validation is recognized as an important step in the computer acquisition of data.[41] An analysis of data validation methods detects four levels of validation (Table II):

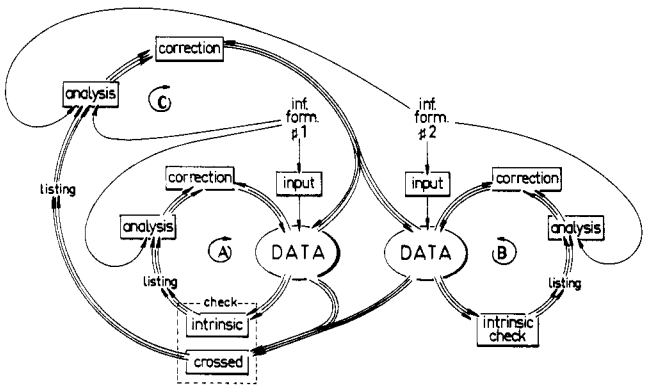


Figure 8. Validation loop. Each type of information form (e.g., #1 and #2) is input via a device, thereby leading to a type of raw data that must be validated. Each type of data is processed through an intrinsic check loop (A and B): this involves running a check program, an analysis of diagnoses made by this program (with, when needed, consultation of the corresponding information form), and computer-aided data correction. Each pair of intrinsically validated data types (corresponding to #1 and #2) is run through a cross-check loop (C); this yields new corrections, which can require new runs through loops A–C. The process is over when the various programs cease to diagnose errors.

item validation, interitem validation, batch validation, and data base dependent validation.[42]

We have combined these four levels into a basic three-step cycle involving computer-aided data checking, analysis of resulting diagnoses, and interactive correction (Figure 8). Checking procedures, depending on whether they involve one or two types of information forms, correspond either to an intrinsic check or to a cross-check, respectively. The elements of a symmetrical four-dimensional (corresponding to the four types of information forms) square matrix can be used to represent all the potential checks (Table III): the diagonal gives the four intrinsic checks; the other elements of the matrix correspond to the six possible cross-checks.

(i) Intrinsic Check. *The intrinsic check diagnoses errors contained in the raw data* corresponding to an information form (BI, SI, EI, AI). This type of check is programmed into the BI input software (so that, for example, it can apply the *Chemical Abstracts* algorithm to sort out the possible errors in an input coden) and into the EI input software (so that, for example, a standard reaction temperature range can be used to check the input temperature range). For SI forms, the intrinsic check software is the standard DARC software for the input and validation of a chemical compound. For AI forms, an intrinsic check software based on Orlicky's four levels[42] diagnoses the errors shown in Table II. The procedure, implemented on a PDP-11/35 involves three ordered steps, ranging from a simple syntactic check to more semantic checks. Roughly one-third of the reactions in the KETO-REACT bank have been corrected with the AI intrinsic check software. When all information form data have been checked intrinsically, they can be cross-checked.

(ii) Cross-Check. *Cross-checking* consists in comparing two occurrences of the same information determined from two different types of information forms and *corresponds to batch validation.*[42] Five of the six potential checks shown in Table III are currently implemented on a PDP-11/35, since in the current version of IGRES a SI–EI check is not possible. Three of these checks (AI–SI, BI–AI, and AI–EI) are run directly, whereas two checks (BI–SI and BI–EI) are done by transitivity; i.e., it is assumed that BI–SI has been conducted as soon as BI–AI and AI–SI have been processed. For example, in the BI–AI check, the effective number of reactions, experimental conditions, methods, and the list of methods from bibliographic data and associated data are compared. In the case of the KETO-REACT data bank, application of both the intrinsic check and the cross-check procedures to the data entered from the four information forms results in the correction of half the reactions in the bank.

