in a significantly improved presentation of the material.

REFERENCES AND NOTES

(1) Pauling, L. *The Nature of the Chemical Bond*; Cornell University Press: Ithaca, NY, 1960.
(2) Randić, M. *Proc. Galveston Conf. Math. Chem.* April 1990; *J. Math. Chem.* **1991**, *4*, 157.
(3) Trinajstić, N. *Chemical Graph Theory*; CRC Press: Boca Raton, FL, 1983.
(4) Klopman, G.; Kalos, A. N. *J. Comput. Chem.* **1985**, *6*, 492.
(5) Trinajstić, N.; Randić, M.; Klein, D. J. *Acta Pharm. Yugosl.* **1986**, *36*, 267.
(6) Cramer, R. D., III. *J. Am. Chem. Soc.* **1980**, *102*, 1837.
(7) Hansch, C.; Fujita, T. *J. Am. Chem. Soc.* **1964**, *86*, 1616. Hansch, C. *Acc. Chem. Res.* **1969**, *2*, 232.
(8) Kovâts, E. *Z. Anal. Chem.* **1961**, *181*, 351.
(9) Dirac, P. A. M. *Quantum Mechanics*; Oxford University Press: London, 1958.
(10) Coulson, C. A. *Proc. R. Soc. London, A* **1939**, *169*, 413.
(11) Randić, M. *J. Am. Chem. Soc.* **1975**, *97*, 6009.
(12) Initially the index was named the "branching index", being sensitive to molecular branching. The index, however, applies equally to linear chains and cyclic structures and a better name, suggested by L. B. Kier, is the "connectivity index": Kier, L. B.; Hall, L. H.; Murray, W. J.; Randić, M. *J. Pharm. Sci.* **1975**, *64*, 1971.
(13) Klein, D. J.; Schmalz, T. G.; Hite, G. E.; Seitz, W. A. *J. Am. Chem. Soc.* **1986**, *108*, 1301. Klein, D. J.; Seitz, W. A.; Schmalz, T. G. *Nature* **1986**, *232*, 6090. Schmalz, T. G.; Klein, D. J.; Hite, G. E. *J. Am. Chem. Soc.* **1988**, *110*, 1113.
(14) Montroll, E. E. *J. Chem. Phys.* **1941**, *9*, 706. Klein, D. J.; Hite, G. E.; Schmalz, P. G. *J. Comput. Chem.* **1986**, *7*, 443.
(15) Essam, J. W.; Fisher, M. E. *Rev. Mod. Phys.* **1970**, *42*, 272.
(16) Cyvin, S. J.; Gutman, I. *Kekule Structures in Benzenoid Chemistry*; Lecture Notes in Chemistry, Vol. 46; Springer: Berlin, 1988.
(17) Ham, N. S.; Ruedenberg, K. *J. Chem. Phys.* **1958**, *29*, 1215, 1229. According to K. Ruedenberg (private communication), it was J. R. Platt who recognized the novel bond orders of Ham and Ruedenberg as Pauling bond orders which had already been described.
(18) Dewar, M. J. S.; Longuet-Higgins, H. C. *Proc. R. Soc. London, A* **1952**, *214*, 482.
(19) Platt, J. R. In *Encyclopedia of Physics*; Flügge, S., Ed.; Springer Verlag: Berlin, 1961; Vol. 37, Part 2. W. C. Herndon, who, in a series of papers, advocated this approach of enumeration of Kekulé valence structures, also drew attention to Platt's early work.
(20) Ruedenberg, K. *J. Chem. Phys.* **1954**, *22*, 1878.
(21) Heilbronner, E. *Helv. Chim. Acta* **1962**, *45*, 1722.
(22) Cohn, M. C. *Bull. Math. Biol.* **1986**, *48*, 417.
(23) Randić, M. *New J. Chem.* **1991**, in press.
(24) Hosoya, H. *Bull. Chem. Soc. Jpn.* **1971**, *44*, 2332.
(25) Hosoya, H.; Kawasaki, K.; Mituzani, K. *Bull. Chem. Soc. Jpn.* **1972**, *45*, 3415. Narumi, H.; Hosoya, H. *Bull. Chem. Soc. Jpn.* **1980**, *53*, 1228.
(26) Gutman, I.; Polansky, O. *Mathematical Concepts in Organic Chemistry*; Springer Verlag: Berlin, 1986.
(27) Balaban, A. T.; Motoc, I. *Match* **1979**, *5*, 107. Motoc, I.; Balaban, A. T. *Rev. Roum. Chim.* **1981**, *26*, 593. Basak, S. C.; Magnuson, V. R.; Niemi, G. J.; Regal, R. R.; Veith, G. D. *Math. Modeling* **1987**, *8*, 302. Motoc, I.; Balaban, A. T.; Mekenyan, O.; Bonchev, D. *Match* **1982**, *13*, 369. Razinger, M.; Chrétian, J. R.; Dubois, J.-E. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 23.
(28) Motoc, I. *Topics Curr. Chem.* **1983**, *114*, 93.
(29) Kier, L. B.; Murray, J. W.; Randić, M.; Hall, L. H. *J. Pharm. Sci.* **1976**, *65*, 1226.
(30) Platt, J. R. *J. Chem. Phys.* **1947**, *15*, 419. Randić, M. *Match* **1979**, *7*, 5. Randić, M.; Wilkins, C. L. *Int. J. Quantum Chem.: Quantum Biol. Symp.* **1979**, *6*, 55. Wilkins, C. L.; Randić, M. *Theor. Chim. Acta* **1980**, *58*, 45. Wilkins, C. L.; Randić, M.; Schuster, S. M.; Marklin, R. S.; Steiner, S.; Dorgan, L. *Anal. Chim. Acta* **1981**, *133*, 637. Randić, M.; Kraus, G.; Jerman-Blazic, B. In *Chemical Applications of Topology and Graph Theory*; Elsevier: Amsterdam, 1983; p 192. Jerman-Blazic, B.; Randić, M. *Proc. Conf. Modeling Sim.* **1983**, *5*, 161.
(31) Randić, M. *Int. J. Quantum Chem.: Quantum Biol. Symp.* **1984**, *11*, 137. Randić, M. In *Molecular Basis for Cancer. Part A: Macromolecular Structure, Carcinogens and Oncogens*; Rein, R., Ed.; Alan R. Liss Inc.: New York, 1985; p 309. Randić, M.; Jerman-Blazic, B.; Grossman, S. C.; Rouvray, D. H. *Math. Modeling* **1986**, *8*, 571. Grossman, S. C.; Jerman-Blazic, B.; Randić, M. *Int. J. Quantum Chem.: Quantum Biol. Symp.* **1986**, *12*, 123. Randić, M.; Jerman-Blazic, B.; Rouvray, D. H.; Seybold, P. G.; Grossman, S. C. *Int. J. Quantum Chem.: Quantum Biol. Symp.* **1987**, *14*, 245. Randić, M.; Grossman, S. C.; Jerman-Blazic, B.; Rouvray, D. H.; El-Basil, S. *Math. Comput. Model.* **1988**, *11*, 837.
(32) Randić, M. *J. Comput. Chem.* **1980**, *1*, 368. Knop, J. V.; Müller, W. R.; Szymanski, K.; Randić, M.; Trinajstić, N. *Croat. Chem. Acta* **1983**, *56*, 405. Bogdanov, B.; Nikolic, S.; Sabjlic, A.; Trinajstić, N.; Carter, S. *Int. J. Quantum Chem.: Quantum Biol. Symp.* **1985**, *14*, 325.
(33) Rouvray, D. H. In *Mathematical and Computational Concepts in Chemistry*; Trinajstić, N., Ed.; Horwod: Chichester, 1986. Buckley, F.; Harary, F. *Distances in Graphs*; Addison-Wesley: Reading, MA, 1990.
(34) Wilson, E. B. *Introduction to Scientific Research*; McGraw-Hill: New York, 1952.
(35) Topliss, J. G.; Costello, R. G. *J. Med. Chem.* **1972**, *15*, 1066. Topliss, J. G.; Edwards, R. G. *J. Med. Chem.* **1979**, *22*, 1238. Stouch, T. R.; Jurs, P. C. *Quant. Struct.-Act. Relat.* **1986**, *5*, 57.
(36) Randić, M. *New J. Chem.* Submitted for publication. Randić, M. *Croat. Chem. Acta* **1991**, in press. Randić, M. *J. Mol. Struct.* **1991**, in press.
(37) Courant, R.; Hilbert, D. *Methoden der Mathematike Physik*; Springer Verlag: Berlin, 1931.
(38) Geisser, S. *J. Am. Statist. Soc.* **1975**, *70*, 328.
(39) Diaconis, P.; Efron, B. *Sci. Am.* **1984**, 116.
(40) Balaban, A. T.; Motoc, I.; Bonchev, D.; Mekenyan, O. *Top. Curr. Chem.* **1983**, *114*, 21.

