Gatekeeper Functions for Domain Specific Information

Yuzuru Fujiwara
(Tsukuba University)

One of the most important roles of gatekeeper is to support smooth and intelligent transfer of information and it is quite useful if the function is operated automatically by intermediary systems.
This paper describes such systems implemented, most of which run on personal computers with CD-ROM drivers handling large amount of data and knowledge necessary to the sophisticated systems. Three examples are shown: the first one is interchange of chemical information, the second one is a multilingual dictionary to interpret key words twelve languages, and the third one is use of thesaurus which was antomatically compiled by using information in the database.

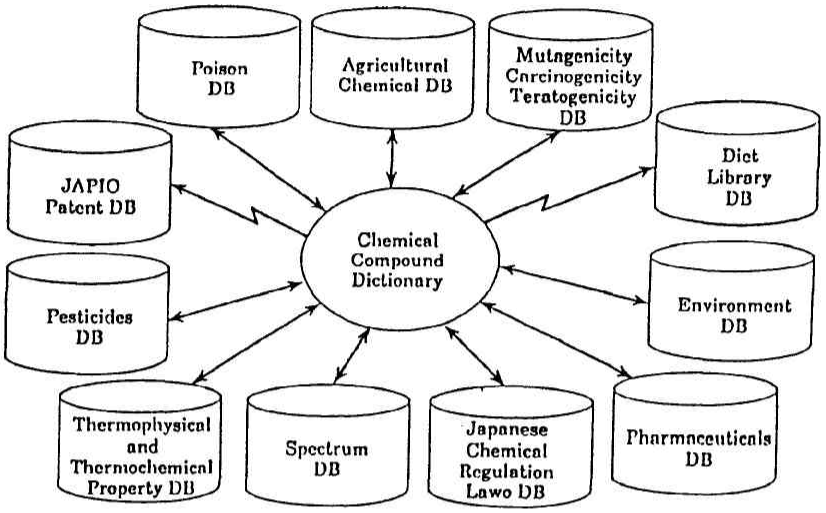
1. Introduction

Since chemical information is used widely in many disciplinary fields,in industries, in business, as well as in daily fife, there are many different representation of chemicals, such as names, structures, substructures graphical display and son on, conversion of data of chemical structures is an example of a gate keeper in the field of chemistry. Next example is a multilingual dictionary which is of wider use. The other one is a thesaurus compiled automatically on the basis of internal information of the database concerning polymer.

2. A Data Conversion System for Distributed DB Systems

Science and technology agency supported a national project to develop integrated chemical database from 1980 to 1985, and ten governmental institutions joined the project as shown in Fig.1. All database systems are designed to be searched in independently and to be navigated via the central chemical compound dictionary from other database upon user's request.

Fig. 1 Chemical Databeses Developed in Japan



The contents of each database and the responsible institutions are shown in Table 1. These databases have chemical names, structural information in common, where as they have different data concerning chemical, physical, and biological properties according to the purposes and the use of the databases. Therefore, the database are maintained separately by responsible institutions as is the case of construction, and hence the systems are scattered geographically and distributed DB system

Table 1 Chemical Database in Japan

Database	Abbreviation	Producing Organization	Main Data Elements Included
Chemical Substance Database	DC	The Japan Information Center of Science and Technology; Japan Association for International Chemical Information	Substance names, structures, database locators
Agricultural Chemical Substances Database	BC	National Food Research Institute. The ministry of Agriculture. Forestry and Fisheries	Biological, Chemical, physical and industrial data of fertilizers, food additives, enzymes and so on
Mutagenicity, Carcinogenicity, Teratogenicity Database	BL	National Institute of Hygienic Sciences	Test data of chemical substances such as mutagenicity, carcinogenicity and teratogenicity
Environment Database	EN	National Institute for Environmental Studies	Analytical methods and data and physical and chemical property data of environmental chemicals
Pharmaceuticals Database	PH	Japan Pharmaceutical Information Center	Name (trade name, common name, abbreviated name), Ingredient, composition, usage, manufacturer and distributer of drug
Japanese Chemical Regulation Laws Database	SF	Japan Chemical Industry Ecology-Toxicology and Information Center	Regulatory and legal information of Japan for potentially hazardous chemical substances
Spectrum Database	SP	National Chemical Laboratory for Industry, Board of Industrial Technology	IR, ^{1}H-NMR, and ^{13}C-NMR spectral data of fundamental standard substances
Thermophysical and Thermochemical Property Database	TH	The Japan Information Center of Science and Technology	Thermophysical and thermochemical data of substances in three or less component systems
Pesticides Database	PE	Japan Agricultural Chemical Industry Association	Active ingredient, usage, formulation type, dosage, production and shipment data of pesticides
Poison Database	TX	Japan Pharmaceutical Information Center	Ingredient, toxicity, clinical effect, treatment case report of poisoning for commercially available substances which may cause acute poisoning

