From Handbooks to Databases on the Net: New Solutions and Old Problems in Information Retrieval for Chemists

Engelbert Zass

Chemie-Bibliothek, ETH Zürich, CH-8092 Zürich, Switzerland

Received December 21, 1995[⊗]

Sources for chemical information are becoming ever more powerful and varied: besides the still important printed sources, there are public databases, large in-house systems, and databases on PCs/CD-ROMs. The price to pay for this information cornucopia, however, is increased complexity for users. Improved front-ends and the change from terminal-mainframe to client-server systems ease the burden of searching, but such means are not yet sufficient to make chemical information retrieval a reliable routine operation of every chemist. We need even more improved database quality, better goal-oriented marketing and training by producers and hosts, and problem-oriented education for chemical information retrieval as an obligatory part of chemistry syllabi.

INTRODUCTION

The 1995 Herman Skolnik Award of the Chemical Information Division of the American Chemical Society[1] honored the achievements of Clemens Jochum and Reiner Luckenbach in making the traditional Beilstein Handbook of Organic Chemistry[2] available in electronic form, first as a public database[3,4] and recently as the in-house system CrossFire.[5,6] The beginnings of this large project can be dated precisely to March 16th, 1982 when the *Stiftungsrat* (governing board) of the Beilstein Institute held an extraordinary meeting with *Beilstein and electronic data processing* as the sole item on the agenda. At that time, all the primary literature for the printed handbook was still being excerpted manually on paper slips, the famous *Zettel*, and the information was also further processed by conventional means. Only at the final stage of typesetting and index production, computers were used. As a consequence of this board meeting, the *Beilstein EDV Planungsgruppe* (electronic data processing planning committee) was founded which from April 1982 to October 1984 examined the feasibility of improving information processing in Beilstein by the use of computers throughout the process and thereby not only provide Beilstein in electronic form but also, more important, reduce the backlog which at that time was more than 20 years.[7] Then, the new *Computing Division* at the Beilstein Institute, created in 1983 under the direction of Clemens Jochum, took over. At the same time, fortunately for this large project, public funding by the *Bundesministerium für Forschung und Technologie* in the context of its current *Fachinformationsprogramm*,[8] which had then one of its main funding areas in creation of factual databases, became available.[9] From October 1983 until December 1992, establishment of the Beilstein database and the infrastructure necessary for its production were funded with a total of 56 Mill.DM; the first online implementation of Beilstein at STN (i.e., necessary changes in the system and in the STN command language *Messenger*) were funded in 1987−1989 with additional 4.4 Mill.DM.[10]

As the largest factual database in organic chemistry, Beilstein in an indispensable source not only for organic chemists, but also for all scientists needing data about organic compounds—at the ETH Zürich, for example, this encompasses chemical engineers (for thermodynamic data), pharmacists, food scientists, and even researchers in the field of quantum electronics.[11] Databases like Beilstein, CA, or others that address such information needs of end-users must be easily searchable. The present implementations of these databases on public hosts, however, are in our experience not user-friendly enough for reliable, cost-effective retrieval by occasional searchers; only recently client-server search systems like Beilstein CrossFire[5,6] and Chemical Abstracts Service SciFinder[12] are basically capable of fulfilling this need. User-friendliness is very often only discussed in the context of user interfaces to databases, i.e., handling information retrieval. This is not sufficient, as systems with good user interfaces like CrossFire and SciFinder also insinuate to end-users a corresponding quality of the information, both with respect to completeness within the defined coverage of the database and absence of errors. Therefore, *user-friendliness* and *database quality* are closely related, practically inseparable aspects for judging chemical information sources as a whole.[13]

Unfortunately, the quality of major chemical databases is presently found to be still wanting. Information that users can reasonably expect to be present according to the stated purpose and coverage of a database often cannot be retrieved even by specialists using established query formulation techniques. For the purpose of the following discussion, we define the cause for any such "nonretrieval" as an *error*.[14] Among these errors, *errors of omission* are present when information was *unintentionally* either missed altogether or misrepresented (misspelled or misplaced within the database).[14] *Errors of commission* occur when information is presented inconsistently or left out intentionally according to a selection policy that, however, does not meet the user's expectation. Traditionally, these *errors of commission* have not been considered as such but blamed on the user and his lack of knowledge about databases. If we intend to proceed to routine end-user searching in chemistry, we need not only good user-interfaces for handling searches but also databases much lower in both types of errors defined here than exist at present. The burden and blame for training and informing

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

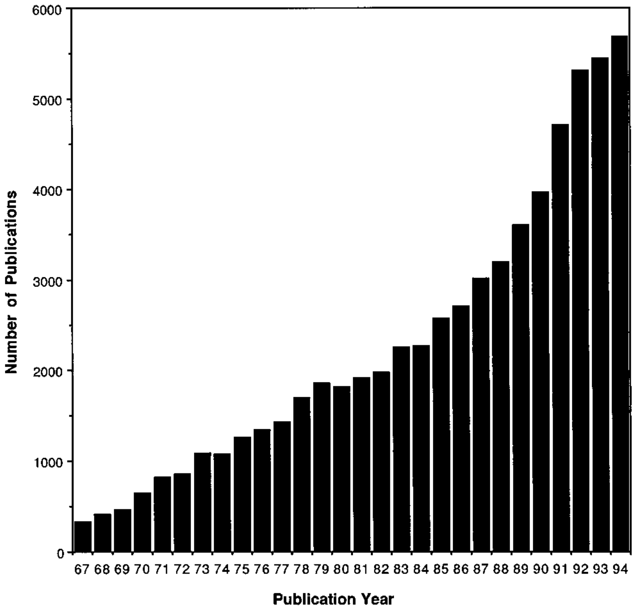


Figure 1. Publications with more than 10 authors ("et al.") in Chemical Abstracts 1967−1994.[20]

users about database content cannot be left to themselves and those educating them in chemical information retrieval. Database producers must meet or change user's expectations, taking appropriate measures[15] so that user's views (*mental representations*) of database content and the real content match as closely as possible. *Good (end) user interfaces like CrossFire or SciFinder make improvements in the quality of databases queried by such interfaces even more mandatory.*

