Economic Aspects of Chemical Information[†]

Wolfgang T. Donner

Scientific Information and Documentation, Central Research, Bayer AG, D-51368 Leverkusen, Germany

Received November 22, 1995[⊗]

Important chemical information is contained in a unique way in the Beilstein handbook as well as in the database with its own search engine (CrossFire). This electronic form is one important component of a chemical information management system needed in industrial R&D. The necessary investment for such a system also needs an economic justification. Effective use of any such system containing chemical information from different sources will be enhanced considerably by the use of CAS Registry Numbers.

Twenty-two years ago a conference took place in Noordwijkerhoud which many of us see as a milestone in the development of what is now called chemical information management. The conference dealt "with the fundamental issues of how chemical structural information can be represented and how this choice of representation affects the types of manipulations that can be performed".[1]

The difference between now and then becomes obvious when looking into the proceedings of that conference: It then was primarily a field of research remote from any practical use. The proper methods for this purpose were still in development, but the first steps were already done. The then existing hardware was nearly inadequate for this purpose: The text was written by typewriter and chemical formulae were drawn by hand. Altogether, in Noordwijkerhoud in 1972 vision and enthusiasm were predominant: there was no need for economic considerations.

This has changed in the meantime, as shown by the following milestones:

- Development of the BASIC Fragment Search System based on the experimental CAS substructure search system by the Swiss chemical companies Ciba-Geigy, Hoffmann-La Roche, and Sandoz together with Chemical Abstracts Service during the late 1970s and early 1980s which finally ended up in CAS-Online on one hand and an in-house system on the other.[2−5]
- Prof. J. E. Dubois developed DARC already in the early 1970s. in the early 1980s this system was applied to the full CAS Registry Structure File and performed—for the first time—online substructure searches on 5 million compounds. This EURECAS had an automatic link to the textual database (CA SEARCH) to deal with both structural and textual aspects of a query.[6] DARC aimed at the online market as well as on the market for in-house systems.
- In the early 1980s W. T. Wipke developed MOLEX as a predecessor of MACCS. This is only one example of the many in-house systems that evolved at this time in many chemical and pharmaceutical companies.
- At this time the development of the Beilstein database started.
- Within the second half of the 1980s a race for the fastest search engine for structures and substructures started. Here the Beilstein search engine gained a leading position.

GENERAL CONSIDERATIONS ON THE INFORMATION MARKET

Today economic considerations are important also in the field of Chemical Information. Looking more closely into this market, it becomes obvious that there are only vague ideas about its size.

FORTUNE gave in its edition of July 10, 1995 some estimates for the global market on hardware, software, and services. The size—according to different sources—is between $650 and $900 billion annually. About $20 billion are spend on hardware alone by U.S. corporations. Software producers will take about $37 billion in 1995, according to an estimate by Data Corporation.

Much within this market are untallied sums, since (according to an estimate of Salomon Brothers) about 85% of bank's software spending was done in-house in 1992. Can this proportion also be assumed for chemical information?

What share does Chemical Information have in this huge market? If we assume it takes between 1% and 10% we arrive at a market of about $6.5−90 billion. This range can be specified somewhat by considering the market size of producers like publishers, software developers, database producers, etc. Together, they will have a size of about $2−10 billion. We have to add the amount spent by chemical and pharmaceutical companies for chemical information, which will be on the order of about $10 billion. This consideration leads to an estimated market share of about $20 billion!

All these numbers can only indicate an order of magnitude as the terms used here are rather vague: What does chemical information comprise? Are we counting items such as advertising or all newspaper information on events connected with chemistry? Even if we confine ourselves to scientific chemical information as published in scientific journals, scientific books, scientific databases, etc. and managed in chemical companies, the difficult question arises on how the available information technology should be accounted.

To quantify the market share for chemical information for a single company, again one ends up very soon: You can consider the annual investments in information technology (hardware, software, external services) which will be on the

[†] Presented at the 1995 Skolnik Award Symposium, ACS Meeting Chicago, IL, U.S.A., August 22, 1995.