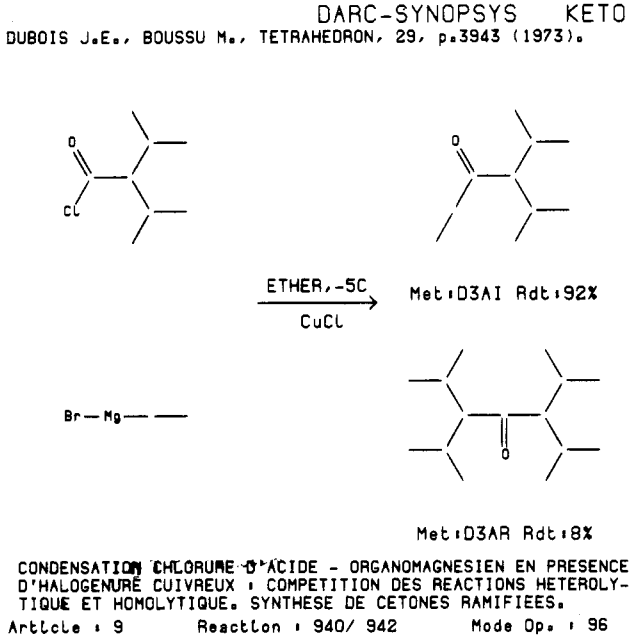


Figure 9. Display of one of the 39 reactions answering the question in Table V (hardcopy obtained by reaction entity output request).

RETRIEVAL OF REACTION DATA

The various existing retrieval systems fall into two categories: (i) those giving access to cards describing reactions[19] and (ii) those giving on-line access to reaction-related information either in the framework of computer-aided synthesis systems[9] or via documentary systems currently being developed.[43] In the DARC-SYNOPSYS expert system, interactive on-line access to a specific reaction data bank built by IGRES has been achieved by implementing an interpreter named RECRE[44] onto a PDP-11/35 and a VAX 11/780. Most of the standard questions occurring to a chemist can most often be processed by being broken down into elementary requests. Access to the different RECRE functions is gained by a 13-command menu, which assists the user in formulating questions in his natural language.

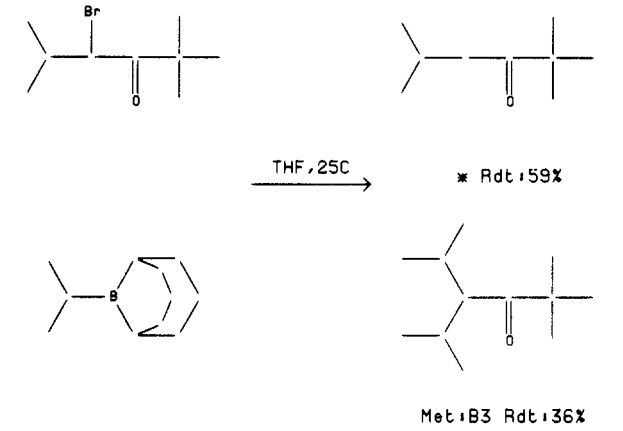
Elementary Request Concept. An example of a standard question submitted to KETO-REACT is "which reactions contained in the bank yield type (iPr)2CHCOC* compounds?" (C* designates a carbon atom with any kind of environment). Retrieval unfolds in two stages: (i) the search for all relevant reactions in the bank; (ii) output of the relevant reactions therein.

The set of *output* procedures is a particular type of elementary request yielding displays of a list of keys, e.g. bibliographical references, experimental conditions, structures, or reactions. The OPUT command processes output requests corresponding to the entities and relationships of the logical scheme (cf. Figures 9–11 for output examples). Use of this command will result in the graphic or alphanumerical display of the attributes of an entity or a relationship. However, if an entity has no attribute, as in the case for the reaction entity in the KETO-REACT bank, all the relationships containing the entity will be output.

The *search* for relevant reactions in the KETO-REACT bank is broken down into two types of elementary requests. In the first type, the *standard DARC substructure search*[15] provides the list of bank compounds belonging to the popu-

DARC-SYNOPSYS KETO

KATZ J.J., DUBOIS J.E., LION C., BULL.SOC.CHIM.FR., p.683 (1977).



ACTION DES ISOPROPYL-9 ET TERTIOBUTYL-9 BORA-9 BICYCLO (3.3.1) NONANES SUR QUELQUES CETONES ALPHA BROMEES. SYNTHESE DE CETONES SUBSTITUEES.

Article : 16 Reaction : 372/ 372 Mode Op. : 164

Figure 10. Display of the reaction yielding a type (iPr)2CHCOC* compound containing a synthesis with a yield of over 50% but not answering the question in Table V (hardcopy obtained by reaction entity output request).