CHEMICAL ABSTRACTS: ERRORS OF OMISSION AND COMMISSION

On the title page of every issue, Chemical Abstracts claims to be the "key to the world's chemical literature", and it is perceived, but also judged, by most chemists along these lines. While a policy of coverage for compounds and subjects is outlined in Appendix II, §14 and Appendix III within the CAS Index Guides, users are confronted with errors of omission that cannot be rationalized by those published policies: e.g., of the four reviews by D. Seebach et al. in *Modern Synthetic Methods* from 1976−1986, only the second one (1980) was abstracted although CASSI indicates full coverage of this source; *Organic Syntheses* is expected to get complete coverage, but there are no records with publication years 1967 or 1968 in the CA File (cf. 1966, 38; 1969, 32 records); from a total of 153 journal publications by A. Eschenmoser 1967−1994, 138 were abstracted by CA,[16] and only 11 of the 15 missing could be accounted for by editorial policy.

Often, omissions are explainable: according to the policy stated in the introduction to the CAS Author Indexes, a maximum of 10 authors are entered into CA; if there are more than 10, only the first nine and "et al." are quoted. Figure 1 illustrates that by adhering to these limits, obviously a leftover from printed indexes which cannot grow in size indefinitely, CAS loses more and more author information every year. While Figure 1 shows absolute numbers, relative figures (with respect to the increase of the annual number of abstracts in CA) are almost equally disquieting: in 1970, only 2.3% of all records had "et al." and therefore authors missing, while in 1980 and 1994, the percentages were 4.2 and 8.7, respectively. This increase can be easily explained by an increase in research cooperation, often interdisciplinary.

As author searching is a very common practice, particularly with end-users, and particularly when a subject is difficult to describe by keywords, this loss of author names in the single most comprehensive source for chemical information is a major problem for the chemical community. Arguments heard in discussions about problems with data structures and field length are not acceptable, as other large databases like the Science Citation Index cover all authors and their addresses from the primary literature.[17]

Modern electronic chemistry information sources must not be forced into a *Procrustes Bed* of rules and limitations once necessary for printed sources. The author is not wantonly suggesting to drop the identity by content of databases and corresponding printed secondary publications, but if we continue losing valuable information from the primary literature like author names because of that, this must be critically re-evaluated. While errors in assignment of keywords are corrected by CAS in both printed CA and the database if brought to their attention, assignments of a CA section, even if admittedly wrong, were not corrected because that would be easy in the database, but not feasible in printed CA.[18]

Another example illustrates that author names and addresses are not only a problem with CA: a joint publication from two organic chemists at the ETH Zürich and two microbiologists in Marburg (Germany)[19] had all four authors indexed by CA, BIOSIS (Biological Abstracts), and the Science Citation Index. Only CA gave full first names, while these were reduced to (less discriminative) initials in the other two databases. While the German Umlaut "ü" in one name was transliterated correctly in CA and BIOSIS, the Science Citation Index converted this to a simple "u". But only this source gave both addresses, while CA, in line with its stated policy, only gave the first one (ETH), and BIOSIS quoted only the second one (Marburg). It is necessary to remember in this context that such address information can be very valuable—but only if it is complete—to differentiate between different authors with common names, particularly so if first names are abbreviated to initials, either by secondary or primary publications.

Other *errors of omission* show up in the Publication Year field in the STN CA database where on Dec 12th, 1995, 84 publications from 1996, 3 from 1997, and 1 from 1998 were present. These are, of course, very small numbers compared to the 360 890 abstracts with publication year 1995, but such errors can be (and must be) easily corrected by checking programs at the time of data entry. In an earliest test (1991), Chemical Abstracts had shown the same kind of publication year errors, while Beilstein online was found containing more than 50 publications that seemed to antedate the Battle of Hastings 1066.[21] Later on, they were found to be corrected.

Among the *errors of commission*, problems of keyword spelling[22] and the use of abbreviations and acronyms often lead to significant losses in recall, particularly with less experienced searchers. CA is considered to be a database with U.S. spelling throughout, but Table 1 shows the reality. The examples in this table are chosen as representative for the most important types of different U.S./British spellings, and the number of references retrieved exclusively by British spelling is unacceptable (Table 1, column 5). Contrary to

