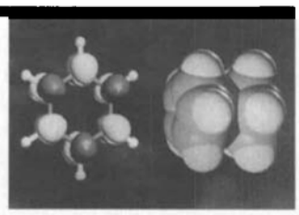
CHIMIA 48 (1994) Nr. 4 (April)

COMPUTATIONAL CHEMISTRY COLUMN

Column Editors:
Prof. Dr. J. Weber, University of Geneva
Prof. Dr. H. Huber, University of Basel
Dr. H. P. Weber, Sandoz AG, Basel



Chimia 48 (1994) 109–111

ISSN 0009–4293

Chemical Information from Public Databases: Recent Changes and Current Trends

Engelbert Zass*

Recent Changes

Not too many years ago, chemical information retrieval from public databases was a relatively straightforward exercise, as there was hardly any choice other than the *Chemical Abstracts* (CA) structure and literature files when looking for compounds and their preparations/reactions/data. Both the number of chemical databases and the search facilities have been increased significantly in what could be termed horizontal growth of sources within the last five years or so: *Beilstein* online and *Chemical Abstracts Service* (CAS) reaction database CASREACT since 1988, patent databases for *Markush* structures (*Markush DARC* 1989, STN MARPAT 1990), *Gmelin* online and the spectroscopy information system *SpecInfo* in 1991, the *Materials Property Data Network* (MPD) available at *STN International* (*Scientific & Technical Information Network*) since April 1991, and in 1992, the reaction database *ChemInform RX* produced by *FIZ Chemie/Bayer*, to name only a few important ones, make up an impressive array of chemical information sources.

The concomitant increase in complexity for users is at least partially offset by improved user interfaces, *e.g.*, the front-end software *STN Express* for structure input, running locally at the user's personal computer (*Macintosh* or *Windows* PC), menu-driven search systems like those offered by DIALOG, long awaited links between in-house and public databases.

Access to the already bewildering variety of chemical databases is also enhanced by 'meta databases' like STN's *Numeriguide* (a master file on the physical properties and their units in STN databases), or DIALOG's *Finder* files to locate journal title, product, and company name information across DIALOG's databases. The disadvantage of these useful meta files, of course, is their limitation to the databases of the respective vendor only – the all-embracing chemistry master index is not yet in sight.

These changes in the chemical information scene necessitate changes in search strategies. Old, long-engrained searching habits die only slowly, but may lead to search results that are suboptimal with respect both to cost and comprehensiveness. The buzzword is multi-file searching, in parallel or sequential fashion, with use of the 'synergies' across databases. The retrieval tools to achieve this are mostly with us already: field/data-type standardization across databases, *e.g.*, STN uses the same data field name HVAP for 'enthalpy of vaporization' in databases as diverse in content and origin as *Beilstein*, *Gmelin*, DIPPR (*Design Institute for Physical Property Data*), TRCTHERMO (*Thermodynamic Research Center*), HSDB (*Health & Safety Database*); the same structure can be used for searching in *CAS Registry*, *Beilstein*, and *Gmelin* (but not in *SpecInfo* as this system uses conventions and a search process that differs from the aforementioned databases; however, at least the commands for entering a structure are the same); search terms can be extracted from records (citations, compounds) retrieved previously, and simply re-used in searching within the same database or across to others.

As an example and an important type of problem, let us take the search for the existence of a compound (structure). Traditionally (*i.e.*, since 1981), compounds were searched for in the *CAS Registry File*, available either at STN (formerly *CAS Online*), or at *DARC/Télésystèmes Questel*. A structure search for a compound costs presently 61 DM at STN, and covers all 12.8 million compounds from the literature since 1957. Crossing over the search result (in form of the *CAS Registry Number(s)* of the compound(s) retrieved) into the *STN CA File* gives either the complete literature for the compound, or, if desired, only that about preparation or a certain topic. For the literature prior to 1967, one may turn to the *CAOLD File* which provides only the *CAS Abstract Numbers* for references from 1957 to 1966 which than have to be converted to literature references using printed CA in the library. It is more convenient to cross over the search results into the *Beilstein* database which, fortunately enough, has *CAS Registry Numbers* assigned to all compounds that are also in the *Registry File* (4.4 out of 5.7 million), and covers literature since 1779. So, it seems suffi-

**Correspondence*: Dr. E. Zass
Chemie-Bibliothek
ETH-Zürich
Universitätstrasse 16
CH-8092 Zürich

cient to search the *Chemical Abstracts* databases for compounds (since 1957) and literature (since 1967), and to turn to *Beilstein* only if one needs pre-1967 literature, or to check whether a compound not found in the *Registry File* is in *Beilstein* – in theory, this should only be the case when it was published before 1957 and not ever again since. Practice, however, shows things to be different: while it comes as a surprise to nobody that CA covers more journals and particularly patents than *Beilstein*, and that some organic compound classes are not covered by *Beilstein* at all (polymers, peptides and nucleotides, organometallics – the latter, however, are to be found in *Gmelin*), we noticed all too often that CAS missed compounds and/or references that should have been covered, but were indeed not. This observation led to the recommendation of searching both CAS and *Beilstein* databases when a comprehensive result is desired. This, of course, means spending a lot more money to get, quite often, but not reliably enough, the same compounds and references. A small compensation for the extra money spent lies in the fact that *Beilstein* usually gives more details on preparation, like starting materials and reaction conditions, while CA mostly just states that a compound was prepared.