The chemical compound dictionary supported a system named STARS (stereo chemically accurate register systems) which serve as a gatekeeper in the distributed database systems. The features of STARS are shown in Table 2 and it is to be noticed that it has a functions to convert chemical names to connection Tables and to generate stereo graphics of chemical compounds

Table 2 Functions of STARS

1. Registration and Management of Compounds
2. Automatic Construction of Connection Tables
3. Generation of 3D Chemical Structures
4. Referential Records to Chemical DB Linked by Networks

Names of chemical compounds are given under the direction of the IUPAC nomenclature standards and they are systematic. This seems to mean that analysis of chemical names is straightforward. However, the IUPAC names are not unique and includes confusion inside. Therefore, the procedure of analysis is not simple as shown in the Table 3 and the following example.

Table 3 General Procedure of Stereochemically Accurate Registry System of Substances (STARS)

1. Morphological Analysis
   - 1.1 Dictionary Matching
   - 1.2 Identify Locants
2. Separation of Compound Words
3. Certification of Ambiguous Structures
   - 3.1 Deciding Positions of Multiple Bond
   - 3.2 Deciding Positions of Free Valent Atomes in Substituenes
4. Formation of Spiro Structures
5. Processing by Special Function Names
   - 1) Bi, Ter, Quater, 2) Homo 3) Nor 4) Seco
   - 5) Cyclo 6) Replacement of Atomes in Skeleton
   - 7) Indicated Hydrogen : H 8) Hydro 9) Dehydro
   - 10) Hydrogen : Hydride 11) De, Des
   - 12) Anhydro 13) Bridge
6. Isotope Labelling

7. Processing of Suffices

   1) ene, yne 2) yl 3) ylidene 4) ylidyne 5) ylene
   6) ium, ylium, cation, ylio, ide, ylide, ion, anion, radical, etc

8. Connection of Skeleton with Functional groupes

9. Connection of Substituents together and with Skeletons

10. Assignning Stereodescriptors to CT

11. CT Check, Bond Check

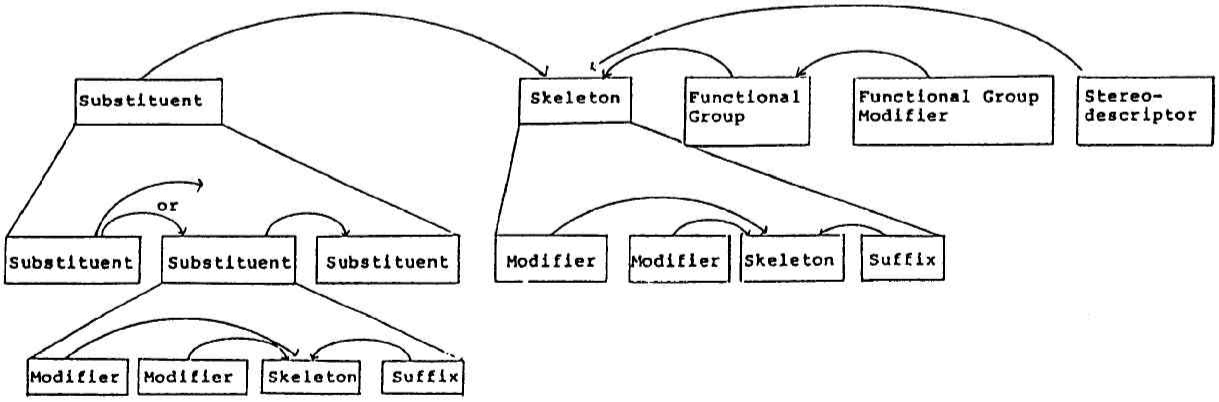
12. Aromatization, Tautomerization

13. Stereochemical Modification of CT

14. CT Standardization

IUPAC names consist of subnames corresponding to subcomponents, numbers of them, location of subcomponents placed and so on as shown in Fig.2. This is similar to composition of sentences consisting of words and the procedure of analysis is similar to analysis of sentences which is the first step of natural language translation.

