IDC-Inorganic Chemicals Data Base. Processing and Storage of Information in the Field of Inorganic Chemistry

HORST ROSCHKOWSKI* and WALTER SIMMLER

Chemie-Information und -Dokumentation Berlin, Gesellschaft Deutscher Chemiker, 1000 Berlin 12, West Germany, and Umweltschutz/Entwicklung und Information, Bayer AG, 5090 Leverkusen, West Germany

Received February 8, 1980

This paper describes the input system for the IDC Internationale Dokumentationsgesellschaft für Chemie mbH Fachinformationszentrum Chemie (German Chemistry Information Center) documentation system for inorganic chemistry. The system allows the storage and retrieval of bibliographic data, inorganic compounds and their reactions, and concepts as derived from papers and patents. Index terms are completed by text modifications, inorganic compounds are indexed as they appear in the literature ($Ca_3(PO_4)_2$ etc.), and symbols of the elements are transformed by computer into artificial symbols which show the logical relationship of the elements within the Periodic System and ease the searching for classes of compounds.

Worldwide, inorganic compounds play an important role in chemical research and production of chemicals. At present, about 40% of the 500 000 publications covered by Chemical Abstracts Service (CAS) annually are of interest to inorganic chemists. More than half of the publications in the field of organic chemistry contains information on inorganic compounds.[1]

In a system which deals with inorganic chemistry, certain peculiarities have to be taken into consideration; for example, in many cases inorganic compounds cannot—in contrast to low-molecular organic compounds—be described by stoichiometric formulas. Especially in inorganic solid-state chemistry, general terms, such as "phase", "system", "doped material", "alloy", and "nonstoichiometric compounds", are used. Often, solid solutions with ranges of homogeneity (e.g., $Zn_xCd_{1-x}S$; $0.15 < x < 0.85$) are mentioned in the literature. Frequently compounds with variable or without defined content of water of crystallization are described. Depending upon the place of occurrence, the composition of minerals may vary considerably. Even synthetic minerals may show differences in their composition.

Furthermore, nonstructural information, e.g., physical and chemical properties or applications, is of high value in inorganic chemistry documentation systems. Mostly a combination of nonstructural with structural information is of importance. A documentation system for inorganic chemistry has to answer questions asking not only for defined and unique compounds but also for compound classes (alkaline earth metal compounds, peroxodisulfates, etc.). It also has to consider peculiarities from fringe areas such as ceramics, metallurgy, etc., and has to solve all the above-mentioned storage and retrieval problems.

A documentation system meeting these requirements was developed by Chemie-Information und -Dokumentation Berlin about 10 years ago and is now used in a modified form by the IDC Internationale Dokumentationsgesellschaft für Chemie mbH[2] to build up a data base for inorganic chemistry. The user-oriented system which handles bibliographic data of patents and papers, inorganic substances (defined compounds, compound classes, fragments), and reactions as well as concepts has been in operation since October 1973 at the Bayer AG[3-5] following a two-year experimental period at the Chemischer Informationsdienst.[6] In addition to index terms, text modifications or explaining contexts are put in the machine-readable file. These contexts are made available to searchers or customers of the documentation system by printout or at terminal screens and enable the questioner to decide whether an original publication has to be consulted or not.

Another special feature of the system is the introduction of artificial but systematic symbols for chemical elements which reflect the relationship of the elements within the periodic system.

Since 1973 patent information derived from abstracts of the Central Patents Index (CPI) is fed routinely into the IDC-Inorganic data base. Starting with the CAS tape service CA Search Volume 89 (1979), information from other types of publications (preferably papers) is also covered. By intellectual analysis of the corresponding CA abstracts, the CAS data base is supplemented.[7] Display screen units are used for processing the input data. Online input is under consideration.

INPUT TO THE IDC-INORGANIC DATA BASE

1. Bibliographic Data. The bibliographic data of the abstracts as well as that of the original document are recorded. The data describing the abstract are as follows:

- year or volume of the secondary service
- source (e.g., CA = *Chemical Abstracts*)
- abstract number or accession number
- CA section and subsection number.

The bibliographic data of primary publications consist of the code identifying the type of publication (journal, thesis, report, monograph), the ASTM-CODEN, author names, abbreviated title of the publication, volume, year, issue, first and last page of the paper, and title of the paper or document.