[⊗] Abstract published in *Advance ACS Abstracts,* August 15, 1996.

S0095-2338(95)00248-4 CCC: $12.00

order of a few hundred million DM for a company like Bayer. This amount corresponds to the above estimates. It becomes more difficult if you want to look into the details, simply by the fact that information is the raw material for the researcher. By searching for, collecting and reviewing of information, and adding one's own experience and experiments one ends up with new ideas and finally new products.

Apart from the size of the market for chemical information there are other aspects, also rather diffuse: This concerns the value of information. Just to mention three examples:

(1) The same fact might be in a book, a CD-ROM, or an online database—each differently priced. Does that mean, we pay for the bottle, not the wine?

(2) Some information might be valuable to somebody and completely irrelevant to others. And this might hold only for a certain time.

(3) Differently from other production factors, the same information can be used at the same time by different persons. It might lose its value for you if others have the same information—or it might gain in value if it is shared by many!

All in all information is a very particular production factor making it extremely difficult to estimate its value.

CHEMICAL INFORMATION TODAY

Today, we observe a clear division between online systems running on public hosts and in-house systems.

Chemical information management within a chemical company covers the following:

- Chemical structures
- Chemical reactions
- Spectra
- Screening data
- Internal reports
- Patent information
- Bibliographic information

The management of chemical structures by the computer is just one component of the management of chemical information. As it is an important one, it is worthwhile to recount some of its applications: The necessary internal registration of compounds by the computer became less labor intensive. The access from each laboratory to most of the information within the company improved considerably the transfer of know-how. This concerns screening data, spectra, internal reports, patents, literature, etc. It also can be used to locate and order a needed compound.

Besides these functionalities we observe that the same techniques are used by the bench chemists to substitute the former card-file. Here all information is collected from internal and external sources that is relevant to the bench chemist's projects.

However, before one could use the advantages of chemical information management with the computer, considerable investments are necessary. The larger a company, the broader the research activities are. The potential gain by using an elaborate information management system is certainly dependent on the size of the company. The necessary investment in hardware, software, and communication also increase with size. The consequence is obvious: Questions concerning return of investment become the more urgent the higher the investment.

The fact that most chemical and pharmaceutical companies use now internal chemical information management systems and the necessary infrastructure indicate the following:

- There is a great demand for information accessible by the end-user.
- Technology is there for its daily use.
- The cost-efficiency analysis is convincing.

Looking deeper into these in-house systems one discovers that they are used in a considerable broader way than thought of in 1972. This can be described by integrated information systems where structure retrieval in different databases is just one component; others are reaction retrieval, data retrieval, word processing, and so on.

This development was considerably driven by technological developments such as PCs and computer networks. Client-server technology supports this company-wide and broader information management.

The user of such a system is primarily the research chemist who needs this way of information management as a support for his daily work. To fulfill this demand, the user-interface of this system has to be intuitive enough for the nonexpert. The chemist at Bayer—as an example—has access to the following sources:

- Internal information (compounds, reactions, reports, spectra, screening data, etc.)
- Available chemicals (in stock internally or externally)
- Patents (on CD-ROM)
- Patent information (WPI file and documentation abstracts)
- Beilstein Database
- Crossover to the most important public online information (e.g., CAS, SpecInfo, etc.)

To support other aspects of his work the internal structure retrieval system has to be connected with other internal systems. Just to mention two examples: The structure editor used for information retrieval should also be connected with the word processor used for preparing his reports. The information of these reports finally should be managed also by the different retrieval components of the in-house system: for structures, reactions (ORAC), data (ORACLE) and full text (TRIP).

The end-user access to information is a major goal for information management at Bayer. We are convinced that information contributes considerably to the efficiency of R&D. Apart from hardware and communication, user-friendliness of software is an important requirement. An end-user oriented system based on the former terminal-host technology was in operation at Bayer since the early 1980s. This was used by more than 1000 users within the company. The modern client-server technology is now replacing this well established system. To keep this Integrated Chemistry System (ICS) as easy as possible and offering even more functionality we had to decide on some general standards, such as just one editor for chemical structures in all the applications within the system (Figure 1). Here we decided on the ISIS product line by MDL which is used for drawing structure diagrams for internal reports or for structure queries in our internal databases.