Some of the money that thus must be spent can be saved in this context by a judicious use of Ca and *Beilstein*: with a time coverage 1779–1993, *Beilstein* contains 5.7 million organic compounds and therefore, a significant portion of all organic compounds known (comparing just numbers with the *Registry File* that seems to be twice as large is not meaningful, not only because of different coverage, but more so because of different registration policies which do not permit simple 'compound counting' in both files). A structure search for a compound presently costs only 18 DM in *Beilstein*, however. A more cost-effective and appropriate strategy for organic compounds is, therefore, to start in *Beilstein* for compounds and literature, and supplement the literature by crossing over the *CAS Registry Numbers* in the CA literature file. The *CAS Registry File* is then only needed for additional isomers or mixtures not covered by *Beilstein*. As one does not start from scratch in this situation, it is often possible to use the 'dictionary search' facilities in the *Registry File* for molecular formulas, name (fragments), and ring system descriptors. Dictionary searching is certainly more complicated and thus potentially more risky than structure searching, particularly for occasional users, or those not well versed in CAS naming policies; but, if used appropriately, costs are often more than halfed compared to the standard approach *via* structure searches in both *Registry* and *Beilstein* or even *Registry* alone.

For data and spectra, *Beilstein* online is a 'must' anyway, as CAS – admittedly, but obviously not too well known among users – does not index 'routine' spectra and data – for a total of 526 hexopyranoses registered in both CA and *Beilstein*, there was information about preparation for 46% of the compounds in CA, and for 47% in *Beilstein*; the respective figures for data were, melting point 0%/26.5%, optical rotation 3%/30%, NMR 32%/59.5%.

Current Trends

For further developments, one can recognize a general trend of vertical growth of sources in addition to the horizontal growth mentioned above. The multi-media, multi-system availability has been made possible by recent developments in hardware and software, and (hopefully from the point of view of the producers) economically feasible by a growing end-user market: databases that were only accessible publicly in very large computer centers become available in-house *via* client-server systems (when computing power is a primary factor) or on CD-ROM. *Beilstein* is a good example on both accounts. The *Current Facts* CD-ROM holds a year's worth of organic compounds from the primary literature (*ca.* 300000 compounds plus data and references out of *ca.* 80 journals) searchable by (sub)structure and/or data on a PC; it is updated quarterly with a lag of *ca.* nine month behind the primary literature. The *Windows* version of *Current Facts* which just appeared uses modern hypertext-like features to link, *e.g.*, starting materials in a description of the preparation of the product to their structures and database entries; this feature enables one to 'roll back' or navigate through entire reaction sequences just by mouse clicks. The new *XFIRE* software developed by *Beilstein* allows to search their entire structure file of more than five million compounds on an *IBM Risc System/6000* as server and *Windows* PCs as clients in-house; an extended version containing all the data and references from *Beilstein* online is under development.

In this context, it is interesting to come back to our discussion about structure searching. The cost argument given above for a '*Beilstein* first' strategy is of course not relevant for those 'happy few' in Basle that have the *CAS Registry Structure File* searchable in-house at fixed cost; while the availability of this large file is, at least at present conditions, limited to a few large companies, the more than five million structures in *XFIRE* look like being affordable for smaller companies and even universities. One can only speculate about the consequences that such a development might have.

There is a somewhat similar situation with reaction databases. In-house reaction database systems like REACCS, SYNLIB, and ORAC are limited to a relatively small group of large and medium-sized companies, and an unfortunately small number of universities that took advantage of the academic programs for these systems. The major reason for this was of course cost, and hardware demands. With PC-based database software like MDL's *ISIS/Base* or *Chemical Design*'s *ChemRXS*, this situation may change drastically in the near future, provided, of course, that the database producers adjust their prices to a potential mass market.

Another development that bears relevance on this topic is the family of reaction databases produced by *InfoChem*: starting from a structure database containing reaction information compiled by VINITI (All-Union Institute for Scientific and Technical Information, Moscow) and ZIC (Central Information Processing Unit for Chemistry, Berlin, former GDR) for the period 1975–1988, they produced a reaction 'parent file' with 1.8 million reactions. Using a proprietary algorithm, a subset (reaction type) database *ChemReact* with 370000 reactions was produced by grouping together reactions with the same reaction centers and immediate environment across the entire database (not only within the same publication), and selecting only one example from such a group based on the successive criteria 'spectral information available for product/publication in leading journal/yield/most recent publication'. *ChemReact* is available both as in-house database for MDL's REACCS, and publicly at STN. Further subsets produced along similar lines are *ChemSynth* (80000 reactions for REACCS), and *ChemSelect* (10000 reactions for REACCS, or ChemBase or ISIS/Base on a PC). The individual reaction types are linked *via* an accession number to all examples in the 'parent file' that is available as a (display-only) file for REACCS, or as *CD-React* CD-ROM (a single disc with 1.8 million reactions!) to accompany the PC version of *ChemSelect*. The interesting aspect beyond this particular product is certainly the algorithmic (*vs.* expensive, not strictly reproducible intellectual) production of

subsets, and their multi-system availability.

