CSIRO PUBLISHING

Essay

www.publish.csiro.au/journals/ajc

Aust. J. Chem. **2004**, *57*, 387–392

The Challenges with Substance Databases and Structure Search Engines

Helen Cooke[A] *and Damon D. Ridley*[B,C]

[A] Chemistry Department, UMIST, Manchester M60 1QD, UK.
[B] School of Chemistry, University of Sydney, Sydney NSW 2006, Australia.
[C] Author to whom correspondence should be addressed (e-mail: d.ridley@chem.usyd.edu.au).

Structure connection tables, which are commonly used for the representation of chemical structures in electronic databases, are valuable for substances where specific valence bond structures are known or can be drawn. However there are many classes of substances (for example alloys, catenanes, polymers, or salts) which cannot be fully represented by valence bond structures. There are also issues of definition (such as when a substance is a co-ordination compound, or hydrate, or salt), and of bonding types (resonance, donor complexes, π-complexes). Producers of chemical substance databases may address these issues in different ways and generally need to introduce concepts (for example multicomponent substances) with which chemical scientists may not be familiar. In addition to these aspects of database content, the searcher needs to understand the algorithms behind the structure search engines. For example, the *SciFinder* search engine has considerable in-built chemical intelligence at the initial search level and then has many tools to mine the data once obtained; the *CrossFire* search engine also employs several algorithms by default and allows further options to vary them.

Manuscript received: 7 May 2003.
Final version: 19 February 2004.

Introduction

Currently over 30 000 new substances are described each day, and only computers can keep track of this volume of data. With today's electronic storage capabilities the challenge is not with the storage of information, but in how the data is stored.

Systematic ways to digitize chemical structure information is one of the challenges being addressed by IUPAC Division VIII (Systematic Nomenclature and Structure Representation) that is responsible for maintaining and developing standard systems for designating chemical structures including both conventional nomenclature and computer-based systems.

Members of that Committee have a lot of material with which to work. For example, between the excellent 1974 review 'From van't Hoff to unified perspectives in molecular structure and computer oriented representation' by Gasteiger et al.[1] and his most recent article with Engel 'Chemical structure representation for information exchange',[2] there have been many suggestions on how to address the issues.

Most of these reports focus on computing aspects,[3–17] or on solutions to specific problems (such as stereochemical representations,[18–20] Markush structures,[21] or polymers[22,23]) but there are no reports on the general and practical issues confronted by everyday scientists, who simply want to extract structural information.

Each method of searching for information through electronic media has its own pitfalls. For example, because of our knowledge of the variations in the ways scientists write their manuscripts, we recognize that any keyword search should be approached with caution. We are aware of the need to search for synonyms and to allow for variations through truncated search terms. Accordingly we know that our initial keyword search may neither be comprehensive nor precise, but generally speaking we get enough reasonable answers with which to work or else we can relatively quickly adopt alternative terminologies to be used as better search terms.

However the situation may be different when we search for structures or sequences. We are aware of non-electronic structure and sequence conventions, and mostly we assume they are handled in a similar way when it comes to computer storage. In many instances this is not the case and the dangers of this assumption are considerable. If the information has been stored in the computer in some different way, then given that computers are unforgiving and will only return the information for which we have asked, we may completely miss what we wanted.

We attempt here to explain the complications, and of what we need to be aware as we seek substances through structure-based queries. Issues with input and searching of nucleic acid and protein sequences present different challenges and have been discussed in part elsewhere.[24]

Structure Storage in Substance Databases

Structure connection tables are widely used to store substances in computer databases. They are basically valence

 10.1071/CH03315 0004-9425/04/050387

Table 1. Substances classes not easily represented by valence bond structures

Alloys	Catenanes and related species	Donor complexes
π-Complexes	Co-ordination compounds	Host–guest complexes
Hydrates and solvates	Mixtures	Polymers
Resonance	Salts	

bond representations of molecules and fail for species that cannot be well represented by a single valence bond structure. Engel and Gasteiger[2] say this is true for 'many organometallic complexes such as ferrocene, for electron deficient compounds such as boranes, for singlet–triplet states such as carbenes, and for radical cations such as the species observed in mass spectroscopy'. In fact this is a gross oversimplification of the problem! Valence bond structures show deficiencies in a host of situations and the more common ones are listed in Table 1.

There are other issues with building substance databases, including for example how the allotropes, combinatorial libraries, isotopic forms, physical states (solid/liquid/gas), post-treated polymers, stereoisomers, or substances incompletely described should be handled.