Fig. 2 General Pattern of Systematic Name



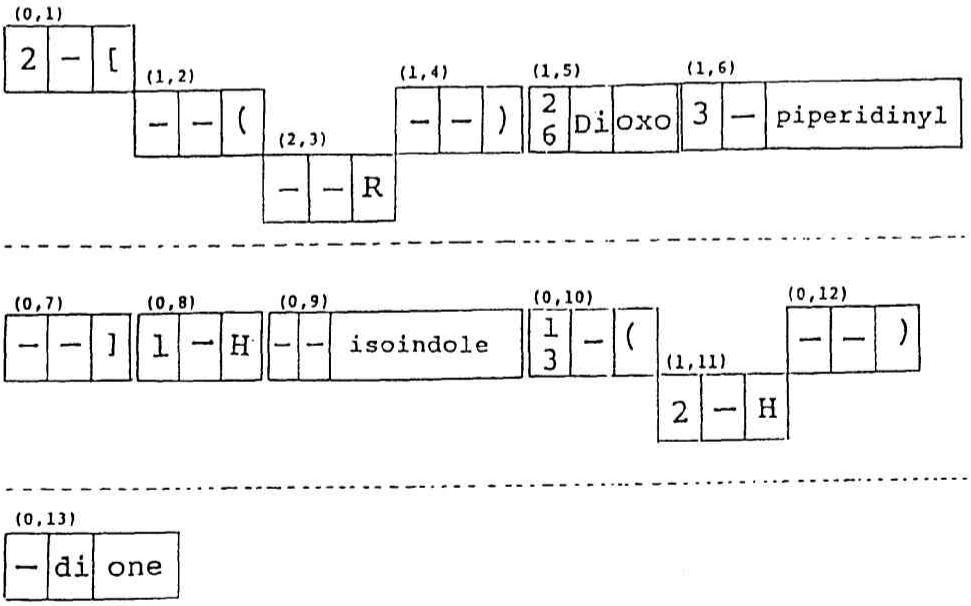
As analysis of chemical names resembles that of natural languages, the conversion systems requires an analysis dictionary, the contents of which is shown in the Table 4.

Table 4 Contents of Analysis Dictionary

Element Name	Subclass	Total Number	Examples
Skeletons	11	31,000	Methane, Ethane, Benzene Acetic acid, etc
Substituents	28	1,500	Methyl, Ethyl, Phenyl, Methlene Alanyl, Glucopyranosyl, etc
Functional Groups	11	1,000	-oic acid, carboxylic acid -ol, -al, nitril, etc
Suffices	3	20	ene, yne, yl, ylidene, ylidyne ium, cation, etc
Bridges	1	230	Methano, Ethano, Epozy, Epimino Benzeno, etc
Modifiers	10	1,150	Oxa, Aza, Siloxane, Homo, nor Cyclo, Seco, Lactone, Anhydride, etc
Multipliers	4	100	Di, Tri, Tetra, Bis, Tris, etc

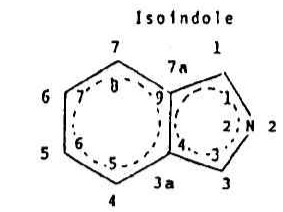
Fig.3 shows an example of morphological analysis of subcomponents in chemical names and the there are levels of composition according to the nested structures of names. This example is a case of English name and STARS can also handle Japanese name.

Fig. 3 Morphological Analysis (English)



When the morphological analysis of chemical names are complete, each subcomponent corresponding to a substructure in the compound in converted to a connection table representing the substructure, and the these subcomponents are connected to each other as designated by the locants and by other information, if any such as stereo chemical designated e.g. R in this particular example

Fig. 4 Results of Analysing Chemical Names / Generating the Connetion Table and Displaying the Structures graphically



Bond: Single = 1
Double = 2
Triple = 3
Aromatic = 1.3

Atom No	Locant 1	Locant 2	Et	CV	HN	St	Connectivity B_1	A_1	A_2	A_3	B_2
1	1		C	3	1			9 1.3	2 1.3		
2	2		N	3				1 1.3	3 1.3		
3	3		C	3	1			2 1.3	4 1.3		
4	3a		C	4				3 1.3	5 1.3	9 1.3	
5	4		C	3	1			4 1.3	6 1.3		
6	5		C	3	1			5 1.3	7 1.3		
7	6		C	3	1			6 1.3	8 1.3		
8	7		C	3	1			7 1.3	9 1.3		
9	7a		C	4				8 1.3	1	4 1.3	