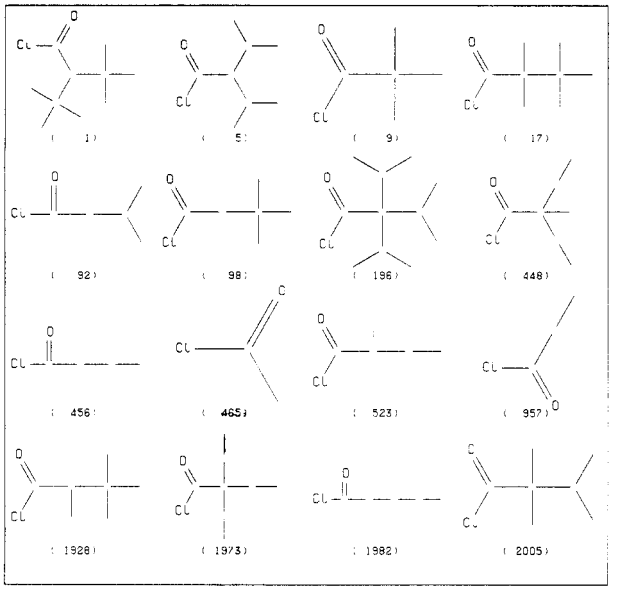


Figure 11. Display of some acid chloride compounds obtained by a substructure search for *CCOCl in KETO-REACT (hardcopy obtained by compound entity output request).

lation defined by (iPr)2CHCOC*. This search, which has been incorporated into RECRE as an elementary request, corresponds to the REST command. This command accepts the input of the graphic or alphanumerical specification of a question and yields a list of relevant compounds, which is then processed to yield the list of bank reactions in which they are end products. Such a nonstructural search corresponds to elementary requests called "*simple questions*". Such questions accept the input of an ARGument list (ARG) of items such as compounds and lead to an ANSwer list (ANS) of items such as reactions. Of all the available simple questions (Table IV), question 10 is the one with which this step can be carried out. It is noteworthy that a simple question can be completed by a selection criterion expressed by means of a TUning list (TU) (e.g., question 4 in Table IV). This initial standard question can be broken down into two simpler requests (here, two elementary requests) with the answers to the first of these

Table IV. Set of Available RECRE Simple Questions for KETO-REACT Information Retrieval

(1) Which compounds (ANS) are involved in the reactions (ARG)?
(2) Which compounds (ANS) are involved as starting products in the reactions (ARG)?
(3) Which compounds (ANS) are involved as end products in the reactions (ARG)?
(4) Which compounds (ANS) are prepared by the reactions (ARG) using the methods (TU)?
(5) Which compounds (ANS) are prepared by the methods (ARG)?
(6) Which compounds (ANS) are prepared by the methods (ARG) in the reactions (TU)?
(7) Which compounds (ANS) are prepared by the syntheses (ARG)?
(8) Which reactions (ANS) involve the compounds (ARG)?
(9) Which reactions (ANS) involve the compounds (ARG) as starting products?
(10) Which reactions (ANS) involve the compounds (ARG) as end products?
(11) Which reactions (ANS) lead to the compounds (ARG) using the methods (TU)?
(12) Which reactions (ANS) are cited in the articles (ARG)?
(13) Which reactions (ANS) involve the methods (ARG)?
(14) Which reactions (ANS) use the methods (ARG) to prepare the compounds (TU)?
(15) Which reactions (ANS) involve the syntheses (ARG)?
(16) Which articles (ANS) cite the compounds (ARG)?
(17) Which articles (ANS) cite the compounds (ARG) as starting products?
(18) Which articles (ANS) cite the compounds (ARG) as end products?
(19) Which articles (ANS) cite the reactions (ARG)?
(20) Among the articles (ARG), which (ANS) were written by the authors (TU)?
(21) Which articles (ANS) cite the methods (ARG)?
(22) In the data bank, which articles (ANS) were written by the authors (TU)?
(23) Among the articles (ARG), which (ANS) appeared between the dates X1 and X2?
(24) Among the articles (ARG), which (ANS) are in-house?
(25) Among the articles (ARG), which (ANS) are not in-house?
(26) Which articles (ANS) appeared in the periodicals (ARG)?
(27) Which articles (ANS) appeared in the periodicals with the CODENS (ARG)?
(28) Which methods (ANS) serve to prepare the compounds (ARG)?
(29) Which methods (ANS) are used to prepare the compounds (ARG) involved in the reactions (TU)?
(30) Which methods (ANS) are involved in the reactions (ARG)?
(31) Which methods (ANS) are used in the reactions (ARG) to prepare the compounds (TU)?
(32) Which methods (ANS) are cited in the articles (ARG)?
(33) Which experimental conditions (ANS) are used in the syntheses (ARG)?
(34) Which syntheses (ANS) are involved in the reactions (ARG)?
(35) Among the syntheses (ARG), which (ANS) have yields >X?
(36) Among the syntheses (ARG), which (ANS) yield the compounds (TU)?
(37) Among the syntheses (ARG), which (ANS) use the experimental conditions (TU)?
(38) In which periodicals (ANS) were the articles (ARG) published?
(39) What are the CODENS (ANS) of the periodicals having published the articles (ARG)?

used as input by the second (*rule 1 in the breakdown process*).

