In Search of Organic Zeolites – Does Modern Information Retrieval Inevitably Become a 'Sieving-the-Desert' Exercise?

by **Engelbert Zass***

Informationszentrum Chemie Biologie, ETH Hönggerberg, HCI, CH-8093 Zürich
(phone: +41 1 632 29 64, fax: +41 1 633 1297, e-mail: zass@chem.ethz.ch)

and

Dietmar A. Plattner, **Albert K. Beck**

Laboratorium für Organische Chemie, ETH Hönggerberg, HCI, CH-8093 Zürich

and

Markus Neuburger

Institut für Anorganische Chemie, Universität Basel, Spitalstrasse 51, CH-4004 Basel

Dedicated to Professor *Dieter Seebach* on the occasion of his 65th birthday

Information retrieval for planning and executing research projects and for publishing results is considered a routine task that is usually neither mentioned explicitly in a scientific publication nor described in any detail. In the information searches for the preceding publication ('Building an Organic Zeolite from a Macrocyclic TADDOL Derivative or How to Teach an Old Dog New Tricks'), we were confronted with so many problems during retrieval of the desired information about related work that we decided to deviate from this tradition. We had to use the *Cambridge Structural Database*, the *Chemical Abstracts* structure and literature databases, and the *Beilstein* database to the full extent of their contents, indexing, and search facilities to retrieve the necessary information about 'organic zeolites'. In the process, we found important limitations and deficiencies in any one of these databases, and we had to conceive search procedures that we considered rather unusual even after more than 20 years of experience in searching chemistry databases. The results and, particularly, the problems encountered underline the necessity for enhanced integration of individual compound and property databases and improved standardization as a prerequisite for this.

1. Introduction. – The quest for organic nanoporous materials, *i.e.*, organic molecules composing frameworks that do not alter upon ingress or egress of guest compounds, has often been claimed to be of general importance [1][2], but has, so far, not been posed in a systematic fashion [3]. In the context of a purely organic system with the desired properties of a zeolite, we were eager to know how many precedents are documented in the literature, always with the possibility in mind that the host and host/guest structures were published without note or reference to each other. The problem was defined as follows: all organic molecules characterized by single-crystal X-ray or neutron-diffraction analysis have to be matched to counterpart structures containing the same molecule ('main component') but differing in composition. If both structures are isomorphous, such a system is a possible candidate for an organic zeolite.

During the preparation of the preceding publication entitled 'Building an Organic Zeolite from a Macrocyclic TADDOL Derivative or How to Teach an Old Dog New

Tricks' [4], we were not satisfied with just a few examples of organic zeolites, but desired to retrieve as many examples of this phenomenon as feasible. Regarding the availability of large, powerful databases of chemical compounds and their properties, and, particularly regarding the kind of advertising these commercial information sources get, we assumed that we would be able to satisfy our information needs. The numerous difficulties encountered in this search will be described herein, as well as the deficiencies of the above databases, which, *inter alia*, hindered us from achieving the desired comprehensiveness.

The question stated above can, in principle, be perceived as an information retrieval problem in two different ways:

A) Searching for *compounds* that have the described properties, a prerequisite for this approach being appropriately searchable crystal structure data

B) Searching for the *phenomenon* ('organic zeolites'), a prerequisite being appropriate and consistent author terminology or indexing.

Problem *A* involves compounds and their properties and must, thus, be searched in suitable compound/property databases, while *B* involves describing the problem with appropriate keywords in a suitably indexed literature database.

Generally speaking, keyword searches are simple to execute, but almost impossible to run in a comprehensive way because of the immense variation in chemical terminology used by different authors and by the abstracting or indexing services that produce the databases. Compound searches can usually be phrased more precisely, particularly when full or partial structures are involved. In our case, however, structures were the desired output of our search and could be defined only indirectly by their properties, not *via* structural formula. This simple analysis indicates the complexity of our type of question, a complexity that was borne out in full in the course of our endeavors.

2. Preliminary Searches. – After this formal analysis of the information retrieval problem, we first tried to look for all compounds with the appropriate crystal-structure data. A thorough investigation of the kind of isomorphism we were interested in involves a comparison of the cell parameters of every compound for which a crystal structure has been reported and which contains at least two components with their cell parameters. Candidates are those compound pairs (a multi-component compound and one of its components) where the cell parameters are similar within a certain degree of deviation. Regarding the phenomenon of 'organic zeolites', only those cases of isomorphism with an identical unit-cell choice were considered.

2.1. Cambridge Structural Database. When looking for crystallographic data of organic compounds, the first and foremost information source is, of course, the *Cambridge Structural Database* (*CSD*) [5]. Although this database contains, in principle, all the information needed to answer our question (*cf. Chapt. 4*), the standard user interfaces *ConQuest* or *Quest* do *not* permit the kind of data comparisons outlined above. For the time being, we, therefore, had to look for different ways to retrieve the desired compounds.

2.2. Chemical Abstracts. Since the first attempt at approach *A* failed, we tried to search for the phenomenon '*organic zeolites*'. The most comprehensive literature

source for chemistry and related fields is the *Chemical Abstracts* (*CA*) database[1]), which roughly corresponds to the traditional printed *Chemical Abstracts*. This database is offered in several variants by several public hosts and by the *Chemical Abstracts Service* (*CAS*) [6] itself. For our keyword searches, we used the *CAplus* [7] version of this database and the *SciFinder Scholar* [8] interface for searching it. This selection was based on *CAplus* being the most comprehensive and up-to-date version of this database and *SciFinder Scholar* providing a natural-language interface to facilitate keyword searches by automatically taking care of different grammatical forms (singular, plural *etc.*) of search terms, abbreviations, different spellings (US/UK), and, albeit to a limited extent, also of synonyms.

As a starting point for our search, we already knew of a review article by *Lee* and *Venkataraman* [1] about organic zeolites. When looking for articles on a phenomenon or other topic that needs to be described with keywords in a database search, it is standard procedure to 'retrieve' the articles already known to be relevant and analyze their indexing in the database for any useful clues to describe the topic at hand. The indexing of [1] is reproduced in *Fig. 1*.

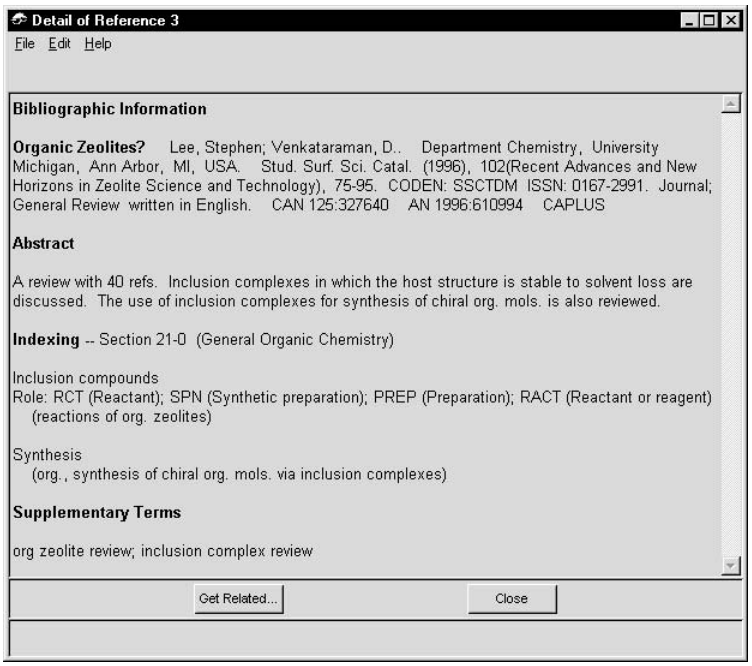


Fig. 1. *Record in* CAplus *database for the review by* Lee *and* Venkataraman [1], *showing the* CAS *indexing for this article* (© American Chemical Society)

---

[1]) For a sensible discussion about information retrieval, one must clearly differentiate between *a*) the original databases like *CAS Registry*, *CA*, *CAplus*, *Beilstein*, *Cambridge Structural Database* (*CSD*), *Science Citation Index* (*SCI*), *b*) the versions of these databases publicly available *via* hosts (*e.g.*, *STN Registry*, *STN CA*, *STN CAplus*, *STN Beilstein*), and *c*) the interfaces used to search these databases at hosts or *via* client-server systems, *e.g.*, *STN Messenger*, *SciFinder Scholar*, *CrossFire*, *Quest*, *Web of Science*, etc.

Unfortunately, the information shown in *Fig. 1* was not very useful. We, therefore, attempted to search for 'organic zeolites' and related topics, using appropriate phrases and entering them into the natural-language interface of *SciFinder Scholar*. This interface in the 'explore by topic' mode analyzes user input and automatically identifies important keywords and their relationships. The search terms are then combined and automatically permuted in different fashion, as shown in *Fig. 2*, *a* and *b* for two searches used here[2]).

In some searches, we made use of the excellent 'refine' and 'analyze' features in *SciFinder Scholar* that are typical for this search interface for the *CAS* databases. With 'refine', additional keywords or other searchable information, like publication year or document type (*e.g.*, review), are used to narrow the number of references retrieved. 'analyze' can be used to identify important information like author names or indexing terms in the references found for further refinement.

We performed a whole series of searches for a variety of search phrases that we considered as potential descriptions of our topic in *SciFinder Scholar*. The most important ones are summarized here:

a) 'Organic zeolites': 1877 references

b) 'Organic zeolites (organic zeolithes)': 1881 references[3]), stepwise refined by the additional terms 'structure' (351 references) and 'solvent' (38 references). Inspection of the bibliographic data of these references (title, abstract, indexing) showed only one article to be marginally relevant to our question

c) 'Structures that are isomorphous upon removal of solvent': 18 references with 'all terms present anywhere in the reference' (none relevant); 86 references with 'structures', 'isomorphous', 'solvent' closely associated with one another (*i.e.*, leaving out 'removal'), further refined after 'analyze by index term' ('crystal structure', 'isomorphism') and then by *CA Section Title* ('Crystallography and Liquid Crystals', 'Phase Equilibriums' *etc.*)[4]) to give four remaining references, one being marginally relevant

d) 'Crystal structure unchanged upon solvent removal': one reference, not relevant

e) 'Inclusion of solvent with isomorphous structure': 10 references, none considered relevant

---

[2]) The detailed processing in *SciFinder Scholar* is not transparent to the user and is considered proprietary by *CAS*. Obviously, the term 'anywhere in the reference' refers to the *Boolean* AND operator, and the term 'closely related' to a proximity operator on the sentence level like the L operator used in our *STN* searches (*cf. Chapt. 3.1*). The term 'as entered' refers to the exact phrase, while 'as concept' implies, *inter alia*, substitution of the original search terms with different spellings, grammatical forms (singular, plural), and some synonyms, and relaxing the restriction that the terms of the phrase must be in the order given with no intervening words.

[3]) We included here the different spelling 'organic zeolithes' in parentheses. In principle, *SciFinder Scholar* is expected (and advertised in this sense) to automatically take care of such different spellings and even real synonyms. The case shown here, however, is one of many examples where this is obviously incomplete. Due to the immense complexity of chemical terms, this is not too surprising; *CAS* produced, in our experience, a very good interface, but users must be aware of its unavoidable shortcomings.

[4]) *CA Sections:* General fields of chemistry (at present: 80) used to arrange the literature references abstracted in the print version of *Chemical Abstracts*, useful in the *CA* literature database for refining searches. A reference can only be in one section (*i.e.*, that of its main topic), but may be cross-referred to further sections relevant to its content.

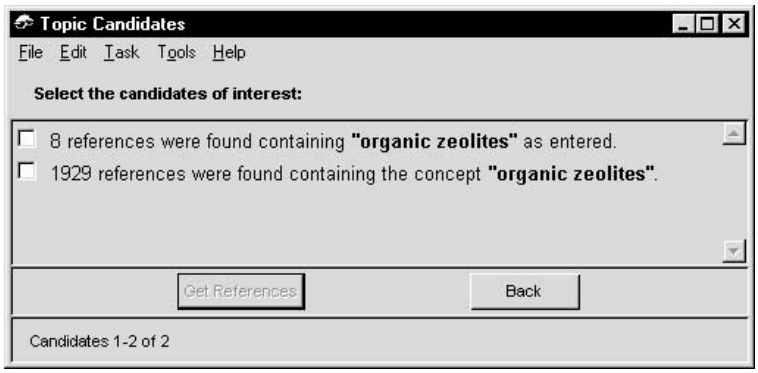


Fig. 2,a. *Result screen for keyword search* ('Explore by Topic') *of the term 'organic zeolites' in* SciFinder Scholar (May 23, 2002, © American Chemical Society)

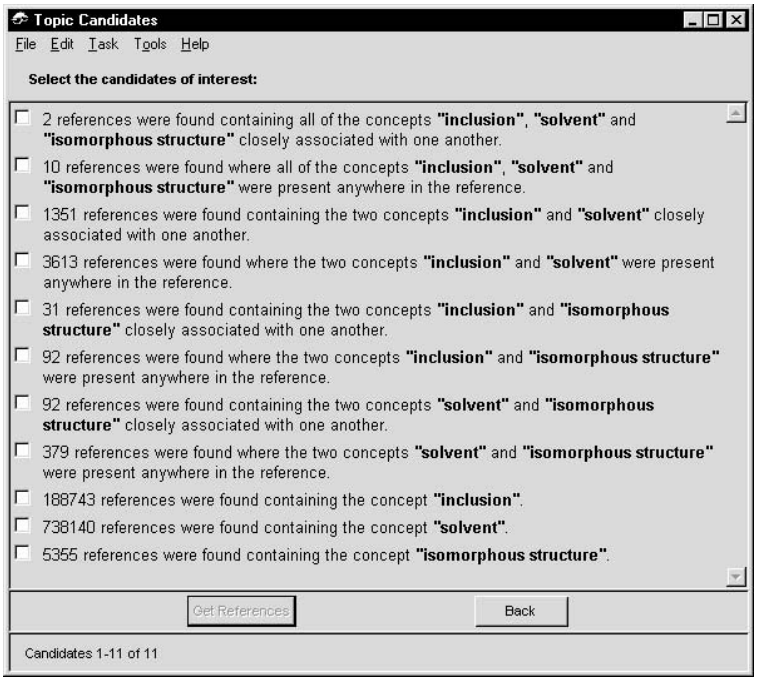


Fig. 2,b. *Result screen for keyword search* ('Explore by Topic') *of the term 'organic zeolites' in* SciFinder Scholar (May 23, 2002, © American Chemical Society)

f) 'Inclusion compounds with isomorphous structures': 18 references, none considered relevant

g) 'Isomorphous structure upon solvent removal': five references, not relevant

h) 'Isomorphous structures': 5304 references. When additionally analyzed by 'Index Term' and refined by 'isomorphism': 98 references; when analyzed by 'Index Term' and refined by 'crystal structure', 'molecular structure': 28 references, four being of interest

i) 'Single crystal to single crystal transformation': 10 references for phrase as entered; 2348 references for the concepts 'single crystal' and 'single crystal transformation' closely related, after refining with 'isomorphous': 6 references (none relevant).