Table 1. British and U.S. Spelling of Keywords in Chemical Abstracts (STN CA File, 8/9/95)

keywords (U.S./British spelling)	1 total number of records	2 in TI (Title) field	3 in Titles of non-English publications	4 in IT (indexing term) field	5 exclusively retrieved by British spelling
CENTER	124 342				
CEN**TRE**	460	204	38	13	365
SULFUR	166 912				
SUL**PH**UR	303	167	36	99	90
COLOR	143 113				
COL**OU**R	90	38	13	0	48
PROGRAM	94 262				
PROGRAM**ME**	132	100	15	1	70
HOMOLOG	9 333				
HOMOLOG**UE**	256	52	6	13	189
ANESTHETIC	14 450				
AN**AE**STHETIC	110	62	10	4	42
LEUKOCYTE	34 732				
LEU**C**OCYTE	271	158	34	21	80
PYROLYZED	6 043				
PYROLY**S**ED	16	7	1	2	12
LABELED	199 618				
LABE**LL**ED	1 859	853	148	88	928
ENZYMIC	144 041				
ENZY**MAT**IC	2 819	1873	356	149	1458

our expectation before the facts, this presence of British spelling cannot be attributed to original titles in (British) English, as often less than half of the cases are found in the title (column 2), and of those, a significant portion has titles translated from foreign languages into English by CAS (column 3). British spelling also appears in index terms assigned by CAS (Table 1, column 4). Only in the case of *sulphur*, this could be justified to some extent at least by common names like *Sulphur Springs or Sulphur blue 7*. If one takes a look at the distribution of records *retrieved exclusively by British spellings* over time for all the examples from Table 1 combined (OR logic, then *SmartSelect* on the publication year field), these were particularly common in the period 1969−1977 (peak year 1975, with 398 records exclusively retrieved by British spellings), but even in recent years, there were 50−80 such records per year. These figures are admittedly small, both in absolute and relative sense, but still, as a consequence, both U.S. and British spellings have to be used throughout searches, and, in our experience, chemists too often forget that, even if they once were trained accordingly. Not all such cases (see examples in Table 1) can be taken care of technically by using appropriate truncation/masking to accommodate both spellings in one search term; then, in addition, the search term charge levied in the STN CA and other files punishes users for a lack of consistency at CAS.

The similar problem of acronyms and corresponding full phrases is significantly more severe: e.g., in a search for *nuclear magnetic resonance* versus *nmr* in August 1995 in the STN CA File, the number of records was 96 015/101 437 in the IT (Index Term field), 13/102 439 in the ST field (Supplementary Terms, i.e., phrases from the CA Issue Keyword Indexes), and 295/178 519 in the Abstracts. Of those records in the IT field, 35% were retrieved with the full phrase only, and 39% were retrieved with the acronym only. The full phrase was rarely used in the ST field and in the Abstracts, but then it was often the only form found: 92% of the ST entries, 79% of the abstracts that had the full phrase did not have the acronym. In a recent search across all data fields (12/11/95), 131 912 records were retrieved exclusively with *nmr* (58% of the records retrieved with *nmr*, 55.5% of the total records retrieved with both search terms), and 9427 exclusively with *nuclear magnetic resonance* (9%, 4%). Students being trained in chemical information at ETH Zürich find this situation unacceptable and keep asking why spellings and abbreviations are not standardized by computer[23] or at least cross-referenced automatically in the database. In addition to such *errors of commission*, there also exists too high a number of simple misspellings of keywords in the CA database.[22]

Table 2. Examples for Indexing of Preparation and Data for "Hexoses" ($C_6H_{12}O_6$) in Chemical Abstracts and Beilstein 1967−1994[24]

	Chemical Abstracts	Beilstein
total no. of compd records	584 (100%)	551 (100%)
preparation	52.5%	35%
melting point	<1%	15.5%
optical rotation	3%	18%
NMR	32.5%	47%

Most users of CA are not aware of the fact that routine spectroscopic data are not indexed, while Beilstein and Gmelin index all such data from the primary literature. This important fact is illustrated by the example in Table 2. Even with a method as simple as that used for generation of the data in Table 2, the results not only convincingly show the advantages of Beilstein for property and data searches but also show that *both* databases need to be searched for comprehensive results. In addition, Beilstein gives numerical data directly for melting points and optical rotations without recourse to the primary literature.

Users need such practical information about differences among databases, not only abstract policy statements. We obviously have to ask not only for more standardization of spellings and better compliance of database producers with their own selection and indexing rules but also for more information in the form of "boiler plates" for databases that make content (coverage and indexing and particularly lack of it) almost obtrusive for less-experienced users. For example, scope notes in the CAS Index Guides tell us that indexing of classes of compounds by appropriate (controlled) index headings is rather restricted and not nearly as general as we found our users assuming to be the case. Our experience showed that users, although advised in strong terms to consult the CAS Index Guides before searching CA online, cannot be relied upon at all to do so. Therefore, such kind of scope information must be presented in an appropriate way[25] in the database itself or via a really user-friendly search interface like SciFinder.[12]

BEILSTEIN: ERRORS OF OMISSION AND COMMISSION

CA is certainly not the only database to be found wanting in the aspects discussed above. Informed rumor has it that Beilstein stopped covering patents somewhere in 1980,[26] but to the author's knowledge, there is no solid information on this in any of the documentation provided by Beilstein—the information given on pp 2−10 of the *STN 1993 Beilstein File Database Description* is not very helpful in this respect: "The Beilstein file cannot be considered as a patent file, nevertheless patents may be searched". In fact, the *overall* percentage for compounds records with patent citations in them is only 23% for the entire period 1800−1981.[27] But Beilstein is indeed an important source for compounds in patents, as no less than 98.5% of all records from 1800−1949 had patents; this figure then drops to 94% (1950−1959), 20.5% (1960−1969), and 21% (1970−1979).[27] After 1981 (2014 records with patents), only a few stray examples can be found. How are chemists looking for organic compounds from patents to know that, and what alternatives do they have to Beilstein for the years before large, compound-coded patent databases like IFI Claims or Derwent's World Patent Index appeared? An equally uninformed situation exists regarding data on toxicology or biological function that were regularly excerpted from 1980 onwards,[26] but for 1789 (TOX) and 8888 (BF) compounds (5% and 3.5% of all compounds with such data, respectively), there exists such information even before that date.[28] Again, the database documentation is silent on time coverage.