As examples of just the last case, imagine we were creating a database for substances and wished to store substance/structure information where the document described 'xylene', or 'alanine', or 'calcium (in the blood)', or 'deuterated ethanol', or even 'dideuterated ethanol'. There is no single substance called xylene (instead, *o*-, *m*-, or *p*-xylene), and is the 'alanine' described one of the enantiomers or the racemate (or some other mixture of the enantiomers)? What is meant by the chemical species 'calcium in the blood'? Meanwhile most of us can readily work out there are five valence bond structures for 'dideuterated ethanol' and two of them have the further complication of a stereogenic carbon, so how would we enter the substance referred to 'dideuterated ethanol' in the database?

Consider also the case of a newly discovered enzyme. If we know its function we may call it 'superbug destructase', but its structure will vary slightly from one organism to the next. There are additional issues of primary, secondary, tertiary, and quaternary structures. Consider also the cases where substances are in equilibrium, for example tautomers and carbohydrates (and the mutarotation issue).

Clearly database producers must formulate several policies, of which we may well need to be aware before we perform searches.

Substance Databases and Structure Representations

Most chemists are familiar with the substance databases produced by Beilstein and Gmelin,* and with the Registry database[25] from the Chemical Abstracts Service,† end-user desktop versions of which are available through *CrossFire*[26,27] and *SciFinder*.[28] Many new substance databases are now appearing on the Web, and some are

Table 2. Substance listings for multicomponent substances in major substance databases

Collated April 2003. Data was obtained from the STN versions of the databases

Number of components	Beilstein	Gmelin[A]	Registry
1	7 800 000	1 000 000	44 000 000
2	740 000	55 000	2 700 000
3	46 000	19 000	600 000
4	4600	5000	400 000
5	500	1700	250 000
6	47	1000	160 000
7	7	800	100 000
8	1	400	64 000
9	0	270	37 000
10	0	160	18 000
>10	0	300	18 000

[A] Gmelin differentiates between the number of components and number of fragments.

structure-searchable. It would be convenient if each database addressed the issues of structure storage and query input in similar ways, as potentially this would overcome differences between structure query interpretation in different systems. The benefits of a common interface for chemical database searching have been discussed by Cooke and Schofield,[29] and the IUPAC Division VIII is working on a new notation for designating structures which may overcome some of these differences.

While these rules may help new providers, it is questionable if the traditional database producers will change existing policies. Accordingly we all would be wise to make sure we understand how substances of interest to us are entered.

Multicomponent Substances

Fortunately many substances may be represented with unambiguous valence bond structures, and database producers handle these with structure connection tables. So we need only to address here the special situations in Table 1. In turn, these are often handled through the concept of 'multicomponent substances'.

The concept of multicomponent substances has been introduced into electronic structure databases to represent substances where a single connection table of valence bond structures cannot easily be used to completely describe the substance. Substances are thus represented as components made up of the separate valence bond structures.

For example, borax (a hydrate in Table 1) is indexed in Registry as a two-component substance in which one of the components is the anhydrous form ($B_4Na_2O_7$) and the second component is water. In turn the molecular formula of borax is entered as B4Na2O7.10H2O.

Multicomponent substances in computer substance databases always will have 'dot disconnected' molecular formulae of this type, and there may be many components listed. Table 2 summarizes the numbers of listed substances with

* www.mdl.com; www.beilstein-institut.de

† www.cas.org

various numbers of components in three of the major substance databases. The point is that over 10% of substances are listed in this way, so we certainly are not talking about an insignificant issue.

Classes of Multicomponent Substances

Alloys, catenanes (and rotaxanes), host–guest complexes, hydrates and solvates, mixtures, polymers, and salts usually cannot be represented by single valence bond structures and hence are entered as multicomponent substances.

Even so, there may be further issues to consider. For example water molecules may be present in some substances as 'true' hydrates whereas in other substances they may be attached to metal ions as part of co-ordination complexes. Different registrations result and different databases treat the issues differently.

The position becomes more confusing with salts where the basic indexing policy is that salts are represented through their acid and base components. However the nature of the acid or the base components may vary depending primarily on the periodic table group of the atoms involved. Thus an ammonium salt is generated from an acid with the base ammonia, but what is the base involved in the formation of sodium salts? We can all have our own opinions on this but basically what we need to know is the database policy and here it is easiest to look at a few examples.