The Beilstein Structure Registry System. 1. General Design

LÁSZLÓ DOMOKOS

Beilstein Institute, Varrentrappstrasse 40-42, D-6000 Frankfurt/Main 90, Germany

Received November 16, 1990

The rapidly growing Beilstein Online Database contains already more than 3.4 million organic compounds. Since it is a compound-oriented factual database, the structure registration plays a central role. The understanding of the basic features of the registration is important for the effective usage of the database. This paper describes the software and the philosophy of the structure-registry system embedded into the data processing from the data acquisition through to the dissemination of the data.

INTRODUCTION

The registration of chemical structures is fundamental to each large structural database and structure retrieval system. All larger systems have their own registration software. The best known is the CAS Chemical Registry System developed in the early 1960s. Registry III, its latest version, has been used since 1974.[1,2] These systems were designed to be optimal for the class of compounds to be processed, for the data structure, and for the software and hardware environment. As a consequence, these systems are usually not transportable and are not commercially available. To meet the requirements of the Beilstein database the Beilstein Structure Registry System was developed between 1986 and 1990.

The Beilstein Online Database receives the structural and factual data from three different sources: the Beilstein Handbook of Organic Chemistry, the file cards abstracted from the literature for the Handbook production, and the original literature covering all important journals of organic chemistry.[3] These three sources correspond to the literature periods 1830–1959, 1960–1979, and 1980 onwards, respec-

0095-2338/91/1631-0320$02.50/0

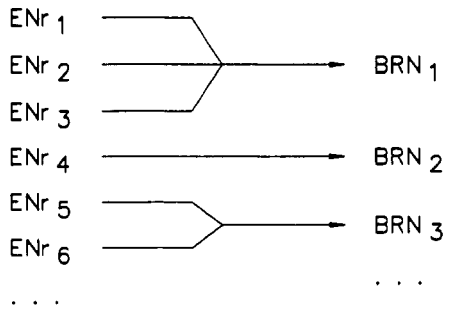


Figure 1. ENr–BRN relationship.