The new pricing scheme introduced by Beilstein for the public database inJuly 1995 has been hotly debated and criticized a lot.[29] Predictability of costs and price structure contribute not only to the user-friendliness but also to the quality of a database in the wider sense discussed here. This new pricing policy is an example of a principle being basically sound but causing problems because of a rather radical implementation that seemingly did not take some major consequences into prior consideration. The principle of paying only for the information retrieved, not for the time spent at the computer, or for the number of search terms, makes perfect sense, and a degressive pricing scheme for the information output looks convincing, at least in a price list. But at which rates shall information brokers or information departments charge their customers under such a pricing scheme—the actual rates, thereby punishing first customers of the month, or the maximum rate, cheating the later customers? This problem is aggravated by the fact that the STN cost display alway shows the maximum rate, not the actual one, and that printing individual data fields is often more expensive than using predefined formats, punishing less experienced users. The fact that both connect time and (sub)-structure searches are free encourages liberal searching without undue regard for cost, certainly desirable in a public database, as this cost-independence is one of the attractive features of in-house systems. It is to be feared, however, that it also encourages "playing around" excessively in an undesirable sense,[30] using up resources at the host and hampering other users.

One of the strengths of large hosts like STN or DIALOG compared to isolated in-house systems is the powerful software for cross-file searching and the synergistic effects between databases created thereby. But this also implies that individual database producers introducing radical changes can be confronted with undesirable side effects: the *cheapest* (See Note Added in Proof) way to literature about a group of organic compounds at STN goes via a (now free) substructure search in Beilstein, *SmartSelect* of CAS Registry Numbers, search them in the Registry File (look at a few samples in a free format), followed by a literature search and print-out of the results in the CA File—although the Beilstein File provided most of the computing power used, with the exception of the *SmartSelect* charge, all the revenue goes to CAS. The seemingly simplest way out, disabling such a crossover (See Note Added in Proof), or eliminating CAS Registry Numbers in Beilstein databases, is definitely not an acceptable solution. This would amount to hindering transfer of information and crippling existing search technology for "political" reasons, something that would not and must not be tolerated by the chemical community. The reason for this search tactic, of course, is *both* the free substructure search *and* the lack of any free format in Beilstein to check the compounds retrieved or the amount of information available for them—under the old price structure for Beilstein, random structure display was free, and display of the FA (Field Availability) table was very cheap. Therefore, an appropriate solution could be to levy a (small) charge on substructure searching and provide free (or cheap) random display formats for structure and FA to evaluate search results. Likewise, if Beilstein intends to make the information relatively cheaper for high-volume users, a bonus or reimbursement system at the end of the month or year would be easier both for customers and the host accounting than the degressive pricing used at present.

The problem with the variations in indexing when searching for NMR data in CA occurs in a different form also in Beilstein: here, such information is assigned to the different data fields NMRA (NMR absorption), NMRS (NMR spectra), and CTNMR (controlled terms NMR, for special methods like NOE, use of shift reagent, etc.). In addition, literature references about NMR from the time period 1960−1979 that are not yet published in the printed handbook appear in the CTUNCH (controlled term unchecked) field. The 2 078 155 compound records with NMR information in STN Beilstein in August 1995 had it distributed as follows: NMRA 75.5% (70% exclusively in this field), NMRS 2.8%, CTNMR 2.4%, and CTUNCH (NMR) 26% (22.6% exclusively in this field). To most users, this differentiation that is used for all major spectroscopic methods does not make sense, and, worse, they may not even realize the necessity to include all these fields in a search profile for comprehensive retrieval of spectroscopic data as no appropriate examples are given in the database documentation.[31]

The somewhat mysterious CTUNCH field can be explained by the fact that the present information in Beilstein originated from three different sources: the printed handbook covering the literature to 1959 for all compounds and to 1979 for heterocycles ("full file" in Beilstein parlance),[4] the as yet unpublished literature ("short file" in Beilstein parlance), consisting of quite different data material, excerpts from handwritten *Zettel* for the remaining compounds 1960−1979 keyed into the database, and the material excerpted already in electronic form from the primary literature after 1980 which was produced and checked according to the standards and data structure of the database.[26] The resulting differences in quality and completeness of the data and their consequences for searching are not clear to many users, and Beilstein has in our judgment not taken sufficient measures to convey this important information to them.