The molecular formula for sodium acetylide in each of the Beilstein, Gmelin, and Registry databases is C2HNa, although the former two databases consider it is made of two 'fragments'; in Registry it is a single compound with structure Na—C≡CH. On the other hand, the molecular formulae for sodium acetate are C2H3O2.Na (Beilstein), C2H3NaO2 (Gmelin), and C2H4O2.Na (Registry).

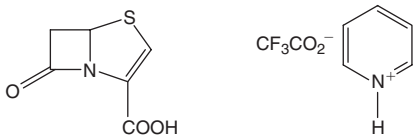
Whereas the sodium salt of the acid **1**, Scheme 1, is listed in Beilstein with molecular formula C6H4NO3S.Na and in Registry as C6H5NO3S.Na, both databases give the same molecular formula (C5H5N.C2HF3O2) for pyridinium trifluoroacetate, which a chemist would draw as structure **2**. The point is that the entries for salts in these two databases are similar in one case but different in a subtle way in another case.

Co-ordination compounds provide further challenges, and not only in relation to whether a compound is a co-ordination compound or a salt. For example, consider the various registered substances in Fig. 1, and consider how we may find the substances through name, formula, or structure searches.

Bonding Issues

The most common bonding issue is resonance. Although substances are represented graphically with valence bond structures the structure connection table may easily be made to handle alternate bonding arrangements, for example in aromatic rings.

Donor bonding (or, more specifically, when one of the atoms provides both electrons to the bond) has some interesting twists. Thus we do not really worry about whether we draw dimethyl sulfoxide as $Me_2S{=}O$ or $Me_2S^+{-}O^-$, but computers may not like this inconsistency. Beilstein, Gmelin, and Registry choose the former representation, and then the



1 **2**

Scheme 1.

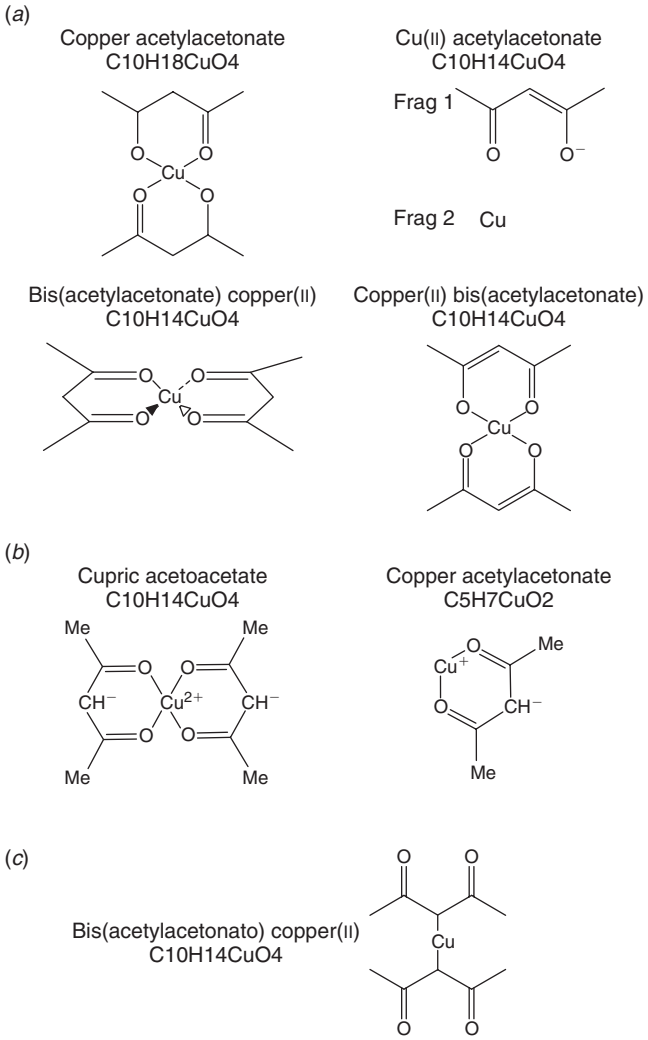
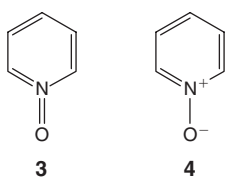


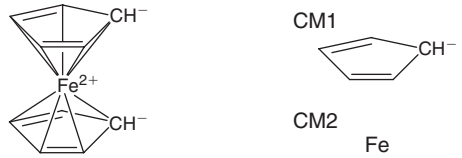
Fig. 1. Different registrations for copper acetylacetonate and related substances in (*a*) Gmelin, (*b*) Registry, and (*c*) Beilstein.



