Online User Needs in Chemical Information

Peter Langer*

FIZ Chemie GmbH, Franklinstrasse 11, D-10587 Berlin, Germany

Arthur J. C. Wilson

St. John's College, Cambridge CB2 1TP, U.K.

Received December 16, 1993®

The user needs of chemical online databases expressed in the literature are reviewed in this article. Accuracy, comparability, completeness, consistency, and timeliness of the information are high on the list of desiderata of these users. Data from various sources should also be merged more easily. Users showed great concern for the price level and especially for the online pricing metrics. Furthermore, properties of compounds, including stereochemistry, three-dimensional structure information, economic information, toxicity, environmental behavior, *etc.*, must be made more searchable. User friendliness of the services is of high priority.

1. INTRODUCTION

The information needs of database users vary considerably and depend on the field of interest or on the specific problem to be solved. The primary subject of this paper concerns user needs of public online databases in the chemistry field. In the secondary literature, most of the primary literature consisting of reports of original research are processed and offered in printed or machine-readable form. The units of the primary literature are classified, summarized, indexed, and made accessible. Only rarely, however, is there an attempt to judge the quality of the information processed or to relate the units to each other. The latter function, plus assimilation and synthesis, are the province of the tertiary literature. In parallel with the secondary and primary printed sources, many bibliographic databases and some full-text databases have been established. Factual databases, and particularly numeric databases, have grown in number and importance in recent years.

2. GENERAL USER NEEDS

A preliminary discussion of general database user needs follows. Although several examples are related to chemical databases, this section encompasses a wider perspective. Pötzscher and Wilson[1] required that information should not be lost through failure of the transmission process from the primary literature to the secondary. Besides this completeness of coverage they designate accuracy, timeliness, costs, and cost effectiveness as general needs. Schreieck[2] also designates completeness of coverage and timeliness as information demands and added the demands of comparability and flexibility. Flexibility during data input, and then the flexible presentation of retrieval answers, are discussed. Besides the question of accuracy, Wales[3] considers whether consistency is paramount. Additionally, he demands better integration of the many sources of information.

2.1. Completeness of Coverage and Flexibility. Inflexible data input can lead to incompleteness. Zass[4] made this clear with the example of CAS ONLINE (Chemical Abstracts Service), which indexes only up to 10 authors and stores only the first address even if it is a joint publication of different working groups. He showed that 39 authors from the publication of Woodward *et al.* on the total synthesis of erythromycin were thus neglected. CAS cites *e.g.* only the first address and the database BIOSIS[5] only the second address from a joint publication on the factor F430 from methanogenic bacteria.[6] The experienced online searcher is aware of this deficiency and can identify the missing authors, locating the missing workplace by searching the Science Citation Index. This hindrance, naturally, does not permit efficient, user-friendly retrieval. The inexperienced online searcher and the end user are not aware of this deficiency. As a minimum, there should be a reference like "*etc.*" in the affected fields, suggesting that further available information (authors, addresses, and so on) had not been indexed. In addition to omissions caused by inflexibility there are also errors of omission which occur when a taken-for-granted data element (publication year, document type, language code, classification code) is absent from a record. Jacsó[7] laments that such omissions, also called blank fields, often result in the nonretrieval of relevant records and that most of the databases choose to remain silent about this type of error. His article on errors of omission also covers the techniques of "defensive searching" that will help the searcher compensate for these errors.

Besides the desire for completeness of data fields, we also have to consider whether databases could cover a specific science or branch completely, like organic chemistry. Respondents of a worldwide survey[8] on the information needs and use in organic chemistry did not associate criteria of completeness with individual sources. Instead, they had a tendency to consult all available sources even if sufficient and credible information had been obtained from an earlier source. A steady increase in the number of new compounds, new developments in information technologies, and the constant increase, online and print, in new information sources were detected as main factors that lead to the need for multiple sourcing. Because these factors (new information, new information technologies, and new information sources) have an effect on all sciences, users are willing to accept sources that are not necessarily complete, as long as they can be searched quickly and are reliable enough to contribute to the process of developing one's own picture.

The situation is different for reaction databases. ChemInform RX[9] as well as CASREACT do not intend to cover literature comprehensively. Their purpose is to reflect new synthetic methods in organic chemistry with focus on preparations and transformations not on compounds. Even

® Abstract published in *Advance ACS Abstracts,* May 15, 1994.

0095-2338/94/1634-0707$04.50/0

from the documents selected for coverage only significant example reactions are included into the databases.[10] Databases should be as comprehensive as possible, omitting irrelevant data, trivial data, and duplicate records.

