HETEROCYCLES, Vol. 82, No. 1, 2010, pp. 63 - 86. © The Japan Institute of Heterocyclic Chemistry
DOI: 10.3987/REV-10-SR(E)8

CHEMICAL INFORMATION RETRIEVAL – A SHORT DISCUSSION ABOUT THE STATE OF THE ART, PROGRESS, AND PITFALLS

Engelbert Zass

Chemistry Biology Pharmacy Information Center, ETH Zurich, Wolfgang-Pauli-Strasse 10, CH-8093 Zurich, Switzerland. E-Mail: zass@chem.ethz.ch

Abstract – Examples of author, keyword, structure, and reaction searches related to the scientific achievements of A. Eschenmoser were analyzed to illustrate the power, but also the limitations of modern database systems like SciFinder, Reaxys, and Web of Knowledge.

Gratefully dedicated to Prof. Dr. Albert Eschenmoser on the occasion of his 85th birthday.

INTRODUCTION

Since the early fifties of the last century, when Albert Eschenmoser started his scientific career, availability and accessibility of chemical information and thus chemical information retrieval have undergone the most dramatic and influential changes since the beginnings of chemistry. The literature of chemistry as a means to propagate research results was essential in shaping chemistry as a science. Quite early in its history, key elements of the information chain were started: scientific journals in 1665 (Philosophical Transactions of the Royal Society, London; Journal des Scavants, Paris), later journals devoted to chemistry (Crells Chemische Annalen 1778) as primary literature, handbooks (Gmelin Handbuch der theoretischen Chemie 1817), and the Pharmaceutisches Zentralblatt in 1830. These types of chemical information resources are still around today, but almost everything else has changed.

In more than 30 years of searching databases, and about 25 years of teaching and supporting chemists to do their own searching, the author has learned that all too often users of electronic information sources are mislead by the now standard graphic user interfaces (GUIs) which are easy to use; chemists assume that the content of the underlying databases is likewise easy to utilize. This, unfortunately, is definitely not true.[1] The complexity of the database content reflects by necessity the complexity of chemistry itself, and this will not (and to a large extent cannot) be solved by user interfaces. Problems are not limited to

the user side, as selection, indexing and abstracting policies of database producers since the time of paper have changed much less than computer technology in general, and user interfaces in particular. Some established traditions from the "paper world" no longer make sense in the age of electronic information, they can be and must be improved. It does not make sense anymore, and it cannot be explained to students, for example, that Manganese(II)sulfate still needs to be searched in the CAS Registry database[2] via the chemically incorrect "dot-disconnected" molecular formula $H_2O_4S.Mn$, which was created in order to show all salts of sulfuric acid together under the formula of the acid in the molecular formula index of Chemical Abstracts in print, ceased at the end of 2009. Therefore, the following discussion will concentrate more on pitfalls, which are often not obvious (but which users need to know in order to obtain information reliably), than on the more obvious progress in chemical information retrieval.

RESULTS AND DISCUSSION

In the long established publication and processing chain of primary (journal articles, patents, conference and research report, theses), secondary (abstracting & indexing sources), and tertiary (handbooks, encyclopedias, monographs, textbooks) chemical information, only secondary and tertiary sources in their publicly available electronic version (databases) are discussed here. Information retrieval is almost always a two-step process: searching in databases, identifying/selecting the appropriate primary publications, then procuring the important ones in full text, electronic or print.

Searches related to the research of Albert Eschenmoser and his coworkers and colleagues are used here as examples – as the problems discussed are quite ubiquitous, this approach will not limit the applicability of the arguments given. These searches were executed in CAS databases[3] with the interface SciFinder (all searches),[4] in Reaxys[5] (joint databases Beilstein,[6] Gmelin, Patent Chemistry Database[7]) for compound and reaction searches, and in the Science Citation Index[8] via the interface Web of Knowledge[9] for author searches. Other interfaces and databases were used additionally when the problem at hand demanded this.

Author Searching

Author searching is very often considered the easiest way to retrieve information, as many achievements in chemistry are attached to the names of the chemists who led the research. While it lost the tediousness involved with printed author indexes, it is by no means easy, and fraught with a lot of problems: inconsistent use of middle names/initials by authors, large variety of addresses/corporate names, limits in the number of authors excerpted from the primary source, and abbreviating full first names to initials. The German Umlaut is either transcribed (ü to ue), or incorrectly converted to the parent vowel; here not only the (secondary) databases, but already the primary publications have different policies which in a search must be taken care of. Even Chemical Abstracts Service (CAS[10]) as a leading source in chemistry is not

always consistent here: the coauthor Schöning is written correctly with Umlaut in *Science* and *Helv. Chim. Acta*,[11] but transcribed by CAS as "Schoning", while his colleague Schlönvogt is transcribed to Schloenvogt.[12,13a] Transliterations from non-Latin alphabets are another problem: when a Russian author publishes in a German journal, his name will be transliterated differently than by CAS from a publication in the original cyrillic alphabet. Measures to bring different name versions of the same author together (including changes in the last name due to marriage!) have become available,[14] as well as authority lists/codes for company names, important in patent databases (Derwent Company codes[15] in the World Patent Index[16]).

The personal publication list of A. Eschenmoser, containing 274 publications, was compared in detail with the results of two author searches. In SciFinder[4] (17.6.2010[17]), 311 records in the bibliographic databases CAplus[18] and Medline[19] were found,[20] including 44 duplicates; only the 231 database records from CAplus (Chemical Abstracts) remaining after exclusion of patents[21] were considered for the comparison. The search in the Science Citation Index (SCI)[8] gave 245 records as of 16.6.2010.

Not recorded in any of the two databases were (a) A. E.'s Diploma and Ph.D. theses, and his Habilitationsschrift[22]; (b) 3 scientific articles in newspapers; (c) 3 scientific reports published locally in Switzerland; (d) 9 abstracts, forewords, comments, laudatios, or commemorations, including two in *Chimia*,[23] although similar material from this journal was excerpted in SCI[8] and in three instances even in CAplus.[18] A focus article was recorded in CAplus, but not with A. E. as contributing author.[24] These omissions in what is considered to be the two most important databases for literature searching in chemistry may well be regarded as relatively unimportant, but the following references also found missing are considered to be more critical: an article from the introductory issue of *Chemistry & Biology*,[25] and no less than 11 chapters from books and conference proceeding,[26] most of which are useful reviews of the work of A. E. and his colleagues.

Retrieved only in Science Citation Index[8] (Web of Knowledge[9]), but missed in SciFinder CAplus[18] were a total of 34 publications, despite the fact that SCI covers only about 6'600 journals for all sciences, while CAplus excerpts no less than 10'000[27] journals for chemistry and related fields alone. Some of the publications missed are indeed recorded in CAplus, but were not retrievable by the author search,[28] others were not covered by CAplus at all. Until 1997, CAS indexed only ten authors at most (for publications with more than ten authors, only the first nine followed by "*et al*." were listed); a total of 4 publications[13] fell victim to this author limit in CAplus, two of them had even only the first author of 11 and 12 authors, respectively, listed.[13b] Three more publications are similarly missed in CAplus, one with Eschenmoser misspelled to "Eschmenmoser",[29] one with the correct last name displayed and printed, but for unknown reasons not retrievable,[30] and the third one with "Anon." given as author instead of A. Eschenmoser.[31]

Retrieved in the Science Citation Index but missing in CAplus were also 4 other publications which one

would definitely expect in the latter database.[32]

Others items exclusively retrieved in SCI[8] were those principally not covered by CAS,[10] among them a letter against the proliferation of journals, signed by many scientists and published 1974 in three different journals. Abstracts from conferences are certainly important if the work reported there is not (yet) published elsewhere; they may be even of interest if this is the case in order to follow the history of a research project. Chemical Abstracts did not cover abstracts for most of the time, and at present only very selectively. Nevertheless, CAplus contains 3 of the 7 *Abstracts of Papers of the American Chemical Society*[33] found in SCI, one conference abstract missed by SCI,[33] while another abstract was missed by both databases.[34] SCI exclusively covered 14 conference abstracts from different journals (7 in *Chimia*), and a discussion from conference proceedings.[33]

Among the publications missed in CAplus[18] is one (R.A. Welch Award Address[35]) of three papers in the *Proc. Robert A. Welch Found. Conf. Chem. Res.*, while the other two from 1968 and 1993 were retrieved in CAplus (SCI does not cover these proceedings). Two of three articles from *Pure Appl. Chem.* were retrieved in CAplus, but not in the Science Citation Index (where coverage for this journal started only in 1974), while the third from the *18th Int. Symposium on the Chemistry of Natural Products* about Hexose Nucleic Acids[36] was found in SCI, but not in CAplus; this is not understood because both the preceeding (Deslongchamps) and the following paper (Hanessian) from this journal issue are in CAplus.

