The Beilstein Information System Is Not a Reaction Database, or Is It?

C. Jochum

Beilstein Institute, Varrentrappstrasse 40-42, 60486 Frankfurt/Main, Germany

Received June 28, 1993®

The Beilstein Information System is the world's largest collection of chemical properties of organic compounds. The Beilstein database contains over 5 million compounds with associated properties, covering the literature period from 1779 to 1991. The gap to the present literature will be closed in 1994. The properties covered most thoroughly within the Beilstein Information System are physical data and chemical behavior. Chemical behavior data contain preparations and reactions of almost all compounds in the database. For many compounds more than one preparation or reaction is contained in the database. Thus more than 8.7 million preparations and reactions are described. On-line access to these preparations and reactions is still limited. This paper describes the current access and future plans for the development of a full Beilstein Reaction Database.

1. INTRODUCTION

After being available on line for more than 3 years, it is well-known that the Beilstein database is the world's largest structure based organic factual database.[1-3]

What is a factual database? In a factual database well-defined facts are organized in numeric and keyword fields. Text fields are rarely used and are usually limited to literature references, comments, or titles of documents. The advantage of such a data structure is obvious: flat file design which lends itself well for an implementation as a relational database; language independence, since numeric fields and keyword fields are by definition language independent, (an English language citation can be understood by practically any scientist worldwide (which is not always true for English language text)).

To closely understand the design of the Beilstein File, it is important to have a close look at the sources of this information system.[1]

2. SOURCES OF THE BEILSTEIN DATABASE

2.1. The *Beilstein Handbook* 1779–1959. The articles of the *Beilstein Handbook*, i.e. the complete factual descriptions of a compound, have always been written according to a very well defined structure. Naturally, since the analytical methods and chemical preparations have changed over the past fifty years, the instructions and definitions for the Beilstein manuscript writers have been altered slightly over this period of time. These changes had to be taken into consideration for the definition of a computer-optimized data structure.

After a very thorough analysis of the article structures of the main volume and of all supplementary series, a data structure was defined which allowed the computerized input, storage, and retrieval of the *Handbook* data without loss of information. However, some compromises had to be made: For most organic compounds described in the primary literature, only very little factual information is known. In many cases, only the boiling point, melting point, refractive index, and one or two methods of preparation have been described in the literature.

A comparatively small percentage of all known compounds (less than 5%) is very important for chemical reactions or pharmaceutical purposes and has therefore been published widely in the chemical primary literature. Therefore many physical data, preparations, and other factual data are known for these large information compounds. The final electronic data structure constitutes a very sophisticated compromise between the information contents of these different classes of compounds.

2.2. Data to the Fifth Supplementary Series 1960–1979. The literature of the Fifth Supplementary Handbook Series (literature time frame: 1960–1980) has been completely abstracted. This factual information which is contained on 7.5 million file cards (one card per compound per literature citation) is the basis for the *Handbook* articles of this series. Since this *Handbook* series will only be completed by the end of this decade, the on-line user will be provided with access to the "raw" information.

The input of the *Handbook* and the file cards has been completed and is available on line.

2.3. Literature Period from 1980 to Present. The Beilstein Institute is currently indexing the third source of information: the factual data of the primary literature from 1980 onward. The compounds and associated literature data are abstracted in a completely new electronic and paperless way using off-line microcomputers. The abstractor enters the structure graphically and inputs the factual data with a menu-driven program. The design of the data structure took new developments of analytical and synthetic methods into account.

The structure of the database can be divided into two parts: the numerical factual file and the structure file. These two parts are subsequently described in more detail.

3. DESIGN OF THE BEILSTEIN DATABASE

3.1. Chemical Structures. The structures are the cornerstone of the Beilstein database, so it was of vital importance to design a registry program which would be able to recognize like structures as like and discriminate between different structures. Such a system has been developed at the Beilstein Institute, the structures being stored and registered in the Beilstein registry connection table format (BRCT) which form subunits, one per component of the structure distribution format (SDF).[4] This system has several special features. All multiple bonds are described in terms of the atom centered descriptor of individual valence π-electrons; this removes the problems associated with coding and searching for aromatic bonds. All tautomers are individually registered without normalization, thus preserving full information integrity; the coding and normalization of the tautomers is stored in nonregisterable lists, only used for the search system.