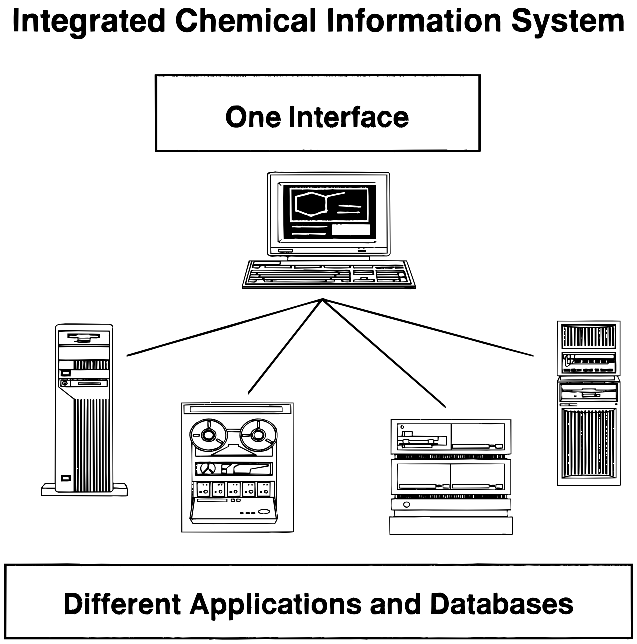

Integrated Chemical Information System
One Interface
Different Applications and Databases

Figure 1.

Today, more than 750 000 internal compounds are registered. A further 5.7 million compounds comprise the database of the former IDC. In most cases chemical structures are just the key for other information as properties, screening data, or research reports concerning this particular compound or class of compounds. The documentation abstracts of IDC are a valuable source for patent information due to the high quality of indexing. The WORM technology is used for these documents.

The chemist can use the same editor to perform searches in important external chemical databases, simply by formulating his query within the internal system, then using our crossover routine (Cross-Over) that submits the query to the selected databases at STN and brings back the hits. To avoid unexpected surprises, this crossover tool gives an estimate before starting the external searches for the probable costs. Another tool such as our View-Client is of great help to analyze the hits.

The same editor is used for Available Chemicals. By this the chemist searches for a compound in the needed purity and quantity in an inventory containing available chemicals within the company and at external suppliers. As soon as the chemist has made a decision, the compound can be ordered electronically by the same system. The order is thus submitted to the suitable supplier (Aldrich, Fluka, Merck, etc.)

The current management of reactions is still performed with ORAC. Altogether, more than 2.7 million reactions are available, the majority of them licensed from external sources such as the CIRX reactions (by MDL) based on ChemInform; several collections on heterocyclic reactions, the ORAC box, and Theilheimer (by ORAC); high quality databases on special topics such as protecting groups and others (by Synopsis); and the reactions from ZIC/VINITI (by Infochem). We expect a better integration of the reaction databases into our ICS with the Reaction Library Browser announced by MDL.

Based on TRIP (by PSI, now Fulcrum) there are several applications for full text searches. Screening data are managed by many applications based on ORACLE. In addition to that, more than 1.5 million patent documents of the most important patent offices are of easy access to the bench chemist via the internal network. This application has been realized together with Holthaus & Heinisch using jukebox technology.