On the other hand, the author search in SciFinder[4] retrieved no less than 17 publications missed in Science Citation Index.[8] Besides the five examples already mentioned above, others appeared in conference proceedings (3) or journals (1) not covered by SCI at all, or not at the time of publications (3). Five publications, however should be in SCI, as they are from journals covered at that time. Again, like in the cases for CAplus, these may be simple mistakes, or editorial decisions obviously hard to understand. The Science Citation Index would have been less comprehensive regarding his publications if A. Eschenmoser himself had not noticed that some publications were not covered in the database; upon his personal request, some (but not all) of those missing were added belatedly to the Science Citation Index database.[37]

It follows from this and other examples analyzed that every author search in chemistry should include at least both SciFinder[4] and the Science Citation Index.[8] The same is true for citation searches: while the Science Citation Index had this search feature right from its beginnings and extended it later back to about 1900, CAS started only in 1997 to include citation data in its bibliographic database. Besides time coverage, the different coverage of primary sources already mentioned is also a differentiating factor. The highest cited publication by A. Eschenmoser in CAplus is a 1999 article in *Science*[38] cited 223 times (31.8.2010[17]), while in SCI, the top cited publication is about the isoprene rule from 1955[39] (489 citations as of 31.8.2010). Of 6'159 publications after 1973[40] which cited one or more publications from A.

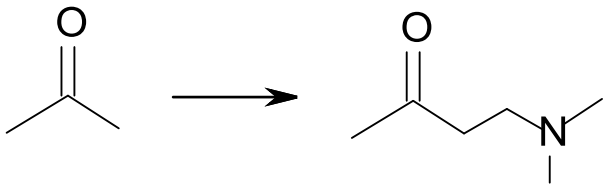
Eschenmoser (31.8.2010[17]), 3’439 (56 %) were found only in the Science Citation Index, and 498 (8 %) exclusively in CAplus; the rest was retrieved in both databases.[41]

An author search in SciFinder is normally carried out also[20] in the medical literature database Medline;[19] 47 publication by A. Eschenmoser were retrieved from the latter database. The same search via the freely accessible PubMed[42] Web interface in Medline found an additional reference[43] – this is not the first time a difference in search results between both interfaces for the same database is noticed. There is another problem here: when searching with “Eschenmoser Albert” in PubMed,[42] only 24 references were found, compared to a search with the initial only – users beware! Medline is the major and sometimes the only database used by many students of pharmacy or biology. This can be dangerous, as an analysis for publications of Cornelia Halin[44] on 15.1.2010 showed: out of a total of 30 publications found, PubMed[42] (21 publications) and Google Scholar (15),[45] both favorized by many students, came out least, followed by SciFinder CAplus (23),[46] Science Citation Index[8] and BIOSIS (Biological Abstracts)[47] with 27 each. The best choice for this search was Web of Knowledge[9] which searches across Science Citation Index, BIOSIS,[46] Derwent Innovations Index,[48] and some other databases, and retrieves all 30 references found here in one search.

Searching for Topics

A topic search is the most complex type of search, one which cannot be at all comprehensive in any database due to the great variety and variability of scientific language and terminology, despite a lot of effort by database producers to support keyword searches with thesauri and systematic indexing. The classical way to search involves keywords, truncation symbols to include singular, plural, and other grammatical forms, and Boolean or proximity operators. This approach is powerful, but may be demanding for the user since it requires knowledge of terminology including spelling and grammatical variants, synonyms, abbreviations and acronyms. All that must be part of a search query that can only try to approximate a high degree of comprehensiveness in all but orientational searches. In order to make this easier at least for routine searches, interfaces with “natural language processing” were developed, the most important being at present SciFinder[4,49] and PubMed.[42] In these interfaces, phrases are entered “as spoken”, e.g., “total synthesis of colchicine”, instead of “total (w) synthes? and colchicine” in traditional retrieval systems. The “natural” query of the user is then translated into a query of the traditional type by the interface, as this is the only one which the retrieval system understands. This parsing of the natural language query is the decisive step. PubMed[42] and SciFinder[4,49] at present differ much in how they handle it: in PubMed, the user can check the system translation of his original query and may modify it, while SciFinder does not offer this option and hence is a “black box”.[50] As proclaimed by CAS, the interface should take care automatically of singular/plural and also synonyms;[50] however, this is not always done,

not even for simple cases of singular/plural: with "total synthesis", both singular and plural form are indeed retrieved; "corrins" retrieves also "corrinoids", but not "corrin", and vice versa.[51] One can enter alternative forms into parentheses in SciFinder's interface, like "porphins (porphin, porphyrin, porphyrins)", but this is limited to a maximum of three terms in parentheses, sometimes not enough. Here, all four closely related terms give quite different results, and must thus be entered explicitly in the search. Keywords are used in CAplus[18] for methods, processes, biological species, compound classes, and reaction types. The latter two are critical due to indexing policies by CAS. Name reactions are indexed rather as an exception, and compound class terms are used for indexing only if the emphasis in the publication is on the compound class, and not on individual compounds from that class: for example, of 12'911 publication records for (+)-estrone, only 5'792 had the keyword "Steroids" or the more specific "19-Norsteroids" in their record (22.8.2010[17]). Thus, comprehensive compound class searches can be done only via substructure. The same is true for reaction types; only 727 of 9'788 references found for the general Mannich reaction (substructure query in Scheme 1) in SciFinder reaction database CASREACT[52] (31.8.2010) contained the keyword "Mannich". The CAS Index Heading is a descriptor (controlled vocabulary) used to index the main topics of publications; here, only 419 references carried one of the four Headings *Mannich bases*, *Mannich reaction*, *Mannich reaction catalysts*, *Mannich reaction kinetics*.



Scheme 1. Mannich-type reaction query

When searching in "Explore References – Research Topics" with "sulfide contraction", a synthetic method developed by Eschenmoser *et al.* for the synthesis of corrins,[53,54] in SciFinder (3.8.2010[17]), the plural of this phrase, and the spelling "sulphide" were automatically searched by the natural language interface to give 246 records[20] (224 of those from CAplus); when refining with the author "Eschenmoser", six remain, corresponding to five publications (one is a duplicate record from Medline[19]). Although Reaxys[5] is not suitable for keyword searches as it lacks the indexing important in Chemical Abstracts, the query "sulphide contraction* or sulfide contraction*"[55] found 73 references containing no less than 5'380 reactions which were refined by the Filter "Reaction Type" to only 41 classified as sulfide contractions – but 5'304 reactions do not have this reaction type classification, so certainly a lot of relevant information is missed (31.8.2010). Instead of using the large databases in SciFinder[4] and Reaxys[5] in reaction keyword searches, for a query of this type, more appropriate sources are available: the reagent database e-EROS,[56]

or the electronic Ullmann's Encyclopedia of Industrial Chemistry[57] gave more concise information about this reaction type with key references like ref. 54. An even better source here is Science of Synthesis,[58] the new edition and electronic version of the well-known Houben-Weyl handbook: with the phrases "sulfide contraction*" or "sulphide contraction*", three highly relevant chapters from Science of Synthesis, and another hit from the old Houben-Weyl 4th ed.[59] were retrieved. Again, it is important to use in this query different spellings (sulfide/sulphide) as well as truncation (with the symbol*) to find singular/plural, and to search for the exact phrase (in quotation marks), as otherwise, one gets irrelevant hits.