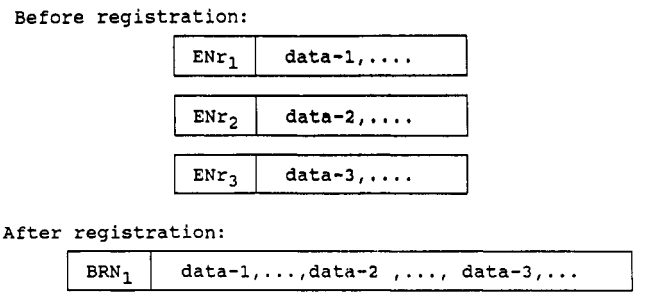


Figure 2. Referencing the data by ENr and BRN.

tively. The data are abstracted by several hundred chemists using menu driven PC programs tailored specially for the different sources. After a careful quality check in the Beilstein Institute, the data are processed and loaded into the Institute's internal database. The database is installed on an IBM 3090/200J mainframe and managed by the database management system ADABAS. The major processing steps of the database generation are

conversion of the input data formats of the different input programs to the data structure of the database[4]
structure registration
loading of factual data into the database
data maintenance
distribution to the on-line hosts and in-house customers

The Beilstein database is a compound-oriented database, i.e., the basic entry is the compound. All structural, physical, chemical, and bibliographical data associated with a compound are stored in a single logical record. Of course, because of the different types and of the amount of information, the data are actually stored in different physical records in 35 different ADABAS files.

The three data sources supply documents corresponding to Handbook articles, to file cards, and to descriptions of a compound in a publication. These input documents are uniquely identified by a 13-character so-called excerption number, ENr, which is assigned by the excerption programs. The ENr serves for the management and bookkeeping of the input documents.

DISCUSSION

Generally, a compound is described in several different documents, in cases of popular compounds even in several hundreds. In order to load and handle the data in the compound-oriented database, it is necessary to recognize which ENr's belong to the same compound. This task is carried out by the structure registry system. The purpose of the registry system is to recognize identical compounds and to distinguish different ones. This is done on the basis of the chemical structures. Each input document, ENr, contains a well-defined chemical structure. The structures are recorded by a graphics software program using a mouse. This program is based on the structure editor MOLKICK.[5] Beside the manual input of structures, about 85% of the Handbook structures without stereochemical information was generated by the program VICA.[6] This program was developed in the Beilstein Institute for translating chemical nomenclature names into the corresponding atom-bond structure.

The result of the registration process is the **Beilstein Registry Number**, BRN, which is unique for each compound. The BRN is a sequential number starting at 1001. Generally, the registry system can be viewed as an *n*:1 function which assigns to each ENr a BRN. To each ENr belongs only one BRN, but a BRN may be associated with several ENr's (Figure 1).

After the registration, the factual and bibliographical data belonging to the same compound, i.e., to the same BRN, will be merged and referenced not by the ENr but by the BRN (Figure 2).

The data structure has been designed so that the process is reversible, i.e., all data which originate from a certain ENr document can be recognized and extracted from a compound record. This is necessary if structures have to be corrected. For example, if the structure of ENr_2 has been erroneously input then the registration assigns to the corresponding "data-2" facts an incorrect BRN_1. After recognition of the error and correction and reregistration of the structure, the factual data "data-2" have to be extracted and moved to the new correct BRN.

The requirements for the registry system are

reliability
capability to handle stereochemical information
reconciliation with the usage of the database
concordance with the widely accepted compound definition
concordance with other well-known systems and with Beilstein traditions
highly automatized
speed
possibility for error correction

The registration software can be divided into two large parts, namely, the chemical part and the technical part. This paper will focus mainly on the technical part. Some of the chemical concepts have appeared in previous publications,[8,9] further details will be published in the future.

The steps of the registration process are

conversion
normalization
assigning the BRN
control
correction and update

Conversion. The purpose is to convert the input data format into an easy-to-handle internal working format. The system accepts and stores the structures in the form of connection tables, CT. Different CT formats are supported and converted to an internal working format. The basic input structure description format is the ROSDAL[7] string, which is the output of the MOLKICK graphical structure editor. The format conversion routine is not only the first part of the registry system but also a general purpose CT conversion program. It accepts also other well-known formats like MACCS, developed by Molecular Design Ltd (MDL), CAS Private Registry, DARC, etc. The same program is able to generate the above-mentioned non-Beilstein CT formats for data exchange.

Normalization. The normalization step represents the chemical intelligence of the registratioin, hence it is crucial for the correct registration.[8,9] Since the Beilstein Information System covers an extensive literature period (1830–1990), structures and chemical names belonging to the same chemical substance have been published in various forms and according to different conventions. Consequently, the CTs generated from this heterogeneous material with different MOLKICK and VICA program versions may look fairly different. The aim of