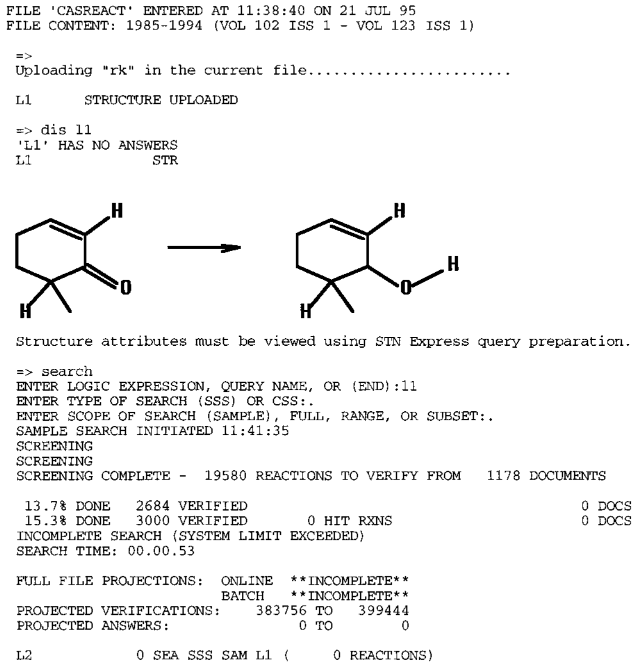
These differences can be illustrated by using NMR data again as an example:[32] entries in the data fields NMRA and NMRS were accompanied by the (searchable) information about the nucleus examined for 98% of the literature before 1960, and practically for all references after 1979,[26] but only for 49.5%[33] of the references 1960−1979; the percentage increased to 76% for the period 1960−1979 if only heterocycles were considered (then to a significant extent already published in printed form). With Beilstein information now available via three routes—printed handbook, public databases at STN and DIALOG, and in-house system CrossFire—any differences in content and data structure of these sources should be publicized to a larger extent than now.

Although all major hosts have done a lot of work in standardizing *field names* across their databases to enable crossfile searching with the software they developed, this is not yet complete (e.g., UV/vis spectra in STN Beilstein in fields EAM and EAS, in STN Gmelin in field UVS). Much less standardized are *data formats* in corresponding fields of databases (e.g., author names or multiword descriptors), and there is almost no standardization across hosts, even for those offering the same databases. This is also true regarding units of numerical data: for density, one finds in STN databases lb/in^3, mg/m^3, kg/m^3, and g/cm^3, but fortunately, the STN unit conversion takes care of that problem.

For chemical structures, standards like SMD or MolFile exist and are used[37] but probably not yet universally enough: the database front-end/structure query drawing software STN Express imports and exports both SMD and MolFile, exports drawings as HPGL, PICT (Macintosh), or PostScript, and generates the SMILES linear notation but not the ROSDAL[38] string used by Beilstein. Structure drawing programs like MDL's ISIS/Draw or CambridgeSoft's ChemDraw Plus also import/export besides several graphics formats the standard structure formats SMD and MolFile, and ISIS/Draw can also export both SMILES[39] and ROSDAL[38] linear notations.

END-USER SEARCHING FOR CHEMICAL INFORMATION

Deficiencies in database quality discussed above in a wider sense of the term clearly contribute to the problems of chemists retrieving chemical information by electronic means directly themselves without intermediacy or intervention of an information specialist. The major problem caused by these and other deficiencies is a lack of reliability in such end-user searchers; particularly so when rather complete retrieval is desired, or the (non)existence of a certain compound, reaction (type), or data has to be ascertained. A second problem is the cost incurred by wide-spread end-user searching. Within the current price structure of public databases, this is difficult to estimate and to budget. In addition, errors in searching (e.g., a substructure search with a wrong bond definition) or displaying (e.g., display all hits in full format of the wrong results set) can at present be costly in the extreme, without appropriate prior warning.



```
FILE 'CASREACT' ENTERED AT 11:38:40 ON 21 JUL 95
FILE CONTENT: 1985-1994 (VOL 102 ISS 1 - VOL 123 ISS 1)

 =>
 Uploading "rk" in the current file........................

 L1       STRUCTURE UPLOADED

 => dis l1
 'L1' HAS NO ANSWERS
 L1               STR

 Structure attributes must be viewed using STN Express query preparation.

 => search
 ENTER LOGIC EXPRESSION, QUERY NAME, OR (END):l1
 ENTER TYPE OF SEARCH (SSS) OR CSS:.
 ENTER SCOPE OF SEARCH (SAMPLE), FULL, RANGE, OR SUBSET:.
 SAMPLE SEARCH INITIATED 11:41:35
 SCREENING
 SCREENING
 SCREENING COMPLETE -  19580 REACTIONS TO VERIFY FROM   1178 DOCUMENTS

  13.7% DONE   2684 VERIFIED                                      0 DOCS
  15.3% DONE   3000 VERIFIED       0 HIT RXNS                     0 DOCS
 INCOMPLETE SEARCH (SYSTEM LIMIT EXCEEDED)
 SEARCH TIME: 00.00.53

 FULL FILE PROJECTIONS:  ONLINE  **INCOMPLETE**
                         BATCH   **INCOMPLETE**
 PROJECTED VERIFICATIONS:    383756 TO   399444
 PROJECTED ANSWERS:               0 TO        0

 L2           0 SEA SSS SAM L1 (     0 REACTIONS)
```

Figure 2. Sample search of a reaction query created in *STN Express for Macintosh 3.12* and uploaded to STN CASREACT.

Nevertheless, the desire and trend is clearly toward end-user searching by chemists, heralded often as "going back to the old times before the advent of databases when we searched Chemical Abstracts ourselves in the library". Sometimes, it is forgotten that these "old times" were not as good as they now seem to be due to the fact that many chemists lacked appropriate instruction in doing these manual searches as they now often lack knowledge and experience to make the best use of present databases. Users are encouraged for obvious reasons by hosts and producers advertising that they can have all the information "at their fingertips" or at the "touch of a button". Problems like those discussed in this article are not known to many users, and unfortunately often belittled if not ignored altogether by hosts and producers. As a consequence, many users are not willing to invest the necessary amount of training and exercise to become proficient and critical users of databases, they have too much confidence, both in themselves and in the quality of sources, and they are, according to our observation often not flexible enough.

