SPECIALIZED SERVICES

The heart of the CAS services remains *Chemical Abstracts* itself, the printed Abstract Issues with their attendant indexes. But in addition, there is a wide array of services made available in various formats for specialized purposes.[2] Several of these pertain to the field of fossil fuels.

Among the services available in print, the Section Groupings provide the total content of the Abstract Issues in five packages: the biochemical, organic, macromolecular, applied and engineering, and physical and analytical groupings. Section 51 as a whole is included in the grouping that is designated "CA Applied Chemistry and Chemical Engineering Sections". Complete bibliographic information is included, as well as the Keyword Index entries for the entire issue of CA from which the particular grouping is extracted.

"Coal Science and Process Chemistry" is representative of one of the newest family of services, *CA Selects*. This is a printed current-awareness service, issued biweekly, which includes the bibliographic information and abstracts that appear throughout CA dealing with coal—its liquefaction, gasification, carbonization, properties, analysis, composition and combustion—mine gases, and brown coal. Thus all abstracts that have anything to do with coal, in whatever section they appear, are collected into a conveniently scanned publication. This publication, along with more than 70 others in various specialized subject areas, was recently introduced to provide an inexpensive current-awareness alternative to the many such computer-readable services. The information provided by the various *CA Selects* topics is extracted from the CAS database by computerized retrieval that is based on custom-designed search profiles.

CA Condensates (CACon) and *CA Subject Index Alert* (CASIA) are computer-readable tape services, issued biweekly. The former is a compilation of the bibliographic information, section numbers, and keywords from the Abstract Issues; abstracts are not included. CASIA provides the index entries that are subsequently published in the semiannual Subject and Formula Indexes; these index entries are associated with the corresponding abstracts in the printed issues through the abstract numbers. Section and subsection numbers and Registry Numbers for specific substances are included in CASIA.

Lastly there is the tape file *Energy*, which includes abstracts, bibliographic information, section and subsection numbers, keywords, and index entries for a selected group of energy-related CA sections. Section 51 appears in this file as a whole. The other sections included are those that deal with propellants and explosives (50); with electrochemistry (72) and with electrochemical, radiational, and thermal energy technology (52); with thermochemistry and thermal properties (69); and with nuclear phenomena and technology (70 and 71).

REFERENCES AND NOTES

(1) "Subject Coverage and Arrangement of Abstracts by Sections in *Chemical Abstracts*," 1975 edition, Chemical Abstracts Service, Columbus, Ohio.
(2) The CAS database concept is described by R. E. O'Dette, *J. Chem. Inf. Comput. Sci.*, **15**, 165–9 (1975).

A Problem-Oriented Analysis of Database Models

NADIA THALMANN*†
Section Systèmes d'Information, Faculté des Sciences de l'Administration, Université Laval, Québec, Canada

DANIEL THALMANN
Département d'Informatique et de Recherche Opérationnelle, Université de Montréal, Montréal, Canada

Received January 17, 1978

Which is the most convenient database model considering specific applications? The goal of this paper is to try to answer this question by the use of a chemical example. Examples of requests describe the problems of insertion, deletion, and updating; these requests are analyzed for the hierarchical model and are expressed in a relational language defined by the authors and in Socrate for the network model.

1. INTRODUCTION

An increasingly important aspect of commercial data-processing activities is database management systems (DBMS). First, two questions have to be answered: what is a database and why are DBMS necessary?

Engles[9] defines a database as a collection of stored operational data used by the application systems of some particular enterprise. We may accept this definition if the word "enterprise" means an organization such as a bank, a school, or a hospital. A DBMS is necessary for the following reasons: (i) the operations of a file system will refer to entities in the physical part of the database description, (ii) logical errors have to be checked, (iii) protection against misuse has to be assured, and (iv) standardization should be assured.

The DBMS world often seems very confusing because of the variety of options available to implement such systems. However, most authors[2,7,15] consider three general database models: (i) the hierarchical model, (ii) the network model, and (iii) the relational model.

Two other models have been proposed: the entity set model by Senko[12] and the entity-relationship model by Chen.[3] We shall not discuss these two models in this paper as their qualities are similar to those of the relational model.

If some authors[2,7] have discussed the advantages and disadvantages of each model, a major tendency in database management literature and systems is to choose always examples of employees, ages, departments, and so on. It is not surprising because DBMS were first designed for management applications, but today, databases take an important place in all science areas.

†This work is partly supported by the Swiss National Fund for Scientific Research, while on leave of absence from the Department of Chemistry of the University of Geneva, Switzerland.