A keyword search for literature about the important Chinese medicinal plant *Artemisia annua*[60] further exemplifies the problems: when searching with "Artemisia annua" in the natural language interface of SciFinder (CAplus + Medline,[20] duplicates not removed, 31.8.2010[17]), the interface offered as usual a choice of the phrase searched exactly "as entered" (1'904 records), and "as concept", i.e., including synonyms etc. (2'180). Our tests showed in the latter case that the English name "sweet wormwood" is automatically included as a synonym, which retrieved a total of 84 records, 19 of those carried only this species name, and were not retrieved with "Artemisia annua as entered". Also automatically included is the Latin Genitiv "Artemisiae annuae" (194 retrieved, 174 exclusively so). On the negative side of the result are 73 non-relevant records about other well-known *Artemisia* or wormwood species, many of them retrieved because "annual", or "annua" associated with another Genus appeared in the same record. When the abbreviated name "A. annua" was searched "as entered", among the total of 471 records retrieved, there were 17 records not found with "Artemisia annua as concept"; i.e., the abbreviated species name is not automatically included. Examination of these records showed that 9 of these were relevant, concerning *Artemisia annua*, but the remaining 8 dealt with *Adonis annua*.

Compound Searches

From the primary literature, author and other bibliographic information is excerpted as well as information about compounds and reactions information in a publication; this information is stored in separate bibliographic, compound, and reaction databases (e.g., CAS databases[3]), or in one combined database as in Reaxys.[5] This permits in principle to retrieve all compounds or reactions from a publication, and then limit the results by structure or otherwise. If one is interested in all nucleotide sequences from A. Eschenmoser's publications, Reaxys is not a good choice, as author information in this database is incomplete,[61] and nucleotides and peptides are not systematically covered. CAS has a rather comprehensive coverage for such sequences, but the 4'026 compounds and 10'747 reaction records found in SciFinder (18.8.2010[17]) from Eschenmoser publications cannot easily be restricted to nucleotides.

This raises the question of database interfaces as a limiting factor in accessing database content. Using the

retrieval language STN Messenger[62] at the host STN International,[63] i.e., a command-line interface to the same CAS databases[3] as available under the SciFinder GUI,[4] 4'007 compounds retrieved out of 259 publication records (30.7.2010) were limited with "NUCLEIC/FS" to 467 nucleotides which might have been further narrowed down by sequence length. In a similar operation, 75 references from a literature search with the spider genus name *Atrax* gave 682 compounds which were first limited to those indexed by CAS with the role[64] OCC = occurrence (implying a natural product), then with PROTEIN/FS to 18 peptides which had four references about their isolation from *Atrax* species (18.8.2010). As a consequence, libraries must not only license SciFinder for routine searches by end-users, but also the powerful but complex command-line interface STN Messenger[62] for information specialists to do the more demanding searches in CAS databases.

Structure searches. Searches for structures (exact compounds) and substructures (compound classes) are of absolutely central importance in chemical information retrieval. In substructure searches, the massive size of present structure databases[65] and their fast growth demand tools to enable a user to specify just the compounds he desires, and not too many false hits among them. One needs specific bonds and atoms, as well as general bond types ("any bond") and "super atoms" like M = any metal, X = all halogens, A = any atom except hydrogen, Q = any atom except carbon or hydrogen, or user-defined ones like "atom = N, S, P". Such features are present in virtually all substructure search systems. The situation is not so favorable for Markush structure features as "any alkyl group", "any branched alkyl group", etc., or user-defined groups like R = Me, Et, i-Pr, OH). Another problem area is the topology of rings and bonds: in a powerful retrieval system, a user should be able to specify the topology (chain, ring, chain or ring) for every individual bond or atom; for rings drawn in a query, and for every atom/bond, one should be able to specify not only whether that position might be substituted or not, but also whether and where another ring (of any size) may be anellated. An exact/minimum/maximum number of rings for the entire structure should also be definable.

We illustrate such features and problems in searches for α-aminonitriles, a compound class which played an important role in A. Eschenmoser's research in prebiotic chemistry.[66]

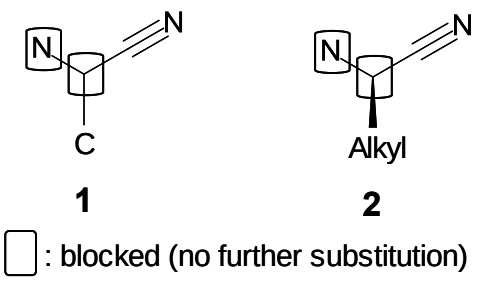


Figure 1. α-Aminonitrile substructure queries

The substructure search with query **1** retrieved 573 compounds (salts, mixtures, and labeled compounds excluded; 753 compounds without this limitation) in Reaxys[5] (19.8.2010[17]), and 921[67] in SciFinder CAS Registry.[2] If only aminonitriles derived from simple D-amino acids are desired, the α-carbon in **1** which may be substituted in any way must be limited to being part of an alkyl group, at the same time specifying chirality. The modified query substructure **2**[67] retrieved 3 (R)-aminonitriles from Reaxys, all included in the 9 found analogously in SciFinder.[68] Without the chirality requirement, we obtained 30 in Reaxys, and 75 in SciFinder (60 of those had no stereochemistry specified in the structure[69]).

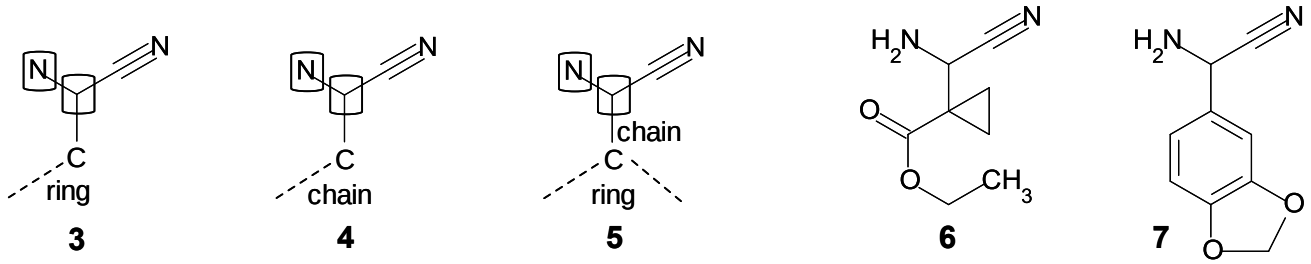


Figure 2. α-Aminonitrile queries and compounds: topology

In query substructure **1**, the α-carbon atom may be in a chain or in a ring. If one needs to specify the topology of such a single atom, tricks have to be used sometimes. In Reaxys, substructures **3** (20.8.2010: 175 compounds[67]) or **4** (286) will serve their purpose, but 3 compounds of type **6** (only α-carbon in a ring) were retrieved in both searches; when specifying the topology of the α-carbon as "chain",[70] only 283 "all chain" nitriles were obtained. If, on the other hand, only compounds like **7** are desired, substructure query **5** is to be used: 150 compounds.[67] If no further rings anywhere in the structure but the one sketched in the query are wanted, the total number of ring closures may be set to 1, which reduced the result to 114 in Reaxys.

Topology for individual bonds or atoms cannot be defined explicitly in SciFinder. One has to use for such a search in CAS Registry[2] the command-line interface STN Messenger[62] which permits all the specification of atom and bond topologies discussed here. As alternative, instead of narrowly defining substitution and ring systems before the search, the problem of specific answers may also be solved by providing analysis features for search results afterwards. SciFinder Scholar[49] is an example here: when searching CAS Registry with this client software, the ring systems in the retrieved compounds may be analyzed and categorized on three different levels: skeleton only, with atoms, with atoms and bond types. In the same vein, one may analyze the actual composition of variable atoms (e.g., which halogens at position X ?), user-defined R-groups, or the first atom of the substituents connected to any position of the

query structure where substitution was permitted ("real-atom attachments"). Unfortunately, the Web version of SciFinder[4] (which will replace the dedicated client software SciFinder Scholar[49] in the near future) provides at present only analysis for "real-atom attachments" (named "Refine by atom attachment"), and lacks the other useful structure analysis features.