Et: Element
CV: Connected Valance
HN: Number of Hydrogens
St: Stereodescriptor

Et: Element
CV: Connected Valance
HN: Number of Hydrogens
St: Stereodescriptor

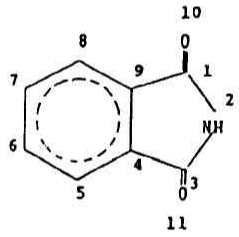
Atom No.	Locant 1	Locant 2	Et	CV	HN	St	Connectivity B_1	A_1	A_2	A_3	B_2
1	1		C	4				9 1	2 1	10 2	
2	2		N	3	0			1 1	3 1	14 1	
3	3		C	4				2 1	4 1	11 2	
4	3a		C	4				3 1	5 1.3	9 1.3	
5	4		C	3	1			4 1.3	6 1.3		
6	5		C	3	1			5 1.3	7 1.3		
7	6		C	3	1			6 1.3	8 1.3		
8	7		C	3	1			7 1.3	9 1.3		
9	7a		C	4				8 1.3	1 1	4 1.3	
10			C	2						1 2	
11			C	2						3 2	
12	1		N	2	1			17 1	13 1		
13	2		C	4	0			12 1	14 1	18 2	
14	3	*	C	3	1	R	2 1	13 1	15 1		
15	4		C	2	2			14 1	16 1		
16	5		C	2	2			15 1	17 1		
17	6		C	4	0			16 1	12 1	19 2	
18			C	2						13 2	
19			C	2						17 2	