0095-2338/79/1619-0086$01.00/0

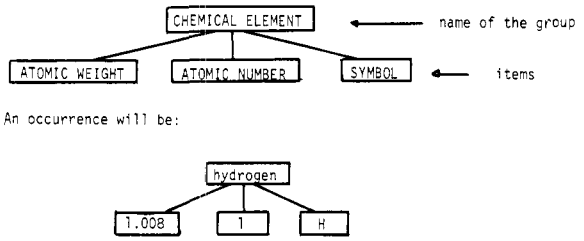

Figure 1. A group with an occurrence.

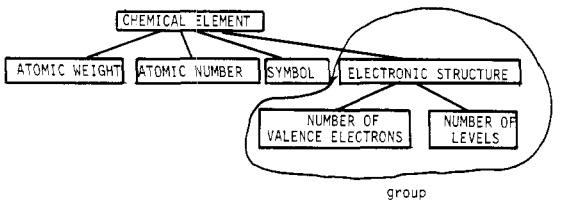

Figure 2. A compound group.

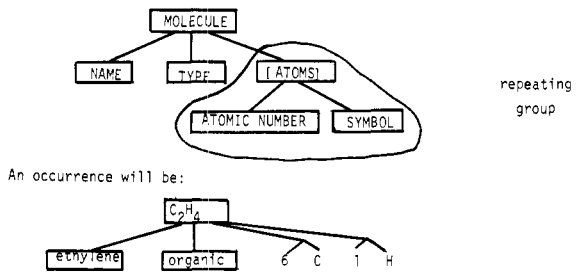

Figure 3. The group MOLECULE and the group ATOM.

The use of databases in chemistry,[8] clinical research,[13] meteorology,[11] and many other fields is now a reality. We have to consider this situation, and it is necessary to study each model with a broad viewpoint of the applications. That is the reason why we have decided to choose a chemical example to describe briefly each of the data models and to show their advantages and disadvantages.

2. THE THREE GENERAL DATA MODELS

1. The Hierarchical Model. A hierarchical structure (or tree structure) always starts with a *root segment*. *Dependent segments* may be added on a succeeding level and have only one entry point. A segment and its dependents are called a *group*. Figure 1 shows a group with an occurrence: the segment is the CHEMICAL ELEMENT, it is the name of the group and the dependents are the ATOMIC WEIGHT, the ATOMIC NUMBER, and the SYMBOL. We have chosen the hydrogen atom as the occurrence.

A group is called *simple* if none of its dependents has dependents; otherwise, it is called a *compound* group. In Figure 2, a compound group, ELECTRONIC STRUCTURE, has been added to CHEMICAL ELEMENT as a new dependent. If a group is an element of a compound group and has more than one realization, it is called a *repeating group*.

In Figure 3, we consider two groups: the group MOLECULE, including the segment MOLECULE with the dependents NAME, TYPE, and SYMBOL; the group ATOM including the segment ATOM with the dependents ATOMIC NUMBER and SYMBOL. For one realization of the group MOLECULE, we can associate more than one realization of the group ATOM; the two groups represent a hierarchy.

2. The Network Model. If the hierarchical model can be represented by a tree structure, the network model can be defined in *graph* terms. A network structure represents a collection of related segments and no segment has a special function like the root segment in the hierarchical model. We can derive the network structure from the hierarchical by adding the notion of *reference*, which can be defined as a logical pointer. So it is possible to link two trees without the necessity to have a hierarchy between them. The link can be made by a set of references; this set is considered as a group of one of the trees. In Figure 4, [COMPONENT] is a set of references which are pointers to ATOMS.

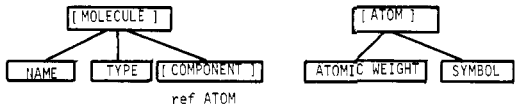

Figure 4. A link by a set of references.

SYMBOL	ATOMIC NUMBER	ATOMIC WEIGHT	SYMBOL	ATOMIC NUMBER	ATOMIC WEIGHT
H	1	1.008	F	9	18.998
He	2	4.0026	Ne	10	20.18
Li	3	6.94	Na	11	22.99
Be	4	9.012	Mg	12	24.30
B	5	10.81	⋮	⋮	
C	6	12.011	Cl	17	35.453
N	7	14.0067	⋮	⋮	
O	8	15.9994			

Figure 5. The relation ATOM.