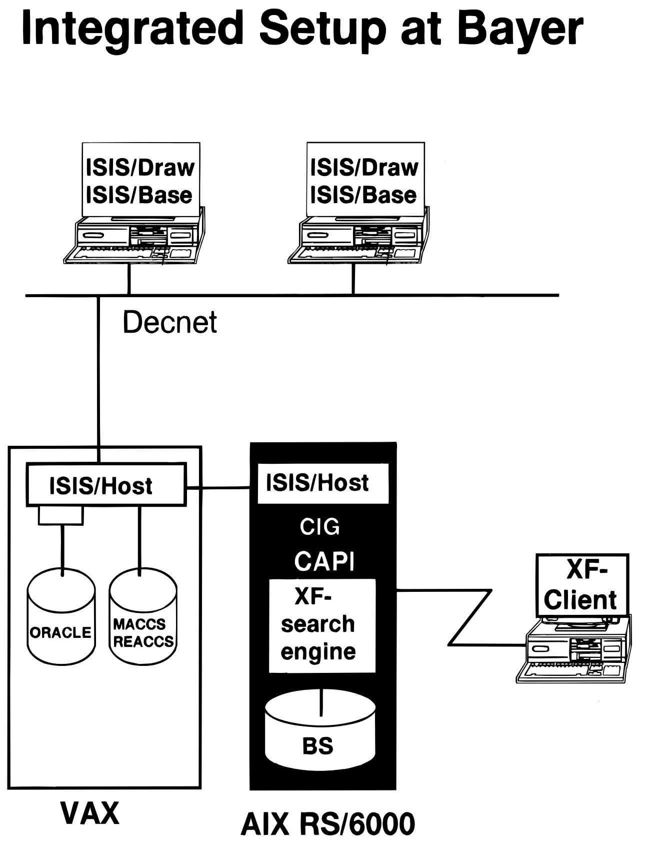


Figure 2.

Even if all these products come from the market, much effort goes into the connections and communications between the different products to end up with applications that conveniently support the needs of the users.

BEILSTEIN

In such a situation one always has to consider the pros and cons of an in-house installation of a database that can also be accessed on a public host. Beilstein is an excellent example of this, as the Beilstein information is on the market as

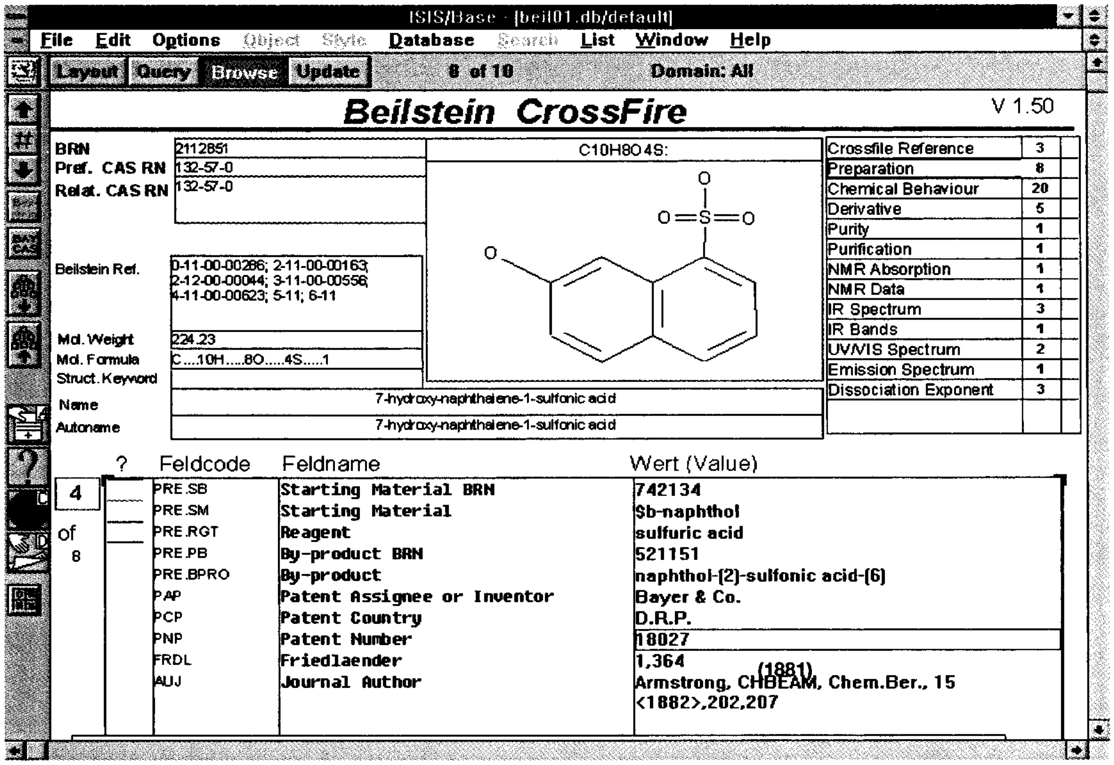
- •Handbook
- •Online on DIALOG and STN
- •Inhouse Database
- •Current Facts on CD-ROM