For a question such as "which articles by J. E. Dubois contained in the bank mention ketone alkylation reactions?", the user must apply *logical operations* to lists of answers. The initial question first breaks down into two simple questions. For the one, the argument is the set of articles contained in the bank, the tuning list is J. E. Dubois, and the answer is those articles authored by him (question 22 in Table IV); for the other, the argument is the B code methods (Table I), and the answer is the articles in the bank mentioning ketone alkylation (question 21 in Table IV). Both lists of answers are then subjected to logical operation AND, which will yield the answer to the initial question. The available logical operations (AND, OR, EXCEPT) are accessible via RECRE by means

of the OLIS command. This command can be used to choose an operation as well as two lists (already constituted during the retreival session) to be subjected to this operation. This initial question can be broken down into two simpler requests (here, two elementary requests), the answers to which have been combined by a logical operation (*rule 2 in the breakdown process*).

In case a question subjected to breakdown rules 1 and 2 yields no elementary requests, *the breakdown process is iterated until elementary requests are finally obtained.*

Retrieval Management. Aside from the four above-mentioned types of elementary requests (commands REST, QUES, OLIS, and OPUT), the main RECRE menu has *nine other commands for assisted retrieval.* These commands, which involve retrieval management, are as follows: information commands DADI, INFL, and INFO; control commands PRIN and DISP; search-end commands FINI and PANN; miscellaneous commands ALIS and BTCH.

The information commands can (i) display the dictionary of available data (DADI), (ii) display the directory of lists in the memory and the nature of the variables contained in these lists as well as the number of constituent elements (INFL), and (iii) access operations already run or under way (INFO). The control commands can (i) constitute and print the search listing (PRIN) and (ii) display all elements contained in a memorized list (DISP). With search-end commands, the user can save files produced during the retrieval session (FINI) and/or keep a record of the circumstances surrounding an error managed by the software before exiting from a session (PANN). The other miscellaneous commands available with RECRE are command ALIS for the a priori user-defined constitution of lists and command BTCH for the processing of a batch of RECRE commands.

Simple Questions. Access to available simple questions (Table IV) is gained by means of the QUES command through which the types of elements in argument, answer, and (if necessary) tuning lists can be defined interactively. The types available are compound, reaction, article, experimental conditions, synthesis, yield, author, method, coden, periodical, and date, numbered from 1 to 11, respectively. To gain access to a simple question, the user interactively inputs the number corresponding to the type of answer (e.g., 2 to designate reactions) and the number corresponding to the type of argument (e.g., 1 to designate compounds). Thereafter, RECRE produces the natural language text of all simple questions corresponding to this choice (e.g., for a type 1 argument and a type 2 answer, the user can choose a text corresponding to any question in Table IV numbered from 8 to 11).

Table IV contains various types of simple questions: those leading from one entity to another (1–4, 8–11, 12, 16–18, 19); those binding the synthesis relationship to the three structurally natured entities (reaction, compound, experimental conditions) (7, 15, 33, 34); those binding the article entity to two of its attributes (periodical and coden) (26, 27, 38, 39); those bearing on the method attribute (5–6, 13–14, 21, 28–29, 30–31, 32) [included in Table IV because such questions can optimize access to predefined types of structural changes (Table I)]; those involving selections (i.e., questions in which both the argument *and* the answer can be a list of elements of the same type), as exemplified by a few questions related to the article entity (20, 22–25) and the synthesis relationship (35–37).

The set of all available simple questions is open: depending on use-related statistics or on user demand, the bank manager can easily introduce new questions (e.g., when a question cannot be broken down through the use of existing simple questions) and/or modify or delete existing questions. *The description of a simple question is dual.* It breaks down into a text in a natural language and into a description in terms of files, records, and fields, the constitution of which does not involve any programming.