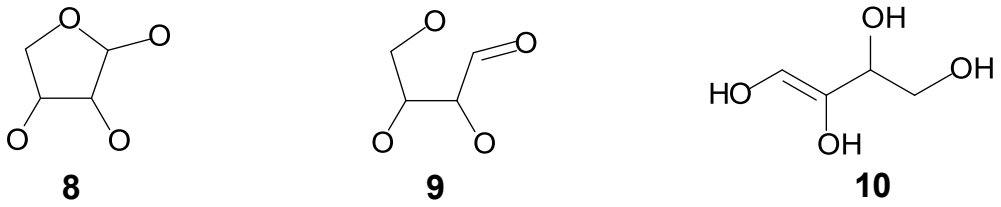


Figure 3. Tetrose substructure queries

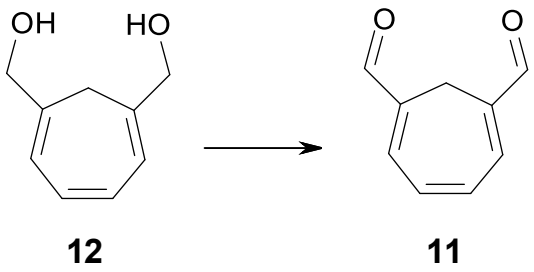
Hexoses, pentoses, and tetroses play an important role in Eschenmoser's research about the etiology of nucleic acids.[38] Carbohydrates are difficult to search, as one has to take care of both the cyclic hemi-acetal/ketal as well as the acyclic isomers. Even more demanding are substructure searches, as it is virtually impossible to differentiate all existing protected derivatives of a specific carbohydrate from oligosaccharides.

In Reaxys,[5] both isomeric structures **8** and **9** (Figure 3; in an exact structure search, all free valences are automatically filled with H) have to be searched separately; searching for all diastereomers[67] gave 20 records for the furanose form **8**, and 9 for the corresponding aldoses **9** (25.8.2010[17]). SciFinder[4] has a more general approach to structure searching, taking care automatically of tautomers, and other double bond isomers, as well as of the dichotomy of metal compounds represented either as coordination compounds or as salts with charge separation. Structure search results are categorized in a "precision analysis"[71] into "Conventional Exact", "Closely Associated Tautomers and Zwitterions", and "Loosely Associated Tautomers and Zwitterions". This retrieval system can in principle handle both sugar structure isomers in a single search[67] (25.8.2010): when entering the acyclic structure **9**, 10 such aldoses were retrieved in the category "Conventional Exact", while "Closely Associated Tautomers and Zwitterions" showed 9 hemiacetals plus the enol form **10** of query structure **9**.[72] Conversely, the structure search with the hemiacetal structure **8** gave 9 records[67] under "Conventional …", and 10 under "Closely …".

The fact that in both databases more records are retrieved than actual stereoisomers exist is quite common for chiral compounds: these databases contain, besides stereoisomers with known absolute configuration, and racemates, additional records where only the relative configuration is given, or where some chiral centers are unspecified; ions, radicals, and (in CAS Registry[2]) oligomers may also be retrieved.

Reaction Searches

Preparations for the cycloheptatriene-1,6-dialdehyde (**11**) from any starting material can be searched in Reaxys,[5] and within SciFinder,[4] both in the reaction database CASREACT[52] and in the bibliographic database CAplus[18] that contains indexing for compounds and their preparation.



Scheme 2. Cycloheptatriene-1,6-dialdehyde

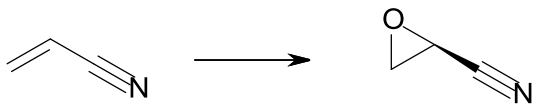
Reaxys retrieved (25.8.2010[17]) 7 reactions which originated from 2 publications,[73] involving seven different starting materials, among them the parent hydrocarbon cycloheptatriene, the corresponding diacid and its dichloride. The corresponding search in CASREACT showed only 2 reaction records, describing the conversion of the diacid to **11**; in contrast to the identical reaction in Reaxys, no detailed reaction conditions were shown, the single publication retrieved[73a] was the same as in Reaxys. Although CASREACT contains some reactions going as far back as 1840, its coverage over time is quite uneven[74] (see Table). Therefore, CAS recently added the search feature "Find Additional Reactions" in CASREACT which searches in the CAplus[18] bibliographic database for additional preparations; in our case, this led to one patent, and a publication also found in Reaxys.[73b] Alternatively, we searched **11** by structure in CAS Registry,[2] and continued with "Get references - Limit results to: preparation" which turned us to CAplus, retrieving again the patent and the two articles[73] already known. Unfortunately, such searches in CAplus for preparation often give a lot of irrelevant results, because they operate on the "preparation role"[64] in indexing which is assigned in our experience much too generously by CAS: included in such results are not only preparations of the compound itself, but also for analogs and unspecified derivatives, and preparation is interpreted in an extremely wide sense - our most startling example so far was the indexing of the isolation of the anesthetic lidocaine from horse urine as a preparation of this compound!

Table. Time Coverage of Reaction Databases[74]

Time Period	Reactions in Reaxys [a]	Reactions in CASREACT [a]
2000-2009	11'398'250 (6'763'681)	18’367’740 (5’393’787)
1990-1999	7'163'548 (3'316'045)	2’542’996 (1’053’985)
1980-1989	4'871'249 (2'565'724)	4’450’075 (1’638’592)
1970-1979	2'323'455 (2'073'726)	387’012 (244’379)
1960-1969	1'631'571 (1'441'370)	52’492 (45’724)
1950-1959	892'040 (694'190)	38’873 (31’859)
1940-1949	404'484 (320'309)	9’530 (8’115)
1930-1939	384'744 (315'469)	3’870 (3’446)
1920-1929	251'384 (223'510)	1’838 (1’665)
1910-1919	139'857 (124'619)	570 (547)
1900-1909	171'596 (154'147)	528 (508)
before 1900	206'933 (190'203)	404 (386)

[a] The numbers of reactions include both single-step reactions and multi-step sequences; single-step reactions shown in parentheses (20.8.2010/7.9.2010[17]);. they indicate the different time coverage, but are not directly comparable, as Reaxys contains reactions not covered by CASREACT: inorganic reactions, and “half reactions” where only the product or the starting material is given (similar to “Find Additional Reactions” in SciFinder).

The exact transformation (a fully defined reaction instead of the previous “half-reaction”) of diol **12** to dialdehyde **11** should have been retrieved in the previous search, but it was not: this reaction was only accessible indirectly via refs. 73 which both cite an earlier paper by Vogel *et al.*[75] describing the oxidation **12** → **11** - unfortunately, this publication is not in CASREACT at all, both CAplus and Reaxys contain it, but only index further reactions of **11**, not its published preparation.[76]



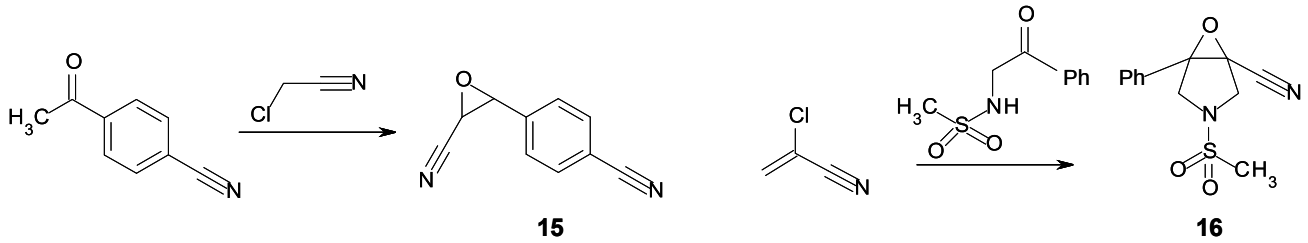
13 **14**

Scheme 3. Epoxidation of acrylonitriles

In the epoxidation of acrylonitriles **13** (Scheme 3) to the chiral epoxides **14**, a reaction search with substructures in Reaxys retrieved 14 reactions, all relevant, from 6 publications (1.9.2010). When the topology of the double bond in **13** was limited to “ring”, only 4 reactions remained. The same reaction

substructure search without specifying stereochemistry at the center bearing the nitrile group gave 155 reactions; some of these had the "wrong" configuration, others had no stereochemistry specified at that center. Also retrieved were "false hits", e.g., **15** and **16** (Scheme 4).