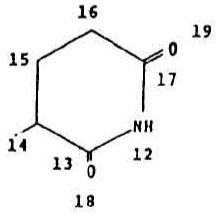
Isoindole Part

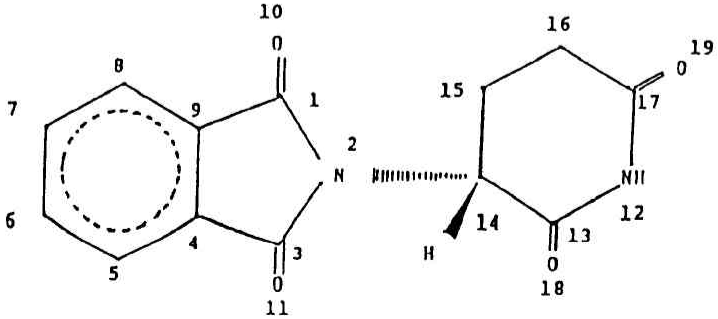
dione Part

Piperidinyl Part

Dioxo Part

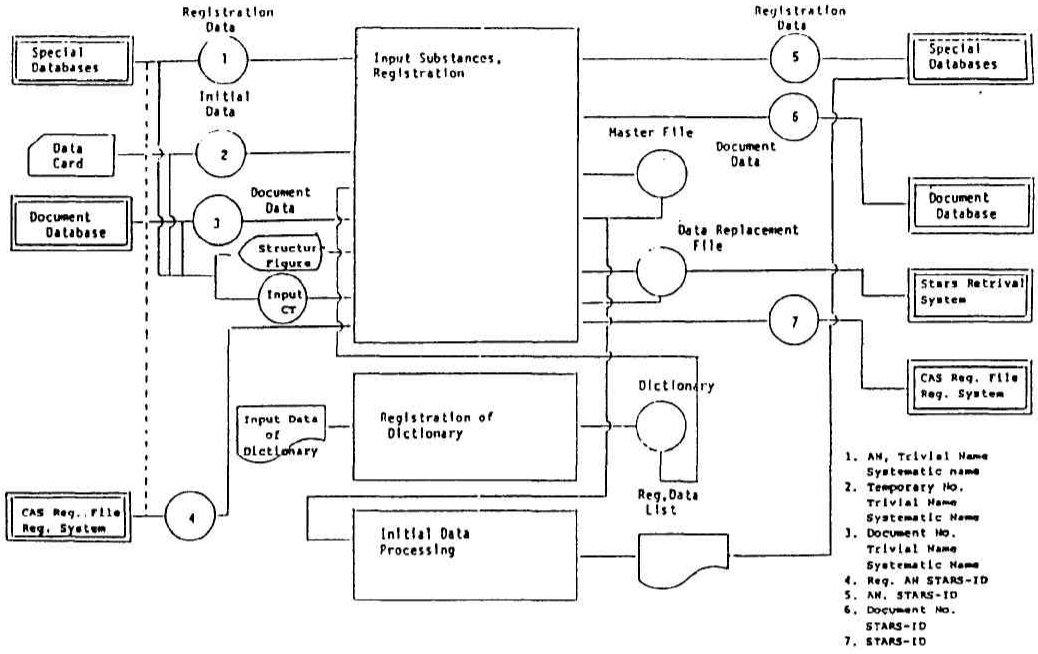






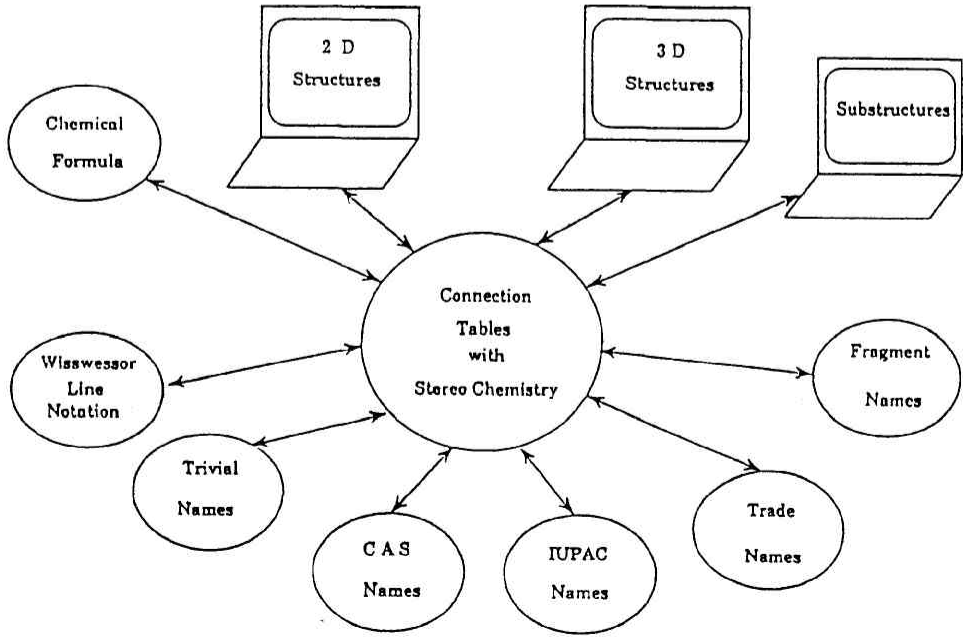
The system of STARS is considered as an expert system to convert linearly coded data i.e.names into two dimensional data as well as three dimensional one and hence it is a large complex system, the configuration of which is shown in the Fig.5.

Fig. 5 System Configuration of STARS



It is not easy to convert data of two or more dimensional data into those of linear dimension without loosing information. Chemical compounds are concrete entitiles of the reae world and have three dimensional structures. Therefore, interconversion among the various representation of chemical structures requires high level of expertise and possible systematic ways of conversion are shown in the Fig.6.

Fig. 6 Representation of Chemical Structures



If standardized names are given, the situation in much simpler than the actual case. On the contrary, there are many nomenclature systems allowed by IUPAC with some priority among the systems as shown in the Table 5. Although STARS set an standard naming criteria to avoid the confusion, it is designed to accept non-standard expression for flexibility and extensibility.

Table 5 IUPAC Nomenclature Systems

1. Substitutive Nomenclature
2. Additive Nomenclature
3. Subtractive Nomenclature
4. Radico functional Nomenclature
5. Conjunctive Nomenclature
6. Replacement Nomenclature
7. Nomenclature of Assemblies of Identical Units
8. Trivial Name
9. Semi - Trivial Name
10. Trade Name
11. Others

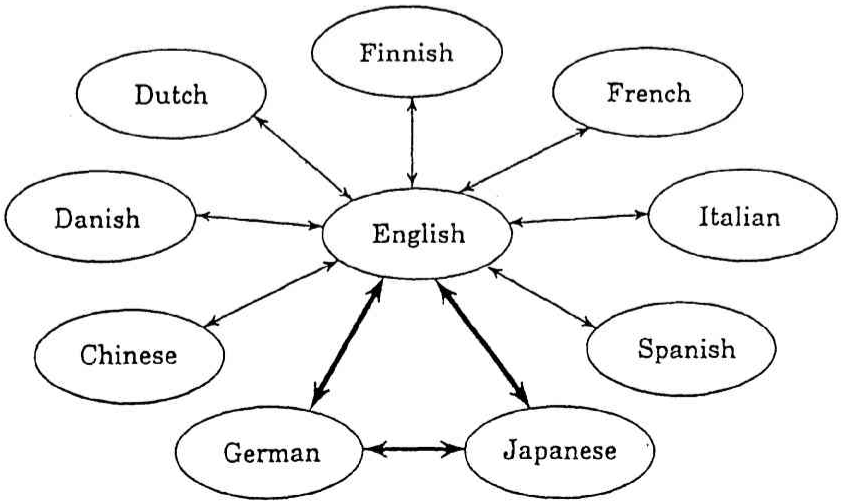
e.g. Source - Based - Nomenclature
cf. Sturcture - Based - Nomenclature
Mixtures
Generic Representation
Unknown Structures

3. A Multilingual Access System to Distributed DB Systems

Many databases contain information written in languages used in countries where the databases are produced. On the other hand, interpretation of technical terms are very difficult except specialists in relevant fields. This requires an appropriate intermediary to undertake this time consuming task and expertise demanded to access databases scattered all over the world.