The *SciFinder Scholar* interface permits searches *via* structures (not used here), *via* topics (phrases), or author names. This interface was designed to be readily used by chemists not specifically trained in information retrieval and does not, therefore, permit the sort of complex searches that are possible in the *CAS* databases. The *STN Messenger* [9] command-driven interface is more difficult to use than *SciFinder Scholar*, but it provides the power of combining terms with *Boolean* logic and other operators to generate complex queries. Several attempts to rephrase the queries used above in *STN Messenger* did not give results more useful than the ones retrieved before.

The scarcity of relevant references found in these searches, as well as the failure to identify further relevant terminology for our question in the retrieved references, convinced us that keyword searching would not answer our question in an appropriate way because we could never be certain about what we had found and what we had missed.

2.3. Beilstein. In a first attempt to locate compounds of interest in the *Beilstein* database [10], we used the *CrossFire* version to search for crystalline compounds with more than one component and for which isomorphism had been reported. The following query was used for the data ('fact') search in this database: (csg or bisub = 'crystal structure determination') and nf >1. The data field 'csg' stands for 'crystal space group', 'bisub' for 'basic index for substances', *i.e.*, an index for all compounds in the *Beilstein* database that contains information about compound properties as keywords, in contrast to the properties that are reported in individual data fields[5]).

The above query retrieved 19223 compounds. It was then refined with the command 'bisub = polymorph*'[6]) to give 119 compounds. Manual inspection showed that most of them were simple metal or hydrogen halide salts, but 30 compounds, mainly solvates or charge transfer complexes, were considered to be of interest, but not followed up in detail at this stage of the search process[7]).

2.4. *Citation Searching.* The fact that we already knew relevant articles on organic zeolites enabled us to use a third approach to answer our question: citation searching.

---

[5]) Regarding the enormous number of properties known for chemical compounds, the editors of factual databases like *Beilstein* and *Gmelin*, which cover the entire range of chemical and physical properties (in contrast to specialized factual databases, *e.g.*, with only spectral or thermodynamic data), have to decide which properties to represent in specific data fields, allowing precise and detailed searching, and which properties to treat in a more generic fashion by describing them simply with appropriate keywords that are grouped together. For many properties, it is advisable to search for them both ways, as in our example.

[6]) The asterisk (*) is a wildcard for 'any number of any characters', to take care of the spelling variants we found inspecting the 'bisub' index in *Beilstein*. By checking this index, we also found that the German term 'isomorph' is not used in this database; this is in clear contrast to the *Chemical Abstracts* database and shows that the terminology used in different databases cannot be assumed beforehand, but must be checked individually.

[7]) With four exceptions (three due to missing space-group information in the *Beilstein* database, one being a three-component system), all of these compounds were later retrieved in our improved *Beilstein* search (*cf. Chapt. 3.2*).

Thus, looking for all later publications that cite the paper by *Lee* and *Venkataraman* [1] should lead to related publications.

Citation searches are very often a useful alternative in searching for phenomena or other topics to the keyword search used above. While the latter is based on matching the terms used by the author(s) and by the abstracting and indexing service to describe the content of an article by the searcher – no mean feat, and almost impossible to execute in a comprehensive yet precise way – citation searching is based on 'links' already established by the author(s) from the publication at hand to earlier publications related in content. Both approaches have their obvious weaknesses, although they are, to some extent, complementary. Citation searching is completely dependent on the authors citing prior work in a way that is correct not only by content, but also formally (no errors in the literature references given).

The 'classic' source for citation searching is the *Science Citation Index* (SCI) [11] produced by *ISI* [12] extending back to 1945. This database is available *via* several different providers and user interfaces, including *Web of Science* [13]. Recently, *CAS* has also added citation data to the *Chemical Abstracts* (*CA*) database, starting in 1999. We used both databases to search for publications that cite the review [1] or a paper by *Dianin* [14] who, in 1914, had synthesized a compound that should later become the classic example for an organic zeolite. In the *Web of Science* [13] version of the *SCI*, we found 50 references citing *Dianin*'s work [14]. None of them was considered relevant upon inspecting the titles. The recent review article by *Lee* and *Venkataraman* [1] was cited by a publication about organozeolite materials that was also not considered relevant. With the command 'get related information: citing references' in *SciFinder Scholar* [8], we retrieved only two nonrelevant papers citing *Dianin* [14] plus two publications citing [1], which were also of no use to us[8]).

2.5. *Author Searching.* Again starting from the relevant paper at hand, we performed author searches in *SciFinder Scholar* with the command 'explore by author' for *A. V. Dianin* [14] as well as for *S. Lee* and *D. Venkataraman* [1]. Among the twelve papers co-authored by *Lee* and *Venkataraman* that we retrieved, several were of interest and partially relevant to our topic.

2.6. *Evaluation of Preliminary Results.* Both the direct approach in the *Cambridge Structural Database*, as well as keyword, citation, or author searches for the phenomenon 'organic zeolites' in *Chemical Abstracts* and *Science Citation Index* fell short of our expectations, qualitatively as well as quantitatively. The relatively few relevant references retrieved rather whetted our appetite for more than fulfilling our need for a comprehensive collection of organic compounds possessing the desired property. As we saw no chance to modify the keyword-based approach to search for this collection in a useful way, we returned to the compound approach for the rest of our searches.

---

[8]) The large difference of citing publications for the *Dianin* article (50 in *SCI* [11] *vs.* only two in *CAplus* [7]) is explained by the different time coverage of these databases for citation data: *SCI* since 1945, *CAplus* only since 1999 (this coverage refers to the citing publication, not the cited one, as, otherwise, we would have found nothing for *Dianin*'s publication from 1914!). The *SCI* missed one of the two citing references found in *CAplus* because of a database error in the reference list taken from the citing publication in *Tetrahedron*.

3. Identification of Desired Compounds. – To identify potential candidates for 'organic zeolites', we relied on the following general approach, restricting our search in all databases to *two-component systems* (*i.e.*, one host with one guest):

A) identify all two-component systems where X-ray-structure analyses are known

B) restrict these compounds to those where an X-ray structure is also known for at least one of the components

C) check the detailed crystal-structure data for these candidates in the *Cambridge Structural Database* and then turn to the original literature for verification.

We were very well aware that steps *A* and *B* would involve handling a very large number of compounds, taxing the abilities of commercially available retrieval systems to the utmost, and that the key problem in this approach would be to reduce the number of compounds entering the labor-intensive step *C* without losing too many relevant candidates: 'sieving the desert' indeed[9])!

3.1. Chemical Abstracts *Databases. Preliminary Searches.* The *CAS Registry* [15] structure database, by far the largest compound database available, was used as a starting point. One of the purposes of this attempt was also to find out whether it was feasible at all to run a rather general type of search in a very large database with the retrieval systems publicly available at present.

The common approach for searching by structure or substructure in the *CAS Registry* database is not possible in our example, since structures are the desired results, not the query for the search, which is the property 'crystal structure'. Our example is, thus, data-driven, not structure-driven.

Fig. 3 shows a typical entry (record) for an organic compound in the *CAS Registry* database [15]. Besides structure, nomenclature, ring description, and calculated physical properties, this record in '*STN* Files' shows other databases at the host *STN International* [16] that contain information about this compound, which is searchable *via* a unique compound identifier, the *CAS Registry Number*.

Our problem would be easily solved if the '*STN* Files' field contained a flag for all organic compounds for which crystal structures have been reported in the *Cambridge Structural Database*. In contrast to, for example, biomedical databases (*e.g.*, *BIOBUSINESS*, *BIOSIS*, *DDFU*, *DRIUGU*, *EMBASE*, or *MEDLINE* shown in *Fig. 3*), this important reference information is missing entirely for crystal-structure databases. This made it necessary to use the rather roundabout, tedious routes described below. These strategies were also forced upon us by the fact that structures and literature in *Chemical Abstracts* are in two separate databases (in contrast to the *Beilstein* database, *cf. Chapt. 3.2*[10]).

For the above reasons, the following laborious protocol was followed:

1) Search for references to publications in the *CA* literature database [17] containing crystal-structure analyses (our desired property)

---

[9]) We found this appropriate methaphor in *Sir Arthur Stanley Eddington*'s book 'New Pathways in Science', Cambridge University Press, Cambridge, 1935, p. 263.

[10]) For historical correctness and fairness, it should be mentioned here that the *CA* literature database and the *CAS Registry System* were developed during the introduction of electronic data processing in *ca.* 1965–1970. While the literature database was made publicly available already in 1972, the more complex and demanding *CAS Registry* structure database became accessible in 1980. The *Beilstein* database was started only in 1984 and became publicly available in 1988, thus profiting from dramatic advances in software and hardware technology in the meantime.

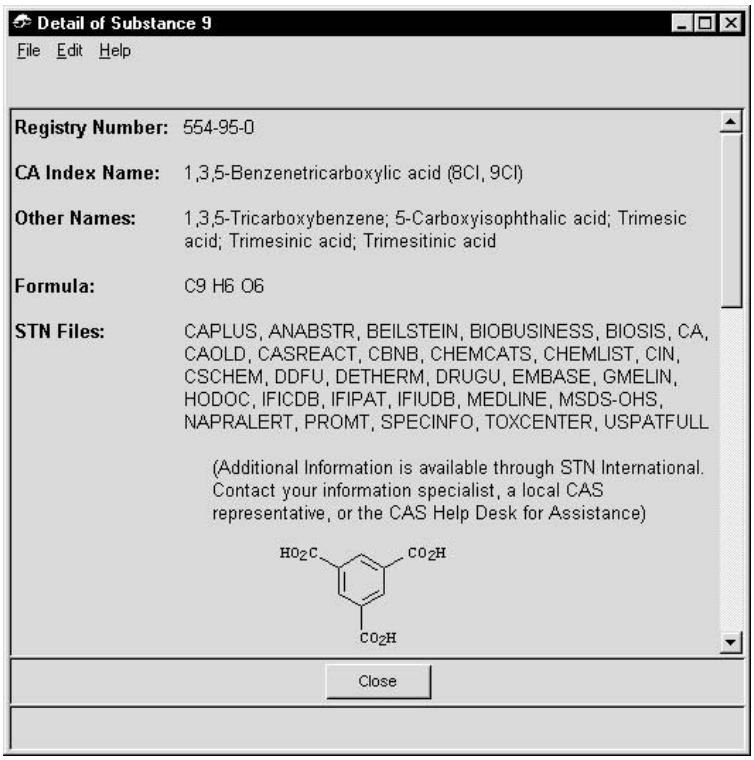


Fig. 3. *Database record for trimesic acid in* CAS Registry (first part of the record as shown in *SciFinder Scholar*, © American Chemical Society)

2) Extract *CAS Registry Numbers* for all compounds in these literature references (*i.e.*, compounds for which the crystal structure was determined as well as those that were indexed from a publication in a different context)
3) Transfer these *CAS Registry Numbers* to the *STN Registry* compound database to form a set and apply some constraints
4) Cross the set back to the *CA* database and search the 'raw' candidates in the context of crystal-structure analysis
5) Extract the 'hit' *CAS Registry Numbers* (*i.e.*, only numbers for those compounds that had crystal-structure data according to the indexing in the *CA* database)
6) Transfer the *CAS Registry Numbers* for the desired compounds to the *CAS Registry* compound database to form a set for further processing.

After several search sessions, we developed the search strategy shown in *Fig. 4*.

In an application of our basic strategy outlined above, we started with a set of 639509 literature references in the *STN CA* literature database (*Fig. 4:* L1: FILE = CA). We set out to deal with crystal-structure analyses in a broad sense by fulfilling the search condition L2, which contains the word stem 'cryst?' (truncation, with the '?' functioning as a *wildcard* for 'any number of any characters', taking care of the *CAS* abbreviation 'cryst' as well as 'crystal', 'crystals', 'crystallographic', *etc.*) or the word

```
L1  (   2572353)SEA FILE=CA STRUCT?
L2  (    639509)SEA FILE=CA (CRYST? OR MOL?) (2A) L1
L3  (     19276)SEA FILE=CA ISOMORPH?
L4  (      6817)SEA FILE=CA L2 AND L3
L5              SEL  L4 1- RN :   19579 TERMS
L6  (     12535)SEA FILE=CA L2 (L) (SOLVENT? OR SOLVAT?)
L7  (      7832)SEA FILE=CA L6 AND PY > 1990
L8              SEL  L7 1- RN :   46265 TERMS
L9  (      4703)SEA FILE=CA L6 NOT L7
L10             SEL  L9 1- RN :   19411 TERMS
L11 (     19520)SEA FILE=REGISTRY L5
L12 (     46241)SEA FILE=REGISTRY L8
L13 (     19377)SEA FILE=REGISTRY L10
L14 (     69567)SEA FILE=REGISTRY (L11 OR L12 OR L13) AND C/ELS
L15 (     47050)SEA FILE=REGISTRY L14 AND 1/NC
L16 (     22517)SEA FILE=REGISTRY L14 NOT L15
L17 (   2572353)SEA FILE=CA STRUCT?
L18 (     19276)SEA FILE=CA ISOMORPH?
L19 (    829001)SEA FILE=CA SOLV?
L20 (     43120)SEA FILE=CA L16 (L) (L17 OR L18 OR L19)
L21 (     29662)SEA FILE=CA L20 AND L17
L22             SEL  L21 1- RN HIT :   13280 TERMS
L23 (     13106)SEA FILE=REGISTRY L22
L24       12502 SEA FILE=REGISTRY L23 NOT PMS/CI
```

Fig. 4. *Preliminary search strategy for candidate compounds in the* STN *databases* CA *and* Registry (February 13, 2002)

stem 'mol?' (*CAS* abbreviation 'mol', 'molecular' *etc.*) in close proximity to 'structure' (proximity operator (2A) in L2: 'a maximum of two intervening words' between 'mol? or cryst?' and L1 (='struct?'), in the order given or, as common in *CAS* indexing, in inverted sequence). As this number of 'hits' was far too large[11]) for the extraction of *CAS Registry Numbers* for the compounds indexed for these publications, we narrowed the number of literature references by the demands that (L4) either the term 'isomorph?' (again, truncation for 'isomorphous', 'isomorphic', 'isomorphism'[12]), *etc.*) appears anywhere (*Boolean* operator AND) in the *CA* database record for the publication (*i.e.*, either stated by the author(s) in publication title or original abstract, or by *CAS* document analysts in indexing or abstract) or, alternatively, that (L6) at least one of the terms 'solvent?' or 'solvat?' appears in close context with 'cryst? or mol?' (L2), with the proximity operator (L) specifying 'same sentence' (*i.e.*, either the same sentence in the abstracts, or the same index entry in the *CAS* indexing, *cf. Figs. 4* and *5*).