Making the handling of databases easier and going from command-driven mainframe-terminal to client-server systems with appropriate graphic user interfaces (windows, menues, and icons) is certainly a necessary, albeit not a sufficient condition for successful end-user searching. Some existing user interfaces, although looking "friendly", cannot yet fully cope with the complexity of the databases they address: Figure 2 shows a rather simple reaction query for STN CASREACT drawn in STN Express. When uploaded to STN, a sample search shows that the search will not run to completion, not even as a batch job. STN Express made too many generalizations regarding the bond types. Here, a user needs to know the bond conventions (normalization) of the CAS structure/reaction databases in order to get the search going. This is just the kind of problem a really user-friendly interface should prevent.

However, even more important than user interfaces, and a *conditio sine qua non*, are improvements in database quality and meeting user expectancies in the areas of coverage, indexing consistency, data structure, and particularly help systems making the user perception of the database a close match of reality. Impediments and traditions from the printed medium must not hamper database design: indexing problems that can at present only be solved intellectually should not routinely be solved algorithmically,[34] and "machine intelligence" should not be wasted on inadequate database content without proper warning to users, like the powerful author search in SciFinder[12] (retrieving not only exact name matches, but also similar names) on incomplete CA author data (cf. Figure 1).

On the user side, education in chemical information must be an obligatory part of every chemistry syllabus.[35,36] Improvements in user surfaces and database quality will reduce the need for training and permit the instruction to concentrate on search strategies instead of handling of search programs and their commands, but education and support by information specialists in nonroutine cases will still be necessary in the forseeable future.

ACKNOWLEDGMENT

Searches in public databases quoted in this work were funded by the *Eidgenössische Drucksachen- und Materialzentrale*, Bern. Use of the academic programs of STN International (CAS databases and Beilstein online) and Beilstein Information Systems (Beilstein CrossFire) is gratefully acknowledged. The author thanks S. A. Bizzozero for helpful comments.

Note Added in Proof. Since Jan 1996, crossover from Beilstein into Registry is punished by a charge of $8 per CAS Registry Number.

REFERENCES AND NOTES

(1) This publication is based on a presentation at the Herman Skolnik Award Symposium, 210th American Chemical Society National Meeting, Chicago, IL, Aug 22nd, 1995.

(2) Luckenbach, R. The Beilstein Handbook of Organic Chemistry: The First Hundred Years. *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 82−83. Luckenbach, R.; Sunkel, J. Problem Solving with the Beilstein Handbook. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 271−278. Chrzastowki, T. E.; Blobaum, P. M.; Welshmer, M. A. A. Cost/Use Analysis of Beilstein's *Handbuch der Organischen Chemie* at Two Academic Chemistry Libraries. *Serials Librarian* **1991**, *20*(4), 73−84.

(3) *The Beilstein Online Database. Implementation, Content, and Retrieval*; Heller, S. R., Ed.; ACS Symposium Series 436; American Chemical Society: Washington, DC, 1990. Buntrock, R. E.; Palma, M. A. Searching the Beilstein Database Online: A Comparison of Systems. *Database* **1990**, *13*(6), 19−34. Bucher, R.; Jochum C. Beilstein Developments: New Components of the Beilstein Information System. *IATUL Quart.* **1991**, *5*, 94−107. Hicks, M. G. Beilstein Current Facts in Chemistry: A Large Chemical Database on CD-ROM. *Anal. Chim. Acta* **1992**, *265*, 291−300. Hicks, M. G. CD-ROM Chemical Databases: The Influence of Data Structure and Graphical User Interfaces on Information Access. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 32−38.

(4) Jochum, C. Building Structure-Oriented Numerical Factual Databases: The Beilstein Example. *World Pat. Inf.* **1987**, *9*, 147−151.

(5) Lawson, A. J.; Swienty-Busch, R.; Crossfire (XFIRE) Substructure Retrieval of Megafiles: A Comparison of Online and In-House Performance Using the Beilstein File. *Online Inf.* **1993**, *17*, 187−194.

(6) Zirz, C.; Sendelbach, J.; Zielesny, A. Utilization of Beilstein Information at Baeyer. In *Proc. 17. Online-Tagung DGD*; Neubauer, W., Schmidt, R., Eds.; Dtsch, Ges. Dok.: Frankfurt/Main, 1995; pp 247−257. Zass, E.; Donner, W.; Sendelbach, J.; Zirz, C. Experience with the use of Beilstein Inhouse (CrossFire) in an academic and industrial environment. In *Proceedings of the 1995 International Chemical Information Conference Nimes 1995*; Collier, H. R., Ed.; Infonortics, Ltd.: Calne, UK, 1995; pp 138−144.

(7) Published *Main Volumes* and *Supplementary Series* I−IV cover the literature up to 1959, while the first subvolume of the 5th Supplementary Series, covering heterocycles from the primary literature 1960−1979, appeared only in 1984.