This is a common problem in reaction substructure searches, because the match of both of the partial reaction partner substructures is a necessary, but, in contrast to reaction searches with exact structures, not a sufficient condition to retrieve only the desired reactions. This needs marking of corresponding atoms at or near the reaction centers in both reactant and product substructure, the so-called atom-atom mapping. Databases use different proprietary mapping algorithms for reactions. These are not without problems, particularly in Reaxys: when mapping is added to the query, only 121 (instead of the original 155) reactions were retrieved – but in eliminating the few "false hits", a much larger number of correct hits which lacked mapping in the database were also eliminated. Using mapping in the previous search for **13** → **14**, where no "false hits" were present at all, would have eliminated half of the relevant reactions !



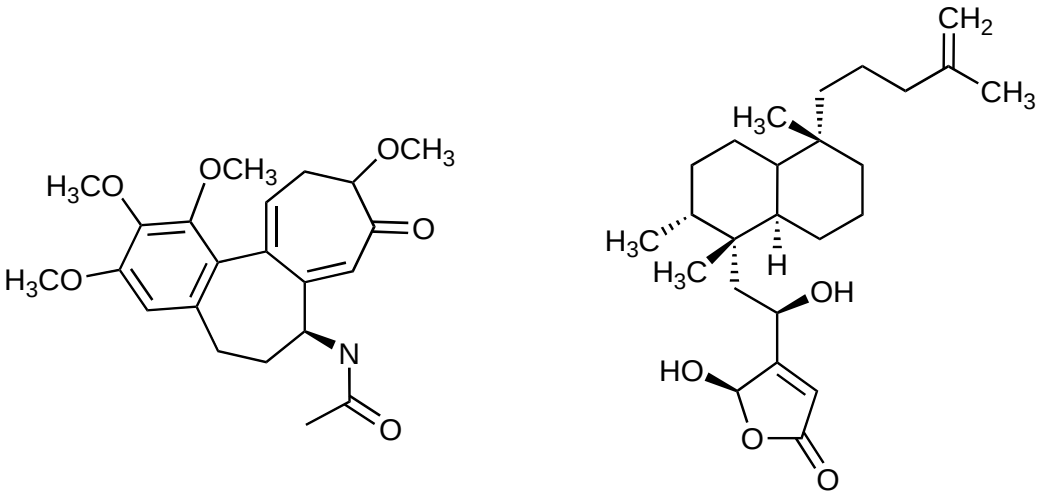
Scheme 4. False hits from Reaxys for reaction **13** → **14**

When attempting to search the reaction **13** → **14** in SciFinder CASREACT,[52] one is informed that stereochemistry is ignored in the search. Without specifying stereochemistry, 227 reactions in 82 publications were retrieved (24.8.2010[17]); with reaction center mapping included, seven non-relevant reactions (2 references) were excluded. Being unable to search for the stereochemistry of reactions is a grave disadvantage in this large and otherwise very important database. Fortunately, there is a way to bypass this limit. But it is probably not obvious to the routine user: we start with a product structure search in SciFinder CAS Registry[2] (in this database, stereochemistry is searchable), and we selected those compound records with "Absolute stereo match" from the automatic stereoanalysis:[69] 59 compounds.[77] For these, 94 reactions were retrieved with "Get Reactions – Product", preparations of the desired chiral oxiranecarbonitriles from **any** starting material. They must be refined with the original reaction **13** → **14** (stereosearch now not necessary, as already implied for the product) to give 16 reactions out of 7 publications (only three of those were also found in Reaxys); in this case, mapping of reaction centers (in the Refine step) made no difference.

The reaction examples shown here need large databases like CASREACT[52] or Reaxys[5] to give a useful number of results; for more general questions, functional group transformation like "methods for oxidizing 1,3-diols to 1,3-dialdehydes", too many reactions are retrieved, and post-processing options for results (filters) are still not powerful enough. Therefore, for synthetic methods, Science of Synthesis,[58] e-EROS[56] and similar sources are preferred.

Total Syntheses (Reaction Sequences). Early reaction databases concentrated on individual transformations, single-step reactions, regardless of the fact whether the transformation described was executed individually (i.e., as application example for a new reagent), or as a key step in a multi-step synthetic sequence. CASREACT[52] was the first database to address the problem of reaction sequences, followed later by the Beilstein database.[6] When this database became available in 2009 as part of a larger database (including Gmelin and the Patent Chemistry Database[7]) under the completely redesigned user interface Reaxys,[5] a new feature, "Synthesis Plans", was available which utilized the reaction sequence information to its full extent. Multistep reactions in both Reaxys and CASREACT are *per se* limited to a single publication. Synthesis Plans in Reaxys permits a user to overcome this limit and retrieve any synthesis for a given compound in the database, be it in the same publication, or in several others. Several preparations of a compound can such be compared in this tool (reagents, conditions, or yield if given), and the most useful transformation may be selected to repeat then the search with the new starting material, thus building synthetic trees across publications.[78]

Using the review by Graening and Schmalz[79] as a guide, we examined whether important colchicine total syntheses, beginning with those by Eschenmoser and van Tamelen in 1959, are retrievable in Reaxys and SciFinder.[76] Reaxys retrieved with structure **17** (1.9.2010[17]) only two papers about the total synthesis by Cha (1998); when "Ignore stereo" was set, in addition, total syntheses by Banwell and Graening, and the first total synthesis by Eschenmoser appeared - but without any details, and with "Schreiber *et al*." as author.[61]



(-)-colchicine (17) (-)-dysidiolide (18)

Figure 4. Total Syntheses

In SciFinder CASREACT (1.9.2010), the situation was equally unsatisfactory: only one of the 15 total syntheses 1959-2004 summarized in ref. 79 was found: again 2 publications by Cha (same as Reaxys) out of a total of 3. This result for natural (-)-colchicine (**17**) was enlarged if stereochemistry was again ignored and labeled compounds were permitted: 8 publications, now including syntheses by Banwell ((±)-colchicine), Pontikis ([7-14C]colchicine), and Graening (this was missed in the first search because (±)-colchicine was indexed, not the (-)-enantiomer **17**). When the result in CASREACT was augmented by "Find Additional Reactions" from CAplus, further total syntheses by Tobinaga, Evans, and Banwell were found – but from all these publications, not a single reaction showed up in CASREACT. A "brute force" approach, looking for all preparations of any colchicin isomer in CAplus[18] and Medline[19] gave 1863 references (including duplicates; 1.9.2010), with a majority of non-relevant publications, particularly from Medline, but also from CAplus;[80] even so, of the total syntheses listed in ref. 79, those by Kaneko and Kato were not retrieved. Refine with "total synthesis" left only 57 references, a manageable number, but then the syntheses by Eschenmoser, Nakamura, and Toromanoff were also missed.

Problems concerning total syntheses in reaction databases are not limited to older syntheses like that of colchicine (**17**): the natural product (-)-dysidiolide (**18**), a compound of potential pharmaceutical interest because of its cytostatic effect, has been the target of seven total syntheses 1997-2002 by Corey, Boukouvalas, Danishefsky, Shirai, Forsyth, Yamada, and Waldmann. For **18**, its racemate and two epimers (June 2009), Reaxys[5] retrieved in principle all total syntheses, but not every relevant publication about them. SciFinder CASREACT[52] retrieved only five of them, and missed the ones by Boukouvalas

and Danishefsky, although their publications do have CASREACT records – but they contain only an intermediate single step from each synthesis.[81] The content even for those total syntheses found in both databases is different, as CASREACT sometimes reports only part of it: e.g., longest reaction sequence from Corey[82] (CASREACT 3 *vs.* Reaxys 24 steps), and from Yamada[83] (22 *vs.* 1). When searching for preparations of dysidiolides in CAplus,[18] all total syntheses and their publications are found, plus publications by Kaliappan, Piers, Jung, and Waldmann about partial or formal syntheses, the latter not to be expected in reaction databases.

Most disquietening is the fact that, at least for these two examples, Google Scholar (2.9.2010) kind of beats SciFinder and to some extent Reaxys: already on the first results page for the search "dysidiolide total synthesis", all seven of them were retrieved, plus two formal syntheses. For colchicine, the first results pages gave the total syntheses by Graening, Evans, Boger, and Woodward. Given the ease of use in these and many other examples in getting at least partial results, how to convince students to search SciFinder, Reaxys etc., and how to justify the considerable expenses for these databases in the future ?