3 **4**

Scheme 2.

database producers usually apply the rule in related cases. However, while pyridine *N*-oxide is represented as structure **3**, Scheme 2, in Beilstein and Registry, it is represented as structure **4** in Gmelin. We may feel uncomfortable about the 'pentavalent' nitrogen but the computer doesn't worry.



5 **6**

Scheme 3.

π-Bonding situations are represented in Gmelin and Registry by additional connections between the atoms involved and in these databases ferrocene appears as structure **5** (Scheme 3; molecular formula $C_{10}H_{10}Fe$ as expected). However ferrocene in Beilstein looks like **6** and the molecular formula is 2 C5H5.Fe.

Polymers

Only Registry includes organic polymers, and the basic rule is that the polymer is indexed through the individual monomers. So ABS (CAS Registry Number 9003-56-9) is a three-component entity where the components are acrylonitrile, butadiene, and styrene. The molecular formula is (C8H8.C4H6.C3H3N)x.

In this case we don't know the precise structure, because the individual monomers may be distributed relatively randomly throughout the polymer backbone. In other cases, for example the nylons, the precise structure is known. Such polymers are said to have 'structure repeating units' (SRUs) and these are given connection tables which may be searched by structure.

Other policies may be applied to post-treated polymers, block polymers, and so forth, and it is helpful if the searcher has at least a basic understanding of indexing before trying to find the substances required.

Equilibrium Issues

Still other cases involve substances in equilibrium, and the most common examples are the tautomers. Here database producers are very much guided by the authors so if the document specifically mentions (or characterizes) the enol form then it will be indexed. Both Beilstein and Registry have different registrations for the keto and enol forms of ethyl acetoacetate.

Of course tautomers are not limited to ketones/enols. They may also occur with most situations H–X–Y=Z ⇌ X=Y–Z–H, but database policies apply in the registrations. Thus 2-hydroxypyridine and 2-pyridone have separate registrations in Beilstein and Gmelin, but appear under a single registration in Registry.

In other cases, for example the equilibria between cyclic and acyclic forms of carbohydrates, the substances involved are indexed separately.

Structure Searching

The final consideration is the way in which the search engine works. In particular, a key issue is to what extent the search engine has been programmed to help overcome the very considerable complications in data storage discussed above. Does the search engine have 'chemical intelligence', can it interpret the user's real requirements, and, if so, how does it do it?

The three major substances databases, Beilstein, Gmelin, and Registry, are available on the STN Network[‡] and may be searched by structure through the software *STN Express*[§] or through plug-ins associated with STN on the Web.[||] The interfaces provide some chemical intelligence, for example resonance issues (but not issues with tautomers) are covered automatically.

The Beilstein and Gmelin databases are also part of MDL *CrossFire*,[¶] while the Registry database may be searched through *SciFinder*[**] or *SciFinder Scholar*.[††] The search engines behind these, and the STN structure search engine, all are different so again it helps to have knowledge of what happens 'behind the scenes'.

Using STN terminology, there are three types of searches possible: exact, family, and substructure. If the structure drawn is the acid **1** then an exact structure search will add hydrogens (or hydrogen isotopes or charges—but no other non-hydrogen atoms) automatically to vacant positions. Answers will include the exact substance, plus isotopic and stereoisomeric forms.

A family search (also called 'related' on *SciFinder*) will retrieve all substances from the exact search, plus substances in which these exact structures are part of multicomponent registrations. So a family search on the acid **1** will additionally retrieve all the various salts. In all instances family (rather than exact) searches should be considered from the outset since, for example, the adverse (or beneficial) effects of an acid or base may be reported only through its salts. A challenge in pharmaceutical science is to obtain biologically acceptable forms of the drug and this may be possible only through the salts or through various types of encapsulation. These derivatives will nearly always be registered as multicomponent substances which will be found in the family/related structure search.

The default search type on Beilstein and Gmelin is an exact search. Searches which allow for retrieval of compound families as defined above are not possible by setting a single search option, but instead several options may be set individually or in combination by checking the appropriate boxes which allow retrieval of compounds related to the query structure, for example multicomponent structures, charged compounds, and radicals.