Extraction of the *CAS Registry Numbers* for all compounds indexed for these publications (step *2*) in our general strategy) with the 'SMARTselect' command in *STN*

---

[11]) The 'SMARTselect' command, now superseded by the improved commands 'ANALYZE' and 'TRANSFER' in the retrieval language *STN Messenger*, was used for extraction of the *CAS Registry Numbers*, and again proven to be absolutely indispensable for this kind of search. These operations are, at present, limited to an extraction of a maximum of 50000 terms from database records (literature references or compounds). Most operations were successful within this limit, and in the few cases beyond this, 'slicing' of the records sets was applied.

[12]) As an example, the following variations for this term, with many obvious misspellings, appeared in the *STN CA* database [17] (number of references in parentheses: valid for May 30, 2002): ISOMORHIC (1), ISOMORMORPHOUS (1), ISOMOROPHIC (1), ISOMOROPHOUS (1), ISOMORPBOUS (6), ISOMORPH (389), ISOMORPHAL (1), ISOMORPHES (1), ISOMORPHEUS (1), ISOMORPHI (1), ISOMORPHIC (3731), ISOMORPHICAL (1), ISOMORPHICALLY (328), ISOMORPHICITIES (1), ISOMORPHICITY (10), ISOMORPHICLY (4), ISOMORPHICM (1), ISOMORPHIES (1).

Messenger (abbreviated SEL in *Fig. 4*) went smoothly for the 6817 literature references in L4 to give 19597 *CAS Registry Numbers* (L5); for the 12535 literature references in L6, however, we hit the system limit for the maximum number of extracted terms in a first attempt. We, thus, split this set into two by introducing the publication year 'PY' as an arbitrary criterion[13]). The first slice (L7, L8) gave 7832 references, and 46265 extracted *CAS Registry Numbers*, the second slice (L9, L10) 4703 references, and 19411 *CAS Registry Numbers*. In the next operation, both slices were transferred to the *STN Registry* database (FILE = REGISTRY), combined in one result set, limited to carbon-containing compounds (L14: C/ELS = carbon/element symbol), and split into one-component systems (L15: 1/NC = number of components) and the remaining multi-component systems (L16: 22517 compounds).

For step *4*) of our general strategy, all these multi-component systems were searched again in the *CA* literature database in the context of either 'structure' (L17), 'isomorphism' (L18), or 'solvent/solvate' (L19) to yield 43120 literature references fulfilling these condition (L20). In order to reduce this large number and to increase the relevance of these references, we next demanded the presence of the term 'structure' (L17) not as an alternative, but as a requisite; this makes the full search condition read as follows: any publication where the term 'structure' appears *anywhere* in the database record, and where any compound from the 22517 multi-component systems specified before was indexed in *close context* with either 'structure', 'isomorphous', or 'solvent/solvate' (or, of course, any of the variants/spellings of these terms as specified by truncation). These constraints reduced the result to 29662 references. Now, the *CAS Registry Number* extraction was used again, but this time more specifically (L22: SEL RN HIT) for only those numbers that were 'hits', *i.e.*, which appeared in the required context[14]). 13280 Numbers then gave 13106 compound records in *CAS Registry*[15]). After eliminating polymers (L24: NOT PMS/CI = polymeric substance/class index), 12502 compound records for multi-component systems related to crystal-structure data remained for further processing. Of these, 8775 were two-component systems. This number was significantly reduced to 2867 by eliminating all metal-containing compounds (NOT M/ELS = metal/element symbol; simple salts as well as coordination compounds). Further elimination of simple salts with hydrogen halides, halides, tetrafluoroborates, perchlorates, and hexafluorophosphates (which were assumed to be of no relevance to our investigation) left us with 2002 two-component systems.

---

[13]) This kind of 'slicing' often helps to bypass system limits set to prevent one user monopolizing too much processing power in a public, commercial system. The price to pay for this is, in a literal sense, additional charges for two 'SMARTselect' queries and additional processing time.

[14]) SELECT permits one to extract either all *CAS Registry Numbers* for a given reference or only those that were searched for, which is the reason for the rather roundabout search strategy applied here – before we could look for specific registry numbers, we had to take them all and re-search them in the desired context!

[15]) In the *CAS Registry* database, every record carries a *CAS Registry Number* that uniquely identifies the compound described in this record. The obvious discrepancy between the total number of *CAS Registry Numbers* extracted from the literature references (L59: 13280) and the corresponding records in the *STN Registry* generated from them (L60: 13106) is that the indexing in the literature databases contains also deleted and alternate *CAS Registry Numbers* that are extracted and counted, while they are grouped together with the main registry number in the same record in the *STN Registry* database.

From these, we extracted the *CAS Registry Numbers* of the individual components (1903), sorted them by occurrence, and looked at the structures of the 100 most-common components to identify solvents and other interesting candidates for guest compounds in organic zeolites among them.

To get an idea about the utility of this rather large intermediate search result, we refined it by again using an important feature in the *STN Registry* compound database: for multi-component systems, the database record contains not only the *CAS Registry Number* assigned to this system itself, but also the registry numbers of its individual components. We extracted 11061 *CAS Registry Numbers* for the different components present in the 12502 multi-component systems retrieved before (*cf. Fig. 4*). In a stepwise refinement, with keywords used in the same fashion as shown in *Fig. 4*, we reduced these to 2177 single components from our multi-component systems that had indexed crystal-structure data in the *CA* database. When we restricted the literature to those publications that contained both at least one of our 12502 multi-component systems and at least one of our corresponding 2177 single components in the context of 'crystal/molecular structure', 'isomorphism' or 'solvent', we retrieved 1844 literature references. These were further restricted by demanding that the *Chemical Abstracts Index Heading*[16]) 'Crystal Structure' must be present, which gave 974 references. Extracting again the 'hit' *CAS Registry Numbers* in this context led to 3425 numbers, which were restricted to single components (1249), then to compounds without metals (658; *e.g.*, no salts, no complexes), and, finally, to those with fewer than ten C-atoms (233; among them solvents we viewed as candidates for inclusion in multi-component systems). Restricting the literature to those publications that contained structure information on one of the 233 single components as well as on one of the 1315 multi-component compounds containing any of the 233 single compounds, we finally retrieved 86 references. These were printed out in a free display format (DISPLAY SCAN TI SC HIT) that showed title, *CA Section*, and that part of the indexing containing any of the terms we had used in searching. Manual inspection showed 19 of those to be potentially relevant. The full bibliographic data for these references were then retrieved in *SciFinder Scholar*[17]).

---

16) *CA Index Headings* are standardized terms or phrases used by *CAS* to index the main topics of a publication; these headings are searchable in the printed *CA General Subject Index* and, of course, also in the database. Using these headings usually enhances the relevance (precision) of a search, but the 'price to pay' may be loss of relevant articles that were indexed differently.

17) This gave us 'the best of two worlds': we used the *STN CA* [17] literature and compound databases with a command-driven interface to realize our very complex search strategy that was absolutely impossible to execute in the *SciFinder Scholar* interface [8]. Using the powerful *STN CA*, however, not only demands more knowledge about and experience with *CAS* databases than does *SciFinder Scholar*, it also carries costs for connect time, every search term, and every reference/structure displayed in a format other than the free DISPLAY SCAN in *STN Messenger* [9]. For obvious reasons, the output of DISPLAY SCAN gives no bibliographic data to access the original literature, and it displays the information in random order. This prevents 'separating the chaff from the wheat', and, in our situation, we would either have to search each of the 19 interesting references (both tedious and expensive because of search term charges) or to print out all 86 references in a paying format, thus wasting money on 67 irrelevant references. Entering the titles of the 19 references of interest from the capture file of the *STN* search to the natural language interface of *SciFinder Scholar* retrieved the desired information easily, though, and at no additional cost (copy-and-paste instead of typing).

In evaluating these publications, we came across two new search terms that were obviously of importance for our topic. We, therefore, searched in *SciFinder Scholar* for all publications about 'pseudopolymorphism' (209 refs.), and 'clathrand' (7 refs.; March 5, 2002). Among several relevant publications retrieved, there was a particularly interesting one by *Nangia* and *Desiraju* [3] about pseudopolymorphism and the H-bonding of organic solvents in molecular crystals.

Possible Strategies. Although, at this stage of our investigation, we had to expect that the list of candidates we would retrieve from any search in the CAS databases might be both too large and not precise enough to be really useful for our purpose of identifying organic zeolites, we nevertheless decided to continue our searches as a general kind of 'feasibility study'[18]). Based on the results of the preliminary searches, we envisaged three different approaches as starting points for further selections:

1) All two-component systems (2604812[19])) in the *STN Registry* compound database containing C and H (2419357), except biopolymers (peptide or nucleotide sequences, leaving 2392672), polymers (2162999), metal salts (1489884), salts of hydrogen halides HX (X = F, Cl, Br, I), protonated forms (component = H), and simple salts with the anions of I, I_2, Cl, Br, F, H, BF_4, ClO_4, PF_6 (597799). These remaining two-component systems had 24146 references indexed with 'crystal/molecular structure', and 60066[19]) compounds occurred in this context.

2) All two-component systems with the second component from a list of 92 solvents and other candidates for inclusion (see *Appendix*). This particular approach was inspired both by the analysis of the most-common components of the multi-component compounds mentioned above, and also from the publication of *Nangia* and *Desiraju* who had investigated the *Cambridge Crystallographic Database* (*CSD*) for common solvent inclusions in organic clathrates [3]. With this list, we retrieved 182601[20]) two-component systems in the *STN Registry* database, excluding biopolymers (171794), polymers (165211), metal-containing compounds (145365), compounds with HX (X = F, Cl, Br, I), I, Cl, Br, F, H, BF_4, ClO_4, PF_6. The remaining 143687 two-component entries had a total of 72430[20]) references in the *STN CA*, 6805 of those were indexed with 'crystal/molecular structure'. The 'SMARTselect'[11]) command extracted 13191 two-component systems from the starting total of 143687.

3) Take the publications in the *STN CA* literature database indexed with the phrase 'crystal structure' or 'molecular structure' (taking care of abbreviations and different spellings as well of inversion of terms and intervening words by '(cryst? or mol?) (2A) struct?' as described above): 644107[21]), restrict these publications to those in the *Organic Sections*[4]) of *CA* (116907), but eliminate references from the *Organometallic Section* (76146 references). This approach was not pursued further, as the extraction of *CAS Registry Numbers* from these

---

[18]) At this time in our investigation, we had not yet decided on the intensive processing of data from the *Cambridge Structural Database* described in *Chapt. 4*.

[19]) Numbers given were taken from searches in *STN Registry* and *STN CA*, respectively, on April 8, 2002.

[20]) Numbers given were taken from searches in *STN Registry* and *STN CA*, respectively, on April 11, 2002.

[21]) Numbers given were taken from searches in *STN Registry* and *STN CA*, respectively, on April 4, 2002.

publications would give numbers too large to handle with reasonable effort within system limits. Extraction of the first 5000 references alone gave more than 50000 compounds!

All sets of compounds retrieved according to these approaches do need further refinement by indexing terms, both to enhance the precision (relevance) of the candidate compounds and to reduce them to a number that can be handled. Detailed analysis of the indexing showed that relevant compounds were indexed with the phrases 'crystal structure', 'mol. structure', or 'crystal and mol. structure'. Most, but not all of the indexed compounds also had the role[22]) 'PRP' (properties) assigned to them. This led us to the improved strategy shown in *Fig. 5:*

First, the 60066 two-component systems resulting from strategy *1*) were recalled from storage (L21 in *Fig. 5*) and restricted to those publications where they were indexed in the close context of 'crystal structure', as expressed by the search phrase '(crystal or mol?) (W) structure?', with the '(W)' proximity operator specifying that the terms 'crystal' or 'molecular' (*wildcard* '?' to take care of abbreviations, singular/plural, *etc.*) is immediately followed by 'structure' in the word order given (in contrast to the less-specific, nondirectional '(A)' operator used in *Fig. 4*). The remaining 8819 references were further restricted to those present in the *CAS Organic Sections*[4]) or those either present in or cross-referred to the section *Crystallography and Liquid Crystals* ('cryst?/SC,SX' = section code, section cross-reference). Extracting the pertinent *CAS Registry Numbers* from the 8402 references (L26, *Fig. 5*) retrieved 10947 compounds that consisted of 8434 individual components (L31). These were again restricted to those 3514 for which crystal structures had been determined for the two-component systems (L32 – L35). These systems were reduced to only those 8091 cases in which there was information available on both the total system and one of its components. Unfortunately, this number was much too large to be useful, and attempts to restrict it further failed; limiting the 6779 references to those with the *Index Heading* 'Crystal Structure' still gave 2580 references, while using 'Crystal Structure Determination' produced only one.

3.2. Beilstein. Regarding the relative failure of our strategy in the *Chemical Abstracts* database – basically due to the lack of *CAS Registry Numbers* in the *Cambridge Structural Database*, and also to shortcomings of the indexing in *CAS* databases, we turned to the *Beilstein* database [10] as a kind of 'last resort' for our problem. Looking at the *Beilstein* database from our point of view, one can say that it is not nearly as comprehensive as the *CAS Registry* regarding the number of compounds[23]), and not nearly as detailed in crystal-structure data as the *Cambridge Structural Database*. However, it turned out to be the only database available where we could directly execute the search strategy for the candidate identification outlined above.