CONCLUSIONS

The examples shown here illustrate that very often one database (information source) is not sufficient to solve a problem, and for the major databases, even more than one interface may be needed for some searches. Apart from the detailed problems discussed here, the major problem of present systems is that the modern graphic user interfaces hide the complexity of the underlying databases. This was quite different with printed sources like Chemical Abstracts or the Beilstein handbook, as well as with the first, command-line retrieval database systems becoming available after 1972. In addition, end-users do not recognize that interfaces like SciFinder are limited to routine searches.

According to our experience at ETH Zurich, there is only one remedy to these problems, training in the use and critical usage of all information sources - including Google and Wikipedia[84] so popular among the current student generation - as integral, obligatory part of the education of chemists,[85] and ongoing support for chemistry students and researchers by information specialists who are also to handle those searches that cannot be dealt with by a routine user. Chemical information problems may be classified into three categories: those solvable by the chemists with the experience and sources at hand, others which in principle a chemist could do himself, but where the query formulation is not obvious and needs to be shown to him (e.g., how to bypass the lack of stereochemistry in a reaction search in CASREACT[52]), and finally those questions where neither the search tools nor the experience of a chemist are adequate for the problem at hand. In this category fall some data searches, and many topic searches in Chemical Abstracts where a certain degree of comprehensiveness is an issue. The latter category demands, besides in-depth experience to handle different retrieval systems/interfaces, a thorough knowledge of database

coverage and indexing policies that are the domain of the specialist. In this context, it is also very important that specialists keep in close contact with database producers/providers, both for meta-information about sources, and to give feedback for improvements regarding database content (coverage, indexing) and user interfaces.

Chemical information retrieval has been going through more major changes in the last 30 year than in the almost 300 years before, and we can expect further major changes in the near future. The classical role model of primary, secondary, and tertiary literature has survived so far despite the very extensive substitution of paper by electronic media. In contrast to other areas of science, not only publishers, but also chemists, both as authors and users, seem to be still quite conservative in their publishing habits. There are presently no replacements in sight in the primary literature for electronic journal articles and patents. Tertiary literature, i.e., monographs, textbooks, encyclopedias, is even more important than ever, given the intellectual work that goes into these information sources to reduce information overload. The category which seems to be in some danger, however, is the secondary literature with its abstracting & indexing services. While in the era of print, they were absolutely indispensable to access the primary literature, this is nowadays open for change when a very large part of this primary literature is available and directly searchable in electronic form. Primary information, however, is distributed among many publishers, with different user interfaces and retrieval systems. Federated searching across all major publishers could solve this problem, it is not so much a technical but a political and organizational one. In such a system, we would above all need (sub)structure searching for both compounds and reactions, a *sine qua non* in chemical information. With existing solutions for converting structure drawings into searchable structure representations (connection tables), or name-to-structure conversion, this is also already feasible in principle. The biggest problem so far lies in another area: keyword searching in the full-text of journal articles and patents is definitely no substitute for indexing and other controlled, standardized terminology used in Chemical Abstracts and similar sources. Similar problems exist with physical data in the primary literature. Tagging of data, and keywords assigned by authors in a publication will most probably never be accepted by them because of the sizeable extra effort involved. Other approaches, using ontologies and semantic analysis, look more promising here.[86]

"Information at your fingertips", as producers make us believe in their adds, is only true for either very simple questions, or a first quick orientation. "The search is over" only after several sources have been used, cross-checked, and after the primary literature has been acquired, read, and evaluated. Given such self-presentation of publishers on one hand, and the licence fees we are made to pay for their products, we demand even higher quality than exists now – but this cannot be just put on the doorsteps of producers. We as authors and users have to contribute our part, in being more consistent ourselves when publishing (names, addresses), and in checking how our publications appear in databases,[87] giving the producers a

critical feedback from our searches, as they cannot really improve their products without our assistance.

ACKNOWLEDGEMENTS

The author thanks Dr. M. Brändle for helpful suggestions and corrections for this manuscript.

REFERENCES (AND NOTES)

1. For a discussion of search problems in specific research projects, see E. Zass, D. A. Plattner, A. K. Beck, and M. Neuburger, *Helv. Chim. Acta*, 2002, **85**, 4012 (data search); D. Seebach, A. K. Beck, S. Capone, G. Deniau, U. Groselj, and E. Zass, *Synthesis*, 2009, 1 (substructure search); D. Seebach, E. Zass, W. B. Schweizer, A. J. Thompson, A. French, B. G. Davis, G. Kyd, and I. J. Bruno, *Angew. Chem.*, 2009, **121**, 9774; *Angew. Chem. Int. Ed.*, 2009, **48**, 9596 (topic search).
2. CAS Registry (compound database): http://www.cas.org/expertise/cascontent/registry/regsys.html; http://www.cas.org/expertise/cascontent/registry/; D. W. Weisgerber, *J. Am. Soc. Inf. Sci.*, 1997, **48**, 349.
3. CAS Databases: http://www.cas.org/expertise/cascontent/ataglance/.
4. SciFinder (Web): http://www.cas.org/products/scifindr/; http://www.cas.org/support/scifi/; D. D. Ridley, 'Information Retrieval: SciFinder' (2nd ed.), J. Wiley & Sons Ltd., Chichester, 2009.
5. Reaxys: http://www.info.reaxys.com/; http://www.info.reaxys.com/training-center; J. Goodman, *J. Chem. Inf. Model.*, 2009, **49**, 2897.
6. S. R. Heller (ed.), 'The Beilstein System: Strategies for Effective Searching', American Chemical Society Publications, 1998.
7. Patent Chemistry Database (PCD): http://www.patentchemistrydatabase.com/; http://www.elsevier.com/wps/find/bibliographicdatabasedescription.cws_home/713992/description#description.
8. Science Citation Index Expanded (SCI): http://thomsonreuters.com/products_services/science/science_products/a-z/science_citation_index_expanded.
9. Web of Knowledge: http://thomsonreuters.com/content/PDF/scientific/Web_of_Knowledge_factsheet.pdf.
10. CAS (Chemical Abstracts Service): http://www.cas.org/.
11. K.-U. Schöning, P. Scholz, S. Guntha, X. Wu, R. Krishnamurthy, and A. Eschenmoser, *Science*, 2000, **290**, 1347; K-U. Schöning, P. Scholz, X. Wu, S. Guntha, G. Delgado, R. Krishnamurthy, and A. Eschenmoser, *Helv. Chim. Acta*, 2002, **85**, 4111.
12. I. Schlönvogt, S. Pitsch, C. Lesueur, A. Eschenmoser, B. Jaun, and R. M. Wolf, *Helv. Chim. Acta*, 1996, **79**, 2316.