It is noteworthy that, from a formal viewpoint, simple questions can be expressed in relational terms:[31] a list is considered as a unary relationship table. For example, for simple question 3 in Table IV, starting from the list of reactions ARG, the answer ANS is expressed by

```
JOIN REA_CO_A AND ARG OVER NREAC# GIVING T
PROJECT T OVER NCO# GIVING ANS
```

Table V. Using RECRE To Break Down the Question "Which Reactions Result in Type (iPr)2CHCOC* Compounds in Yields of Over 50%?"[a]

elementary request	formulation	result	list
structural search	(iPr)2CHCOC*	37 compounds	1
simple question 10	Which reactions (2) involve the compounds (1) as end products?	61 reactions	2
simple question 34	Which syntheses (3) are involved in reactions (2)?	173 syntheses	3
simple question 35	Among the syntheses (3) which (4) have yields >50%?	40 syntheses	4
simple question 36	Among the syntheses (4) which (5) yield the compounds (1)?	39 syntheses	5

[a] Samples of output for lists 5 and 4 are given in Figures 9 and 10, respectively.

Table VI. Using RECRE To Break Down the Question "Which Are All the Reactions Making It Possible To Go from an Acid to Its Chloride?"[a]

elementary request	formulation	result	list
structural search	*C–CO–OH	115 compounds	1
structural search	*C–CO–Cl	54 compounds	2
simple question 9	Which reactions (3) involve the compounds (1) as starting products?	132 reactions	3
simple question 10	Which reactions (4) involve the compounds (2) as end products?	49 reactions	4
logical operation	list (3) INTER list (4)	39 reactions	5

[a] An output sample for list 2 is given in Figure 11.

Combinations of Elementary Requests. Tables V and VI show question-to-elementary request breakdown sequence for two common types of complex searches. The first two breakdown stages of the question "which reactions result in type (iPr)2CHCOC* compounds in yields of over 50%?" (Table V) have already been described above. Like the breakdown of these first two stages, the breakdown in the next three stages involves the application of a simple question to the last list obtained by the previous simple question. It should be noted that the third stage (simple question 34 applied to list 2) is included because several syntheses can be involved in a reaction; moreover, stages 4 and 5 (simple questions 35 and 36) can be applied interchangeably to list 3.

While seemingly more difficult, because the KETO-REACT data bank has optimized access to structural changes involving ketone end products, a question such as "which are all the reactions making it possible to go from an acid to its chloride?" (Table VI) can also be answered through the breakdown process. Instead of relying on optimized access through an attribute, a structural search for the starting moiety (step 1) and end moiety (step 2) must be conducted (interchangeable steps), and the reactions in which they are involved must be determined (steps 3 and 4—also interchangeable). The ensuing answer list is the intersection of lists 3 and 4. The 39

reactions resulting from this intersection correspond to all the transformations of an acid to its chloride contained in the KETO-REACT bank. However, in certain cases, the reaction list can also contain reactions involving (i) structural changes not localized on the same atom and (ii) structural changes other than the one being sought.[6] Such noise is inconsequential and can be reduced by various methods, notably by enhancing the RECRE software for a more efficient retrieval of structural changes.[50]

CONCLUSIONS

The KETO-REACT data bank illustrates a methodology for the collection and organization of specific reactions corresponding to a given synthon (e.g., C–CO–C). The acquisition (IGRES) and retrieval (RECRE) procedures applied to this data bank can also be used to process reactions according to other criteria such as reaction type (e.g., oxidation and reduction), key problems in synthesis (e.g., functional group protection and orientation), and existing corpus of inhomogeneous reactions. The general applicability of these procedures stems from the flexibility of the bank architecture devised specifically for the easy handling of other types of data.

Specific reaction banks are necessary and essential components in the design of the DARC-SYNOPSYS expert system in chemistry.[6] Aside from their documentary interest, these banks can be used notably to (i) assess the scope of application of a reaction[14] and (ii) facilitate the extraction of rules for production systems in chemistry.[50] The controlled passage from specific reaction data banks to generic reaction data banks will be made easier by other DARC system components, mainly at the level of the expression of the changes involved in a chemical transformation.[46-49]

ACKNOWLEDGMENT

We are very grateful to Dr. C. Lion for his valuable aid in collecting and indexing KETO-REACT data, to R. Attias for his substructure retrieval software design, and to O. Bruno, G. Carrier, P. Chambert, and F. Mostaghimi for their technical assistance.