---

[22]) *CAS* uses roles in indexing for compounds and compound classes to categorize the type of information found in the publications, *e.g.*, preparation, analysis, properties; see http://www.cas.org/ONLINE/QR/casroles.pdf.

[23]) Database statistics for compound databases like *CAS Registry* [15] or *Beilstein* [10] state only the number of records. These numbers are not really comparable because of the formal principles involved in computer registration of structures, and certainly not identical with the number of compounds a chemist would count.

```
FILE 'REGISTRY' ENTERED AT 09:07:24 ON 31 MAY 2002
                ACT EZNC2/A
               ---------
L1  (   2162999)SEA FILE=REGISTRY (C AND H)/ELS AND 2/NC NOT SEQUENCE/FS NOT PM
L2  (   1489884)SEA FILE=REGISTRY L1 NOT M/ELS
L3  (   1487315)SEA FILE=REGISTRY L2 NOT UNSPECIFIED/MF
L4  (    596799)SEA FILE=REGISTRY L3 NOT (CLH OR BRH OR HI OR FH OR H OR F OR C
L5  (    233555)SEA FILE=REGISTRY L4 AND C < 4
L6  (    119303)SEA FILE=REGISTRY (L4 NOT L5) AND C > 20
L7  (    243941)SEA FILE=REGISTRY L4 NOT (L5 OR L6)
L8  (     24146)SEA FILE=CA ((CRYST? OR MOL?) (2A) STRUCT? AND (L5 OR L6 OR L7)
L9              SEL  L8 20000- RN HIT :   11154 TERMS
L10 (     11100)SEA FILE=REGISTRY L9
L11 (   2162999)SEA FILE=REGISTRY (C AND H)/ELS AND 2/NC NOT SEQUENCE/FS NOT PM
L12 (   1489884)SEA FILE=REGISTRY L11 NOT M/ELS
L13 (   1487315)SEA FILE=REGISTRY L12 NOT UNSPECIFIED/MF
L14 (    596799)SEA FILE=REGISTRY L13 NOT (CLH OR BRH OR HI OR FH OR H OR F OR
L15 (    233555)SEA FILE=REGISTRY L14 AND C < 4
L16 (    119303)SEA FILE=REGISTRY (L14 NOT L15) AND C > 20
L17 (    243941)SEA FILE=REGISTRY L14 NOT (L15 OR L16)
L18 (     24146)SEA FILE=CA ((CRYST? OR MOL?) (2A) STRUCT? AND (L15 OR L16 OR L
L19             SEL  L18 1- RN HIT :   50548 TERMS  (TERM LIMIT EXCEEDED)
L20 (     50423)SEA FILE=REGISTRY L19
L21       60066 SEA FILE=REGISTRY L10 OR L20
               ---------

     FILE 'CA' ENTERED AT 09:09:41 ON 31 MAY 2002
L22      183262 S L21
L23      573347 S (CRYSTAL OR MOL?) (W) STRUCTURE?
L24        8819 S L22 (L) L23
                SET TERM L#
••••••••••••••••••••••••••••••••••••••••••••••••••••••••••••••••••••••••••••••
L26        8402 S L24 AND (ORG/FS OR CRYST?/SC,SX)
••••••••••••••••••••••••••••••••••••••••••••••••••••••••••••••••••••••••••••••
L28         SEL L26 1- RN HIT :   10973 TERMS

     FILE 'REGISTRY' ENTERED AT 09:22:54 ON 31 MAY 2002
L29       10947 S L28
L30         SEL L29 1- CRN :    8435 TERMS
L31        8434 S L30/RN

     FILE 'CA' ENTERED AT 10:44:31 ON 31 MAY 2002
L32     4536614 S L31
L33       11994 S L32 (L) L23
L34        8091 S L33 AND (ORG/FS OR CRYST?/SC,SX)
L35         SEL L34 1- RN HIT :    3514 TERMS

     FILE 'REGISTRY' ENTERED AT 10:51:09 ON 31 MAY 2002
L36     1674914 S L35/CRN
L37        8838 S L36 AND L29
                SAVE L37 EZNC21X/A

     FILE 'CA' ENTERED AT 10:53:38 ON 31 MAY 2002
L38        6779 S L37 (L) L23

     FILE 'REGISTRY' ENTERED AT 10:54:53 ON 31 MAY 2002
                ACT EZRN/A
               ---------
L39 (         1)SEA FILE=REGISTRY 100-01-6/RN
••••••••••••••••••••••••••••••••••••••••••••••••••••••••••••••••••••••••••••••
L131         92 SEA FILE=REGISTRY (L39 OR L40 OR L41 OR L42 OR L43 OR L44 OR L4
                ---------
L132        SEL L131 1- RN :      92 TERMS
L133     278581 S L132/CRN
L134       3391 S L133 AND L37

     FILE 'CA' ENTERED AT 10:56:39 ON 31 MAY 2002
L135       2720 S L134 AND L38
                E CRYSTAL STR/CV
L136       2580 S E4 AND L135
L137          1 S E10 AND L135
```

Fig. 5. *Improved search strategy for candidate compounds in the* STN *databases* CA *and* Registry (May 31, 2002)

Searching for properties and physical data in *Beilstein* is the *raison d'etre* of this database, and is, thus, straightforward. As we needed crystal-space-group data to identify candidates for organic zeolites, we searched for the presence of such data with the data-field abbreviation 'csg' (crystal-space-group) and combined this with 'nf' (number of fragments) for both two- and single-component systems to retrieve 11229 and 43255 compounds, respectively (*Table 1*).

Table 1. *Statistics for Search Results in* CrossFire Beilstein (BS0201, update 1st quarter 2001)

	Two-component systems	Single components
Query	nf = 2 and csg	nf = 1 and csg
Compounds retrieved	11229	43 255
Export data format[a])	BRN, MF, FBRN, CSG; CSG.L	BRN, MF, CSG
Entries (lines) in export file	11384	44633
Unique entries	11284	44603
Unique compounds	11229	43255
Processed lines	11229	44633
Systems with space-group data for at least one **component matching** space group	3919 **609**	
nonmatching space group	3310	

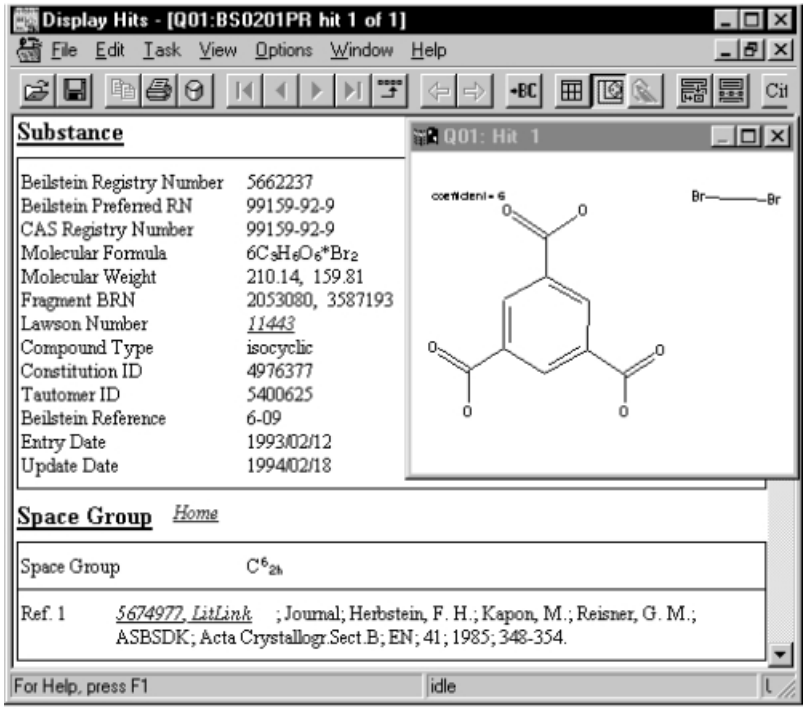
[a]) Data-field codes: BRN = *Beilstein Registry Number*, MF = molecular formula, FBRN = *Beilstein Registry Number* of fragments (components), CSG = crystal space group, CSG.L = crystal-space-group literature (added in the *Beilstein Commander* 'export wizard' in 'export – define data' with 'additional data to export: incl. references')

One of the results is shown in *Fig. 6*, but since we were not interested in the individual display of that many compounds and just needed their data for further post-processing to identify candidates for organic zeolites, we had to use the export facility in *CrossFire*. We defined the appropriate data fields for export, in particular the crystal-space-group data and the *Beilstein Registry Numbers*, essential in linking two-component systems to the data for their individual components.

Fig. 6 shows several different types of entries with exported data for two-component systems. The individual data fields (*cf. Table 1*) for each compound entry are separated by a vertical line that facilitates parsing. As a prerequisite for automatic postprocessing, we identified different types of entries in the export files by a preliminary analysis of the exported data using the 'awk', 'grep', and 'sort' commands on a UNIX server. While compounds with only one set of crystal-structure data were not problematic, giving rise to only a single entry (line) in the export file, there were 137 instances of compounds among the two-component systems with more than one data set, and, thus, more than one entry in the export file, three compounds with four entries, 12 with three, and 122 with two. These additional entries had the same *Beilstein Registry Number* and additional space-group information, but they lacked both the molecular formula and the *Fragment Beilstein Registry Number* (*Fig. 7, c*).

3.3. *Postprocessing of Search Results.* The primary search results needed a significant amount of postprocessing to produce the desired list of candidates, which

a)



b)

```
5662237|6C9H6O6*Br2|2053080//3587193|#C!%6&&2#h!%|5674977; Journal;
Herbstein, F. H.; Kapon, M.; Reisner, G. M.; ASBSDK; Acta
Crystallogr.Sect.B; EN; 41; 1985; 348-354;
```

Fig. 6. *Crystallographic data for trimesic acid from* CrossFire Beilstein (BS0201, update 1st quarter 2001). a) Display in *CrossFire*. b) Part of data as exported for processing (© *Beilstein-Institut zur Förderung der chemischen Wissenschaften*, © *MDL Information Systems GmbH*).

had to be compared with the data in the *Cambridge Structural Database*. The following processing steps were executed:

1) Integration of the fitting raw data for the single components into the two-component results (limiting all two-component compounds with crystal-structure data in *Beilstein* to those where at least one of the components also had crystal-structure data, as outlined in our general strategy)

2) Differentiation between the cases where *a*) the space groups for the two-component system and at least one of the components are identical and *b*) the remaining compounds where this is not the case

3) Manual selection and transformation of the information for the candidates in a format suitable for input/comparison with the *CSD*.

For the first step, we eliminated the 155 lines corresponding to the 137 two-component systems with more than one entry in the export file (*cf.* examples in *Figs. 7,c*

```
3583392|2C26H22O2*C4H8O2|3597060//506104|#C!%5&&2#h!%|5507200;
Journal; Harmata, Michael; Barnes, Charles L.; TELEAY; Tetrahedron
Lett.; EN; 31; 13; 1990; 1825-1828;

8652919|C18H20O2*2C2H6O|229912//1718733|#C!%2&&3#i!%|6251706; Journal;
Sandstedt, Christian A.; Michalski, Darek; Eckhardt, Craig J.; TETRAB;
Tetrahedron; EN; 56; 36; 2000; 6625 - 6632;
```

a) examples for special cases regarding CODENs

```
4881395|C6H10O4*2H3N|1209788//3587154|#C!%5&&2#h!%|5600098; Journal;
Teslya, I. A.; Tursina, A. I.; Iskhakova, L. D.; Avdonina, L. M.;
Marugin, V. V.; JSTCAM; J.Struct.Chem.(Engl.Transl.); EN; 31; 6; 1990;
956-960; ZSTKAI; Zh.Strukt.Khim.; RU; 31; 6; 1990; 123-127;

5712257|C33H35NO5*C3H2N2|5322562//773697|#C!%5&&2#h!%|5700157;
Journal; Eerden, Johan van; Grootenhuis, Peter D. J.; Dijkstra, Pieter
J.; Staveren, Catherina J. van; Harkema, Sybolt; Reinhoudt, David N.;
JOCEAH; J.Org.Chem.; EN; 51; 20; 1986; 3918-3920;5761678; Journal;
Grootenhuis, Peter D. J.; Eerden, Johan van; Dijkstra, Pieter J.;
Harkema, Sybolt; Reinhoudt, David N.; JACSAT; J.Amer.Chem.Soc.; EN;
109; 26; 1987; 8044-8051;
```

b) examples for special cases regarding crystal space groups

```
3897071|C36H60O30*CH4O|79627//1098229|#P!2&1%2&1%2&1%|543259; Journal;
James et al.; ACCRA9; Acta Crystallogr.; 12; 1959; 385;

5674176|C12H24O6*2C2H4O2|1619616//506007|#C!%3&&2#h!%|6095223;
Journal; Albert, Antje; Mootz, Dietrich; ZNBSEN; Z.Naturforsch.B; GE;
53; 2; 1998; 242-248;

5880542|C5H9NO2*ClH|80810//1098214|P2&1%/a (=C%5&&2h%)|1071702;
Journal; Mitsui et al.; ACCRA9; Acta Crystallogr.; <B> 25; 1969; 2182;
```

c) example for a two-component system with several data entries (exported data before processing)

```
4172386|C14H10*C10H2N4|1905429//1875027|#C!%3&&#s!%|5989107; Journal;
Stezowski, John J.; JCPSA6; J.Chem.Phys.; EN; 73; 1; 1980; 538-547;
4172386|||#C!%5&&2#h!%|5989107; Journal; Stezowski, John J.; JCPSA6;
J.Chem.Phys.; EN; 73; 1; 1980; 538-547;
4172386|||#C!%3&&2#h!%|5671806; Journal; Lefebvre, Jacques; Odou,
Gerard; Muller, Michel; Mierzejewski, Andrzej; Luty, Tadeusz; ASBSDK;
Acta Crystallogr.Sect.B; EN; 45; 1989; 323-336;
4172386|||#C!%5&&2#h!%|5671806; Journal; Lefebvre, Jacques; Odou,
Gerard; Muller, Michel; Mierzejewski, Andrzej; Luty, Tadeusz; ASBSDK;
Acta Crystallogr.Sect.B; EN; 45; 1989; 323-336;
```

d) example for missing space group information (see footnote 26)

```
3692530|2C2H3O2*Mg|1901470//3587170|aus dem Roentgendiagramm
ermittelt|1199282; Journal; Walter-Levy et al.; COREAF;
C.R.Hebd.Seances Acad.Sci.; 249; 1959; 1234, 1236 Anm. 3;
```

Fig. 7. *Raw data exported from* CrossFire Beilstein (BS0201, update 1st quarter 2001) *for two-component systems with crystal-space-group data* (see text for explanations; © *Beilstein-Institut zur Förderung der chemischen Wissenschaften*)

and *8, c*) and processed the remaining 11229 compound entries with a UNIX-shell script to substitute the FBRN (*Beilstein Registry Number* of the components, also called 'fragments') with the molecular formula and space group for the component taken from the 43255 single compounds with crystal-structure data (*cf. Table 1*). We next compared the space groups for the two-component system and the single component and wrote each line (representing a two-component system with single-component information) in either one of two files based on the results of this comparison. By manual inspection of the 609 compounds found with matching space groups, salts and other metal-containing compounds were removed, leaving 545 candidates. In order to be able to input these candidates into the *CSD* for an eventual comparison, or to compare these results with those generated by processing of data exported from the *CSD* (see *Chapt. 4*), we had to identify a common denominator with the *CSD*. Both the *Beilstein Registry Number* and the *CAS Registry Number*[24]) are not usable in this context, as they are not present in the *CSD*, and the 'systematic' names found in all compound databases discussed here were not consistent enough to be of any use. This left us with only the molecular formula as the 'common denominator'.