Table I. Overview of the SDF Lists

type[a]	name of list	parent list
P	SDF	
P	BRCT, beilstein Registry CT	SDF MF
P	MF, multicomponent structure	SDF
C*	π-bonding electron	BRCT
C*	from	BRCT
C*	ring closure	BRCT
C*	atom	BRCT
C*	localized hydrogen	BRCT
C*	delocalized hydrogen	BRCT
C*	localized charge	BRCT
C*	delocalized charge	BRCT
C*	localized unpaired valence electron	BRCT
C*	delocalized unpaired valence electron	BRCT
C*	abnormal mass (known location)	BRCT
C*	abnormal mass (unknown location)	BRCT
C*	H isotope (known location)	BRCT
C*	H isotope (unknown location)	BRCT
C*	fragment integer coefficients	MF
C*	undefined fragment coefficients	MF
S*	stereo atom	BRCT
S*	stereo bond	BRCT
S*	stereo axis	BRCT
S*	stereo polyatom	BRCT
S*	stereo polybond	BRCT
S*	stereo polyaxis	BRCT
S*	noninterpreted stereo descriptors	BRCT
D	nondefault valence	BRCT
D	tautomer group mobile	BRCT
D	tautomer group localized	BRCT
D	graph-atom coordinates	BRCT
D	graph-bond orientation	BRCT
D	nongraph-atom coordinates	BRCT
D	graph atom Z-coordinates	BRCT
D	nongraph-atom Z-coordinates	BRCT
D	display formula	BRCT
D	Fischer atom	BRCT
D	bond type	BRCT
D	neutralized charges	BRCT
D	CIP stereo atom	BRCT
D	CIP stereo bond	BRCT
D	CIP stereo axis	BRCT
D	stereocenter ligand priority vector	BRCT
D	structure hash code	BRCT
D	original numbering	BRCT
D	original stereodescriptors	BRCT
D	supplementary descriptors	BRCT
Z	ENr (excerption number)	SDF
Z	molecular formula	SDF
Z	compound name	SDF
Z	Beilstein Handbook citation	SDF
Z	literature citation	SDF
Z	remarks	SDF

[a] Lists marked with an asterisk (*) belong to the registerable part, RP.

the normalization is to convert all different forms into a chemically equivalent, unique, and well-defined form. The major parts are as follows: syntax checking, charge handling, determining the stereo descriptors, and the unique numbering of the atoms by a Morgan type algorithm.[10] The result is the normalized Structure Description Format, called SDF record.[11]

Incorrect structures are recognized and trapped by the software. Trapped structures are manually controlled and corrected by chemists if necessary.

SDF Format. The SDF represents the standard internal and distribution format of Beilstein structure CTs. It has a concise and flexible construction. Each relevant feature of atoms, of bonds, and of the molecule is described by separate optional lists (Table I). The same SDF format and a similar structure registration is used for the Gmelin Database of inorganic chemistry. From the registration point of view, five qualitatively different kinds of lists can be distinguished

C Constitution lists containing information about the constitution of the structure, like "atom type" list, "from", "ring closure", and "π-electron" lists for bond definition, "localized charge" list, "abnormal masses" list, etc.

S Configuration lists containing stereochemical information, like "stereo atom" list, "stereo bond" list, etc.

D Lists for display information like "coordinate" list, "bond type" list, etc.

Z Lists with additional information. The SDF format serves not only for internal CT description but is also the distribution format for Beilstein customers and conveys some non-CT information as well, like reference to the *Beilstein Handbook*, chemical name, literature references, etc. The Z lists are not stored in structure files but in the factual files of the database.

P Parent lists. These lists serve for logical grouping of the above-explained lists. Each parent list has a 12-byte header vector and contains several C, S, D, Z and/or P lists. The complete SDF record itself is a single P-type list. Because it may contain further P lists, the SDF has a recursive type structure which is outlined in Figure 3.

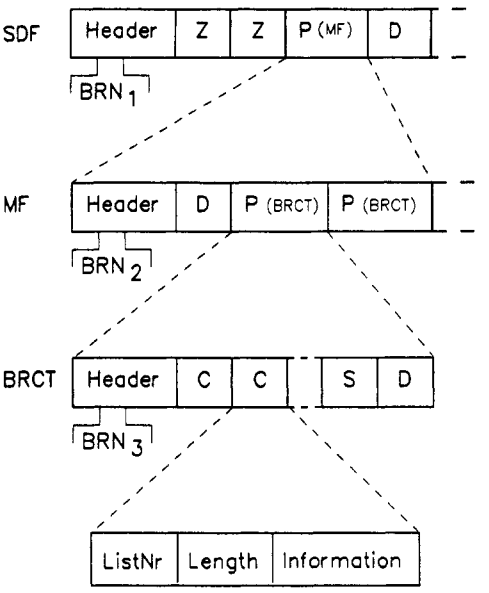


Figure 3. Hierarchy of the SDF record.

An important P-type list is the Beilstein Registry Connection Table (BRCT) list. The BRCT list cannot contain further P lists, and it contains C-, S-, and D-type lists for the definition of a single component. In practice, each BRCT list describes one component of the structure.

The parent list's header vector contains the BRN and some other practical information for data handling, like length of the P list, number of atoms, number of further P lists within the current P list, etc. An important feature of the registration is that it assigns a BRN not only to the SDF list, i.e., to the complete structure, but also to each P list. This means that each component is registered also as a separate structure (which does not mean that it is an existing compound in the Beilstein Online Database).

Without describing all details, a complete description of the SDF lists used with the Beilstein Registry System is given in Table I. SDF lists used only with the Gmelin Database are not included.

Assigning the BRN. This is the central step of the registration process. The strategy applied here is very simple. Each structure is examined whether it is a new one or not. If yes, then a new BRN will be assigned to the structure that is equal to the largest existing BRN plus 1. If not, then the already existing BRN will be assigned. The decision is made on the basis of the C- and S-type lists. Any difference in these lists

means a different compound. Differences in the D, Z, and P lists are not considered. Since the C- and S-type lists are normalized, a byte-by-byte comparison is sufficient.