® Abstract published in *Advance ACS Abstracts,* January 15, 1994.

0095-2338/94/1634-0071$04.50/0

Table 1. Number of Occurrences of the Beilstein Database Fields for the Whole File

	records	occurrences
preparation	4 085 246	6 394 028
chemical behavior	522 590	1 623 721
isolation	50 541	94 882
chemical derivative	407 490	674 789

Tetrahedral stereocenters, double bonds, and allene axes, which correspond to most of the stereochemistry, are stored in terms of a parity code; this is an atom index based descriptor which can unambiguously define the stereochemistry not only for registration but also for substructure searching. Other types of stereochemistry are stored in terms of the standard stereochemical descriptors and are thus not searchable using a structure search.

It is possible to convert from the BRCT format to other formats as required; for example, to the CAS format required to build the file searchable on STN.[5] Test files have also been converted into DARC, MACCS, and HTSS formats.

The BRCT describes structures fully but is not able itself to describe reactions. The vehicle for describing reactions would be a modified version of the SDF format. In the present SDF format the interrelationships between individual compounds in a multicomponent system (e.g. a salt) can be described. It is planned to extend this methodology to describe reactions, with the inclusion of the necessary lists needed to hold the atom–atom mappings, reaction site, and transformation information.

3.2. Factual Data. The factual data in the Beilstein database[6] are stored in three types of fields:

Numeric Fields. There are over 80 different factual fields. Each factual field can be divided into subfields to contain parameter data, temperature, pressure, etc.

Boolean Fields. These fields store the presence of a keyword or a parameter.

String Fields. Chemical names, literature citations, and comments are stored as strings.

3.3. Chemical Reactions. Chemical reactions[7,8] are stored in four fields in the Beilstein database: preparation, chemical behavior, chemical derivative, and isolation from natural products.[2,9]

The *preparation field* is by far the most important field for reactions in the data base. This field contains the description of a compound's formation from starting materials, with reagents and solvents, etc. These fields are the basic requirement for one to be able to build a reaction database.

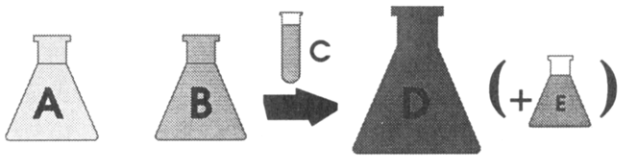
The *chemical behavior field*, sometimes known as the reaction field, is only present when a reaction has been studied for a particular purpose, such as from the mechanism, rate, etc.

The *chemical derivative field* contains information concerning the formation of standard chemical derivatives (hydrazones, etc.) of the compound in question.

The *isolation from natural products fields* contains a description of the method of isolation.

Clearly the most important field is the preparation. The vast number of preparations that we have in the database 6 394 028 make it essential for us to develop a system which will enable the user to get to the information that he requires and avoid being swamped by thousands of hits. Whereas for those reaction databases already in existence, which contain selected general examples, and are thus suitable for general queries, the queries put to a Beilstein reaction database would be sensibly more specialized in nature. The fact that Beilstein is orientated toward preparative organic chemistry is illustrated

The Beilstein Database on STN: Preparations:



What is		Search Code	Searchable via
A	Starting Material	/PRE.SM	name, BRN
B	Starting Material	/PRE.SM	name, BRN
C	Reagent (e.g. Solvent, Catalyst)	/PRE.RGT	name
D	Product, i.e. Beilstein Registry Compound	/CN or /BRN STR or /MF or /RN	name, BRN, structure, MF, CAS Regno
E	By-Product	/PRE.BPRO	name, BRN

Figure 1. Field codes used to represent preparations in the on-line database on STN.