Unfortunately, even for a relatively small database like the *CSD* (at present, *ca.* 250000 compounds *versus* over 8 million in *Beilstein* and almost 40 million in *CAS Registry* [18]), there were too many isomers for a given formula. We, therefore, decided to use a standardized part of the literature reference common to both *Beilstein* and *CSD* as a second criterion to reduce the number of 'false' hits. As part of the literature references associated with any data in *Beilstein*, the internationally standardized CODEN for the cited journal is given, and *CSD* uses an internal code number for journals for which a translation table to CODENs is available [19]. Simple in principle, this conversion, again, was not so easy to apply: together with data in *Beilstein*, only the full literature reference can be exported, not the CODEN alone. The exported references are in a format where individual data for the reference (*Beilstein Citation Number* (*CRN*), author names, *etc.*) are separated by semicolons, but, due to variations in the number of authors, the CODEN can be in quite different positions of this reference string (*cf.* examples in *Figs. 7, a* and *8, a*). To parse this string, we had therefore to disconnect it at each semicolon and check the content of this sub-string for the presence of a CODEN. This parsing procedure[25]) worked for all 545 entries, yet

---

[24]) The assignment of *CAS Registry Numbers* to compounds in the *Beilstein* database was based on constitution, not configuration, since the two databases use different ways to represent stereochemistry. Thus, in many cases, several *CAS Registry Numbers* for diastereoisomers were assigned to a single compound record in the *Beilstein* database. Furthermore, the assignment of *CAS Registry Numbers* for new compounds was terminated in *ca.* 1993, because *CAS* considered *Beilstein* a competitor with its own databases on the information market. This factors reduce the usefulness of *CAS Registry Numbers* in the *Beilstein* database. At present (update BS0102), 52% of all the compounds in *CrossFire Beilstein* have *CAS Registry Numbers*, but for our two-component systems, this was only 26%.

[25]) Our parser looked for three capital letters in a sequence to identify CODENs, and for at least five consecutive numbers to extract the CRN (*Beilstein Citation Registry Number*) that we used for checking. Originally, the entire post-processing was done with a primary results file that did not contain the literature reference, as we only later decided to use this when we needed the CODEN as a second criterion besides the molecular formula. References were then exported separately, concatenated with the first results file on the basis of the common *Beilstein Registry Number*, and then parsed as described. Exporting all the necessary information in one step as described here for parsing was developed later.

```
3583392 2C26H22O2*C4H8O2 #C!%5&&2#h!% | #C!%5&&2#h!% C26H22O2 3597060
// XXXX XXXX XXXX  || 5507200 TELEAY

8652919 C18H20O2*2C2H6O #C!%2&&3#i!% | #C!%2&&3#i!% C18H20O2 229912 //
XXXX XXXX XXXX  || 6251706 TETRAB
```

a) examples for special cases regarding CODENs

```
4881395 C6H10O4*2H3N #C!%5&&2#h!% | #C!%5&&2#h!% C6H10O4 1209788 //
XXXX XXXX XXXX  || 5600098 JSTCAM ZSTKAI

5712257 C33H35NO5*C3H2N2 #C!%5&&2#h!% | #C!%6&&2#h!% C33H35NO5 5322562
// #C!%5&&2#h!% C3H2N2 773697  || 5700157 JOCEAH 5761678 JACSAT
```

b) examples for special cases regarding crystal space groups

```
3897071 C36H60O30*CH4O #P!2&1%2&1%2&1% | #P!2&1%2&1%2&1% C36H60O30
79627 // #D!%17&&2#h!% <=#Cmcm!> CH4O 1098229  || 543259 ACCRA9

5674176 C12H24O6*2C2H4O2 #C!%3&&2#h!% | #D!%15&&2#h!% C12H24O6 1619616
// #Pna!2&1% (= #C!%9&&2v%) C2H4O2 506007  || 6095223 ZNBSEN

5880542 C5H9NO2*ClH P2&1%/a (=C%5&&2h%) | XXXX C5H9NO2 80810 // XXXX
XXXX XXXX  || 1071702 ACCRA9
```

c) example for a two-component system with several data entries after automatic processing

```
4172386 C14H10*C10H2N4 #C!%3&&#s!% | #C!%5&&2#h!% C14H10 1905429 //
#C!%5&&2#h!% C10H2N4 1875027 || 5989107 JCPSA6
```

Fig. 8. *Processed data from* CrossFire Beilstein (BS0201, update 1st quarter 2001) *for two-component systems with crystal-space-group data* (same as in *Fig. 7*, but examples with matching space groups shown in boldface)

produced several special cases needing manual inspection, *i.e.*, five entries for which automatic conversion failed, 17 entries with more than one reference, and 31 with one literature reference but two CODENs. The latter were found to be references to bilingual journals like *Angewandte Chemie*, or Russian journals with the bibliographic data including CODENs for both the Russian original and the English translation. Automatic conversion of CODENs to the *CSD* coding as the next step succeeded only with 421 of the 545 candidates, because, for journals that had changed their name and thus also their CODEN, only the current version was in the list provided by the *CCDC* [20] on the Web [19]; the older CODENs had to be translated manually by means of CODEN lists from older printed versions of the *CSD* documentation.

Thus, the original 11229 two-component systems were limited to those 3919 where a crystal space group was reported for at least one component in the *Beilstein* database (for 439 systems, this information was present in *Beilstein* for *both* components), and these were split further into a file containing 609 two-component systems, with their space groups matching that of at least one of their components, and the remaining 3310 with nonmatching space groups (*cf. Table 1*). *Fig. 8* shows the same examples of two-component systems as *Fig. 7* after processing and incorporation of component information, those with matching space groups being shown in boldface: *Beilstein*

Registry Number, molecular formula, space group of two-component system (separated by spaces), then a vertical line, followed by space group, molecular formula, and *Beilstein Registry Number* if a component had (matching or different) space-group information (for components lacking this information, XXXX was entered), followed by two vertical lines, the *Citation Registry Number*, and the CODEN. Only the ASCII file with the matching space groups was then converted to tables in *Microsoft Word for Macintosh* for inspection and further manual processing.

As the rather complex space group symbols cannot be represented properly in the standard ASCII code used internally in the database, *Beilstein* uses a linearized code for this data. Unfortunately, this is insufficiently standardized[26]). While, in most cases, *Schönflies* symbols are used, *Hermann–Mauguin* symbols are also found or even appear together with *Schönflies* symbols. With *Hermann–Mauguin* symbols, the *Beilstein* database does not differentiate between space groups P1 and $P\bar{1}$ and similar cases. By comparing the screen display for space groups and the linear representations, we could manually 'translate' these exported linear representations into the space group numbers provided by the *International Tables for Crystallography* [21].

In order to keep the scripts for automatic processing relatively simple, we had to accept the following limitations (despite the danger of missing relevant candidates): first, for the 137 two-component systems mentioned above that had more than one entry in the export data file (*i.e.*, more than one space group/reference reported), only the space group for the main entry was automatically compared with that of the single components. The compound data shown in *Figs. 7, c* and *8, c*, for example, were classified as 'no match', because the space group in the first full entry for the two-component system (the only one used in our script) does, indeed, not match that of the components (*Fig. 7, c*), while the one given in the second and fourth entries does! Second, the script was expected to fail to match space groups for those 139 entries for two-component systems that showed both *Hermann–Mauguin* and *Schönflies* symbols. All these problematic entries mentioned here would have to be inspected manually, and this would have been even more time-consuming than identifying them as potential problems in the first place.

3.4. *Comparing Search Procedures in the* Chemical Abstracts *and* Beilstein *Databases.* When comparing the search procedures in *Chemical Abstracts* and *Beilstein* databases, the following general differences are worth discussing. First, the fact that, in *Beilstein*, structures, data, and literature references are in a single database, makes structure-based searches easier and more straightforward than in the *CA* databases, where, for historic reasons, structure, literature, and references are in separate databases[10]). The *SciFinder Scholar* interface partly bridges this gap, but is not applicable to a search as complex as the one presented here. The powerful *STN Messenger* retrieval language [9], in particular the 'SMARTselect' command, which

---

[26]) There were 564 entries for single components and 139 entries for two-component systems that had both *Schönflies* and *Hermann–Mauguin* symbols. As can be seen from examples in *Figs. 7, b* and *8, b*, the format of these entries is not exactly standardized: the second symbol is sometimes *Schönflies*, sometimes *Hermann–Mauguin*, given mostly in parentheses (but in some cases enclosed in <>) preceded by an equal sign, sometimes with, sometimes without spaces. Eleven entries contained text beside the space group designation, and seven contained only text instead of a space group (*cf.* example in *Fig. 7, d*). These inconsistencies made reliable automatic processing of such entries difficult.

was absolutely indispensable in our context, enabled us not only to execute the necessary searches across individual *CAS* databases in a more roundabout way, as compared to *CrossFire Beilstein*, but also enabled us to do some postprocessing already during the search, while, in *CrossFire Beilstein*, all the necessary postprocessing had to be done on exported data outside *CrossFire*.

An even more important difference lies in that *Beilstein* uses standardized data-field designations and/or keywords to describe properties of compounds in a unique way, while, in the *CA* literature database [7][17], the major problem was the lack of a unique and standardized way of indexing crystal-structure data. This is true for most properties of compounds, despite the extensive efforts of *CAS* to use standardized *Index Headings* and a set of defined 'roles'[22]) for compound indexing. In this context, it is important to mention that routine physical data and spectra are not indexed at all by *CAS*, which is, unfortunately, not very well known among chemists and stands in contrast to *Beilstein*. This renders the two databases to a large extent complementary, *CAS* having a distinctly larger coverage of journals and, in particular, patents. This was one of the reasons why we concentrated our efforts first on *CA* databases.

Obviously, 'sieving' the *CAS* databases for compounds with certain properties (like belonging to 'organic zeolites') is technically feasible. For such rather common properties as, *e.g.*, crystal structures, however, it appears that the indexing used in these databases is not specific or precise enough to make the results useful. This is in strong contrast to our results from both the *Beilstein* and the *CSD* databases (see *Chapt. 4*). Comparing these databases, specializing on properties directly to *Chemical Abstracts* (marketed as 'key to the world's chemical literature') may be considered unfair to the latter, but duly reminds us that we need more than one 'key' to chemical information. Looking at *CA* databases in particular, we learned that the *SciFinder Scholar* interface is by far the easiest for straightforward, routine questions. But any complex problem needs the full power of the *STN Messenger* retrieval system, which is much more difficult to handle, though.

4. Cambridge Structural Database. – Regarding the inability of the user-friendly *Quest* and *ConQuest* interfaces to execute the necessary searches for isomorphous compounds in the *Cambridge Structural Database* (*CSD*), we eventually had to take recourse to the following procedures: in a preliminary search with the *Quest V5* interface, we created a subset of 99148 organic compound records[27]) by searching with a combination of bit screens 57 (entry as an 'organic compound') and 153 (atom coordinates field present). These compounds were exported in the form of two result files, one of which contained the structure parameters in the FDAT format, the other the molecular formula and the bibliographic data for the corresponding publication (JNL journal file). In a first step, by means of the *CSD* REFCODE as the common denominator for these two result files, the structure data and the molecular formula from the JNL file were converted to one file for processing. The cell dimensions from each compound record in this file were then compared with the data for every other

---

[27]) This number corresponds to 92423 different compounds (*i.e.*, the first six letters of the *CSD* REFCODE are not identical) to be compared, in the context of our investigation, to 59657 organic compounds in the *Beilstein* database (*CrossFire* BS0201) that included space-group data.

record in the file. Whenever the deviation between the cell dimensions of the records compared was less than 3% for the axes, and not more than one degree for the angles, this pair of compounds was written to a results file. Thereby, standard settings were assumed for the space groups, *i.e.*, we would miss candidates where one structure was solved in $P2_1/c$, and the other one in $P2_1/a$, because the effort to also include such examples was considered to be too great regarding the large number of compounds to be processed. The resulting 58148 pairs of compounds with similar cell parameters were further tested for *a*) identical space group symbols, *b*) elimination of duplicates (non-identical composite molecular formula of the pairs, first six characters of REFCODE not identical), and *c*) for formula strings for one component of each pair being a sub-string of the formula string for the other component.

These criteria significantly reduced the number of candidates for isomorphism to 110, which were inspected individually and searched in the *CSD* to eliminate mixed crystals, metal-containing compounds, and other compounds that were obviously not fulfilling our criteria. This left us with only 49 pairs of *CSD* entries, corresponding to 33 compound pairs (*Table 2*). These were checked each in the original literature, leaving 27 compound pairs (corresponding to ten chemically different host compounds) that had isomorphous crystal structures for both the host alone and the clathrate. The remaining entries, shown in italics in *Table 2*, either contain H_2O as the guest[28]) or show ambiguities in the crystal structure of the host[29]). Not all of the 27 compound pairs turned out to be organic zeolites, because, even after consulting the original publications, we were unable to unambiguously establish that the crystal lattice had remained intact after egress of the guest compound.