The Fig.7 shows a translation pass among the twelve languages supported by a set of dictionaries on CD-ROM called CO-WORD published by Sanshu-Sha, Japan edited by the authors. The latest issue of CD-WORD contains dictionaries of twelve languages.

Fig. 7 Multilingual Keyword Conversion



4. The Thesaurus for the Polymer Database ; CAPDAS

Polymer are used in many advanced technological fields as well as in daily life such as clothes, houses, and foods. Therefore polymer materials are named in diversified ways according to their wide varieties of properties and according to their field in use. Good thesauri are necessary as one of user interfaces for most database systems.

One of the serious issues of thesauri for specific fields in to take a lot of man power of specialists in the fields for compilation and maintenance of thesauri.

A polymer database has been compiled by Japan High Polymer Center and the contents are shown in the Table 6.

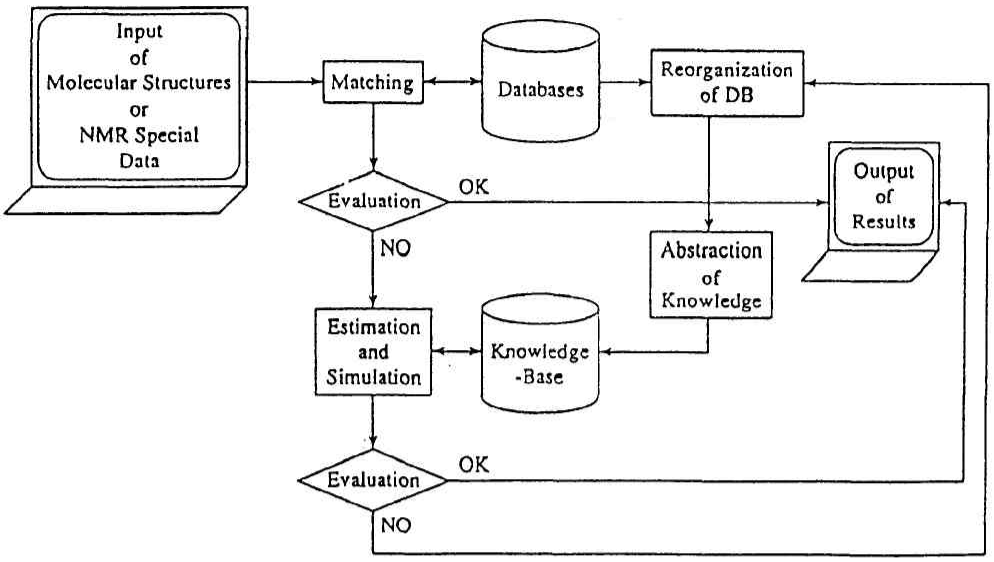
Table 6 Contents in CAPDAS

Code	Data	MB
NMP	Polymer Carbon NMR Papers	18.4
NMS	Polymer Carbon NMR Spectra	7.6
CTC	Technical Reports and Catalogs	4.9
PGB	Guide Book of Commercial Polymers	1.6
DIC	Polymer Vocaburary	0.3
PIC	Image Data in Catalogs	20.0
	Total	52.8

The CAPDAS system has the configuration shown in the Fig. 8. One of the features is the access via the thesaurus and another is simulation of NMR spectral data based on parameters obtained by learning functions. Both features are explained bellow.