Variable packaging of chemical information is also at play in the recently released *CAS Surveyor* CD-ROM containing thematic subsets of the large CA database. A CD-ROM version of the entire printed CAS *12th Collective Index* has been available for some time; searching for compounds is only possible there by name, not by structure. A companion CD-ROM contains the abstracts and the literature references from this time period. The current awareness publication *Current Contents* is produced by the *Institute for Scientific Information* on paper, on disk or CD-ROM for both *Macintosh* and MS-DOS PCs, and as public database on DIALOG. All this, of course, implies that librarians and information managers have to decide on what medium they will offer this information – print, online in-house (CD-ROM stand-alone or in a network, client/server databases), online on public hosts like STN, DIALOG *etc.* If they can afford several or even all media, users must be trained not only in the selection of the media (after prior selection of a source like *Chemical Abstracts*, *Beilstein etc.*) but also in the appropriate search procedures which can be quite different. This is no mean task, particularly, if not only quality of the search result, but also cost-effectiveness plays an important role (as it should). Information retrieval nowadays can be described on three levels borrowed from the military field: it has a strategic level (selection of a source), an operational (selection of the medium), and a tactical level (construction of the appropriate search profile).

Despite these fascinating developments, there remain several wishes yet unfulfilled: paramount among these are data quality and user friendliness. Even in highly reputed sources like *Chemical Abstracts*, coverage and quality still leave something to be desired. Author searching, *e.g.*, is quite often a simple and useful entry point for a chemical topic. The usefulness of this approach, however, is diminished by the fact that only a maximum of ten authors are registered by CAS, and only one address. CA is not the only database to be that restrictive, but contrasts unfavorably in this respect with the *Science Citation Index* that includes all authors and addresses. The myth prevailing, particularly in universities, that searching *Chemical Abstracts* online is an easy way to get complete publications lists of any author since 1967 must be done away with for this (and other) reasons.

Problems concerning coverage of compounds and literature in *Chemical Abstracts* and *Beilstein* were already discussed here (as were their unfortunate consequences for the cost of comprehensive searches). While some of the differences in coverage can be accounted for by different selection and indexing policies, some are obvious violations of their own set of rules, or simply mistakes. Clearly, database quality must be improved further.

Despite significant progress, user-friendliness in public databases is also still insufficient, particularly so for the occasional searcher. In substructure searching, internal conventions for aromaticity/tautomerism and a restrictive formal interpretation of ring/chain – a seemingly acyclic structure fragment cannot be in a ring unless explicitly declared as such – have to be considered by the user instead of taken care of by the computer as with in-house database systems.

Large reaction databases do not cover the time before 1975, so that for comprehensive results, one has to turn to CA and *Beilstein* which as compound-oriented sources are much less than ideal for that purpose. *Beilstein* does contain a wealth of reaction information, and technically, it is easier and more precisely searchable there than in CA; unfortunately, the present price policy of STN which can only be considered an affront to users, makes such searches prohibitively expensive in all but the most simple cases – the author's personal record came up to 3300 DM just for a search of *Baeyer-Villiger* oxidation of norbornanones, giving four reactions in *Beilstein*. In the DIALOG implementation of *Beilstein* online, this is no problem, as a similar search there cost only *ca.* $ 300.

Searching for NMR data, *e.g.*, means using *SpecInfo*, with only *ca.* 100 000 compounds and, therefore, a relatively small chance to find exactly what one is looking for (it is fair to say here that this system has other features and strenghts than the number of spectra stored), and consequently, both *Beilstein* and CA must be accessed, searching data fields NMRS, NMRA, CTNMR, CTUNCH (NMR) in *Beilstein*, and using both the acronym 'NMR' and the phrase 'nuclear magnetic resonance' (or parts thereof) in CA, as this is not standardized with a simple label 'NMR' as it easily could be in both databases. What are computers around for if not particularly to eliminate such stumbling stones?

This short account could not help to be biased by personal experience, and many remarkable developments, like the concentration process among database host, electronic primary journals (*RedSage* project), or the growing importance of *Internet*, had to be left out. Finally, with all these nice 'multis', multi-file, multi-system, multi-media, we must not forget that these are mere vehicles to carry the sole important aspect – information which the user needs, and needs as easy, as fast, and as economic as possible, and he really could not care less where he gets it from – paper, CD-ROM or terminal.