Certainly it holds not alone for Bayer that the attractiveness of the Handbook is decreasing somewhat. One reason is the fact that with an increasing number of volumes it becomes more and more difficult immediately to find the answer to a certain problem within the 480 volumes. Here, the computer offers clear advantages.

An in-house installation of Beilstein, however, needs to be integrated into the in-house system, to ensure *one* user interface allowing for many applications (Figure 2).

The Beilstein database and CrossFire had to fit into our world, where chemists are using the structure editor of ISIS, where the communication within the network runs via TCP/

"Bayer Acid": Patent from 1881


ISIS/Base - [beil01.db/default]
File Edit Options Object Style Database Search List Window Help
Layout Query Browse Update
8 of 10
Domain: All
Beilstein CrossFire
V 1.50
C10H8O4S:
O
O=S=O
O
4
of
8

BRN	2112851
Pref. CAS RN	132-57-0
Relat. CAS RN	132-57-0
Beilstein Ref.	0-11-00-00286; 2-11-00-00163; 2-12-00-00044; 3-11-00-00556; 4-11-00-00623; 5-11; 6-11
Mol. Weight	224.23
Mol. Formula	C....10H.....8O.....4S.....1
Struct. Keyword	
Name	7-hydroxy-naphthalene-1-sulfonic acid
Autoname	7-hydroxy-naphthalene-1-sulfonic acid

Crossfile Reference	3
Preparation	8
Chemical Behaviour	20
Derivative	5
Purity	1
Purification	1
NMR Absorption	1
NMR Data	1
IR Spectrum	3
IR Bands	1
UV/VIS Spectrum	2
Emission Spectrum	1
Dissociation Exponent	3

?	Feldcode	Feldname	Wert (Value)
	PRE.SB	Starting Material BRN	742134
	PRE.SM	Starting Material	$b-naphthol
	PRE.RGT	Reagent	sulfuric acid
	PRE.PB	By-product BRN	521151
	PRE.BPRO	By-product	naphthol-[2]-sulfonic acid-[6]
	PAP	Patent Assignee or Inventor	Bayer & Co.
	PCP	Patent Country	D.R.P.
	PNP	Patent Number	18027
	FRDL	Friedlaender	1,364
	AUJ	Journal Author	Armstrong, CHBEAM, Chem.Ber., 15 <1882>,202,207 (1881)

Figure 3.

IP, and where ORACLE and ISIS-Host are used. In a joint effort of Beilstein and IBM we succeeded with this installation under these named conditions.

The relevance of Beilstein information in comparison with other sources is also important. To answer this question simply by counting about 7 million organic compounds at Beilstein and about 12 million Registry Numbers would neglect the specific profiles of both sources. Both sources are widely complementary as the following aspects show (Figure 3):

- The time range of the Beilstein database goes back into the 18th century.
- The Beilstein database is unique in containing patent information prior to about 1960, although there is no patent information after 1979.
- The Beilstein system of property fields gives precise and easy-to-get data for organic chemists, process engineers, and analytical chemists.

These arguments show the importance of the Beilstein information in addition to the information of CAS. They also explain the strength of the database in comparison to the handbook. But why not use the data at STN via the above mentioned crossover tool?

Here, CrossFire has to be considered: CrossFire is a product finding an excellent reception by the end-user. It offers an intelligent user-front end which we regard as an example for many other products. It not only allows better access to the information contained in Beilstein but also does this in an intelligent and—at least for the chemist—intuitive way.