(8) Czermak, M. J. New Trends in Specialised Information Policy within the Federal Republic of Germany. *Inf. Serv. Use* **1986**, *6*, 27−33. Czermak, J. M. Chemical Information−Promotion of Innovation in Science and Technology. In *Physical Property Prediction in Organic Chemistry*; Jochum, C., Hicks, M. G., Sunkel, J., Eds.; Springer: Berlin 1988; pp 1−6. Czermak, J. M. A Policy for Scientific and Technical Information in Chemistry. In *Chemical Information [Proceedings of the International Conference Montreaux 1989]*; Collier, H. R. Ed.; Springer: Berlin, 1989; pp 13−18. Der Bundesminister für Forschung und Technologie (Öffentlichkeitsarbeit). Specialized Information Program 1985−1988 of the German Federal Government; Bonn, 1985. Der Bundesminister für Forschung und Technologie (Öffentlichkeitsarbeit). *Fachinformationsprogramm der Bundesregierung 1990-94*; Bonn, 1990.

(9) It is interesting to remember in this context that the creation of the *CAS Registry System* was also publicly funded; cf.: Progress in the Chemical Information System Program. I. Initial Computer-Based Techniques for Input and Output of Chemical Information to Create Literature Access Tools. U.S. NTIS, PB Report 1972; PB-221377. Progress in the Chemical Information System Program. III. Ongoing Computer-Based Techniques for Input and Output of Chemical Information. U.S. NTIS, PB Report 1977; PB-273998. For the later *pre-65 registration project*, again external private and public funds had to be raised, cf.: Feasibility study on registration of chemical substances from pre-1965 indexes of Chemical Abstracts. U.S. NTIS, PB Report 1980; PB81-197055.

(10) Data from a search in the STN FORKAT database, 2/11/94.

(11) Looking for gas phase IR spectra in the ETH Chemistry Library, they were shown the (printed) Sadtler IR spectra collection, STN SpecInfo, and Beilstein CrossFire and found the latter the most useful source of these for their purpose.

(12) Williams, J. SciFinder From CAS: Information at the Desktop for Scientists. *Online* **1995**, *19*(4), 60−66. Williams, J. The SciFinder Experience. In *Proceedings of the 1995 International Chemical Information Conference Nimes 1995*; Collier, H. R., Ed.; Infonortics, Ltd.: Calne, UK, 1995; pp 145−159. Arisumi, P. P.; Turner, W. R. The Introduction and Role of SciFinder, a Unique End-User Search Tool, in the Pharmaceutical Company. "http://www.ch.ic.ac.uk/Paper33" or "http://hackberry.chem.niu.edu/Infobahn/Paper33/title-.html".

(13) For a set of criteria to evaluate chemical databases, cf.: Zass, E.; Donner, W.; Leuther, P.; Lockhoff, A.; Römelt, J.; Spanagel, H.-D. Criteria for the Evaluation of Databases. *Mitteilungsbl.-Ges. Dtsch. Chem., Fachgruppe Chem.-Inf.-Comput.* **1993**, (26), 31−40.

(14) This is a broader, strictly user-oriented error definition and not source-oriented as the commonly used one which only refers to data entry errors like misspelled keywords (e.g., "glas" instead of "glass"), missing data fields, or wrong numerical data.

(15) The present state-of-the-art, *small print* in voluminous database descriptions is not considered appropriate in this context.

(16) But two of those publications cannot be retrieved via the name of this author because of the limit of ten authors in CA discussed in the text.

(17) E.g., the 147 authors and 37 addresses from Oliver, S. G., et al. *Nature* **1992**, *357*, 38−46.

(18) Utility of CA sections as general search terms and for differentiation (e.g., *evolution of HCN* in the sense of gas evolution from prebiological evolution) is reduced by this policy which is probably not known to many researchers.

(19) Pfaltz, A.; Kobelt, A.; Hüster, R.; Thauer, R. K. Biosynthesis of Coenzyme F430 in Methanogenic Bacteria. *Eur. J. Biochem.* **1987**, *170*, 459−467.

(20) These figures were generated by doing a *SmartSelect* for Publication Year on the search results for "et al/AU" in the STN CA File (because of the limit for *SmartSelect*, the search had to be split into four ranges, and the select results were then combined). The author takes this opportunity to thank STN and other hosts for the powerful search tools they have developed in the last few years, particularly for term extraction/statistical analysis like this one or cross-file searching.

(21) Zass, K. Future Requirements for Chemistry Information from the Users' Viewpoint. *Mitteilungsbl. Ges. Dtsch. Chem. Fachgruppe Chem. Inf.* **1993**, (27), 26−43.

(22) Problems of keyword *misspelling* were independently addressed by *S. R. Heller* in his presenation at the same symposium.

(23) One student said "I always use a spelling checker. Why don't they do that at CAS?"

(24) Search at STN (12/12/95). Registry File: S C6H12O6/MF AND BEILSTEIN/LC (L1, 584 compound records; *thus, another 264 $C_6H_{12}O_6$ compounds records that had not been correlated automatically with records in the Beilstein File were excluded from comparison*); CA File: S L1/P NOT 1995/PY, S L1 (S) (MELTING OR MP) NOT 1995/PY, S L1 (S) (OPT? (A) ROTAT? OR ORP) NOT 1995/PY, S L1 (S) (NMR OR RESONANCE) NOT 1995/PY; then for each result SEL RN HIT 1- to determine the number of records (compounds) with preparation/data reported in Table 2 as percentage of total compounds (584). Beilstein File: S L1 (L26, 551 compound records; *because of this CAS Registry Number crossover, the additional 81 $C_6H_{12}O_6$ compound records with literature from the period 1967-1994 in Beilstein, but no CAS Registry Number assigned, are excluded from this comparison*); S L26 AND PRE/FA (L) 1966 < PY < 1995, S L26 AND MP/FA (L) 1966 < PY < 1995, S L26 AND ORP/FA (L) 1966 < PY < 1995, S L26 AND (NMR#/FA OR CTNMR/FA OR NMR/CT) (L) 1966 < PY < 1995; number of records (compounds) reported in Table 2 as percentage of total compounds (551). Because of different compound registration procedures, numbers of records in Beilstein and Registry are *not* exactly comparable, and neither is identical to numbers of compounds in a strict chemical sense. In CAS Registry Number crossovers as used here, the automatic assignment of these numbers in the Beilstein database which is incomplete and on the basis of topology only, i.e., assignment fo the CAS Registry Number of *one* diastereomer to *all* diastereomers in the Beilstein file, creates sets of compounds that cannot be considered identical in a strict sense. Nevertheless, this method of comparison was chosen, as it reflects common usage in searching several databases, and manual creation of ("hand-picked") identical sets of compounds in both databases would be both tedious and costly for the large sets that are needed to make a comparison significant.