The citation of patents includes the following data covered by CPI:

- country code
- patent number
- application number
- patentee code
- international patent classification
- type of publication ("P" for patent)
- first date of application
- country code of the first application
- priority number
- applicant or patentee name
- title of the patent.

The above bibliographic data are taken from available files (Derwent SDI Tape, CA SEARCH) and transferred to the IDC-Inorganic data base. For special cases methods have been worked out which allow the input of additional (or corrected) data.

* Author to whom correspondence should be sent at the Berlin address.

0095-2338/80/1620-0154$01.00/0

Table I. Examples of the Classification System

System Number	concept in free text
	E. Chemische Thermodynamik, Gleichgewichte (chemical thermodynamics, equilibria)
E 1000.9	Allgemeine Thermodynamik (general thermodynamics)
E 2000.F	Gase, Flüssigkeiten, Suprafluidität (gases, fluids, suprafluidity)
E 3000.K	Thermodynamische Funktionen, Thermochemie (thermodynamic functions, thermochemistry)
E 4000.P	Gleichgewichte (equilibria)
	F. Elektrochemie (electrochemistry)
F 1000.B	Theoretische Elektrochemie (theoretical electrochemistry)
F 2000.G	Elektrolyte, Elektrolytische Leitfähigkeit (electrolytes, electrolytic conductivity)
F 3000.L	Potentiale, Ketten, Elemente (potentials, cells, elements)
F 4000.Q	Elektrolyse (electrolysis)
F 5000.W	Polarisation, Elektrodenvorgänge (polarization, electrode processes)
F 6000.0	Polarographie (polarography)
F 7000.4	Korrosion (corrosion)
	G. Kolloidchemie, Grenzflächen (colloidal chemistry, interfaces)
G 1000.C	Theorie (theory)
G 2000.H	Kolloide (colloids)
G 3000.M	Gele (gels)

2. Input of Concepts. For the input of concepts (nonstructural index terms), the following kinds of codes for concepts have been introduced:

- codes for terms of a hierarchical ordered classification system (System Numbers)
- codes for concept headings or index terms of a controlled vocabulary (Alpha Numbers)
- codes for general concepts which are used as links in combination with substance entries (SV Codes).

Each substance entry may or (in some cases) must be supplemented by phrases or text modifications which describe the substances more precisely and which can be applied to free text searching.

The System Numbers are also used by the German abstracting service "Chemischer Informationsdienst" (Chemical Information Service) for the mechanized arrangement of abstracts in its weekly issues.[8] In 1970 and 1971, the Alpha Numbers were used to produce the issue indexes of ChemInform mechanically.[9]

The approximately 90 System Numbers of the Classification System (see Table I) serve to characterize broader subject fields (e.g., electrical properties). The System Numbers assigned to a document reflect its essential content and are a useful means to perform searches for broader topics.

Alpha Numbers characterize commonly used, well-defined, specific concepts belonging to the scope of the system. They reflect the content of a document and are not related to individual substances mentioned in a document or abstract. Each Alpha Number is accompanied by a precise text which may appear on the printouts. In most of the cases the context enables the searcher to decide without consulting the abstract or even the document whether a reference retrieved is of relevance or not. The list of Alpha Numbers comprises about 1500 terms (see Table II).

Not all of the concepts mentioned in the list have their individual Alpha Number (second column). For example, the term BRILLOUIN ZONE belongs to the broader term ELECTRON STRUCTURE which has the Alpha Number E1740.7 (first column). Behind the terms separated by "+" are the codes for the general concept. The terms of the list meet the requirements of inorganic chemistry and are a subset of the IDC Thesaurus.[10] General Concept Codes (SV) are

Table II. Section of the List of Alpha Numbers[a]