13. a) S. Pitsch, R. Krishnamurthy, M. Bolli, S. Wendeborn, A. Holzner, M. Minton, C. Lesueur, I. Schlönvogt, B. Jaun, and A. Eschenmoser, *Helv. Chim. Acta*, 1995, **78**, 1621; b) E. Bertele, H. Boos, J. D. Dunitz, F. Elsinger, A. Eschenmoser, I. Felner, H. P. Gribi, H. Gschwend, E. F. Meyer, M. Pesaro, and R. Scheffold, *Angew. Chem.*, 1964, **76**, 393; J. Schreiber, Dorothee Felix, A. Eschenmoser; M. Winter, F. Gautschi, K. H. Schulte-Elte, E. Sundt, G. Ohloff; J. Kalvoda, H. Kaufmann, P. Wieland, and G. Anner, *Helv. Chim. Acta*, 1967, **50**, 2101; c) G. Ksander, G. Bold, R. Lattmann, C. Lehmann, T. Früh, Y.-B. Xiang, K. Inomata, H.-P. Buser, J. Schreiber, E. Zass, and A. Eschenmoser, *Helv. Chim. Acta*, 1987, **70**, 1115.
14. For attempts at identifying/clustering author names, see the Thompson Distinct Author Identification System: http://science.thomsonreuters.com/support/faq/wok3new/dais/; Scopus Author Identifyer: http://www.info.sciverse.com/scopus/scopus-in-detail/tools/authoridentifier; Open Researcher and Contributor ID (ORCID): http://www.orcid.org/news.
15. Derwent Patent Assignee/Company Codes: http://science.thomsonreuters.com/support/patents/dwpiref/reftools/companycodes/.
16. WPI (World Patent Index): http://thomsonreuters.com/products_services/legal/legal_products/intellectual_property/DWPI.
17. Date of search; details about query and results are available from the author on request (zass@chem.ethz.ch).
18. CAplus (bibliographic database): http://www.cas.org/expertise/cascontent/caplus/index.html.
19. Medline (Medical Literature Analysis and Retrieval System Online): http://www.nlm.nih.gov/pubs/factsheets/medline.html.
20. Literature searches in SciFinder are automatically executed in both CAplus and Medline, duplicates can be eliminated, results may be restricted by “Analyze” or “Refine” to either database.
21. Patents are not in A. E.’s publication list, and also not covered as bibliographic records in the Science Citation Index.
22. A. Eschenmoser, Synthese von 2,5-Dimethyl-3-oxymethyl-heptadien-(1,5), Diplomarbeit, Eidgenössische Technische Hochschule Zürich, 1948, DOI: 10.3929/ethz-a-006069807, http://e-collection.ethbib.ethz.ch/eserv/eth:1282/eth-1282-01.pdf; A. Eschenmoser, Zur säurekatalysierten Zyklisierung bei Mono- und Sesquiterpenverbindungen, Promotionsarbeit Nr. 2018 ETH Zürich, Juris-Verlag, Zürich, 1952, DOI: 10.3929/ethz-a-006069807, http://e-collection.ethbib.ethz.ch/eserv/eth:1282/eth-1282-01.pdf; A. Eschenmoser, Zur Kenntnis Säurekatalysierter Cyclisationen von Polyenverbindungen der Terpenreihe, Habilitationsschrift ETH Zürich, 1956, DOI: 10.3929/ethz-a-006078807, http://e-collection.ethbib.ethz.ch/eserv/eth:1364/ eth-1364-01.pdf.
23. A. Eschenmoser, *Chimia*, 1986, **40**, 389; H. Bethge and A. Eschenmoser, *Leopoldina (R.3) 1986*

(1988), **32**, 9; A. Eschenmoser, *Chimia*, 1991, **45**, 397.

24. Nina Hall, *Chem. Commun.* (Cambridge, UK), 2004, 1247.
25. A. Eschenmoser, Towards a chemical etiology of the structure of nucleic acids, *Chemistry & Biology*, April 1994, Introductory issue, iv.
26. A. Eschenmoser, 'Studies on organic synthesis', 23rd International Congress of Pure and Applied Chemistry, Vol. 2, Butterworths, London, 1971, p. 69; A. Eschenmoser, 'Organische Naturstoffsynthese und Vitamin B_{12}', Jahrb. Akad. Wiss. Göttingen, Vandenhoeck & Ruprecht, Göttingen, 1977, p. 29; A. Eschenmoser, 'Chemistry, Cinquantenaire de la Fondation de l'Académie Pontificale des Sciences', Comptes-rendu et Actes de la Session Plénière et des Célébrations 1986, Pontif. Acad. Sci. Scr. Varia, 1988, **73**, 253; A. Eschenmoser, 'Kon-Tiki-Experimente zur Frage nach dem Ursprung von Biomolekülen', 'Materie und Prozesse vom Elementaren zum Komplexen', Verh. Ges. Dtsch. Naturforsch. Ärzte (116. Versammlung, Berlin 1990), ed. by W. Gerok *et al.*, Wissenschaftliche Verlagsgesellschaft mbH., Stuttgart, 1991, p. 135; A. Eschenmoser, 'Zur Frage nach dem Ursprung des Lebens', 'Mensch und Natur (Festschrift zur 250-Jahr-Feier der Naturforschen-den Gesellschaft in Zürich 1746-1996)', ed. by Redaktionskommission der NGZ, Koprint AG, Alpnach Dorf, 1996, p. 62; A. Eschenmoser, 'Chemische Aetiologie des Strukturtyps der natürlichen Nukleinsäuren', Jahrb. 1999 Dtsch. Akad. Naturforsch. Leopoldina (R.3) 2000, **45**, 195; D. Arigoni, J. D. Dunitz, and A. Eschenmoser, 'Vladimir Prelog, 23 July 1906 - 7 January 1998', Biog. Mems. Fell. R. Soc. London, 2000, **46**, 443; A. Eschenmoser, 'Epilogue: Synthesis of Coenzyme B_{12}: A Vehicle for the Teaching of Organic Synthesis', 'Essays in Contemporary Chemistry - From Molecular Structure towards Biology', ed. by G. Quinkert and M. V. Kisakürek, Verlag Helvetica Chimica Acta, Zürich, 2001, p. 391; A. Eschenmoser, 'Design versus Selection in Chemistry and Beyond', 'Science and the Future of Mankind - Science for Man and Man for Science', *Pontif. Acad. Sci. Scri. Varia*, 2001, **99**, 235; A. Eschenmoser, 'Creating a perspective for comparing', Fitness of the Cosmos for Life: Biochemistry and Fine-Tuning (Workshop Templeton Found., Harvard Univ., Oct. 2003)', Part IV, ch. 16, ed. by J. D. Barrow *et al.*, Cambridge Univ. Press, Cambridge, 2007, p. 349; A. Eschenmoser, *Pontif. Acad. Sci. Acta*, 2009, **20**, 181.
27. http://www.cas.org/expertise/cascontent/ataglance/index.html; CASSI (Chemical Abstracts Service Source Index, version 1.01, December 2009) lists 22'257, active serials", i.e., journals and series.
28. All A. E. publications not retrieved in SciFinder by the author search were cross-checked with a reference search (Explore References – Journal) for presence in CAplus.
29. A. Eschenmoser and M. V. Kisakürek, *Helv. Chim. Acta*, 1996, **79**, 1249.
30. A. Eschenmoser, *Orig. Life Evol. Biosph.*, 2007, **37**, 309.
31. A. Eschenmoser, *Chimia*, 1989, **43**, 153.

32. A. Eschenmoser and A. Fürst, *Experientia*, 1951, **7**, 290; I. Felner, A. Fischli, A. Wick, M. Pesaro, D. Bormann, E. L. Winnacker, and A. Eschenmoser, *Angew. Chem.*, 1967, **79**, 863; *Angew. Chem. Int. Ed.*, 1967, **6**, 864; A. Eschenmoser, *Chimia*, 1990, **44**, 1; A. Eschenmoser, *Chimia*, 1993, **47**, 148.
33. Items retrieved in author search, but not included in A. Eschenmoser's personal publication list.
34. R. Scheffold, E. Bertele, M. Pesaro, and A. Eschenmoser, *Chimia*, 1964, **18**, 405.
35. A. Eschenmoser, *Proc. Robert A. Welch Found. Conf. Chem. Res.*, 1974, **18**, 269.
36. A. Eschenmoser, *Pure Appl. Chem.*, 1993, **65**, 1179.
37. A. Eschenmoser, personal communication, 2010.
38. A. Eschenmoser, *Science (Washington, D. C.)*, 1999, **284**, 2118.
39. A. Eschenmoser, L. Ruzicka, O. Jeger, and D. Arigoni, *Helv. Chim. Acta*, 1955, **38**, 1890.
40. SCI and CAplus are routinely searched via their different web user interfaces (Web of Knowledge, SciFinder). In order to eliminate duplicates mechanically, however, we had to do a multifile search at the host STN International, using the database versions STN CA and STN SCISEARCH (starting only in 1974, while SCI under Web of Knowledge extends back to 1899).
41. Similar differences in the results of citation searches were found by K. M. Whitley, *J. Am. Soc. Inf. Sci. Technol.*, 2002, **53**, 1210.
42. PubMed: http://www.ncbi.nlm.nih.gov/pubmed/.
43. R. Krishnamurthy, S. Guntha, and A. Eschenmoser, *Angew. Chem. Int. Ed. Engl.*, 2000, **39**, 2281.
44. Chair for Drug Discovery Technologies, ETH Department of Chemistry and Aplied Biosciences; http://www.pharma.ethz.ch/institute_groups/drug_discovery_technologies.
45. Google Scholar: http://scholar.google.ch/.
46. Again, SciFinder Medline retrieved one publication less than PubMed.
47. BIOSIS: http://thomsonreuters.com/content/PDF/scientific/BIOSIS_Factsheet.pdf.
48. Derwent Innovations Index: Web version of World Patent Index[16] (coverage starting 1963), available via Web of Knowledge; http://thomsonreuters.com/products_services/legal/legal_products/intellectual_property/ Derwent_Innovations_Index.
49. SciFinder Scholar: http://www.cas.org/support/academic/sf/index.html.
50. A. B. Wagner, *J. Chem. Inf. Model.*, 2006, **46**, 767.
51. Some time ago, "acid" retrieved the plural "acids", but this was not true for "ketone", "aldehyde", or "oxime". This was fixed by CAS in the meantime.
52. CASREACT (reaction database): http://www.cas.org/expertise/cascontent/casreact.html; J. E. Blake and R. C. Dana, *J. Chem. Inf. Comput. Sci.*, 1990, **30**, 394; M. Clark, *J. Chem. Inf. Comput. Sci.*, 1999, **39**, 635.
53. A. Fischli and A. Eschenmoser, *Angew. Chem.*, 1967, **79**, 865; *Angew. Chem. Int. Ed.*, 1967, **6**, 866.