A typical registration run would be not just the registration of one or of several structures but of several thousands or in the early phase of the database building of several hundred thousands of structures. To process a large number of structures, a one-by-one update of the registry file in the database is inefficient. Owing to considerations of performance and of data integrity, all updates of the registry file are made as a mass update. This means that the new structures are added to the registry files in the last step of the process after assigning the BRN's to each structure. It follows, that it is not enough to decide whether a structure is identical with an already registered one, it is also necessary to check whether it is identical with one or more other structures within the set of structures to be registered. Therefore a preprocessing step carries out the following: all registerable lists, namely, the C and S lists, are sorted and concatenated in a well-defined order to the beginning of each SDF record. The concatenated lists are called the "registerable part", RP, of SDF. The records are sorted in an ascending order of the RP. The result is that if two structures in the current SDF file are identical then their SDF records are neighbors in the sorted file. Hence the decision whether two structures are identical in the SDF file is limited to a single string comparison of the concatenated RP's of the neighboring records. In order to decide quickly if the structure already exists in the database, an 8-byte code derived from the C lists is used as an index for retrieving the candidates. The code is not unique but selective enough to make false drops seldom.

As a matter of fact, the above procedure is carried out in two steps. If a structure is new, then its constitution will be examined to determine whether a compound with the same constitution already exists in the database or in the SDF file to be registered, i.e., whether a stereoisomer exists or not. The system maintains a file, called EQCONST, with classes of stereoisomers.

The update of the registry file is carried out after the successful registration of the complete SDF update file. Beside an increased performance, the method has the advantage that tests, corrections, and reregistrations can easily be done before the final update, and the whole process can be restarted in case of errors.

Because of the recursive construction of the SDF records, the registration of multicomponent structures also takes a recursive path. First the P lists on the lowest level, i.e., the BRCT's are registered, then the higher level P-type lists, and finally the SDF, i.e., the complete structure itself.

The step produces several files that are then loaded into the database. These files are

"registry" file contains the header vectors of the P-type lists, the C- and S-type lists, i.e., the RP's, and the 8-byte code for retrieval. The registry file is used to decide whether a compound is new to the database

"display" file contains the D-type lists

"brnenr" file contains for each ENr the assigned BRN. This file plays an important role in the data management. It is the link between the input documents and the database entries. All cross-references of compounds within the database, like reactant–product relationships, are handled in the internal database as an ENr reference. The ENr references are kept in the database and will be transformed to BRN references only when distributing the data. This method allows an easy maintenance when structures need to be corrected and reregistered.

"eqconst" file, as mentioned before, defines the classes of stereoisomers.

Control. The possibilities for false registration are manyfold. In spite of the careful input and of its control, errors can always occur. Moreover, the original publication itself may contain erroneous structures or structures which cannot be input exactly. The large number of structures to be registered and the large variety of possible errors make automated error-checking necessary. The purpose is to reveal doubtful cases which are to be subsequently examined by experts.

Unfortunately, the possibilities for such automated control are limited. The Beilstein Registry System contains the following modules for consistency check:

Check between the molecular formula and structure. This is basically done by the input program.

Using the CAS registry number (regno), if available, for consistency check. The fact, that the relationship between the BRN's and CAS regnos is not a 1:1 correspondence, but a *m:n* correspondence makes the control less straightforward. The assignment of the BRN's to the CAS regnos is made on a constitution basis only, hence the CAS regnos of each stereoisomer from the CAS database are assigned to each Beilstein stereoisomer and vice versa.

Using the VICA "name translation program"[6] the CT is generated from the chemical name. The obtained CT is compared with the original one. Since the comparison is made by a program, it is very fast. Deviating structures are checked manually. The method serves for the checking of the chemical name as well. However, it is limited to those names that correspond to the IUPAC/Beilstein nomenclature and that can be translated by the program.

The Beilstein system itself provides an effective control method for structures coming from the Handbook.

A long term but very thorough control is assured by the production of the *Beilstein Handbook*. It means, that all structures and factual data from the sources file cards and primary literature will be reviewed by the Beilstein chemists when compiling the corresponding Handbook article. If any error in the structures or in factual data are recognized, they will be corrected. After this sound control the data will be marked as a "Handbook data" in the online database.[12] A disadvantage is, that it may take several years from the registration to the final publishing of the corresponding Handbook volume.

An overview of the registration concept is given in Figure 4.

Limitations. It must be emphasized that the registry system, like all other similar systems, works with a model that defines molecules in terms of atom-bond CTs, which is far from an exact description of the chemical reality. Consequently, all the molecular properties not included in the model, as in our case the conformation, for example, cannot be used to distinguish between different chemical structures. The more precise the model, the more exact the registration will be. However, the freedom for choosing the features included in the model is limited to those features which are commonly used in the chemical literature. It would be useless to use more sophisticated representations if the corresponding data are seldom known or are not always published. The Beilstein CT definition includes all widely used structure features.

Two basic types of errors can be distinguished:

type 1	two structures are different but are registered with the same BRN
type 2	two structures are identical but are registered with different BRN's