S	S	Concept + General Concept Code
K0570.Z		FUELS (F. NUCLEAR REACTORS) S.U. NUCLEAR FUELS + T3
T5730.W		FUELS (F. ROCKETS) S.U. PROPELLANTS + T1
B6510.D	B6510.D	FUEL CELLS, ELECTROCHEMISTRY + F5 T4
B6530.J	B6530.J	BRIQUETTING + T1
E1740.7		BRILLOUIN ZONE S.U. ELECTRON STRUCTURE + D9
H0850.C		BROMINATION S.U. HALOGENATION + R1
B7020.3	B7020.3	BROWNIAN MOTION + E9
F1000.9		CRACK FORMATION S.U. STRENGTH + D4
C1530.X	C1530.X	CELLOPHANE + S5
C1550.1	C1550.1	CELLULOSE + S2
C1680.B	C1680.B	CERENKOV RADIATION + D7
W0240.J		CHARACTERISTIC TEMPERATURE S.U. HEAT, SPECIFIC + E2
K2380.3		CHARGE TRANSFER COMPLEXES S.U. COMPLEX COMPOUNDS + S2

[a] S = Alpha Number, F. = for, S.U. = see under.

with some exceptions assigned to main headings which are to be considered as broader terms for a group of terms of the Alpha Number list. Thus, SV "D3" links special terms related to the structure of solids like "dislocations", "grain boundaries", "crystal morphology", etc. A few SV stand for terms which do not correspond to any term in the Alpha Number list, e.g., 22 = preparation, 81 = analytically identified substance. Each indexed individual substance is linked with at least one SV. Since the SV may be considered to be broader terms, they connect in a certain sense substance entries with Alpha Numbers. Moreover, SV decrease the amount of "noise of searches" remarkably.

In the following cases, searchable free text entries are attached to substance entries (their presence is compulsory):

- specifications for modifications (rhombic, α-modification, etc.)
- stereochemical descriptors (cis, trans)
- deviations from the normal physical state of matter (e.g., liquid NH_3)
- mineral names (faujasite)
- trade names (V2A steel, Pyrex glass)
- explanation of symbols used to encode substance classes (see 3.)
- compounds with unknown stoichiometry (see 3.)
- compounds with undefined content of water of crystallization (e.g., $Na_2CO_3 \cdot xH_2O$, x undefined).

Substance entries may also be supplemented by the yield and important information not sufficiently enough covered by the various concept codes.

3. Indexing of Inorganic Substances. By definition, all elements and noncarbon compounds are inorganic substances. Additionally, the following carbon containing substances are considered to be inorganic (* with the exception of those containing organic cations):

carbides
carbonates*
pseudohalides*
CO, CO_2, H_2CO_3
carbon containing alloys

In the IDC Inorganic System, inorganic substances (e.g., Na_2SO_4) are indexed in such a manner that they can be retrieved:

- either as defined compound (Na_2SO_4)
- or as compound class (alkali metal sulfates)
- or as fragment (sulfate)

Table III. Artificial Symbols for Elements and Groups of Elements

Artificial symbols for elements	
symbol for any element	QZ
symbol for any metal	QX
symbol for any nonmetal	QY
symbol for organic radical	QT
Symbols for the groups of the periodic system:	
1. main group	QA (H, Li, Na, K, Rb, Cs, Fr)
2. main group	QB (Be, Mg, Ca, Sr, Ba, Ra)
3. main group	QG (B, Al, Ga, In, Tl)
4. main group	QH (C, Si, Ge, Sn, Pb)
5. main group	QO (N, P, As, Sb, Bi)
6. main group	QP (O, S, Se, Te, Po)
7. main group	QQ (F, Cl, Br, J, At)
8. main group	QS (He, Ne, Ar, Kr, Xe, Rn)
1. subgroup	QN (Cu, Ag, Au)
2. subgroup	QC (Zn, Cd, Hg)
3. subgroup	QD (Sc, Y, La, Ac)
4. subgroup	QI (Ti, Zr, Hf)
5. subgroup	QJ (V, Nb, Ta)
6. subgroup	QK (Cr, Mo, W)
7. subgroup	QL (Mn, Tc, Re)
8. subgroup	QM (Fe, Co, Ni, Ru, Rh, Pd, Os, Ir, Pt)
lanthanides	QE (Ce, Pr, Nd, Pm, Sm, Eu, Gd, Tb, Dy, Ho, Er, Tm, Yb, Lu)
actinides	QF (Th, Pa, U, Np, Pu, Am, Cm, Bk, Cf, Es, Fm, Md, No, Lr)

For that purpose each inorganic compound is registered by its linear structural formula (STRUFO) from which per program a standardized and encoded molecular formula is derived. The linear structural formula is the complete formula with groups of atoms in which the subscripts are in line, whereby index 1 is not omitted and a blank or bracket precedes a symbol of an element (e.g., $(NH_4)_2S_2O_8$ transformed into (N1 H4)2 S2 O8). The standardized notation is required to allow searches for fragments such as the peroxodisulfate group.