5. Results and Conclusions. – 5.1. *Results.* Using the procedures described above, we finally retrieved 545 potential organic zeolites from the *Beilstein* database [10] and 110 candidates from the *Cambridge Structural Database* [5]. This striking difference in the number of candidates is, of course, due to the fact that, in *Beilstein*, we could restrict only by requesting identical space groups, while, with the *CSD* data, we could compare unit-cell dimensions to eliminate compound pairs that differed clearly in this criterion. To our surprise, only eight of the 27 isomorphous pairs were also found among the 545 candidates from *Beilstein* (entries in *Table 2* with footnote f). Given this small overlap, we decided to spare the effort to countercheck all 545 *Beilstein* candidates in the *CSD*, but we checked all candidates from *Table 2* in both *Beilstein* and *Chemical Abstracts* to find out why they were missing. There is an obvious tendency in *Beilstein* and, to a lesser extent in *Chemical Abstracts*, to index not the clathrates actually examined, but only the parent compound (entries in *Table 2* with footnote b): seven for *Beilstein*, three for *CA*. *Beilstein* in particular misses clathrates with inorganic molecules like xenon or hydrogen sulfide, which were probably considered outside the domain of 'straightfor-

---

[28]) As outlined in [3], 'the presence of water in organic crystals is so widespread, and the reasons for its inclusion so varied', so that we did not consider such cases organic zeolites.

[29]) For example, the authors of the (*tert*-butyl)tetrahedrane structure found in their 1987 publication that their compound published in 1984 originally contained gas inclusions [22]. Also, comparison of the two crystal structures for *trans*-(3,4-dimethoxyphenyl)[3-(3,4-dimethoxyphenyl)oxiran-2-yl]methanone [23], in our opinion, suggests that, in one case, guest inclusion was overlooked (*i.e.*, the structure reported for the host could actually be already a clathrate).

Table 2. *Candidates for the Search Term 'Organic Zeolites' Identified in the* Cambridge Structural Database. The following entries can be found: *1*) the name of the compound (upper left of each row); *2*) the system-based REFCODE (*e.g.*, 'DAZBIP' in the first row); *3*) the component-based REFCODE (*e.g.*, 'BTCOAC' in the first row); *4*) the *Beilstein* and *CAS Registry Numbers* (*e.g.*, 5662237 *vs.* 99159-92-9 in the first row); *5*) the molecular formula; *6*) the cell dimensions (bottom lines of each row) for both the corresponding *system* and *component* REFCODEs; *7*) the space group(s) (lower right corner of each row).

alpha-1,3,5-Benzene-tricarboxylic acid bromine clathrate	DAZBIP $C_9H_6O_6$ $\times$ 0.16(Br_2)	BTCOAC $C_9H_6O_6$	5662237[a]) 99159-92-9[d])
#BTCOAC 26.520 16.420 26.551 90.000 91.530 90.000 *C*2/*c*			
#DAZBIP 26.510 16.449 26.580 90.000 91.800 90.000 *C*2/*c*			
Salvinorin hydrate	BUJJIZ $C_{23}H_{28}O_8$ $\times$ 0.32(H_2O)	DADMOK $C_{23}H_{28}O_8$	[b]) [b])
#BUJJIZ 6.368 11.338 30.710 90.000 90.000 90.000 *P*212121			
#DADMOK 6.369 11.366 30.747 90.000 90.000 90.000 *P*212121			
4,5-bis(4-Methoxyphenyl)-2-(3-nitrophenyl)-1 H-imidazole ethyl acetate clathrate	CIZJOK $C_{23}H_{19}N_3O_4$ $\times$ 0.33($C_4H_8O_2$)	CIZMUT $C_{23}H_{19}N_3O_4$	7509741[f]) 256653-58-4[d])
#CIZJOK 8.948 27.469 9.111 90.000 90.000 90.000 *Pna*21			
#CIZMUT 8.896 26.771 9.209 90.000 90.000 90.000 *Pna*21			
Tetra-t-butyltetrahedrane argon clathrate	FIGKUB $C_{20}H_{36}$ $\times$ 0.086(Ar)	CUCZUV $C_{20}H_{36}$	[b]) 107271-29-4[i])
#CUCZUV 15.795 15.795 14.056 90.000 90.000 120.000 *P*63/*m*			
#FIGKUB 15.732 15.732 13.923 90.000 90.000 120.000 P63/m			
2,3,4,5-Tetraphenylcyclopent-2-en-1-one monohydrate	DAHXOZ $C_{29}H_{22}O$ $\times$ H_2O	TPCYPO $C_{29}H_{22}O$	[b]) 6177-94-2[i])
#DAHXOZ 24.023 24.023 20.714 90.000 90.000 120.000 *R*-3			
#TPCYPO 23.459 23.459 20.878 90.000 90.000 120.000 *R*-3			
octakis(m-Tolylthio)naphthlene 1,4-dioxane clathrate	DEFCEW $C_4H_8O_2$ $\times$ $C_{66}H_{56}S_8$	DEFCAS $C_{66}H_{56}S_8$	6378653[f]) [b])
#DEFCAS 15.875 15.875 23.654 90.000 90.000 90.000 *P*4/*ncc*			
#DEFCEW 16.040 16.040 23.793 90.000 90.000 90.000 *P*4/*ncc*			
4-p-Hydroxyphenyl-2,2,4-trimethylchroman chloroform	DIANCH $C_{18}H_{20}O_2$ $\times$ 0.167($CHCl_3$)	PEPTIN $C_{18}H_{20}O_2$	Not found 58904-66-8[i])
#DIANCH 27.116 27.116 11.023 90.000 90.000 120.000 *R*-3			
#PEPTIN 26.965 26.965 10.933 90.000 90.000 120.000 *R*-3			

Table 2 (cont.)

4-p-Hydroxyphenyl-2,2,4-trimethylchroman ethanol clathrate	DIANET $C_{18}H_{20}O_2$ $\times$ 0.33(C_2H_6O)	PEPTIN $C_{18}H_{20}O_2$	8652919[f]) 41037-04-1[i])
#DIANET 26.969 26.969 10.990 90.000 90.000 120.000 *R*-3			
#PEPTIN 26.965 26.965 10.933 90.000 90.000 120.000 *R*-3			
2′,4′-Difluoro-4-hydroxybiphenyl-3-carboxylic acid monohydrate clathrate	QOQXAV $C_{13}H_8F_2O_3$ $\times$ H_2O	FAFWIS $C_{13}H_8F_2O_3$	[b]) 340831-23-4[d])
#FAFWIS 34.666 3.743 20.737 90.000 110.570 90.000 *C*2/*c*			
#QOQXAV 34.650 3.730 20.760 90.000 110.470 90.000 *C*2/*c*			
2′,4′-Difluoro-4-hydroxybiphenyl-3-carboxylic acid hexane solvate	YEJWEP $C_{13}H_8F_2O_3$ $\times$ 0.25(C_6H_{14})	FAFWIS $C_{13}H_8F_2O_3$	[b]) 3491941-55-5[d])
#FAFWIS 34.666 3.743 20.737 90.000 110.570 90.000 *C*2/*c*			
#YEJWEP 34.826 3.730 20.703 90.000 110.630 90.000 *C*2/*c*			
4-p-Hydroxyphenyl-2,2,4-trimethylchroman xenon clathrate	GIRBOY $C_{18}H_{20}O_2$ $\times$ 0.71(Xe)	PEPTIN $C_{18}H_{20}O_2$	Not found 134470-76-1[i])
#GIRBOY 27.023 27.023 10.922 90.000 90.000 120.000 *R*-3			
#PEPTIN 26.965 26.965 10.933 90.000 90.000 120.000 *R*-3			
2,8-Dimethyltricyclo(6.2.1.$1^{3,9}$)dodecane-syn,2syn-8-diol bis(2,5,8-trimethyltricyclo(6.2.1.$1^{3,9}$)dodecane-syn,2syn-8-diol) toluene clathrate	GOCWAW $C_{14}H_{24}O_2$ $\times$ 2($C_{15}H_{26}O_2$) $\times$ C_7H_8	NIWGEF $C_{15}H_{26}O_2$	8104530[c]) 212831-72-6[c])[h])
#GOCWAW 13.765 13.765 7.007 90.000 90.000 120.000 *P*3121			
#NIWGEF 13.708 13.708 7.005 90.000 90.000 120.000 *P*3121			
Hexamethylenetetramine carbon tetrabromide	HARWEC $C_6H_{12}N_4$ $\times$ CBr_4	HXMTAM $C_6H_{12}N_4$	6617836[g]) 123301-79-1[d])
#HARWEC 6.956 6.956 6.956 90.000 90.000 90.000 *I*-43*m*			
#HXMTAM 7.021 7.021 7.021 90.000 90.000 90.000 *I*-43*m*			
	HARWEC $C_6H_{12}N_4$ $\times$ CBr_4	HXMTAM01 $C_6H_{12}N_4$	
#HARWEC 6.956 6.956 6.956 90.000 90.000 90.000 *I*-43*m*			
#HXMTAM01 6.931 6.931 6.931 90.000 90.000 90.000 *I*-43*m*			
	HARWEC $C_6H_{12}N_4$ $\times$ CBr_4	HXMTAM02 $C_6H_{12}N_4$	
#HARWEC 6.956 6.956 6.956 90.000 90.000 90.000 *I*-43*m*			
#HXMTAM02 6.910 6.910 6.910 90.000 90.000 90.000 *I*-43*m*			
	HARWEC $C_6H_{12}N_4$ $\times$ CBr_4	HXMTAM03 $C_6H_{12}N_4$	
#HARWEC 6.956 6.956 6.956 90.000 90.000 90.000 *I*-43*m*			
#HXMTAM03 7.021 7.021 7.021 90.000 90.000 90.000 *I*-43*m*			

Table 2 (cont.)

	HARWEC	HXMTAM04
	$C_6H_{12}N_4$	$C_6H_{12}N_4$
	× CBr_4	
#HARWEC 6.956 6.956 6.956 90.000 90.000 90.000 *I*-43*m*		
#HXMTAM04 6.931 6.931 6.931 90.000 90.000 90.000 *I*-43*m*		
	HARWEC	HXMTAM05
	$C_6H_{12}N_4$	$C_6H_{12}N_4$
	× CBr_4	
#HARWEC 6.956 6.956 6.956 90.000 90.000 90.000 *I*-43*m*		
#HXMTAM05 6.910 6.910 6.910 90.000 90.000 90.000 *I*-43*m*		
	HARWEC	HXMTAM07
	$C_6H_{12}N_4$	$C_6H_{12}N_4$
	× CBr_4	
#HARWEC 6.956 6.956 6.956 90.000 90.000 90.000 *I*-43*m*		
#HXMTAM07 7.028 7.028 7.028 90.000 90.000 90.000 *I*-43*m*		
	HARWEC	HXMTAM08
	$C_6H_{12}N_4$	$C_6H_{12}N_4$
	× CBr_4	
#HARWEC 6.956 6.956 6.956 90.000 90.000 90.000 *I*-43*m*		
#HXMTAM08 7.028 7.028 7.028 90.000 90.000 90.000 *I*-43*m*		
	HARWEC	HXMTAM09
	$C_6H_{12}N_4$	$C_6H_{12}N_4$
	× CBr_4	
#HARWEC 6.956 6.956 6.956 90.000 90.000 90.000 *I*-43*m*		
#HXMTAM09 6.956 6.956 6.956 90.000 90.000 90.000 *I*-43*m*		
	HARWEC	HXMTAM10
	$C_6H_{12}N_4$	$C_6H_{12}N_4$
	× CBr_4	
#HARWEC 6.956 6.956 6.956 90.000 90.000 90.000 *I*-43*m*		
#HXMTAM10 6.927 6.927 6.927 90.000 90.000 90.000 *I*-43*m*		
	HARWEC	HXMTAM11
	$C_6H_{12}N_4$	$C_6H_{12}N_4$
	× CBr_4	
#HARWEC 6.956 6.956 6.956 90.000 90.000 90.000 *I*-43*m*		
#HXMTAM11 6.934 6.934 6.934 90.000 90.000 90.000 *I*-43*m*		
	HARWEC	HXMTAM12
	$C_6H_{12}N_4$	$C_6H_{12}N_4$
	× CBr_4	
#HARWEC 6.956 6.956 6.956 90.000 90.000 90.000 *I*-43*m*		
#HXMTAM12 6.942 6.942 6.942 90.000 90.000 90.000 *I*-43*m*		
	HARWEC	HXMTAM13
	$C_6H_{12}N_4$	$C_6H_{12}N_4$
	× CBr_4	
#HARWEC 6.956 6.956 6.956 90.000 90.000 90.000 *I*-43*m*		
#HXMTAM13 6.955 6.955 6.955 90.000 90.000 90.000 *I*-43*m*		
	HARWEC	HXMTAM14
	$C_6H_{12}N_4$	$C_6H_{12}N_4$
	× CBr_4	
#HARWEC 6.956 6.956 6.956 90.000 90.000 90.000 *I*-43*m*		
#HXMTAM14 6.969 6.969 6.969 90.000 90.000 90.000 *I*-43*m*		
	HARWEC	HXMTAM15
	$C_6H_{12}N_4$	$C_6H_{12}N_4$
	× CBr_4	
#HARWEC 6.956 6.956 6.956 90.000 90.000 90.000 *I*-43*m*		
#HXMTAM15 6.984 6.984 6.984 90.000 90.000 90.000 *I*-43*m*		

Table 2 (cont.)