NAME	SYMBOL	NUMBER OF ATOMS
hydrochloric acid	H	1
formic acid	H	2
hydrofluoric acid	H	1
formic acid	C	2
nitrous acid	O	2
hydrochloric acid	Cl	1
hydrofluoric acid	F	1
acetic acid	H	4
acetic acid	C	2
acetic acid	O	2
nitrous acid	H	1
nitrous acid	N	1
⋮	⋮	⋮

Figure 6. The relation MOLECULE.

3. The Relational Model. The concept of a relation was proposed by Codd.[5] A relation R is a subset of the Cartesian product $S_1 \times S_2 \times \ldots \times S_n$ of sets $S_1, S_2, \ldots, S_n$ which are not necessarily distinct and known as the domains. A relation can be represented as a table with each row representing one n-tuple (or tuple). The order of the lines of this table has no significance, and it is not possible to have two identical lines. A typical example of relation is given in Figure 5; we have chosen the relation ATOM (SYMBOL, ATOMIC NUMBER, ATOMIC WEIGHT) which is well known as the periodic classification.

It is evident that we can make associations between relations. In Figure 6, we introduce the relation MOLECULE (NAME, SYMBOL, NUMBER of ATOMS). The link between the relations ATOM and MOLECULE is made by the common attribute SYMBOL.

3. ADVANTAGES AND DISADVANTAGES OF THE DIFFERENT MODELS

To discuss the advantages and disadvantages of the three models, it is necessary to present examples. That is the reason we consider the same example treated by the three models. We want to build a database of the most frequently known molecules, with their names, and the atoms which belong to these molecules. We include specific data for the atoms (atomic number, atomic weight) and specific data for the

NAME	PKAB	MELTING TEMPERATURE	BOILING TEMPERATURE
hydrochloric acid	0	-112	-83.7
formic acid	3.75	8.4	100.7
hydrofluoric acid	3.42	-92.3	19.4
acetic acid	4.75	16.6	118.1

Figure 7. The relation PROPERTIES.

molecules (pK_{ab}, melting temperature, boiling temperature). For the relational approach, requests are expressed in a relational language defined by the authors while Socrate[1] is used for the network model.

1. The Relational Model. In our example, we choose three relations:

ATOM (SYMBOL, ATOMIC NUMBER, ATOMIC WEIGHT)

MOLECULE (NAME, SYMBOL, NUMBER OF ATOMS)

PROPERTIES (NAME, PKAB, MELTING TEMPERATURE, BOILING TEMPERATURE)

The tables are given in Figures 5, 6, and 7.

We defined a relational language influenced by both the Alpha[6] language and the Pascal[16] language; thus we can consider this language to be a structured relational language. Each relation has its representative variable, a kind of "control variable". Each attribute of the relation possesses a type, either integer, real, Boolean, or alpha *n*, where alpha is a string type of *n* maximum number of characters. There are simple variables, too, and sets based on these four types.

In our example, the description of the three relations is the following:

```
RELATIONS
    ATOM(SYMBOL:ALPHA 2; ATOMIC-NUMBER:INTEGER; ATOMIC-WEIGHT:REAL)
                              REPRESENTED BY A;
    MOLECULE(NAME:ALPHA 25; SYMBOL; NUMBER-OF-ATOMS:INTEGER)
                              REPRESENTED BY M;
    PROPERTIES(NAME; PKAB,MELTING-TEMPERATURE,BOILING-TEMPERATURE:REAL)
                              REPRESENTED BY P;
DEFINITIONS
    ATOM=ACCESSIBLE ALTERABLE; MOLECULE=ACCESSIBLE ALTERABLE;
    PROPERTIES=ACCESSIBLE ALTERABLE;
```

Notice: accessible alterable means that the information about each relation is accessible and alterable in the database.

Example of Requests

1. Find the name of the molecules which include hydrogen.

```
WITH M SELECT ALL
    (M.SYMBOL = 'H') WRITE M.NAME;
```

2. Find the symbol and the atomic number of all the atoms of nitrous acid.

```
WITH M,A SELECT ALL
    (M.NAME = 'NITROUS ACID' AND M.SYMBOL = A.SYMBOL)
            WRITE A.SYMBOL,A.ATOMIC-NUMBER;
```

3. Find the formula of the molecule for which we give the name (e.g., hydrofluoric acid).

```
WITH M SELECT ALL
    (M.NAME = 'HYDROFLUORIC ACID')
            WRITE M.SYMBOL,M.NUMBER-OF-ATOMS;
```