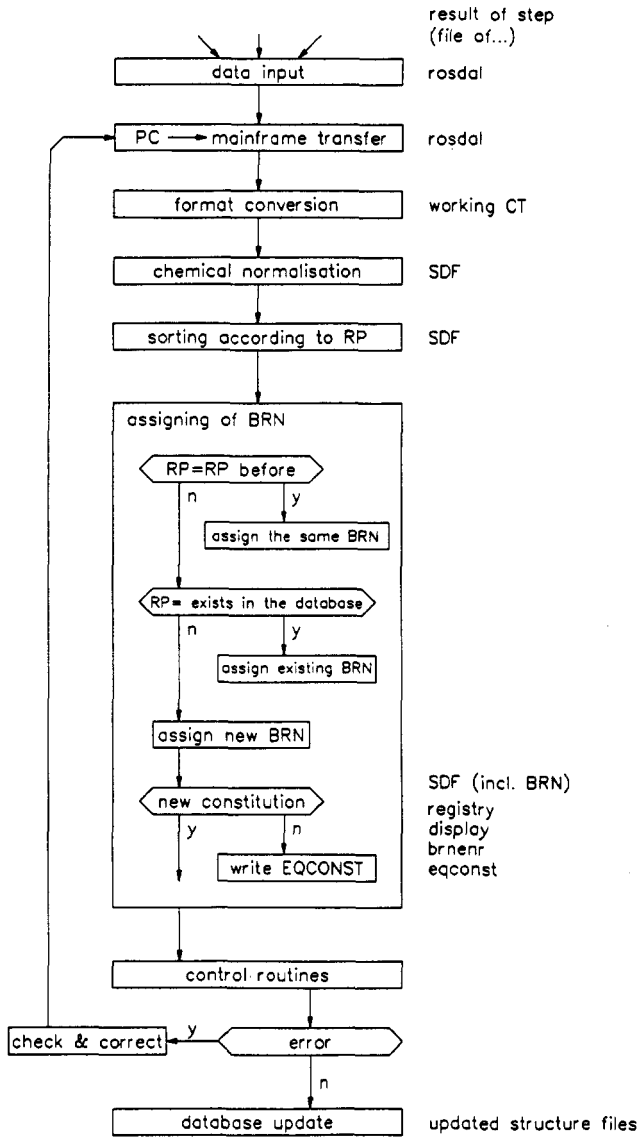


Figure 4. Steps of the registration process.

These errors, except program errors, are influenced by the applied CT model and by the algorithms used with the chemical normalization step. In general, the decrease in one type of error results in the increase of the other type of error. The two absurd extremes are as follows: (1) Registering each CT with the same BRN. In that case the type 2 error is zero but except for one structure all others have a type 1 error. (2) Registering each CT with a different BRN results in zero type 1 error but results in huge type 2 error.

From the chemical point of view, the most acceptable balance has to be found. This decision is certainly influenced by the type of database. There might be different considerations for a compound-oriented factual database, where the chemical and physical properties are stored with the registered form of the structure, e.g., Beilstein Online, and for other registry databases, like CAS, where the factual information can be accessed only in the referenced literature, hence the consequences of false registry number assignments are less severe.

In some cases, depending on the applied rules, different registry systems might give different results; i.e., different structures in one system might be considered as identical in the other system, and vice versa.

The philosophy of the Beilstein Registry System is that it tries to keep the error type 1 low.

A consequence of this strategy is that the Beilstein Registry System registers mostly the published version of the structure. No significant change is made either during the input or by the registration software. The largest, mostly completely equivalent, change made is the evaluation and normalization of the stereochemical descriptors. The first opportunity for a large-scale manual editing of structures by experts is the process of critical evaluation for the Handbook production.

A characteristic example for the above principle is that each input tautomeric structure form is registered separately corresponding to the published version.[9] However, to provide easy access to all registered tautomeric forms, a postregistration step carries out a tautomer normalization giving identical RP's for tautomeric structures. In this way classes of tautomers can be built, and the compounds of these classes can be cross referenced similarly to the "EQCONST" file referencing of stereoisomers. Furthermore, the new version of the S4[13] structure search system is capable of retrieving all registered tautomers regardless of the tautomeric form defined in the search query. This feature can optionally be switched on or off.

The method has a significant advantage. In the case of a change or refinement of the tautomer concept, the normalization and cross referencing can easily be modified without reregistering all structures.

At the moment the normalization is restricted to the most common proton migration tautomerism. No normalization is performed either for ring-chain tautomerism or for valence tautomerism. However, and this is the previously mentioned advantage, this can be incorporated later without changing other parts of the system or the BRN's.

Another example is that in some cases, mostly with early published Handbook structures, the structure is extended with information extracted from the physical data or from the article, like "or mirror image". A controlled vocabulary is used to standardize the descriptors, which are stored in a list of RP. This helps to decrease the type 1 error even when at the time of original publication the correct configuration was unknown or uncertain.

Correction and Update. The basic principle is that a registered structure will never be changed. If an error is detected, the structure will be corrected and reregistered with another BRN. The corresponding data are moved from the old BRN to the new one, but the old structure remains as an unused structure in the registry file. Of course it will be deleted in the Beilstein Online and in-house databases. However, it may happen that an unused structure will be activated again if it turns out to be the correct structure for a new document. Similarly, a structure belonging to a deleted compound will remain in the registry file.

The consequence is that the relation between CT's and BRN's is permanent.

Example. In the following, a simple example demonstrates the registration of a two-component compound. Figure 5, panel a, shows the structure to be registered. Figure 5, panel b, outlines the corresponding SDF description after the registration. Each SDF list is framed. The parent lists have a double-line frame. The lists are identified by their types (C, S, D, Z, or P) and by their names given in quotation marks. The name is followed by the list information.

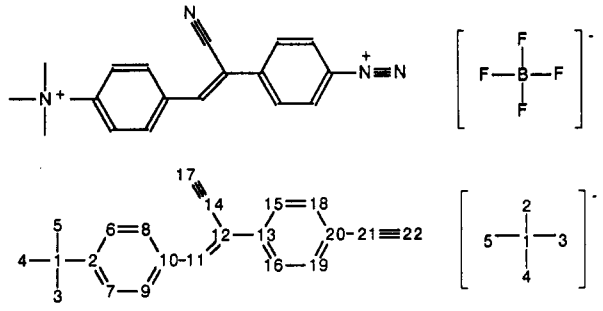
In order to focus on the concept, not all SDF lists are depicted. The following D-type lists are missing in both BRCT's: "bond type" list which describes the bond types used by the input program and D "structure hashcode" list. In the 2nd BRCT: "nondefault valence" list and the "graph atom coordinates" list are missing. The "chemical name" and the "from" list information for the 2nd BRCT are truncated.