A substructure search allows further substitution at all vacant positions, and so a substructure search on acid **1**

[‡] www.cas.org/stn.html
[§] www.cas.org/ONLINE/STN/discover.html
[||] stnweb.cas.org
[¶] www.mdl.com
[**] www.cas.org/SCIFINDER/scicover2.html
[††] www.cas.org/SCIFINDER/SCHOLAR/index.html

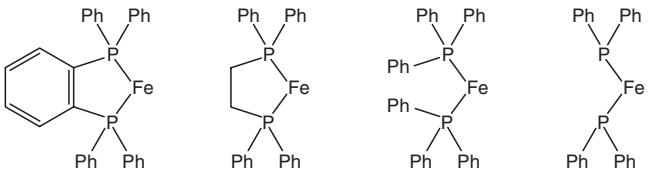
will also include sulfoxide and sulfone derivatives. However all the search engines have options to block substitution at any atom if required. Thus, on *CrossFire* there is the option to undertake substructure searches at several different levels including substitution at all sites in the query structure, at heteroatoms only, or at individual sites specified by the user. In the last option, the exact or a maximum number of substituents at any individual site may be specified.

An important issue with substructure searching on all search engines is the defaults applied, and there are two primary concepts. The concept of 'isolated or embedded rings' relates to whether additional rings fused on rings drawn in the query structure are to be allowed; the concept of 'ring formation' relates to whether atoms in chains drawn in the query structure should only be part of chains in answers or whether the atoms may also be part of rings. For example, the common substructure for part structures **7–9** is **10**, Scheme 4. However whether substances of the type **7** and **8** will be retrieved through a substructure search of structure **10** depends on whether the search engine allows both ring and chain values for the bonds in **10**. Some search engines do assign rings and chains as defaults, others do not—so know your search engine.

Issues with Intelligent Search Engines

Given all the issues with structure input and with indexing, the question is: How intelligent do we want the search engine to be? For example, do we want it to allow automatically for issues with indexing, for tautomerism, for salts or co-ordination compounds, for incompletely described substances? And do we want it to guide us through the various issues in searching, such as for stereoisomers?

Most of us would say: 'Of course!' But when we start to think through how this may be done, several additional considerations arise. For example, we would expect a substructure search of **11** to retrieve **12** and probably **13**, Scheme 5. But if we are to build a 'tautomerism algorithm' into the search engine then questions arise as to whether **14** should be retrieved. To allow for **14**, the search engine may have to allow also for **15**.



Scheme 4.

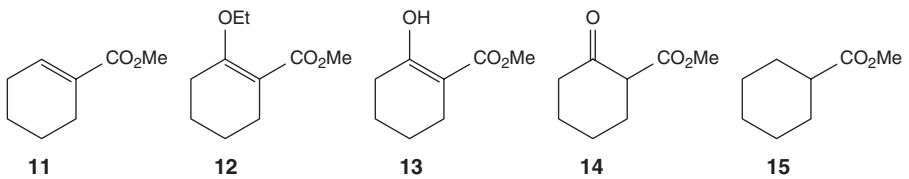
For example in *SciFinder* an exact/related search on either **16** or **17**, Scheme 6, produces the same answer set (currently 66 substances). They include multicomponent substances (the sodium salts and the like), isotopic substances, and *E*/*Z*-stereoisomers (relating to **16**). Use of the various tools under 'Analyze or Refine Substances' then enables the searcher to work through the issues. Thus multicomponent substances, or isotopic substances, may be excluded, and the option 'Analyze Substances by Precision' gives histograms of the tautomeric categories. If we start with the query **16** or **17** then the precision analysis shows 7 and 52 substances respectively under the heading 'Conventional Exact'. We can get exactly what we want, but in the process *SciFinder* alerts us to issues.

With *CrossFire* the same type of search may be done, but the approach is different. Once the structure has been built, several structure options may be selected. For example, the search for compounds **16** and **17** gives the same number of compounds (currently 44) as long as the tautomers box is checked before running the search. Rather than analyzing the results after the search, structure options may be turned on or off in any combination before running the search.

SciFinder guides the user through the issues of stereochemistry. If we draw a query with a certain stereochemistry *SciFinder* provides a histogram of the classes of answers—substances with the exact absolute configuration, the enantiomers, the racemic substances, the diastereoisomers, and those with stereochemistry not specified. The philosophy behind *SciFinder* is to give initially answers in which all database entry issues have been interpreted automatically, and to apply chemical intelligence. Then *SciFinder* provides the tools to obtain more precise answers if required.

CrossFire Beilstein approaches sterochemistry differently. Structures may be built with appropriate wedge bonds and the user then may set the stereo search capability to absolute, relative, or racemic, or to leave the stereo search set to 'off', which retrieves compounds in all configurations. Double bonds in the *cis*- and *trans*- configurations may also be specified and the stereo search option set to 'absolute' or 'off'. Generally it is best to run searches as 'flat' searches in the first instance if a comprehensive search is needed since not all compounds in the database have their stereochemistry specified. If the answer set retrieved is too large or if

Scheme 6.