4. Find the name of all strong acids.

```
WITH P SELECT ALL
    (P.PKAB = 0) WRITE P.NAME;
```

5. Find the symbol and the atomic weight of all the atoms which compose molecules for which the melting point is higher than 10 °C.

```
WITH P,M,A SELECT ALL
    (P.MELTING-TEMPERATURE > 10 AND P.NAME = M.NAME AND
        M.SYMBOL = A.SYMBOL)
            WRITE A.SYMBOL,A.ATOMIC-WEIGHT;
```

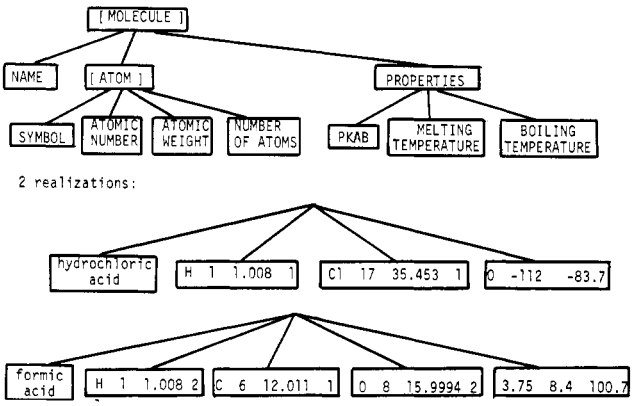

Figure 8. The hierarchical model.

We see it is very easy to manipulate such requests. It is also very simple to insert new tuples; for example, we can add a sulfur atom in the relation ATOM.

```
WITH A INSERT BEGIN A.SYMBOL := 'S'; A.ATOMIC-NUMBER := 16;
            A.ATOMIC-WEIGHT := 32.06
            END;
```

It is possible in the same way to modify or to delete tuples.

2. The Hierarchical Model. The general scheme for our example and two realizations are presented in Figure 8. Let us consider the disadvantages:

(i) It is not possible to access the atomic weight of hydrogen without accessing a molecule which contains hydrogen.

(ii) There exists more than one occurrence of the same atom for different molecules; therefore it is redundant.

(iii) If we want to change the atomic weight of hydrogen from 1.008 to 1.009, the access to each molecule would be required to update this value. This is a consequence of the second point.

(iv) It is not possible to delete the single molecule which contains fluorine without deleting all information about fluorine.

(v) It is not possible to insert the sulfur atom without inserting a molecule with sulfur.

(vi) It is easy to answer the question: "Which atoms constitute the molecule of nitrous acid?" But it is very difficult to answer the question: "Which are the molecules which contain hydrogen?" There is no symmetry in the search program.

Advantages

The model is simple because it gives a good view of the fact that atoms belong to the molecule.

3. The Network Model. We illustrate our problem with the Socrate system[1] in Chart I (with our own translation of the keywords), which was developed at the University of Grenoble by Abrial et al.[1] This system is simple and, we think, a better approach for our discussion than using a big system like CODASYL.[4] We notice that the main disadvantage of the network model is the artificiality due to the introduction of the entity COMPOSITION in the data description. Furthermore, there is no real *data independence*. Why should the user see the links? But we can consider that all the disadvantages of the hierarchical model have disappeared. We can describe very complex structures with the notion of reference. We can delete a link without deleting the whole referred group.

4. IMPLEMENTATIONS

We have discussed the advantages and disadvantages of each model on a theoretical basis. However, we know that various techniques have been used in particular implementations of these models to overcome the different disadvantages.

Chart I

```
entity molecule                          entity atom
    |name                                    |symbol
    |compos reverses each composition        |atomic number
    |entity properties                       |atomic weight
    |      |pkab                             |comp reverses each composition
    |      |melting temperature
    |      |boiling temperature

entity composition
     number of atoms
     at refers atom
     mol refers molecule

Examples of requests

1.  Find the name of each molecule which includes hydrogen.
    I name of each molecule of comp with symbol = 'H'.

2.  Find the symbol and the atomic number of every atom of the
    nitrous acid.
    M  X1 = one composition of one molecule with name = 'NITROUS ACID'
    M  X2 = one at of X1
    for each X2
    W  symbol
    W  atomic number
    end ?
    M  is the abbreviation of Modify and
    W  is the abbreviation of Write; X1 and X2 are two variables.

3.  Find the formula of hydrofluoric acid.
    M  X1 = one composition of one molecule with name = 'HYDROFLUORIC ACID'
    M  X2 = one at of X1
    for each  X2
    W  symbol
    W  number of atoms
    end ?

4.  Find the name of each strong acid.
    W  name of each molecule with pkab of properties = 0

5.  Find the symbol and the atomic weight of the atoms of the molecules
    which have a melting point greater than 10 degrees Celsius.
    M X1 = one composition of molecule with melting temperature of
                                              properties >  10
    M X2 = one at of X1
    for each X2
    W symbol
    W atomic weight
    end ?
```