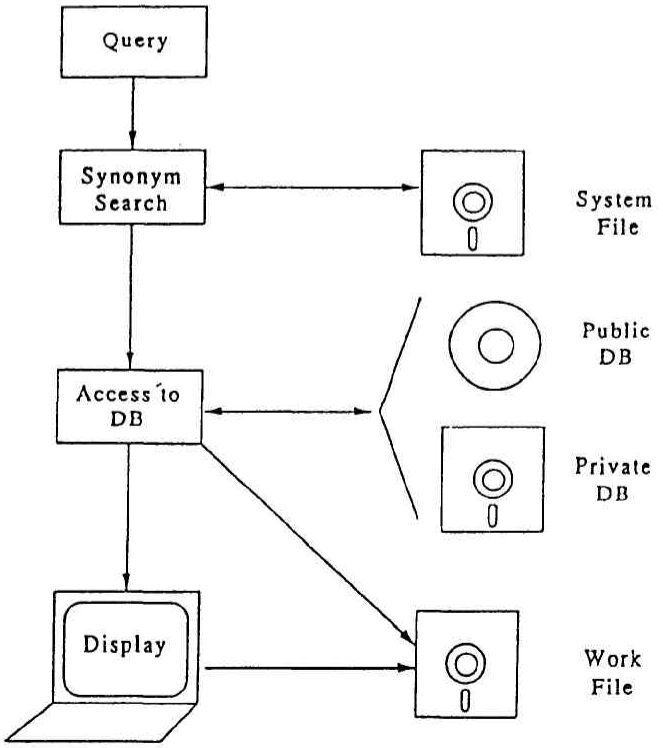
When specialists want to use database systems as a tool supporting research and development directly, they require knowledge and principles together with sophisticated manipulating functions such as simulation, estimation and generation of candidate solutions. The Fig. 8 contains knowledge abstraction function from the database and related application softwares.

Fig. 8 System Configration of IB:PCMR



Automatic compilation of thesauri is important and useful as reported else where(4). The retrieval of data by expanding key words is facilitated in CAPDAS as shown in the Fig. 9. There the thesaurus contain mostly synonymous terms and some of hierarchical terms as well, both of which are used often upon searching data. The list of synonymous terms are stored in a disket in order to add user's terms and to allow for users to update the contents of the thesaurus.

Fig. 9 Retrieval via Thesaurus



The upper half of the Fig. 9 is an example data of NMR spectra stored in CAPDAS and the lower falf in a simulated pattern corresponding to the observed one. The latter one was calculated with a set of parameters stored in a parameter file in advance.

Abstracting shift parameters by solving the additivity equations is one of ways to acquire knowledge from data. This learning function is one of the most required capabilities which are components of a next generation of information systems. The Fig. 10 shows a flow chart of the learning process to obtain shift parameters in addity rules of polymers.

Fig. 9 C-13 NMR Spectra of Polysulfone

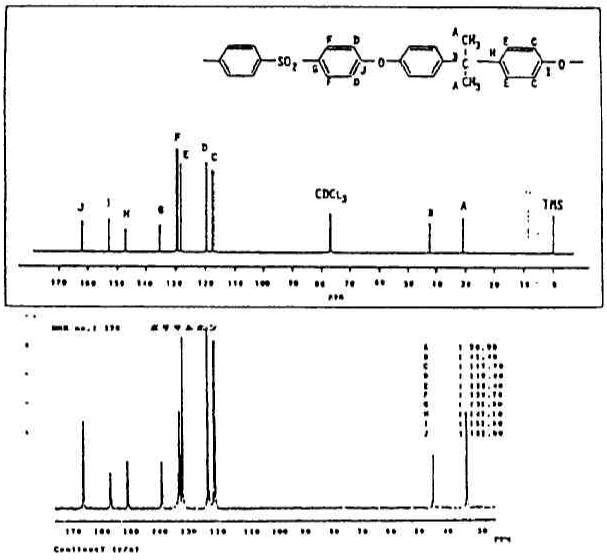
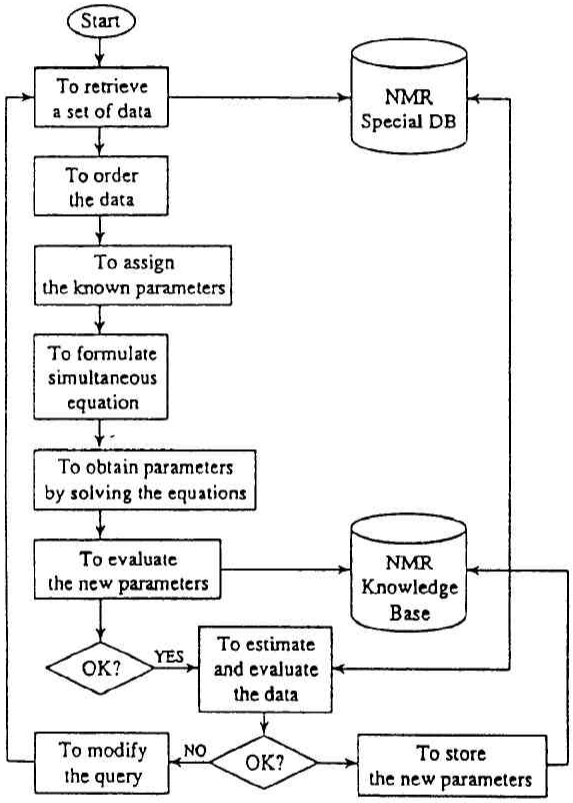


Fig. 10 Learning Process of Shift Parameters in Additivity Rules



Numerical comparison of a simulated spectra with an observed one is shown in the Table 7, where the column designated G-P stands for calculated results using Grant-Paul parameters and where the column of "Modified" has calculated results using modified Grant-Paul parameters to adjust environmental difference in solution between low molecular weight compounds and polymers. The agreement of the latter is better than the former.