There also exists an algorithm to generate a range of figures for compounds with fractional indexes or with inhomogeneities. If there are no specifications, elements are indexed with their normal symbols and the index zero (e.g., S_0 = sulfur vs. S_8 = octasulfur). As mentioned before, other specifications, e.g., modifications, are registered in the context.

Ions are indexed as follows:

positive charge, known index: (1+)...(99+)
positive charge, unknown index: (0+)
negative charge, known index: (1−)...(99−)
unknown charge, unknown index: (1+−)...(99+−), etc.

If the information on the charges in the paper is vague, the entry has to be supplemented by an explaining text (e.g., NØ (O+−) − NØ − ions).

Compounds with known stoichiometry are indexed with their structural formula. The order of the element symbols strictly follows the common chemical presentation of compounds (for example, NH_3 not H_3N). Thus, a character-by-character search yields correct results. Hydrates are indexed in the usual writing manner (e.g., $Na_2CO_3 \cdot H_2O$).

The sign "period" operates as program instruction which terminates the calculation of the coded and standardized internal molecular formula. This permits one to search for all hydrate forms of a given compound using a single molecular formula as a general search term. The search for individual hydrates is also possible.

In contrast to water of crystallization, aqua groups in complexes are included in the calculated molecular formula. Since the sign "period" stops the inclusion of elements behind the period, the use of this sign is allowed only in the case of hydrates. Consequently, $(CaO)_3(Al_2O_3)_2$ and not $3CaO \cdot 2Al_2O_3$ is written. For the codification of compounds with unknown or not exactly known composition, artificial element symbols (Table III) and the subscript zero are used. Thus, the compound class, alkaline earth halides, may be indexed as QB QQ_2 in which QB stands for alkaline earth metals and QQ for halogens. Manganese oxide is indexed as Mn_0O_0. Such kind of entries always are supplemented by contexts.

When stoichiometric and nonstoichiometric parts are found in one formula, the defined part will be described as such, whereas in the undefined part, the index zero has to be used: $(Sr,Ba)(TiO_3)_2$ is transformed into $Sr_0Ba_0(TiO_3)_2$. If a paper reports on groups of elements for which an artificial symbol does not exist, the symbol of the superior element group is used and supplemented by a specifying context; noble metal is indexed: QX_0 (for metal)−noble metal.

Formulas with nonintegral indexes may be handled as decimal fractions. Two digits before and three digits behind the decimal are admissible, whereas the subscript 0.001 stands for ≤0.001 and 99 for ≥99. Figures and even ranges of figures (e.g., $FeO_{0.95}$ or $Zn_{0.2-0.5}Cd_{0.5-0.8}S$) may be used for searches. Instead of decimal numbers, specifications in weight percentages are also accepted by the system. A distinctive mark is introduced to differentiate between the decimal numbers and weight percentage figures. This system-inherent peculiarity allows us to store and retrieve alloys described in variable form. For example, the "alloy with 55.555 wt % Fe, 30–35.445 wt % Ni, 15–20 wt % Cr" is indexed as follows:

$$KFe_{55.555}Ni_{30-35.455}Cr_{15-20}$$

whereas K is the code for alloy (figures in weight percentage).

Table IV. List of Marks for Special Substance Classes

D = doped compounds
E = doped compounds (composition in weight percentage)
G = glasses
F = glasses (composition in weight percentage)
L = alloys, intermetallic compounds
K = alloys, intermetallic compounds (weight percentage)
I = isotopically labeled compounds
H = isotopically labeled compounds (weight percentage)
M = modification, mineral, rocks
N = modification, mineral, rocks (weight percentage)
R = radicals
Q = radicals (weight percentage)
S = systems
T = systems (weight percentage)
V = compound (composition in weight percentage)

MARKS FOR SPECIAL SUBSTANCE CLASSES

Special substance classes such as doped compounds, glasses, alloys, isotopes and isotopically labeled compounds, radicals, minerals, and systems are earmarked by a special letter (see Table IV). Each substance class mentioned above has two marks: one for indexes (decimal figures or unknown stoichiometric composition) and one for specifications in weight percentages.