Each atom of both fragments is numbered by the normalization step uniquely in a well-defined order (Figure 5a). The SDF lists, like the "from", "ring closure", "atom",

a

BRN = 3583201

4-[(Z)-1-cyano-2-(4-trimethylammoniophenyl)-vinyl]-benzenediazonium bis-tetrafluoroborate



b

```
P "SDF"            Header: Length=.., BRN=3583201, #lists=3,...
  Z "CHEMICAL NAME"          4-[(Z)-1-cyano-2-(4-trimethyl....
  Z "ENr"                    PXFCB095-1080
  Z "MOLECULAR FORMULA"      C18H18N4(2+)*2BF4(1-)
  P "MULTICOMPONENT" Header: Length=..,BRN=3583201, #lists=3,
    C "FRAGMENT INTEGER COEFFICIENTS"    2 , 1
    P "BRCT"     Header: Length=.., BRN=3587364, #lists=8,...
      C "FROM"                   1,1,1,1
      C "ATOM"                   1 B, 2 F, 3 F, 4 F, 5 F
      R "NON DEFAULT VALENCE"    1 4
      C "DELOCALIZED CHARGES"    -1
      R "GRAPH ATOM COORDINATES" list of x-y coordinates
    etc.
    P "BRCT"     Header: Length=.., BRN=3563713, #lists=12,..
      C "PI-BONDING ELECTRON" 0100011111111211211122
      C "FROM" 1 1 1 1 2 2 6 7 8 10 11 12 12 13 14 15 16  etc.
      C "RING CLOSURE"           9-10, 19-20
      C "ATOM"                   1 N, 17 N, 21 N, 22 N
      C "LOCALIZED HYDROGEN"  00333111101000011011000
      S "STEREO BOND"            11  1
      C "LOCALIZED CHARGES"      1 +1, 21 +1
    etc.
```

Figure 5. (Panel a) Structure to be registered and its numbering. (Panel b) SDF representation.

"localized charge", etc., identify the atoms by their numbering. Other lists, like the "π-bonding electron" and "localized hydrogen" lists, are positional lists, i.e., if the list is present then the information is listed sequentially for each atom. The hydrogen atoms are not numbered. They are defined in the "localized hydrogen" list where for each numbered atom of the structure graph the number of attached hydrogens is given. A similar construction is used for the hydrogen isotopes.

In reality, the list information is stored in a more compact binary form, for example the "atom" list of the 2nd BRCT looks like:

hexadecimal '04000B040107110715071607'

where the first byte contains the list Id '04'; the next 2 bytes contain the list length '000B', including the current two bytes; the next 1 byte '04' is the number of non-carbon atoms. This is followed by four pair of bytes listing the atom numbering and the atomic number of the hetero atom, which in our case is '07' for each nitrogen. In this example there is only one S-type list describing stereochemical information, namely, the "stereo bond" list. It defines the configuration of the 11th bond which is, according to the from list, the double bond between the atoms 11 and 12.

Figure 5b shows, similar to Figure 3, the role of the P-type parent lists and the recursive structure of the SDF. It can also be seen that each parent list has its own BRN. The BRN's of the two BRCT's are the registry numbers of the two fragments. The BRN in the "SDF" header is the Beilstein Registry Number of the compound. The BRN of the "multicomponent" parent list is identical with that of the SDF, because there is no registerable list, i.e., C- or S-type list, within the "SDF" but outside the "multicomponent" list. The registerable "fragment integer coefficient" list defines the relation between the two BRCT's.

The "ENr" list contains the excerption number, which is a structured string conveying information about the journal of publication, the location, and the excerptor of the data input. As mentioned, the BRN will be assigned to the factual data via the ENr.

Table II. Contents of the Beilstein Registry File (December 1990)

literature period	1830–1979
no. of structures	3 533 922
no. of compounds online	3 415 097
compound class	organic, single component
range of BRNs	1 001–3 534 922
av. no. of non-hydrogen atoms	21.76
av. no. of bonds	23.15
no. of acyclic structures	332 934
no. of sterical structures	823 084
no. of stereoisomer classes	239 561 with 618 995 compds

Distribution of Heteroatoms

O	10 118 389	FE	310	IR	21
N	5 353 083	LI	282	GA	15
S	952 156	SB	276	W	14
CL	862 052	K	198	AT	13
F	545 060	ZN	182	NB	12
BR	274 774	RH	181	RU	12
P	159 409	PB	174	OS	10
SI	102 568	PT	158	Y	10
I	55 526	PD	156	BA	9
B	17 500	AL	79	SC	7
SE	16 764	AG	55	U	7
AS	12 327	AU	46	ZR	5
TE	1 899	TI	38	CA	4
HG	1 296	TI	34	HF	3
SN	1 087	MN	33	RB	3
GE	1 079	V	33	TA	3
NA	1 047	IN	30	BE	2
NI	479	CR	28	CS	2
CO	412	CD	27	TH	2
CU	345	MO	24		
MG	335	BI	21		

SUMMARY

Up to now more than 3.5 million single-component structures have been registered, and ca. 3.4 million of them are available in Beilstein Online. The software and the chemical concept for registering multicomponent substances are under final testing. The software is written in PL1 and Pascal. The average CPU time required for the registration of a structure is 0.1–0.2 s, whereas ca. 90% of the time is spent for the chemical normalization, especially for the unique numbering of highly symmetrical structures.

Table II gives an overview on the present stage of the Beilstein registry file.