54. S. Blechert, *Nachr. Chem. Tech. Lab.*, 1980, **28**, 577.
55. Reaxys has no natural language interface, so one has to take care of different spellings, and of singular/plural with truncation: * = any number of any characters; "sulphide contraction" retrieved only 10 records.
56. e-EROS (Encyclopedia of Reagents for Organic Synthesis): http://onlinelibrary.wiley.com/book/10.1002/047084289X.
57. Ullmann's Encyclopedia of Industrial Chemistry: http://onlinelibrary.wiley.com/book/10.1002/14356007.
58. Science of Synthesis: http://www.science-of-synthesis.com/en/products/reference-works/science-of-synthesis.html; http://www.thieme.de/connect/en/pdf/thieme_science_of_synthesis.pdf.
59. The printed Houben-Weyl old editions were scanned, but they are only browsable and searchable in their tables of contents, not in the full text and with structures / reactions as in Science of Synthesis produced as a database.
60. Artemisia annua (sweet wormwood): D. Greenwood, 'History of Antimalarial Agents', Encyclopedia of Life Sciences, Wiley-Blackwell, London, 15.9.2009, DOI: 10.1002/9780470015902.a0003624.pub2 / http://onlinelibrary.wiley.com/doi/10.1002/9780470015902.a0003624.pub2/full.
61. The Beilstein handbook (from which the majority of records in the Reaxys [5] database originated) did not cover authors completely, as it was compound and not reference-oriented like, e.g., Chemical Abstracts or Science Citation Index: for the primary literature up to 1979, only two authors were excerpted at most, otherwise only the first author was listed with "*et al.*" (until 1959, only last names were listed); for later literature, the limit was 6 authors.
62. Retrieval language STN Messenger: http://www.stn-international.de/command_language.html?&cHash=215d85bfc5.
63. Host STN International: http://www.stn-international.de/index.php?id=123.
64. CAS Roles: http://www.cas.org/ASSETS/EB85B919049C4E448DCF8D391788F0DD/casroles.pdf.
65. E.g., size and daily growth of CAS Registry: [2] http://www.cas.org/cgi-bin/cas/regreport.pl.
66. E. Wagner, Y.-B. Xiang, K. Baumann, J. Gück, and A. Eschenmoser, *Helv. Chim. Acta*, 1990, **73**, 1391, and later publications in this series.
67. In these searches, salts, mixtures, and labeled compounds (isotopes) were excluded in Reaxys and SciFinder CAS Registry; in the latter database, results were restricted to "Conventional exact" after precision analysis (cf. discussion to Figure 3).
68. Also retrieved in SciFinder were two nitrile oxides, because the precision analysis (described later in

this article for tetroses) did not work for this query with a chiral center specified.

69. Whenever chirality is specified in a structure query, SciFinder CAS Registry[2] retrieves all existing stereoisomers, and categorizes them in an automatic stereoanalysis into "Absolute stereo match", "Absolute stereo mirror image", "Relative stereo match", "Stereo that doesn't match query". There are also often records with "No stereo in answer structure"; this refers to compounds where either the configuration was not specified in the primary literature, or records from the early times of the CAS Registry System when a given configuration was not stored in the structure, but in stereodescriptors as part of the systematic name – users have to be aware of this deficiency.
70. While the topology of a bond may be specified explicitly via the "Edit Bond" menu in the Reaxys structure editor, the topology of an atom can only be specified implicitly by defining "rb (ring bonds) = 0" as atom attribute; this feature is not obvious at all and hidden in the Advanced Tab of the "Periodic System" menu item.
71. Unfortunately, this precision analysis after the structure search is no longer done automatically in the SciFinder Web client (as it was in SciFinder Scholar), but must be activated by the user before the search.
72. This compound has no references itself, it was registered by CAS according to its rules because it is a component in a polymer.
73 a) E. Vogel, H. M. Deger, J. Sombroek, J. Palm, A. Wagner, and J. Lex, *Angew. Chem.*, 1980, **92**, 43; b) T. Asu, S. Kuroda, and K. Kato, *Chem. Lett.*, 1978, **7**, 41.
74. Unfortunately, this is not documented appropriately by CAS, and we had to extract the relevant information concerning time coverage from the database ourselves (as we did also for Reaxys). This is one of (too) many examples where producers do not provide the necessary meta-information about their databases which is essential for reliable search results. The worst "offender" here is Google – we know much less about Google Scholar than about any other source mentioned !
75. E. Vogel, R. Feldmann, and H. Düwel, *Tetrahedron Lett.*, 1970, **11**, 1941.
76. The author admits a special reason for choosing this particular example: when he had to prepare **12** from **11** in his advanced organic chemistry lab course in 1971 at Cologne University, his interest was attracted to the colchicine synthesis and other research by A. Eschenmoser. This was seminal in his decision to apply to work for his Ph.D. in A. E.'s group at ETH Zurich after finishing his Diplomarbeit in the group of E. Vogel.
77. About half of the chiral oxiranecarbonitriles found in the substructure search for the reaction product had no stereochemistry specified in their CAS Registry records, cf. 69.
78. Synthesis Plans in Reaxys was conceptually based on earlier features to concatenate reaction sequences as well as single-step reactions in MDL DiscoveryGate (now Symyx DiscoveryGate).

Information for such "synthesis planning" is in principle also available in CASREACT, but SciFinder at present lacks a tool to utilize this easily.

79. T. Graening and H.-G. Schmalz, *Angew. Chem. Int. Ed.*, 2004, **43**, 3230.
80. Another example of the detrimental effect of the too generous assignment of the "preparation" role in CAplus indexing.
81. When the "Find Additional Reactions" feature in CASREACT (which was not yet available in June 2009) was used, these syntheses are retrieved from the CAplus database.
82. E. J. Corey and B. E. Roberts, *J. Am. Chem. Soc.*, 1997, **119**, 12425.
83. H. Miyaoka, Y. Kajiwara, and Y. Yamada, *Tetrahedron Lett.*, 2000, **41**, 911.
84. L. Korosec, P. A. Limacher, H. P. Lüthi, H. Peter, and M. P. Brändle, *Chimia*, 2010, **64**, 309.
85. For examples of chemical information instruction integrated into lab courses and lectures, see http://www.infochembio.ethz.ch/kurse.html (Web pages in German).
86. See, for example, A. H. Renear, and C. L. Palmer, *Science*, 2009, **325**, 828; P. Murray-Rust, H. S. Rzepa, S. M. Tyrrell, and Y. Zhang, *Org. Biomol. Chem.*, 2004, **2**, 3192; J. A. Townsend, S. E. Adams, C. A. Waudby, V. K. de Souza, J. M. Goodman, and P. Murray-Rust, *Org. Biomol. Chem.*, 2004, **2**, 3294.
87. This was suggested already in the early days of database searching by D. Rehm (Univ. Frankfurt/Main, Germany).