Isotopes are identified in the text as corresponding to the index entry for isotopes and isotopically labeled compounds. "^{60}Co" is transformed to "I Co_0 − Cobalt 60".

Mark D and E are reserved for doped as well as doping substances. Dope means trace impurities in the ppm range introduced into ultrapure crystals to obtain substances with special physical properties, e.g., lasers, semiconductors. The impurity is stated in the context (doped with ...).

Radicals and glasses are not only indexed as substance class but also with special Alpha Numbers.

Systems are indexed with their components. The components are described in the context.

H_2O stands only for the compound "water". If concepts such as process water, drinking water, cooling water, natural water, and waste water have to be indexed, only the corre-

Table V. Element Codes for Chemical Elements

	Metals															Nonmetals									
A	--	Li	Na	K	--	--	Rb	--	--	Cs	--	--	Fr	--	--	QA	H	--	--	--	--	--	--	--	--	1. main group
B	--	Be	Mg	Ca	--	--	Sr	--	--	Ba	--	--	Ra	--	--	QB	--	--	--	--	--	--	--	--	--	2. " "
C	--	--	--	Zr	--	--	Cd	--	--	Hg	--	--	--	--	--	QC	--	--	--	--	--	--	--	--	--	2. subgroup
D	--	--	--	Sc	--	--	Y	--	--	La	--	--	Ac	--	--	QD	--	--	--	--	--	--	--	--	--	3. " "
E	--	Ce	Pr	Nd	Pm	Sm	Eu	Gd	Tb	Dy	Ho	Er	Tm	Yb	Lu	QE	--	--	--	--	--	--	--	--	--	lanthanides
F	--	Th	Pa	U	Np	Pu	Am	Cm	Bk	Cf	Es	Fm	Md	No	Lr	QF	--	--	--	--	--	--	--	--	--	actinides
G	--	--	Al	Ga	--	--	In	--	--	Tl	--	--	--	--	--	--	--	B	--	--	--	--	--	--	QG	3. main group
H	--	--	--	Ge	--	--	Sn	--	--	Pb	--	--	--	--	--	--	--	C	Si	--	--	--	--	--	QH	4. " "
I	--	--	--	Ti	--	--	Zr	--	--	Hf	--	--	--	--	--	QI	--	--	--	--	--	--	--	--	--	4. subgroup
J	--	--	--	V	--	--	Nb	--	--	Ta	--	--	--	--	--	QJ	--	--	--	--	--	--	--	--	--	5. " "
K	--	--	--	Cr	--	--	Mo	--	--	W	--	--	--	--	--	QK	--	--	--	--	--	--	--	--	--	6. " "
L	--	--	--	Mn	--	--	Tc	--	--	Re	--	--	--	--	--	QL	--	--	--	--	--	--	--	--	--	7. " "
M	--	--	--	Fe	Co	Ni	Ru	Rh	Pd	Øs	Ir	Pt	--	--	--	QM	--	--	--	--	--	--	--	--	--	8. " "
N	--	--	--	Cu	--	--	Ag	--	--	Au	--	--	--	--	--	QN	--	--	--	--	--	--	--	--	--	1. " "
Ø	--	--	--	--	--	--	--	--	--	Bi	--	--	--	--	--	--	--	N	P	As	Sb	--	--	--	QØ	5. main group
P	--	--	--	--	--	--	--	--	--	Po	--	--	--	--	--	--	--	Ø	S	Se	Te	--	--	--	QP	6. " "
Q	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	F	Cl	Br	J	At	QQ	--	--	7. " "
R	(-)	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	(+)	(+-)	charges
S	--	He	Ne	Ar	--	--	Kr	--	--	Xe	--	--	Rn	--	--	--	--	--	--	--	--	--	--	--	QS	8. main group
Z	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	QT	QX	--	--	--	--	--	QY	--	QZ	unknown
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	Ø	P	Q	R	S	T	U	V	W	X	Z	Z = unknown

sponding Alpha Number is used. Substance terms which cannot be described by formulas, e.g., salts, bases, complex compound, air, glass, and ceramics, are encoded with Alpha Numbers. This method is applied also to some undefined organic compounds such as asphalt, bitumen, mineral pitch, and fats when mentioned in a paper in connection with an inorganic compound.