REFERENCES AND NOTES

(1) Wipke, W. T.; Howe, H. J. "Computer Assisted Organic Synthesis". *ACS Symp. Ser.* **1977**, *No. 61*.
(2) Smith, D. H. "Computer Assisted Structure Elucidation". *ACS Symp. Ser.* **1977**, *No. 54*.
(3) Buchanan, B. G.; Feigenbaum, E. A. "DENDRAL and meta-DENDRAL: Their Application Dimension". *Artif. Intelligence* **1978**, *11*, 4–24.
(4) Wipke, W. T.; Glenn, I. O.; Krishnan, S. "Simulation and Evaluation of Chemical Synthesis". *Artif. Intelligence* **1978**, *11*, 173–193.
(5) Nilsson, N. J. "Principles of Artificial Intelligence"; Tioga: Palo Alto, CA, 1980.
(6) Picchiottino, R.; Sicouri, G.; Dubois, J. E. "DARC System: Transformational Relationships and Hyperstructures in Chemistry". In "Proceedings of the Eighth International CODATA Conference"; Glaeser, P. S., Ed.; North-Holland: Amsterdam, 1983; pp 229–334.
(7) Bersohn, M.; Esack, A. "Computers and Organic Synthesis". *Chem. Rev.* **1976**, *76*, 269–282.
(8) Gund, P.; Grabowski, E. J. J.; Hoff, D. R.; Smith, G. M.; Andose, J. D.; Rodhes, J. B.; Wipke, W. T. "Computer-Assisted Analysis at Merck". *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 83–93.
(9) Corey, E. J.; Wipke, W. T.; Cramer, R. D.; Howe, W. J. "Computer-Assisted Synthetic Analysis for Complex Molecules. Methods and Procedures for Machine Generation of Synthetic Intermediates". *J. Am. Chem. Soc.* **1972**, *94*, 440–459.
(10) DARC: Description, Acquisition, Retrieval, and Computer-aided design; SYNOPSYS: SYNthesis, OPtimization SYStem.
(11) Dubois, J. E. "Structural Organic Thinking and Computer Assistance in Synthesis and Correlation". *Isr. J. Chem.* **1975**, *14*, 17–32.
(12) Dubois, J. E. "Computer Assisted Modeling of Reactions and Reactivity". *Pure Appl. Chem.* **1981**, *53*, 1313–1327.
(13) Mostaghimi, F. "Banque de donnees des cetones aliphatiques. Synthese et proprietes physiques". Thesis, Paris 7 University, Paris, 1978.
(14) Dubois, J. E.; Panaye, A.; Lion, C. "Conception assistee par ordinateur. Notion de Domaine Structural Ordonne d'une Reaction (DSOR)". *Nouv. J. Chim.* **1981**, *5*, 371–380.
(15) Attias, R. "DARC Substructure Search System: A New Approach to Chemical Information". *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 102–108.
(16) H. Gutfreund, A. Bussard, D. Colquheum, J. E. Dubois, M. Kotani, and N. Kurti, unpublished results of the CODATA Bio Sciences Committee, 1979, Paris.
(17) Theilheimer, W. "Synthetische Methoden der Organischen Chemie"; Karger: Basle and New York, 1983.
(18) "Current Chemical Reactions"; ISI: Philadelphia, PA, 1983.
(19) Ziegler, H. J. "Roche Integrated Reaction System (RIRS). A New Documentation System for Organic Reactions". *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 141–149.
(20) Dubois, J. E.; Hennequin, F.; Boussu, M. "Utilisation du systeme topologique DARC a des fins documentaires. Methodes de preparation des cetones aliphatiques saturees". *Bull. Soc. Chim. Fr.* **1969**, 3615–3623.
(21) KETO-REACT reactions have a maximum number of four and six compounds as starting and end products, respectively.
(22) Vleduts, G. E. "Concerning One System of Classification and Codification of Organic Reactions". *Inf. Storage Retr.* **1963**, *1*, 117–146.
(23) Guthrie, R. D. "A Suggestion for the Revision of Mechanism Designation". *J. Org. Chem.* **1975**, *40*, 402–407.
(24) Wipke, W. T.; Braun, H.; Smith, G.; Choplin, F.; Sieber, W. "SECS, Simulation and Evaluation of Chemical Synthesis: Strategy and Planning". *ACS Symp. Ser.* **1977**, *No. 61*, 104.