bis(2,4,6-tris(4-Bromophenoxy)-1,3,5-triazine) hexamethylbenzene clathrate	VEWDOQ $2(C_{21}H_{12}Br_3N_3O_3)$ $\times$ $C_{12}H_{18}$	HEXWIQ $C_{21}H_{12}Br_3N_3O_3$	8668700[g]) 302597-33-7[h])
#HEXWIQ 15.602 15.602 7.050 90.000 90.000 120.000 *P*63/*m*			
#VEWDOQ 15.554 15.554 6.951 90.000 90.000 120.000 *P*63/*m*			
bis(2,4,6-tris(4-Bromophenoxy)-1,3,5-triazine) trinitromesitylene clathrate	VEWLOY $2(C_{21}H_{12}Br_3N_3O_3)$ $\times$ $C_9H9N_3O_6$	HEXWIQ $C_{21}H_{12}Br_3N_3O_3$	8669829[g]) 302597-35-9[h])
#HEXWIQ 15.602 15.602 7.050 90.000 90.000 120.000 *P*63/*m*			
#VEWLOY 15.719 15.719 7.034 90.000 90.000 120.000 *P*63/*m*			
(2,4,6-tris(4-Bromophenoxy)-1,3,5-triazine) 2,4,6-collidine clathrate	WITGEL $C_{21}H_{12}Br_3N_3O_3$ $\times$ $C_{18}H_{24}$	HEXWIQ $C_{21}H_{12}Br_3N_3O_3$	Not found[e]) 265990-27-0[d])
#HEXWIQ 15.602 15.602 7.050 90.000 90.000 120.000 *P*63/*m*			
#WITGEL 15.468 15.468 7.087 90.000 90.000 120.000 *P*63/*m*			
(2,4,6-tris(4-Bromophenoxy)-1,3,5-triazine) 1-methylnaphthalene clathrate	WITGIP $C_{21}H_{12}Br_3N_3O_3$ $\times$ $C_{11}H_{10}$	HEXWIQ $C_{21}H_{12}Br_3N_3O_3$	Not found[e]) 265990-28-1[d])
#HEXWIQ 15.602 15.602 7.050 90.000 90.000 120.000 *P*63/*m*			
#WITGIP 15.569 15.569 7.064 90.000 90.000 120.000 *P*63/*m*			
bis(2,4,6-tris(4-Bromophenoxy)-1,3,5-triazine) mesitylene clathrate	WITGOV $C_{21}H_{12}Br_3N_3O_3$ $\times$ C_9H_{12}	HEXWIQ $C_{21}H_{12}Br_3N_3O_3$	Not found[e]) 265990-29-2[d])
#HEXWIQ 15.602 15.602 7.050 90.000 90.000 120.000 *P*63/*m*			
#WITGOV 15.573 15.573 7.042 90.000 90.000 120.000 *P*63/*m*			
tris(beta-Hydroquinone) xenon clathrate	JAMKEN $3(C_6H_6O_2)$ $\times$ 0.866(Xe)	HYQUIN05 $C_6H_6O_2$	Not found 18932-78-0[d])
#HYQUIN05 16.613 16.613 5.475 90.000 90.000 120.000 *R*-3			
#JAMKEN 16.610 16.610 5.524 90.000 90.000 120.000 *R*-3			
Hydroquinone-hydrogen sulfide clathrate	ZZZVLG01 $C_6H_6O_2$ $\times$ 0.256(H_2S)	HYQUIN05 $C_6H_6O_2$	Not found 60662-39-7[i])
#HYQUIN05 16.613 16.613 5.475 90.000 90.000 120.000 *R*-3			
#ZZZVLG01 16.670 16.670 5.518 90.000 90.000 120.000 *R*-3			
Hydroquinone hydrogen sulfide clathrate	ZZZVLG11 $3(C_6H_6O_2)$ $\times$ 0.87(H_2S)	HYQUIN05 $C_6H_6O_2$	Not found 14342-92-8[d])
#HYQUIN05 16.613 16.613 5.475 90.000 90.000 120.000 *R*-3			
#ZZZVLG11 16.616 16.616 5.489 90.000 90.000 120.000 *R*-3			

Table 2 (cont.)

(+)-Chelidonine monohydrate	VIGFEW $C_{20}H_{19}NO_5$ $\times$ H_2O	JISGIB $C_{20}H_{19}NO_5$	[b]) [j])
#JISGIB 8.964 9.115 10.622 90.000 93.320 90.000 *P*21			
#VIGFEW 8.971 9.120 10.640 90.000 93.430 90.000 *P*21			
trans-1,3-bis(3,4-dimethoxyphenyl)-2,3-epoxy-1-propanone chloroform clathrate	QINFIC $C_{19}H_{20}O_6$ $\times$ 0.33($CHCl_3$)	LIGXUU01 $C_{19}H_{20}O_6$	[b]) [b])
#LIGXUU01 36.048 36.048 8.313 90.000 90.000 120.000 *R*-3			
#QINFIC 36.035 36.035 8.274 90.000 90.000 120.000 *R*-3			
nonakis(2,5,8-Trimethyltricyclo(5.3.1.$1^{3,9}$)dodecane-syn-2,syn-8-diol) bis(diisopropyl ketone) clathrate	YAQQIQ 3($C_{15}H_{26}O_2$) $\times$ 0.67(C7 H14O)	NIWGEF $C_{15}H_{26}O_2$	8663432[f]) 302576-27-8[h])
#NIWGEF 13.708 13.708 7.005 90.000 90.000 120.000 *P*3121			
#YAQQIQ 13.808 13.808 6.999 90.000 90.000 120.000 *P*3121			
tetrakis(2,5,8-Trimethyltricyclo(5.3.1.$1^{3,9}$)dodecane-syn-2,syn-8-diol) benzene clathrate	YAQQOW 3($C_{15}H_{26}O_2$) $\times$ 0.75(C_6H_6)	NIWGEF $C_{15}H_{26}O_2$	8661281[f]) 302576-30-3[h])
#NIWGEF 13.708 13.708 7.005 90.000 90.000 120.000 *P*3121			
#YAQQOW 13.773 13.773 6.998 90.000 90.000 120.000 *P*3121			
hexakis(2,5,8-Trimethyltricyclo(5.3.1.$1^{3,9}$)dodecane-syn-2,syn-8-diol) toluene clathrate	YAQQUC 3($C_{15}H_{26}O_2$) $\times$ 0.5(C_7H_8)	NIWGEF $C_{15}H_{26}O_2$	8662798[f]) 302576-33-6[h])
#NIWGEF 13.708 13.708 7.005 90.000 90.000 120.000 *P*3121			
#YAQQUC 13.729 13.729 7.008 90.000 90.000 120.000 P3121			
nonakis(2,5,8-Trimethyltricyclo(5.3.1.$1^{3,9}$)dodecane-syn-2,syn-8-diol) o-xylene clathrate	YAQRAJ 3($C_{15}H_{26}O_2$) $\times$ 0.33(C_8H_{10})	NIWGEF $C_{15}H_{26}O_2$	8663427[f]) 302576-37-0[h])
#NIWGEF 13.708 13.708 7.005 90.000 90.000 120.000 *P*3121			
#YAQRAJ 13.753 13.753 7.010 90.000 90.000 120.000 *P*3121			
4-(4-Hydroxyphenyl)-2,2,4-trimethylchromane p-xylene clathrate	OBEQUH $C_{18}H_{20}O_2$ $\times$ 0.167(C_8H_{10})	PEPTIN $C_{18}H_{20}O_2$	Not found 132569-46-1[d])
#OBEQUH 27.139 27.139 10.824 90.000 90.000 120.000 *R*-3			
#PEPTIN 26.965 26.965 10.933 90.000 90.000 120.000 *R*-3			
hexakis(4-p-Hydroxyphenyl-2,2,4-trimethylchroman) carbon tetrachloride clathrate	SIHJEY 6($C_{18}H_{20}O_2$) $\times$ CCl_4	PEPTIN $C_{18}H_{20}O_2$	4221079[f]) 154642-96-3[i])
#PEPTIN 26.965 26.965 10.933 90.000 90.000 120.000 *R*-3			
#SIHJEY 27.134 27.134 10.933 90.000 90.000 120.000 *R*-3			

Table 2 (cont.)

	SIHJEY01 $6(C_{18}H_{20}O_2)$ $\times$ CCl_4	PEPTIN $C_{18}H_{20}O_2$	
#PEPTIN 26.965 26.965 10.933 90.000 90.000 120.000 *R*-3			
#SIHJEY01 27.147 27.147 10.939 90.000 90.000 120.000 *R*-3			
	SIHJEY02 $6(C_{18}H_{20}O_2)$ $\times$ CCl_4	PEPTIN $C_{18}H_{20}O_2$	
#PEPTIN 26.965 26.965 10.933 90.000 90.000 120.000 *R*-3			
#SIHJEY02 26.912 26.912 10.901 90.000 90.000 120.000 *R*-3			
bis(2,4,6-tris(4-Chlorophenoxy)-1,3,5-triazine) hexachlorobenzene clathrate	VEWDIK $2(C_{21}H_{12}Cl_3N_3O_3)$ $\times$ C_6Cl_6	VALQEE01 $C_{21}H_{12}Cl_3N_3O_3$	8668967[g]) 302597-31-5[h])
#VALQEE01 15.364 15.364 6.855 90.000 90.000 120.000 *P*63/*m*			
#VEWDIK 15.435 15.435 6.876 90.000 90.000 120.000 *P*63/*m*			
2,4,6-tris(4-Chlorophenoxy)-1,3,5-triazine hexamethylbenzene clathrate	VEWFUY $2(C_{21}H_{12}Cl_3N_3O_3)$ $\times$ $C_{12}H_{18}$	VALQEE01 $C_{21}H_{12}Cl_3N_3O_3$	8668166[g]) 302597-32-6[h])
#VALQEE01 15.364 15.364 6.855 90.000 90.000 120.000 *P*63/*m*			
#VEWFUY 15.411 15.411 6.867 90.000 90.000 120.000 *P*63/*m*			
2,4,6-tris(4-Chlorophenoxy)-1,3,5-triazine 1,3,5-trinitrobenzene clathrate	VEWJIQ $2(C_{21}H_{12}Cl_3N_3O_3)$ $\times$ $C_6H_3N_3O_6$	VALQEE01 $C_{21}H_{12}Cl_3N_3O_3$	8668167[g]) 302597-34-8[h])
#VALQEE01 15.364 15.364 6.855 90.000 90.000 120.000 *P*63/*m*			
#VEWJIQ 15.255 15.255 7.005 90.000 90.000 120.000 *P*63/*m*			
bis(2,4,6-tris(4-Chlorophenoxy)-1,3,5-triazine) hexamethylphosphoramide clathrate	VEWNEQ $2(C_{21}H_{12}Cl_3N_3O_3)$ $\times$ $C_6H_{18}N_3OP$	VALQEE01 $C_{21}H_{12}Cl_3N_3O_3$	8666755[g]) 302597-36-0[h])
#VALQEE01 15.364 15.364 6.855 90.000 90.000 120.000 *P*63/*m*			
#VEWNEQ 15.234 15.234 6.880 90.000 90.000 120.000 *P*63/*m*			

[a]) Compound in *Beilstein* list with nonmatching space groups (*cf. Table 1*). [b]) Crystal-structure data indexed for host compound, but not for the corresponding clathrate. [c]) Not found in *Beilstein* or *CA* candidates because only *two*-component systems were included (in contrast to searches in *CA* and *Beilstein*, we had not limited the processing of the *CSD* data to two-component systems). [d]) Present in *CAS* 'candidate list' with 8838 compounds (L37, *Fig. 5*). [e]) Not found in *Beilstein* database because the journal the compound was published in is not covered by *Beilstein*. [f]) Compound in *Beilstein* 'candidate list' with matching space groups (*cf. Table 1*). [g]) Not in *Beilstein* list: no crystal-space-group data present for host compound in *Beilstein*. [h]) Not indexed by *CAS* as 'crystal/mol structure', only with 'crystallog', and, therefore, not in *CAS* 'candidate list' with 8838 compounds (L37, *Fig. 5*). [i]) Not in *CAS* 'candidate list' with 8838 compounds (L37, *Fig. 5*). [j]) *CAS Registry Number* 6004-04-2 assigned for a hydrate, but no literature references in *CA* database.

ward' organic compounds[30]). Other candidates in *Table 2* proved to be irretrievable with our search strategies in both *Chemical Abstracts* and *Beilstein* because of inconsistent indexing in the former, and lack of space-group data in the latter. It is disquieting that several clathrates of *Dianin*'s compound were, thus, missed (two each in *Beilstein* and *CA*), and that there is no indexing at all by CAS for a publication [23a] that reports nmr and X-ray studies for a certain compound[29]). It is also revealing to look at the tris(halophenoxy)triazine clathrates in *Chemical Abstracts:* two of them in one publication and four in another were indexed with 'crystallog… nanoporous host structure', and three in a different paper with 'crystal structure', although all these publications dealt with the crystal-structure analyses of these compounds. Similarly, for X-ray-structure analyses of certain hydroquinone clathrates, the one containing Xe was indexed by *CAS* with 'crystal structure', while the H_2S clathrate was indexed like this in one publication, but, in another in which the first crystal structure was refined, it was only indexed with 'structure'! Clearly, a consistently assigned role[22]) for 'crystal-structure analysis' is badly needed here.

5.2. *General Conclusions.* Our searches and attempts mercilessly exposed deficiencies and weaknesses in currently available public databases, regarding database design (*e.g.*, lack of *CAS Registry Numbers* in the *Cambridge Structural Database*), database content (*e.g.*, lack of a flag for respective crystal-structure databases in the *STN Registry* data-field listing), and data quality, *e.g.*, insufficient standardization in the *Beilstein* database for crystal-space-group data, and inconsistent indexing for crystal structures in *Chemical Abstracts.* Some of these deficiencies make linking results from structure searches in *CAS Registry* to the *CSD* impossible. The inconsistencies in data formats found in *Beilstein* do not usually cause problems with small search results that are inspected manually by a chemist who immediately discerns any discrepancies and can interpret them correctly in a given context. In our example, however, we depended on machine postprocessing for the very large primary-result sets. Even small inconsistencies are a problem here, demanding either complex scripts, or a large amount of manual inspection/processing.