Finally, a list of the basic features of the Beilstein structure registry system:

Representation of multiple and aromatic bonds by a π-electron formalism, which enables an efficient and elegant handling, SDF list no. 1[14]
Support of stereochemistry
Characterization of the stereocenters by parity vectors
Tautomers are registered separately
Cross referencing of stereoisomers
Cross referencing of tautomeric structures (not implemented yet)
Routines for consistency checking

ACKNOWLEDGMENT

I thank my colleagues who have participated in the design and development of this system, especially S. Welford, T. Cieplak, M. Heinen, and B. Roth (Chemplex GmbH) and all the Beilstein chemists who worked out the chemical concepts and offered continuous help and control. The project was supported by the German Ministry of Research and Technology.

REFERENCES AND NOTES

(1) Dittmar, P. G.; Stobaugh, R. E.; Watson, C. E. The Chemical Abstracts Service Chemical Registry System. 1. General Design. *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 111-121.
(2) Ryan, A. W.; Stobaugh, R. E. The Chemical Abstracts Service Chemical Registry System. 9. Input Structure Conventions. *J. Chem. Inf. Comput. Sci.* **1982**, *22*, 22-28.
(3) Domokos, L.; Jochum, C.; Wittig, G. Data in Beilstein-Online. *Mikrochim. Acta* **1986**, *2*, 423-429.
(4) Data Structure of the Beilstein Database, internal documentation, available from the Beilstein Institute.
(5) The host independent memory resident chemical structure query editor MOLKICK. *Beilstein Brief*, **1988**, 2.
(6) Domokos, L.; Goebels, L. *Der Computer als Nomenklatur-Struktur-Dolmetscher, Tagungsbericht, GDCH 3.* Vortragstagung: Würzburg, 1986.
(7) Rohbeck, H. G. Representation of Structure Description Arranged Linearly. In *Software Developments in Chemistry 5*; Gmehling, J., Ed.; Springer Verlag: New York (in press).
(8) Welford, S. M. Structure Registration for Beilstein Online, Second International Meeting on Chemical Structures, Nordwijkhout, 1990; Warr, W. A., Ed.; Springer Verlag: Berlin (in press).
(9) Welford, S. M. Tautomer Processing in the Beilstein Registry System. In *Software Entwicklung in der Chemie 2*; Gasteiger, J., Ed.; Springer Verlag: New York, 1988; pp 35-43.
(10) Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures. *J. Chem. Doc.* **1965**, *5*, 107-113.
(11) SDF and BRCT, internal documentation, available from the Beilstein Institute.
(12) Jochum, C. Building Structure-Oriented Numerical Factual Databases: The Beilstein Example. *World Patent Information* **1987**, *9*, 147-151.
(13) Hicks, M. G.; Jochum, C. Performance Comparison of the MACCS, DARC, HTSS, CAS Registry MVSSS, and S4 Substructure Search Systems. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 191-199.
(14) Gasteiger, J. A representation of π-systems for efficient computer manipulation. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 111-115.

The STAR File: A New Format for Electronic Data Transfer and Archiving

SYDNEY R. HALL
Crystallography Centre, University of Western Australia, Nedlands 6009, Australia

Received October 2, 1990

A new type of format is proposed for the computer archiving and electronic transmission of text and numerical data. The Self-defining Text Archive and Retrieval (STAR) File uses standard ASCII text to specify both the data structure and the information. The syntax of this file is simple, and it may be easily interpreted visually or by computer. The STAR format is the basis for the Crystallographic Information File (CIF), which has been adopted by the International Union of Crystallography for the submission of data and text to crystallographic journals and data bases.

INTRODUCTION

Many existing computer-archiving procedures use a "fixed format" data structure targeted at specific applications. This approach provides for efficient data access but is inflexible and cannot be changed without reformating existing archived files. These files are unsuitable for long-term archiving of most scientific data where there is a continual evolution of data types.

Another archiving approach is based on "pre-defined free formats". Such formats do not restrict data to specific positions in the file. Often "data keys" are included to aid in data recognition. Examples of this type are the BCCAB archive file[1] used by the Cambridge Crystallographic Data Centre, the Standard Crystallographic File Structure,[2] the JCAMP-DX File[3] for archiving infrared spectra, and the Standard Molecular Data (SMD) Format,[4] a collaborative development of chemical and pharmaceutical laboratories for the global exchange of molecular data. These files, while differing significantly in construction and style, have a common disadvantage: their data syntax is relatively complex and requires careful predefinition to facilitate data access.

The complexity and inflexibility of existing archive files limits the rapid exchange of data, even within disciplines where data requirements are similar. This is a special problem for applications with a continual need for new data items. Some computer-intensive disciplines, such as crystallography, currently support a vast repertoire of specialized and "local" file formats. This was tolerable when electronic data exchange was infrequent and computing considerations required file formats to be finely tuned to specific applications. However, the recent explosion in computer and network performance has signaled an end to this rationale. In an era of increasing global data exchange there is a critical need for a simple but universal archive file.

Table I. A Universal Archive File

·Is used to store *all types* of data
·Is *not necessarily* a data-base file
·Should be *machine independent*
·Should be *simple* to read and to access
·Should be *flexible* to future change

The prerequisites for a universal archive format are simplicity, generality, upwards compatibility, and flexibility (see Table I). Such a file must be machine-independent and portable so that the accessibility of data items is independent of their point of origin. It is fundamental that this file allows data to be incorporated in the future without imposing a need to modify existing files. The Self-defining Text Archive and Retrieval (STAR) format described in this paper is designed to meet these requirements. The STAR file structure is suitable for archiving all types of text and numerical data, in

0095-2338/91/1631-0326$02.50/0