If a required set of parameters is not found in the file, necessary parameters may be obtained by solving a set of simultaneous linear equations in the form of the chemical shift model in the Table 8. That is an additivity rule and it is well known that additivity rules hold for chemical shifts of C-13 NMR. However, the parameters in the rules of polymers are substantially different from those of low molecules because of different molecular environment in solutions especially those of stereochemical sequence effects. Moreover, other factors shown in the Table 8 must be taken into account to obtain satisfactory simulation results.

Table 7 Chemical Shift Estimated by Modified Parameters

NMR No. 123	Ethylene-Butene-1 Copolymer						
	-CH2(b)-CH2(x)-CH2(c)-CH2(r)-CH(g)(CH2(m)-CH3(n))-CH2-						
Part	Obsd.	C-P.	Error	%	Modified	Error	%
A	11.14	14.35	3.21	28.8	11.08	-0.06	0.54
B	26.75	27.16	0.41	1.53	27.51	0.76	2.84
C	27.35	30.21	2.86	10.5	27.33	-0.02	0.07
D	30.00	29.96	-0.04	0.13	29.96	-0.04	0.13
E	30.49	29.96	-0.53	1.74	30.49	0	0
F	34.11	34.47	0.36	1.06	34.75	0.64	1.88
G	39.75	37.05	-2.7	6.79	38.25	1.5	3.77

Table 8 C-13 NMR Knowledge in the Learning System

1. The Chemical Shift Model :

$$(\nu_I) = (\Sigma \alpha_{IJ}\, C_{IJ})$$

2. The Shift Parameters : (α_{IJ})

3. The Higher Order Interactions : (α_{IJk})

4. The Relation Times : T1 , T2

5. The Fine Structures :

(JIJ) Spin-Spin Coapling Constants
Anisotropy

6. Others : Solvent Effects
Reference
Temperature Effect
Experimental Methods
Experimental Conditions
etc.

Since CAPDAS contains a large quaintly of data from diversified sources, same kind of properties are often expressed in different units. In order to overcome this issue, CAPDAS has internal conversion functions of units and may compare numerical values in the database with those of queries. The Table 9 shows some of unit conversion in CAPDAS.

Standardization is necessary as shown by the above mentioned example. However, it is time consuming and always comes late. Nevertheless, standardization should be and is being discussed by many relevant organizations. The Table 10 shows various levels of standardization concerning databases and knowledgebases. Unfortunately none of the level was settled with satifisfaction and it must be emphasized that these are easy from the view point of the present technology.

Table 9 Unit Conversion

yard	m
pound	kg
k cal	J
k gf	N
deg F	deg C , K
radian	degree
ASTM	JIS
DIN	JIS
—	—
—	—

Table 10 Standardization

1. Symbols —— Two to Five Byte Codes
2. Data Formats
3. Data Representation —— Structural Data
4. Attribute Description
5. Terminology
6. Documentation

5. Conclusion

Gatekeepers are expected to play important roles for dissemination and use of information accumulated day after day all over the world. Some of the functions can be implemented in database systems or/and in front end terminals as exemplified above. System supported gatekeeper functions are of wide use and facilitate uniformly improved services of information beyond capabilities of individual information specialists in straight forward services, while information specialists may devote themselves to services of elaborated and intellectual ones which can not given by systems.

Reference

1. Y. Fujiwara, K. Araki et al : "Stereochemically Accurate Registry System : STARS"
   Proc. of 25th Symposium of Scientific and Technological Information (1986. Oct.)
2. Y. Fujiwara et al : "Multilingual Access System for Distributed Databases : Proc. of 43rd FID Conference 345-356 (1986. Aug. Ottawa)
3. Y. Fujiwara : "CD-ROM/Computer Assisted Polymer Database System : CAPDAS"
   Information Management 30 123-134 (1988)
4. Y. Fujiwara et al : "A Dynamic The saurus for Advanced Research and Development"
   Proc. of 44th FID Conference 223-234 (1988. Aug. Helsinki)