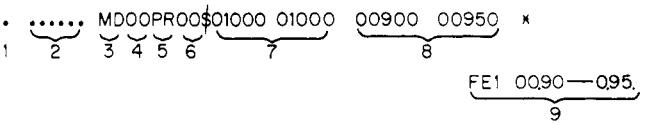
Yet, in most of the cases, partially defined compounds may be coded with the help of the symbol QZ which stands for undefined element(s). Examples are: sulfidic Cu ores = $Cu_0S_0QZ_0$, phosphate rock = $P_0O_0QZ_0$. Of course, in such cases the formulas are supplemented by a context, too.

MECHANIZED ENCODING OF SUBSTANCES. THE STANDARDIZED AND CODED MOLECULAR FORMULA

In the first step, the linear structural formula (STRUFO) is derived from the indexed formula. In the second step, the computer calculates, based on the STRUFO, the standardized and coded molecular formula (SUFO) which is stored in addition to the STRUFO. In the SUFO, all international element symbols are translated into two digit codes (see Table V). Thus, all elements of a certain group of the Periodic System have codes in which the first digit letter is identical (*A*P = alkali metal, *A*B = lithium, *A*C = sodium, *A*D = potassium, etc.). The advantage of such artificial symbols consists in the possibility to search for all elements of a distinct group of the Periodic System in one profile with only one term, e.g., for alkali metals the term consists of "*A*." ("period" stands for any letter). The reply refers to the relevant individual compounds as well as to the relevant compound classes.

In order to calculate a SUFO, all expressions in parentheses within the STRUFO are multiplied by the indexes. The two-digit element codes together with their indexes are ordered alphabetically within the SUFO. Each index consists of two digits (from 00 to 99). If at least one uncertain index or a range of indexes or a decimal fraction appears in the STRUFO, all indexes within a SUFO are converted to zero. In such cases a subsequent set of data is created, automatically, in which the complete index specifications are listed following the order of the element codes in the SUFO. The $ sign indicates the beginning of the data set.

Thus, the index entry Fe $O_{0.90-0.95}$ will appear in the data file as follows:



1 = one digit for the mark code

2 = six digits for three general concept codes (each two digits)

3 = MD is the element code for iron

4 = index for iron

5 = PR is the element code for oxygen

6 = index for oxygen

7 = retrievable index field for Fe

8 = retrievable index field for O

9 = linear structural formula

Although this kind of codification is relatively space consuming, it is nevertheless very economical because it is created mechanically and allows very precise searches for completely described compounds, compounds with nonstoichiometric composition, compounds in which some components are not described precisely, and well-defined structural fragments or groups. The type of search strategy to be employed depends upon the search system established. In another paper the search systems and strategies will be described.

REACTIONS

Inorganic reactions are recorded by encoding the starting materials, intermediates, and the products which are earmarked as such by the concept codes. Additionally, in many

cases the type of reaction, e.g., oxidation, reduction, halogenation, etc., is indexed per Alpha Number. But it should be mentioned that there is no syntax between the various partners of a reaction.

ACKNOWLEDGMENT

Thanks are due to the German Ministry of Research and Technology for supporting developmental work regarding the application of the system of journal literature.