Although the problem discussed in this publication is very specialized, and the size of the information sets to be manipulated in the approaches can be considered somewhat extreme, we think that some general conclusions from our experiences are valid. First, the importance of special databases like the *Cambridge Structural Database* and large factual database like *Beilstein* or *Gmelin* was shown again, and its superiority for a data- or property-driven search over large literature databases like *Chemical Abstracts* proven. The inherent utility of *Beilstein* was in our example enhanced by the lack of precise, reproducible indexing for X-ray-structure analyses of compounds in *Chemical Abstracts*, and, particularly, by the lack of *CAS Registry Numbers* in the *Cambridge Structural Database.* These deficiencies made our first approach – searching for candidates in the largest structure database (*STN Registry* [15]) and transferring the candidates to the database most suited for problems concerning crystal structures – impossible for all practical purposes: these databases lack the precise 'common

---

[30]) A search in *CrossFire Gmelin* for such clathrates was also negative; this seems to be a case of compounds 'falling between two chairs'.

denominator', which is indispensable for such a procedure, particularly for the large numbers at hand.

The first part of the desired procedure in the *STN* version of the *CAS* databases worked surprisingly well. *STN* is to be commended for their powerful software, in particular the 'SMARTselect' command, their data structure, and the powerful hardware in the background, which made the processing of surprisingly large search results feasible within system limits.

SciFinder Scholar proved easy to use due to its natural-language interface and the kind of ranking feature that is inherent in the way it permutes and combines the search terms identified in the query phrase entered by the user. It failed in our example, though, to make a significant contribution due to the complex nature of the problem.

5.3. *Necessary Improvements for Databases.* Enhancements to the databases used here are obviously necessary and of general importance far beyond the specific problem discussed here. First and foremost, the *Cambridge Structural Database* must return to assigning *CAS Registry Numbers* for all compounds in the databases, and this assignment must be done in a reliable way. We suggest a close cooperation between *CCDC* [20] and *CAS*, which would be beneficial for both partners. During assignment of the *CAS Registry Numbers*, the corresponding compounds in the *CAS Registry* database would be flagged out with '*CSD*' in the *STN Files* field (*cf. Fig. 3* from *SciFinder Scholar*; this data field is called 'locator' (LC) in the *STN Registry* version). Beyond this important improvement, we strongly suggest that *CAS* 'connects' compounds in all major crystal-structure databases, *i.e.*, the *Inorganic Crystal Structural Database* [24] and the *Protein Database* [25], to their *CAS Registry*. The flagging of compounds in *CAS Registry* is to be complemented in the *CA* literature database by consistently assigning a role[22]) termed 'crystal-structure analysis' to the indexing of compounds.

Although CODENs are an established standard in the vast majority of databases for the titles of journals, the *CSD* uses its own numeric codes for this purpose [19]. As a consequence, we had to convert the CODENs retrieved from the *Beilstein* databases to the codes used in the *CSD*. Since this had to be done manually in a significant number of cases, a task both tedious and unsatisfactory, we strongly recommend that the *Cambridge Crystallographic Data Centre* [20] replaces its proprietary codes with the official CODENs. Our example in trying to relate information from other databases to the *Cambridge Structural Database* may be considered somewhat extreme regarding the numbers involved, but the fact remains that, even a database as established and specialized as the *CSD* must be willing to permit integration by adopting accepted standards for identifying both compounds and references.

'Compound warehouses' [26] are now designed that link data and properties for compounds from a whole array of different databases for ease of access, and they are in high demand as important tools for problem solution not only in the pharmaceutical industry [27]. With respect to these developments and the general trend towards systems for chemists ('end users'), isolated solutions are no longer justifiable, not even for databases as highly specialized as those containing crystal structures. Although the *CCDC* has improved the user interface significantly, our example illustrates a lack of integration with the 'wide world' of compound databases – certainly a hindrance to broader utilization of this important tool. The 'interoperability' among the specialized

database *CSD*, the 'general purpose' structure and literature databases from *CAS*, as well as the 'general purpose' factual database *Beilstein* need to be improved significantly. Although *Beilstein* and *Chemical Abstracts* were, of course, never intended to replace special databases like the *Cambridge Crystallographic Database*, the consistency and precision of crystal-structure information in these databases is insufficient indeed. It is rather frustrating to note that searches that are technically feasible with today's powerful interfaces fail rather miserably because of insufficient data quality.

While the search facilities and system limits were shown in this example to be useful and satisfactory – we had anticipated many more problems in this respect when we conceived our search procedures – postprocessing of large-answer sets turned out to be again very tedious because of the almost complete lack of suitable postprocessing features in these databases and due to the lack of standardization of crystal-structure and bibliographic data in the case of *Beilstein* (where we had to develop our own post-processing with UNIX-shell scripts). None of the databases used here provides really useful tools for postprocessing of search results, with *STN Messenger* being the most powerful system thanks to the 'SMARTselect' tool, but even this must be considered insufficient. When writing one's own procedures out of necessity, a lot of deficiencies, inconsistencies, and lack of data quality lurking behind the good-looking user interfaces, are mercilessly exposed.

5.4. *Closing Remarks.* Contrary to common practice, we decided to publish the procedures used in information retrieval for this project [4] in some detail, because, in our opinion, it shows both the chances and facilities in modern information retrieval, as well as problems and pitfalls associated with present databases. We are convinced that the procedures we used and, particularly, the stumbling blocks we came across are of some general interest beyond the specific context, because the property 'crystal-structure parameters' we were looking for could be substituted by almost any other physical property to use the same or similar search strategies and procedures. We do also consider this examination as a contribution to the current discussions about database integration and data quality [28].

Several approaches we had attempted failed because of insufficient 'common denominators' and standards as indispensable means for integration of databases. We can only hope that this unsatisfactory situation will be changed in the near future by increased standardization and cooperation among database producers. Likewise, improvements in data quality are certainly less spectacular than new search features and nice-looking graphical-user-interfaces, but they are urgently needed.

The use of databases in the *CAS Academic Program* at *STN International* and of academic licenses for the *Cambridge Structural Database*, *MDL CrossFire Beilstein*, and *CAS SciFinder Scholar* are gratefully acknowledged.

Technical Details

Cambridge Structural Database [5]: April 2002 release, 257162 records. *SciFinder Scholar* [8]: Version 2001. *STN* [16]: Searches were executed with the front-end software *STN Express with Discover! 6.0c for Windows.* For all uses of 'SMARTselect', the mode had been set to 'lists' with 'SET TERM L#'. *Beilstein* [10]: *CrossFire Beilstein* Update BS0201, *Beilstein Commander 2000 for Windows* (Version 5.0 Build 12 SP1 Release 2) on a *Windows NT 4.0* personal computer.

REFERENCES

[1] S. Lee, D. Venkataraman, *Stud. Surf. Sci. Catal.* **1996**, *102*, 75.
[2] A. Nangia, *Curr. Opin. Solid State Mater. Sci.* **2001**, *5*, 115.
[3] A. Nangia, G. R. Desiraju, *Chem. Commun.* **1999**, 605.
[4] D. A. Plattner, A. K. Beck, M. Neuburger, *Helv. Chim. Acta* **2002**, *85*, 4000.
[5] F. H. Allen, V. J. Hoy, 'Cambridge Structural Database', in 'Encyclopedia of Computational Chemistry', Ed. P. v. R. Schleyer, Wiley, Chichester, 1998, p. 155–167; L. N. Kuleshova, M. Y. Antipin, *Russ. Chem. Rev.* **1999**, *68*, 1; *cf.* also http://www.ccdc.cam.ac.uk/prods/csd/csd.html.
[6] *Chemical Abstracts Service* (CAS): http://www.cas.org/.
[7] *CAplus* database: http://www.cas.org/SCIFINDER/SCHOLAR/caplus.html; see also http://www.stn-international.de/stndatabases/databases/caplus.html.
[8] *SciFinder Scholar:* http://www.cas.org/SCIFINDER/SCHOLAR/index.html.; D. Ridley, 'Information Retrieval: SciFinder and SciFinder Scholar', Wiley, Chichester, 2002; cf. also http://www.infochem-bio.ethz.ch/SFS.html.
[9] *STN Messenger:* http://www.stn-international.de/trainingcenter/workshopmaterial/wsmaterial.html.
[10] *Beilstein* database: http://www.beilstein.com/products/xfire; http://www.infochembio.ethz.ch/Xfire.html. 'The Beilstein System. Strategies for Effective Searching', Ed. S. R. Heller, American Chemical Society, Washington, 1998.
[11] *Science Citation Index:* http://www.isinet.com/isi/products/citation/scie/index.html.
[12] *Institute for Scientific Information* (ISI): http://www.isinet.com/isi/.
[13] *Web of Science* (ISI): http://www.isinet.com/isi/products/citation/wos/index.html.
[14] A. P. Dianin, *J. Russ. Phys. Chem. Soc.* **1914**, *46*, 1310. (*Chem. Abstr.* **1915**, *9*, 1903).
[15] *STN Registry* (publicly available version of the CAS Registry (http://www.cas.org/EO/regsys.html) compound database): http://www.stn-international.de/stndatabases/databases/registry.html.
[16] *STN International:* http://www.stn-international.de/.
[17] *STN CA:* http://www.stn-international.de/stndatabases/databases/ca.html.
[18] Current data for *CAS Registry* content in http://www.cas.org/cgi-bin/regreport.pl.
[19] The CSD System Documentation, Vol. 3, Appendix 4: CCDC Journal CODENs, http://www.ccdc.cam.a-c.uk/support/csddoc/volume3/z304.html.
[20] *Cambridge Crystallographic Data Centre* (CCDC): http://www.ccdc.cam.ac.uk/.
[21] 'International Tables for Crystallography', Ed. T. Hahn, Kluwer Academic Publishers, Dordrecht, 1995.
[22] H. Irngartinger, R. Jahn, G. Maier, R. Emrich, *Angew. Chem.* **1987**, *99*, 356; H. Irngartinger, A. Goldmann, R. Jahn, M. Nixdorf, H. Rodewald. K. D. Malsch, R. Emrich, G. Maier, *Angew.Chem.* **1984**, *96*, 967.
[23] a) M. Bardet, M. F. Foray, S. Li, K. Lundquist, R. Stomberg, *J. Chem. Crystallogr.* **1999**, *29*, 1023; b) R. Stomberg, S. Li, K. Lundquist, *J. Chem. Crystallogr.* **1994**, *24*, 407.
[24] G. Bergerhoff, 'Inorganic Three-dimensional Structure Databases', in 'Encyclopedia of Computational Chemistry', Ed. P. v. R. Schleyer, Wiley, Chichester, 1998, p. 1325–1337; http://www.cas.org/ONLINE/DBSS/icsdss.html.
[25] J. L. Sussman, F. C. Bernstein, J. Jiang, M. Libeson, D. Lin, N. O. Manning, J. McCarthy, R. Shea, E. E. Abola, C. E. Felder, J. Prilusky, 'Protein Data Bank (PDB): A Database of 3D Structural Information of Biological Macromolecules', in 'Encyclopedia of Computational Chemistry', Ed. P. v. R. Schleyer, Wiley, Chichester, 1998, p. 2160–2168.
[26] MDL 'Compound Warehouse': http://www.mdli.com/presentations/acs-fall2001/linkingacsfall2001.pdf.
[27] P. Selzer, B. Rohde, P. Ertl, *Nachr. Chem. Tech. Lab.* **2000**, *48*, 1471.
[28] G. Wiggins, *J. Am. Soc. Inf. Sci* **1995**, *46*, 614; http://listserv.indiana.edu/archives/chminf-l.html.

Received June 6, 2002

Appendix. *List of Selected Potential Inclusion Compounds Listed by Current* Chemical Abstracts *Index Name:*

Acetic acid
Acetic acid ethyl ester
Acetic acid, trifluoro-
Acetonitrile
Benzenamine
Benzenamine, 4-nitro-
Benzene
Benzene, (methylsulfinyl)-
Benzene, [(R)-methylsulfinyl]-
Benzene, [(S)-methylsulfinyl]-
Benzene, 1,2-dimethyl-
Benzene, 1,3,5-trinitro-
Benzene, 1,3-dimethyl-
Benzene, 1,4-dimethyl-
Benzene, 2-bromo-1,3,5-trinitro-
Benzene, 2-chloro-1,3,5-trinitro-
Benzene, chloro-
Benzene, methyl-
1,4-Benzenediol
Benzenemethanol
Benzenemethanol, α-methyl-
Benzenesulfinic acid, methyl ester
1,3,5-Benzenetricarboxylic acid
4 H-1-Benzopyran-4-thione
Butane, 1-(methylsulfinyl)-
Butane, 1-[(R)-methylsulfinyl]-
Butane, 1-[(S)-methylsulfinyl]-
1-Butanamine
1-Butanamine, N,N-dibutyl-
1-Butanamine, N-butyl-
1-Butanol
2-Butanol
Carbonic acid, diethyl ester
Carbonic acid, dimethyl ester
2,5-Cyclohexadiene-1,4-dione
Cyclohexanamine
Cyclohexane
Cyclohexanone
Cyclopentanamine
1,4-Dioxane
Ethanamine, N,N-diethyl-
Ethanamine, N-ethyl-
Ethane, 1,1′-oxybis-
1,2-Ethanediol
Ethanol
Ethanol, 2,2,2-trichloro-
Ethanol, 2,2,2-trifluoro-
Ethanone, 1-phenyl-
Formamide, N,N-dimethyl-
Formic acid
Furan, tetrahydro-
Heptane
Hexane
Iodine
Methane, dichloro-
Methane, nitro-
Methane, sulfinylbis-
Methane, tetrachloro-
Methane, trichloro-
Methanol
Morpholine
Naphthalene
Nonane
Pentane
Phenol, 2,4,6-trinitro-
Piperidine
1-Octanamine
1-Propanamine
1-Propanamine, 2-methyl-
1-Propanamine, 2-methyl-N-(2-methylpropyl)-
1-Propanamine, N,N-dipropyl-
1-Propanamine, N-propyl-
2-Propanamine
2-Propanamine, N-(1-methylethyl)-
2-Propanamine, N-methyl-
1,2-Propanediol
Propane, 1-(methylsulfinyl)-
Propane, 1-(methylsulfinyl)-, (R)-
Propane, 1-(methylsulfinyl)-, (S)-
1-Propanol
1-Propanol, 2-methyl-
2-Propanol
2-Propanol, 2-methyl-
2-Propanone
4 H-Pyran-4-thione
Pyrazine
Pyridine
Pyridine, 2-methyl-
Pyridine, 3-methyl-
Pyridine, 4-methyl-
2(1 H)-Pyrimidinone, 4-amino-
2,4,6(1 H,3 H,5 H)-Pyrimidinetrione, 5,5-diethyl-