For example, an IMS[10] database and its physical storage structures are described in the data description language DL/1 as a Database Description (DBD). Thus, our chemical example should be structured as two DBDs, one for the molecules and one for the atoms, with the number of atoms recorded as intersection data. This implementation solves the disadvantages of the complete hierarchical organization; however, it introduces some amount of data dependence, and the use of logical pointers leads to the disadvantages of the network model. Moreover, the aggregate conceptual structure can become very complex.

Concerning the relational model, no matter how it is implemented, the user is not aware of the exact representation and does not really care what it is. This is the major advantage of the model and that is the reason why its implementation is one of the hardest design problems. Nevertheless, we have realized such a relational DBMS[14] with the structured relational data manipulation language which is described in section 3.1.

5. CONCLUSION

It is very easy to ask some requests with the relational language, because it is more natural and near our way of thinking. We can really write structured programs with the relational model and the language defined and implemented by the authors. The use of the hierarchical model is cumbersome and very restricted in the possibilities of requests. In the network model, the numerous hierarchical entities are better, but we have to think of the links between them and it is still very close to the machine description. For the field of chemistry, the relational language is very simple and precise.

We have chosen an original but simple example to present the differences between the three kinds of database models. We think that databases will always take a more important place in chemistry as well as in all other science areas. It is necessary to study the advantages and disadvantages of each model in order to choose the most convenient one for these fields.

ACKNOWLEDGMENT

The authors thank the students who have worked for the implementation of the relational DBMS. They are grateful to the referees for their helpful criticism.

REFERENCES AND NOTES

(1) J. R. Abrial, J. P. Cahen, J. C. Favre, D. Portal, G. Mazare, and R. Morin, "Projet Socrate", Université Scientifique et Médicale de Grenoble, 1972.
(2) M. Adida and C. Delobel, "Les Modèles Relationnels de Bases de Données", Le Chesnay, IRIA, 1976.
(3) P. P. Chen, "The Entity-Relationship Model-Toward a Unified View of Data", *ACM Trans. Database Systems*, **3**, 9–36 (March 1976).
(4) CODASYL, "Database Task Group Report", ACM, New York, 1971.
(5) E. F. Codd, "A Relational Model of Data for Large Shared Data Banks", *Commun. ACM*, **13**, 377–387 (June 1970).
(6) E. F. Codd, "A Database Sublanguage Founded on the Relational Calculus", Proceedings of the ACM-SIGFIDET 1971, Workshop, San Diego, Calif., Nov 1971, pp 35–68.
(7) C. J. Date, "An Introduction to Database Systems", Addison-Wesley, Reading, Mass., 1975.
(8) K. M. Donovan and B. B. Wilhide, "A User's Experience with Searching the IFI Comprehensive Database to U.S. Chemical Patents", *J. Chem. Inf. Comput. Sci.*, **17**, 139–143 (1977).
(9) R. W. Engles, "A Tutorial on Data Organization", *Annu. Rev. Automatic Programming*, 7 (1972).
(10) IBM, IMS/360, "Applications Description Manual", White Plains, N.Y., GH-20-0765.
(11) R. T. Knox, "An On-Line and Batch Meteorological Data Retrieval System for Schools and Colleges", Proceedings of the IFIP 2nd World Conference on Computers in Education, 1975, North-Holland, Amsterdam, pp 583–588.
(12) M. E. Senko, E. B. Altman, M. M. Astrahan, and P. L. Fehder, "Data Structures and Accessing in Data-Base Systems", *IBM Syst. J.*, **12**, 30–93 (1973).
(13) W. L. Sibley, M. D. Hopgood, G. F. Grover, W. H. Josephs, and N. A. Palley, "Data Management for Clinical Research", *AFIPS Conf. Proc.*, 63–68 (1977).
(14) N. Thalmann and D. Thalmann, "Direct Connection Between Compiling Techniques and Databases Courses", Proceedings of the 9th Technical Symposium on Computer Science Education, ACM, Pittsburgh, Pa., 1978, Vol. 10, No. 3.
(15) M. Vetter, "Principles of Database Systems", Proceedings of the International Computing Symposium, 1977, ACM, North-Holland, Amsterdam, pp 555-580.
(16) N. Wirth, "The Programming Language Pascal", *Acta Inf.*, 35-63 (1971).