(25) Hendrickson, J. B. "Systematic Synthesis Design. 4. Numerical Codification of Construction Reactions". *J. Am. Chem. Soc.* **1975**, *97*, 5784–5800.
(26) Ugi, I.; Bauer, J.; Brandt, J.; Friedrich, J.; Gasteiger, J.; Jochum, C.; Schubert, W. "New Applications of Computers in Chemistry". *Angew. Chem., Int. Ed. Engl.* **1979**, *18*, 11–123.
(27) Littler, J. S. "An Approach to the Linear Representation of Reaction Mechanisms". *J. Org. Chem.* **1979**, *44*, 4657–4667.
(28) Roberts, D. C. "A Systematic Approach to the Classification and Nomenclature of Reaction Mechanisms". *J. Org. Chem.* **1979**, *43*, 1473–1480.
(29) Satchell, D. P. N. "The Classification of Chemical Reactions". *Naturwissenschaften* **1977**, *64*, 113–121.
(30) Gold, V. "Glossary of Terms Used in Physical Organic Chemistry". *Pure Appl. Chem.* **1983**, *55*, 1281–1371.
(31) Date, C. J. "An Introduction to Database Systems", 3rd Ed.; Addison-Wesley: Reading, MA, 1981.
(32) Chen, P. P.; "The E/R Model: Toward a Unified View of Data". *ACM Trans. Data Base Systems* **1976**, *1*, 9–37.
(33) "Entity Relationship Approach to System Analysis and Design"; Chen, P. P., Ed.; North-Holland: Amsterdam, 1980.
(34) Sakai, H. "An Unified Approach to the Logical Design of a Hierarchical Data Model". "Entity Relationship Approach to System Analysis and Design"; Chen, P. P., Ed.; North-Holland: Amsterdam, 1980.
(35) Batini, C.; Santucci, G. "Top Down Design in the Entity Relationship Data Model". "Entity Relationship Approach to System Analysis and Design"; Chen, P. P., Ed.; North-Holland: Amsterdam, 1980.
(36) Bersohn, M.; Esack, A. "A Computer Representation of Synthetic Organic Reactions". *Comput. Chem.* **1976**, *2*, 103–107.
(37) Bawden, D.; Devon, T. K.; Jackson, F. T.; Wood, S. I.; Lynch, M. F.; Willett, P. "A Qualitative Comparison of WLN Descriptors of Reactions and the Derwent Chemical Reaction Documentation Service". *J. Chem. Inf. Comput. Sci.* **1980**, *19*, 90–93.
(38) IGRES: Indexing and Generating REactions for Storage.
(39) "Indexing Structures in the DARC System: Users' Manual"; ARDIC: Paris, 1978
(40) Dubois, J. E.; Miller, J. A. "Appareil pour le codage et la visualisation simultanee d'un graphe". ANVAR Patent. With the TOPOCODEUR, the input of voluminous structural and/or textual data corresponding to chemical compounds and/or reactions is possible.
(41) "Information Systems"; Tou, J. T., Ed.; Plenum Press: New York, 1974.
(42) Orlicky, J. "The Successful Computer System: Its Planning, Development and Management in a Business Enterprise"; McGraw-Hill: New York, 1969; pp 162–168.
(43) Lynch, M. F.; Willett, P. "The Automatic Detection of Chemical Reaction Sites". *J. Chem. Inf. Comput. Sci.* **1978**, *18*, 154–159. 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.* **981**, *21*, 137–141.
(44) "RECRE (Retrieval and Edition of Chemical REactions) Users' Manual"; ARDIC: Paris, 1981.
(45) Dubois, J. E.; Laurent, D.; Panaye, A.; Sobel, Y. "Systeme DARC: concept d'hyperstructure formelle". *C. R. Hebd. Seances Acad. Sci., Ser. C* **1975**, *280*, 851–854.
(46) Dubois, J. E.; Picchiottino, R.; Sicouri, G.; Sobel, Y. "DARC System. Block Relationships and Hyperstructures". *C. R. Hebd. Seances Acad. Sci., Ser. 1* **1982**, *294*, 251–256.
(47) Dubois, J. E.; Panaye, A.; Picchiottino, R.; Sicouri, G. "DARC System. Structure of a Reaction Invariant". *C. R. Hebd. Seances Acad. Sci., Ser. 2* **1982**, *295*, 1081–1086.