Scheme 5.

only a few sample answers are needed, then the appropriate stereochemical search option may be selected.

CrossFire Gmelin does not allow searches for specific stereoisomers at the present time.

In addition to structure input and search options, it is helpful to understand that individual compounds may be represented in the literature in different ways by different authors, and these differences may be treated differently by different database producers. So at times several search approaches may be necessary for a comprehensive search.

Conclusions

Many chemical aspects need to be addressed in building structure databases. Either the searcher needs to be aware of these and ask appropriate queries, or else the searcher needs to rely on structure search engines that apply the necessary chemical intelligence. Even here, it is helpful to have an understanding of the issues.

This paper has summarized several factors, and has given illustrations from only the Beilstein, Gmelin, and Registry databases. However there are many other substance databases which have structure search options, and the issues addressed here apply equally well to them.

With the very large, and ever increasing, number of substances we are totally reliant on computer databases and structure search interfaces, and scientists need to understand the content of the databases and the way the searches are executed in order to retrieve comprehensive and/or precise answers.

References

[1] J. Gasteiger, P. D. Gillespie, D. Marquarding, *Top. Curr. Chem.* **1974**, *48*, 1.
[2] T. Engel, J. Gasteiger, *Online Inf. Rev.* **2002**, *26*, 139. doi:10.1108/14684520210432431
[3] I. P. Bangov, M. I. Spasova, *Bulg. Chem. Commun.* **1996**, *28*, 443.
[4] J.-h. Yao, *Jisuanji Yu Yingyong Huaxue* **1998**, *15*, 193.
[5] J.-h. Yao, *Jisuanji Yu Yingyong Huaxue* **1998**, *15*, 65.
[6] J.-h. Yao, *Jisuanji Yu Yingyong Huaxue* **1997**, *14*, 81.
[7] J. Silhanek, *Chem. Listy* **1997**, *91*, 237.
[8] C. A. G. Tonnelier, J. Fox, P. Judson, P. Krause, N. Pappas, M. Patel, *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 117. doi:10.1021/CI960094P
[9] A. Dietz, *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 787.
[10] C. Wentang, Z. Ying, Y. Feibai, *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 604.
[11] A. J. Gushurst, J. G. Nourse, W. D. Hounsell, B. A. Leland, D. G. Raich, *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 447.
[12] E. Meyer, *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 68.
[13] J. M. Barnard, in *Chem. Inf., Proc. Int. Conf.* (Ed. H. R. Collier) **1989**, p. 209 (Springer: Berlin).
[14] G. von Kiedrowski, A. Eifert, *Intell. Instrum. Comput.* **1986**, *4*, 110.
[15] R. C. Haines, *Comput. Chem.* **1989**, *13*, 129. doi:10.1016/0097-8485(89)80005-1
[16] W. Fisanick, A. H. Lipkus, A. Rusinko, *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 130.
[17] W. Fisanick, K. P. Cross, J. C. Forman, A. Rusinko, *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 548.
[18] K. K. Agarwal, H. L. Gelernter, *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 463.
[19] T. Akutsu, *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 414.
[20] L. M. Staggenborg, *Spec. Publ. R. Soc. Chem.* **1993**, *120*, 89.
[21] W. Fisanick, *US Patent 4642762* **1987**.
[22] S. J. Teague, *Polym. Preprints* **2002**, *43*, 821.
[23] W. V. Metanomski, *Polym. Preprints* **2002**, *43*, 785.
[24] R. Austin, D. D. Ridley, *Chem. Aust.* **2002**, *69*(*4*), 13.
[25] D. W. Weisgerber, *J. Am. Soc. Inf. Sci.* **1997**, *48*, 349. doi:10.1002/(SICI)1097-4571(199704)48:4<349::AID-ASI8>3.3.CO;2-H
[26] S. R. Heller, *The Beilstein System: Strategies for Effective Searching* **1998** (American Chemical Society: Washington, DC).
[27] P. Meehan, H. Schofield, *Online Inf. Rev.* **2001**, *15*, 241. doi:10.1108/14684520110403768
[28] D. D. Ridley, *Information Retrieval: SciFinder and SciFinder Scholar* **2002** (John Wiley: Chichester).
[29] F. Cooke, H. Schofield, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1131. doi:10.1021/CI010360L