To integrate CrossFire into our Integrated Chemistry System (ICS) based on components the software market is offering has to be seen in the light of the efforts we undertook to allow the chemist to retrieve all that information needed for daily work. The many casual users prefer to have just one editor for chemical structures rather than to learn different techniques for searching in chemical databases, and especially the Beilstein information is of considerable value for the bench chemist. This effort has been appreciated by our chemists: The use of this source of valuable and unique information (bibliographic data, properties, synthesis) is still increasing. After the facts were also searchable within our Integrated Chemistry System, the usage surpassed the former one at STN. As the user will be charged back for both, this increase of usage indicates the growing appreciation and the demand for the Beilstein information by the bench chemist.

CAS REGISTRY FILE

Finally, I would like to mention another interesting topic in the field of chemical information. Chemical nomenclature is of high importance for the experts but not beloved by all chemists. Often it has been demonstrated how difficult it is to find the correct name for a compound. Universally, chemists prefer to use the structural formula. This is a very convenient characterization beyond any language barrier; it already associates ideas about synthesis or properties. Despite these advantages there are two difficulties: It is difficult to read for a nonchemist and the structural formula can hardly be written (or drawn) by a normal text-editor.

From this the CAS registry number draws it importance. Even if the relation between structural formula and its registry number is just of the same quality as between the name of a person and its telephone number—it simply is an arbitrary number. The directory connecting numbers and the structural formulas is in this case the Registry File.

This registry number in turn is widely used in many applications, such as chemical catalogues and many lists submitted to governmental offices world wide such as EINECS (European Inventory of Existing Commerical Chemical Substances), TCSA Inventory, CHEMIST, NDSL, etc. This kind of usage is not only an advantage to the chemical community but also is so widespread that it implies that the connection between number and formula must be public domain.

This, however, is just one side of the coin. One also has to see that the value of the registry file is strongly connected to the permanent effort of somebody (creating new numbers, updating old numbers, elimination of mistakes). This effort is done by CAS, and there are costs connected to that. At a time when everybody asks for an allocation of costs and talks about the value of information it is just understandable that CAS considers the registry file as proprietory information.

This obviously is the dilemma. The chemical community would prefer an even broader use of the CAS registry number, also using it to connect information in different databases such as Derwent and Beilstein or in in-house systems. It would considerably improve the time consuming task to connect information from different sources. In that sense it could improve the use of information and increase the efficiency. On the other hand, even if each of the databases by CAS, Derwent, Beilstein, and all the others has its own merits, there are overlaps between the databases—and that means competition. Can one of the competitors help the others to improve their competition?

A solution should be found that pays regard to the interests of both CAS and the chemical community. It would be incredible if the only restriction in the general use of the CAS Registry Number could be in the interest of Chemical Abstracts Service or the American Chemical Society.

CONCLUSION

The problems and possibilities of structure manipulations was the topic in Noordwijkerhout in 1972. By then, the tree was just in blossom. Meanwhile, many of the fruits are ripe and used everyday in different applications of information management. Beilstein—and so Prof. Jochum and Prof. Luckenbach—gave important contributions to that development.

My congratulations to Clemens Jochum and Reiner Luckenbach for the honor of receiving the distinction of the Skolnik Award in 1995.

REFERENCES AND NOTES

(1) *Computer Representation and Manipulation of Chemical Information*; Wipke, W. T., Heller, S. R., Feldmann, R. J., Hyde, E., Ed.; John Wiley & Sons: New York, 1974.

(2) 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.

(3) Schenk, H. R.; Wegmüller, F. Substructure Search by the Chemical Abstracts Service Chemical Registry II System. *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 153−161.

(4) Graf, W.; Kaindl, H. K.; Kniess, H.; Warzawski, R. The Third BASIC Fragment Search Dictionary. *J. Chem. Inf. Comput. Sci.* **1982**, *22*, 177−181.

(5) Dittmar, P. G.; Farmer, N. A.; Fisanick, W.; Haines, R. C.; Mockus, J. The CAS ONLINE Search System. 1. General System Design and Selection, Generation, and Use of Search Screens. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 93−102.

(6) Attias, R. DARC Substructure Search System: A New Approach to Chemical Information. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 102−108.

CI950248L