(48) Dubois, J. E.; Sicouri, G.; Sobel, Y.; Picchiottino, R. "DARC System. Localized Operators and Co-structures of a Reaction Invariant". C. R. Hebd. Seances Acad. Sci., Ser. B 1984, 298, 525-530.
(49) Sicouri, G.; Sobel, Y.; Picchiottino, R.; Dubois, J. E. "DARC System. Localizing Variations onto the Reaction Invariant: the Transformation Structure Concept". C. R. Hebd. Seances Acad. Sci., in press.
(50) Picchiottino, R.; Sicouri, G.; Dubois, J. E. "DARC-SYNOPSYS Expert System. Production Rules in Organic Chemistry and Application to Synthesis Design". In "Computer Science and Data Bank"; Hippe, Z.; Dubois, J. E., Eds.; Polish Academy of Sciences: Warsaw; in press.

Question of Data Format in Organic Chemistry

GUIDO SELLO
Laboratorio di Chimica Organica, Universita' degli Studi, Milano, Italy

Received April 13, 1983

Organic chemistry of necessity has developed mechanized information retrieval. The problem of characterizing the format of the compounds' structures and reactivity is examined. Current and future trends are given with comments and critical observations. Information about both informatic and chemical aspects of data collection is treated from both user and producer viewpoints. Techniques and tricks in structuring and maintaining data are given.

INTRODUCTION

The question of database design became pressing when the opportunity to substitute automatic for manual management occurred. Since this option has been available, many remarkable results have been obtained in nearly every operating field. Database design gave useful and accurate solutions for chemical information. A number of works and publications appeared, sometimes starting from quite different positions, that focused specific questions of chemical data management. A number of specific positions have been reported in the field of formating data relevant to chemical arguments.

In this article, we will examine those positions pertaining to Organic Chemistry. There are two kinds of fundamental information available to the user of organic chemical data: molecular structures and reaction records. A problem became apparent to the people who designed databases for organic chemists. A great amount of work, both theoretical and experimental, appeared; all of which pertained to the precision and reliability of the various methods for listing and codifying structures and reactions. An important amount of success was achieved in the field of codification of structures but not in that of codification of reactions. This happened for two strongly correlated reasons: the lack of a real and efficient methodology for analyzing the theoretical aspects of reactions and the necessity of translating them from usual language to code. We will discuss problems and solutions connected to each matter separately.

FAMILY OF MOLECULAR STRUCTURES

This class was treated in several works. From the first algorithms,[1,2] less refined and less general to the latest and most polished solutions, they came to the best level possible in the field of codification; the strictly informatic aspect is of remarkable quality. In fact, we must remember that, dealing with the question of inserting records in a big database, one has to solve both problems of real logic (connected to codification system) and also the usual problems of storage and working feasibility. These last ones, severely limited by hardware available, must be optimized.

The problem of transforming the representation of molecules (usually bidimensional drawings with some tricks for pointing out the third dimension) into computer-readable form (series of binary numbers) arises. We must translate the instinctive information supplied by the drawing into exact machine representation. The separation of essential parts of a visual picture may appear more or less simple, according to the situation; for example, the description of a range of different colors is evident (e.g. one may call the first COLOR 1, the second COLOR 2, etc.), while agreement on the tonality or shape is less clear. In the case of the description of molecules, it is well-known that a drawn representation is built by points, identifying atoms, and by lines, identifying bonds. Actually, what we learn by looking at the graphic representation of a structure is much more: in fact, we may guess that some points-atoms are tetravalent carbons saturated with hydrogens in the valences not displayed, that some bonds are placed above the molecular plane and others below it, that an oxygen atom marked by the symbol O has no hydrogens connected to it if it has already two bonds, and so on. In some way this must be taught to a machine, which (because of its stupidity) knows nothing about organic matter. In addition, it is necessary to prevent the computer from mistaking two similar structures or codifying the same molecule in different ways. These problems must be solved in the best possible way in order to build a system that is efficient, both for management (ease in updating, deleting, inserting) and for retrieval (strictly connected to the data structures), handy, having good interaction with the outside, and reliable, giving nonambiguous answers.

Methods were sought to codify structures that would give canonical formulations in output and allow the user to rely on the correctness of the results. Among the several algorithms, we must consider Morgan's algorithm, which, appropriately modified and extended, is still the most used.[3-5] They are systems to translate the pictorial molecular view into numerical vectors or matrices, whose various components, $a(i(j))$, are the signs of the presence/absence of certain molecular features.[6] Obviously, the trick for getting canonical results resides in solving the points of choice in an absolute manner.

One of the best solutions in the field of structure representation is based on graph theory and, more exactly, on a treelike graph, which is well suited for describing even very complex chemical formulas. When the problem concerning rings or stereoisomers representation (i.e. tridimensionality) has been solved, it is simple to find an algorithm, apply it to a labeled graph, and obtain an univocal solution (e.g., using a prefixed traversal order; preorder or postorder to get only one output).

More complete and exhaustive works in the field are devoted to solving the different questions unsolved by graph theory, mainly concerning the mentioned problems of rings and stereoisomers. In both cases, one has to adapt a general theory, like graph theory, to the reality of organic molecules. For example, in the case of rings it is useless to state the number and kind of all existing rings, while it is appropriate to consider

0095-2338/84/1624-0249$01.50/0