(25) In the oral presentation, a overhead transparency stating "The Director of CAS has determined that relying on keywords for compound class searching can be dangerous for the completeness of your search results" was used to illustrate potential solutions to this serious problem in a jocular fashion.

(26) The literature from 1980 onward was excerpted in electronic form instead of via the *Zettel* used before that time period. Cf.: ref 4 and Jochum, C.; Ditschke, C.; Lentz, J.-P. Universal Input Programm for Chemical Structures. *ACS Symp. Ser.* **1987**, *341*, 88−101.

(27) Search in Beilstein CrossFire (version BS9501): "PC (patent country)*" AND "PY (publication year) 1800−1949" etc.; later searches in version BS9502 using the PY and PYP (publication year patent) fields gave the same results.

(28) Search in STN Beilstein, 11/12/95.

(29) Cf.: chemical information listserver "CHMINF-L@iubvm.ucs.indiana.edu" (operated by G. Wiggins), comments in archived messages of June and July 1995.

(30) This is based on experiences with the very first CAS online academic program, where one could search for a fixed monthly fee as long and as much as one wanted. This problem would be more pronounced in an academic environment, as some related observations within the ongoing German project *Endnutzerf*örderung *Chemiedataenbanken* (ref 36) also confirm, but probably not limited to that user group.

(31) In the CD-ROM Current Facts and the in-house database CrossFire, CTUNCH does not exist, and for the other three fields, a global search field *nmr* was established. Unfortunately, the important *hit* format does not work for such global searches but only for individual fields.

(32) Search in STN Beilstein File (2/8/94): S STEPS NMR#/FA (L) PY < 1960, etc.; S (1#### OR 2#### OR 3#### OR 4#### OR 5#### OR 6#### OR 7#### OR 8#### OR 9####)/NMRA.NUC,NMRS-.NUC (L) PY < 1960, etc. (*NMRA.NUC OR NMRS.NUC)/FA does not work*; S STEPS NMR#/FA (L) 1959 < PY < 1980 AND (5-17 OR 5-18 OR 5-19 OR 5-2#)/SO, etc.

(33) By 12/10/95, this percentage had increased to 86.5% for all compunds 1960−1979. Correspondingly, the number of compound records with NMR literature 1960−1979 in the CTUNCH field was reduced from 652 397 (2/8/94) to 524 177 (12/12/95). This provides proof for the (not too well-known) fact that Beilstein is constantly refining the incomplete data for the period 1960−1979 in the database.

(34) The new roles for compounds in CA are a good example for a sensible approach: roles were assigned intellectually by document analysts since July 1994, the best method available at present, but they were assigned algorithmically for the backfile. Intellectual assignment would have been practically and economically impossible, and such roles are better than no role at all.

(35) Zass, E. Chemical Information Education. In *Chemical Information [Proceedings of the Internatinal Conference Montreux 1989]*; Collier, H. R., Ed.; Springer: Berlin, 1989; pp 55−62. Somerville, A. N. Perspectives and Criteria for Chemical Information Instruction. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 177−181. Zass, E. The Role of Chemistry Information in Chemistry Education. *Mitteilungsbl. Ges. Dtsch. Chem. Fachgruppe Chem. Inf.* **1994**, (29) 13−32. For the chemical information instruction program in the ETH chemistry library, see: "http://www.chem.ethz.ch/chembib/Ausbildung.html".

(36) Schütz, R.; Anwand, D. Chemistry Databases: No Books with Seven Seals. *Mitteilungsbl.-Ges. Dtsch. Chem., Fachgruppe Chem.-Inf.-Comput.* **1994**, (30), 13−28. See, also: Kirste, B. GDCh-BMBF-Projekt Endnutzerförderung Chemiedatenbanken. "http://www.chemie.fu-berlin.de/chemistry/chemdb/db-project.html".

(37) Barnard, J. M. Draft Specification for Revised Version of the Standard Molecular Data (SMD) Format. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 81−96. Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurst, A. K. I.; Grier, D. L.; Leland, B. A.; Laufer, J. Description of Several Chemical Structure File Formats Used by Computer Programs Developed at Molecular Design Limited. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 244−255.

(38) Representation of Structure Description Arranged Linearly (Version 1, January 1991). Beilstein Institut: Frankfurt/Main, 1991. Barnard, J. M.; Jochum, C. J.; Welford, S. M. A Universal Structure/Substructure Representation for PC-Host Communication. *ACS Symp. Ser.* **1989**, *400*, 76−81.

(39) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.

CI950249D