REFERENCES AND NOTES

(1) "Gebietsbeschreibung Anorganika", 2nd ed., IDC Internationale Dokumentationsgesellschaft für Chemie mbH Fachinformationszentrum Chemie, 6000 Frankfurt 90, Germany, March 1979, p 3.
(2) M. Isenberg and H. Kaltenhäuser, "Systematische Speicherung anorganischer Stoffe auf dem GREMAS-Ergänzungsband", Unveröffentlichter Bericht der IDC Internationale Dokumentationsgesellschaft für Chemie mbH, 6000 Frankfurt 90, West Germany, 1969.
(3) W. Simmler, "Wissenschaftlich-technische Dokumentation in der chemischen Industrie—ein Beitrag zum Umweltschutz", In Deutscher Dokumentartag, 1976 (publiziert 1977), Verlag Dokumentation München, pp 164-77.
(4) "Information über die anorganisch-chemische Literatur", *Nachr. Chem. Tech.*, **23**, 524-5 (1975).
(5) H. Grünewald, "Information und Dokumentation auf dem Gebiet der anorganischen Chemie", *Nachr. Dok.*, **28**, 19-25 (1977).
(6) "Magnetbandsystem Anorganische Chemie", *Nachr. Chem. Tech.*, **19**, 182-3 (1971); *Nachr. Dok.*, **22**, 172-3 (1971).
(7) F. Ehrhardt, "Utilization of Chemical Abstracts Service (CAS) Data Bases. Application for the IDC-Inorganica-Documentationsystem", Bundesministerium für Forschung und Technologie, Report BMFT-FB ID 79-03, 1979, p 53.
(8) C. Weiske, "Das Klassifikationssystem des Chemischen Informationsdienstes", DK-Mitt, **21**, 11-5 (1977).
(9) C. Weiske, "Chemical Information Services", *Angew. Chem., Int. Ed. Engl.*, **9**, 550-5 (1970).
(10) E. Meyer, R. Jansen, and E. Sens, "Das IDC-Thesaurus-System", *Nachr. Dok.*, **23**, 203-11 (1972).

Interactive Simulation of Infrared, Mass, and ^{13}C NMR Spectra

M. RAZINGER,* J. ZUPAN, M. PENCA, and B. BARLIČ

Boris Kidrič Institute of Chemistry, 61001 Ljubljana, P.O.B. 380, Yugoslavia

Received October 24, 1979

As a special part of a combined chemical information system,[3] the update of spectral data files is presented and discussed. The amending of infrared spectra in general is much more complicated than that of the MS or ^{13}C NMR ones. The problems of developing an interactive simulation (updating) of IR spectra to make them as similar to the experimental ones as possible are elaborated and solutions suggested. The organization of records in spectral data files and data flow during the update procedure are presented in detail.

INTRODUCTION

Chemical information systems are quite numerous today.[1] They differ in the number of spectrometries upon which they are based, in the extent and quality of the spectral data bases, in the possibility for interactive work, etc. However, one thing is common to all: the quality of answers delivered depends strongly upon the quality of the spectral data on which they are based.[2] The flexibility of manipulation, update, and transformation of the data bases is quite a significant factor in the quality of the whole system.

As the main features of our combined chemical information system are presented elsewhere,[3,4] in this paper mainly the update and manipulation of spectra in the data files and also some aspects of graphical representation of spectral data are discussed. One of the main tasks in constructing the updating procedure was to make possible the manipulation of spectral data with simple instructions. Thus the user can easily modify the spectra already in the data banks, add new ones, or form new files.

Only by constant checking and improving of the quality of the data in the spectral files can we obtain better graphical representation of spectra and better performance of the information system.

GRAPHICAL REPRESENTATION OF SPECTRA

Although the most vital information for the successful operation of an automatic spectral search system and for the graphical representation of spectral data files is the position of peaks in the spectrum, it is by no means the only important information. For the mass and most ^{13}C NMR spectra the other important information needed besides peak position is peak intensity. Spectra of this type can be represented as a set of discrete lines of known positions and lengths which are adequate for obtaining quite useful approximations of the real spectrum. In infrared spectrometry the situation is quite different. Taking a look at an IR spectrum, one can hardly say that the peak positions and intensities completely define it. Band shapes are helpful in recognizing particular group motions, and the judgment of an experienced spectroscopist relies strongly on them.

The best way of solving the problem of adequate representation of real IR spectra is to digitize them.[2,5] This process requires direct connection of the spectrograph to the computer, i.e., online data acquisition, digitalization of spectrograms, or cataloged spectra with a digitizer. Either way requires specialized instruments which are not commonly found in many laboratories, while minicomputers themselves are increasingly used.

Instead of storing the digitized values of the whole spectrographic curve in the spectral data files, only some main parameters of each peak were stored for our system of data files. Peak positions and intensities are stored for all three spectrometries. In the case of ^{13}C NMR spectra, carbon atom assignment number is associated with each peak as the additional parameter. For IR spectra, two additional parameters are stored. The first one is the half-width at half-intensity of the peak (HWHI), and the second parameter determines the shape of the peak in question. The value of this second parameter can be only 0 or 1, defining the function for calculation

0095-2338/80/1620-0158$01.00/0