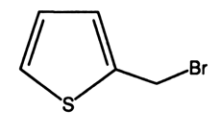
Example: Find preparations starting from 2-methyl-thiophene and bromine

```
=> s 2-methyl-thiophene/pre.sm (p) bromine/pre.rgt
            57 2-METHYL-THIOPHENE/PRE.SM
         22524 BROMINE/PRE.RGT
L6            5 2-METHYL-THIOPHENE/PRE.SM (P) BROMINE/PRE.RGT


=> d 5 brn cn str hit

L6   ANSWER 5 OF 5

BRN  106294  Beilstein
CN   2-bromomethyl-thiophene
     2-Brommethyl-thiophen
```




```
Preparation:
PRE
     Start:  BRN=103734 2-methyl-thiophene
     Reag:   bromine
     Temp:   400.0 Cel
     Reference(s):
     1. Hurd, Anderson, J.Amer.Chem.Soc. 75 <1953> 3517, CODEN: JACSAT
     Note(s):
     2. Handbook Data
```

Figure 2. Search example for a preparation.

by the statistical analysis of the preparation data. It can be seen that 86% of all compounds have at least one preparation associated with them, giving an average over the whole database 1.4 preparations/compound.

Searching reactions is also possible on Beilstein's Current Facts CD-ROM. Each CD-ROM contains all structures, preparations, and factual data which have been published in the literature within the most recent four quarters. Searching for structures/substructures, preparations, and reactions is possible in a way similar to that on STN and Dialog.[10] In addition to a command-oriented search, a menu-driven query formulation is offered by the Current Facts software.

4. SUMMARY AND OUTLOOK

The present Beilstein database allows access to over 8.7 million preparations. Within the next 3 years this number will grow to close to 10 million preparations. While this access is enough to allow retrieval of individual steps of non-aporic synthesis paths, any analogy searching is limited to very similar molecules.

So is Beilstein a reaction database or is it not?

Although all reaction information is present in the Beilstein Information System, searching of complete reaction pathways is still limited. We therefore do not consider the current on-line implementations of the Beilstein file as a reaction database.

To overcome this disadvantage, Beilstein is currently working with two partners on this problem:

(a) An on-line implementation of the Beilstein file on the Host Datastar will give significantly more emphasis to preparation and reaction searching with a very convenient PC-front-end access. The public availability of this implementation is scheduled for the middle of 1993.

It is also planned to emphasize more sophisticated reaction searching on Beilstein's other two hosts, STN and Dialog.

(b) Beilstein currently carries out a research project with IBM Corp. under the name XFIRE. XFIRE I will allow in-house users at academic and industry institutions to load and search very large structure files internally with easy parallel access to public on-line hosts. XFIRE runs on a RS-6000 with a PC-front-end under MS-Windows in a client–server architecture. XFIRE can easily be interlinked with MDL-software through ISIS/Host[11] or Clipboard-Query-Transfer from ISIS/Base.[11]

Subsequent XFIRE II and III will allow systems to handle large in-house reaction databases.

Development of a Beilstein reaction database with the inherent advantages in accuracy and analogy searching will bring another valuable tool to the chemist's bench.

REFERENCES AND NOTES

(1) Heller, S. *The Beilstein Online Database: Implementation, Content, Retrieval*; ACS Symposium Series 436; American Chemical Society: Washington, D.C., 1990.
(2) *Online Searching of Beilstein on STN: How to find preparations and Reactions*; Springer-Verlag: New York, 1990.
(3) Barth, A. *Datenbanken in den Naturwissenschaften*: VCH: Weinheim, Germany, 1992.
(4) Structure Distribution File and Beilstein Registry Connection Table; Version 2.02; Internal Publication; Beilstein-Institute: Frankfurt, Germany, Dec 1990.
(5) STN International; c/o FIZ Karlsruhe, FIZ 4, 7514 Eggenstein-Leopoldshafen.
(6) Data Structure of the Beilstein Database, Version 1.33; Internal Publication; Beilstein-Institute: Frankfurt, Germany, Feb 1992.
(7) Willet, P., *Modern Approaches to Chemical Reaction Searching*; Gower: Brookfield, VT, 1986.
(8) Beilstein Workshop; Computer Reaction Management in Organic Chemistry. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 351–520.
(9) Hicks, M. G. Reactions in the Beilstein Information System: Nonaporic Organic Synthesis. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 352–359.
(10) Dialog Information Services Inc., 3460 Hillview Ave., Palo Alto, CA 94304.
(11) ISIS/Host and ISIS/Base are products of Molecular Design Ltd., 2132 Farallon Dr., San Leandro, CA 94577.