2.2. Accuracy and Consistency. Physical properties and measured values, as well as bibliographic data, should be abstracted precisely.[1] Zass[4] claimed that, in 1991, the CA-File already had publications from the years 1992 to 1999! The database Beilstein[11] cited seveal publications from the years 1012 to 1141! Also in Beilstein, the journal *Angewandte Chemie* is not consistently abbreviated. Zass[4] stresses that such inconsistencies should be avoided by using a masterfile in the input program. Words should be written correctly[12] and consistently. Chemists have proposed[1] to implement an automated transliteration of different spellings in order to search inconsistently spelled words (colour *versus* color, sulfur *versus* sulphur) more easily. Data-Star[13] already introduced such a system called MEDWORD, which automatically accounts for variations in British and U.S. spellings. Besides the search terms "reduction" and "reductions", Zass[4] points out that the abbreviation "redn." is used in the CA-File of STN. He does not want to force the document analysts to avoid these inconsistencies but is repeating the demand of users[1,14] that a thesaurus should be available and filed with the related database. Further inconsistencies often arise from "improvements" like major changes to the index terminology, revised versions of thesauri, or other editorial enhancements.[15] This is because, in many cases, the database producers do not back-index their files or provide an online cross index to refer searchers to previously used terms. In thermodynamic databases much work has still to be done in order to provide consistent data and to standardize the existing data.[16] The Committee on Data for Science and Technology has already formed working groups which are occupied with the standardization and issuing of verified thermodynamic data. A need for standardization also exists for material databases. Westbrook[17] quotes among other things designation of materials, terminology, and test methods. More consistent data or a thesaurus would make the database content more manageable and understandable. Seals[18] stresses that the real answer to online searching seems to lie not in simplifying the protocol but in making database content more manageable and understandable to the end-user.

Inconsistencies within certain databases are criticized, but there is also inhomogeneity in the appearance of different databases. Lack of uniformity in the designation of fields and their order, the designation of the document type (*e.g.*, designations for books: book, book chapter, monograph chapter, monograph review), and the spelling of abbreviations and journal titles, authors, dates, language codes, and company names (Corporate Source) should be resolved.[14,15,19] This lack of standardization among database producers limits a searcher's ability to undertake global or multifile searching, to identify and remove duplicates, and to save search strategies and reexecute them in other databases.[20,21] Bernhart,[19] therefore, proposes obligatory standards, Mintz[15] recommends to use the registered ISSN spelling of a journal title as one of these obligatory standards, and Mintz as well as Norman[22] suggest the formation of an accrediting body for standards in the online industry. Basch,[14] however, writes that searchers are looking for more intelligence from the host CPU, including built-in equivalencies to take these variations into account.

2.3. Rapid Processing. Information must be quickly available.[1] Provisions should be made for rapid ordering and quick delivery of original documents or photocopies when full-text databases are not available. Users are calling for document delivery by Fax[23] from Beilstein.[11] STN has introduced electronic delivery of search results via STNmail.` Databases of library catalogues are also very useful for the localisation and quick ordering of primary literature. Electronic library catalogues are suggested by Russian scholars and scientists.[24] Such on-line catalogues are in fact available for many libraries, mostly academic, in at least 15 countries (Australia, Canada, Finland, Germany, Hong Kong, Ireland, Israel, Mexico, The Netherlands, New Zealand, Spain, Sweden, Switzerland, United Kingdom, United States of America). They are accessible by the "telnet" command. A service called "hytelnet",[25] with a simpler menu-driven version "libs", is installed by major libraries, enabling the user to discover which libraries have publicly accessible on-line catalogues and the procedure for signing on.

2.4. Comparability. In principle, numerical data should be stored in SI units, or, if stored in familiar but non-SI units, the software should offer the option of retrieval in SI units. Information relevant to the data, such as measuring conditions (temperature *etc.*) and the measuring method, should also be stored so that the user can judge whether data are comparable or not. Further details on data are also helpful: data type (experimental, extrapolated, interpolated, and smoothed values, relative values, *etc.*), data field (maximum and minimum value of valid field *etc.*), data origin (diagram, equation, table, *etc.*) and data grade (recommended value, best value, *etc.*).

2.5. Integration. Since none of the files is comprehensive, it is necessary to perform many searches. It is difficult, however, to search several files on different hosts (different retrieval languages *etc.*). For example, the Spanish information market is dominated by a large number of hosts, each offering just a few self-produced databases. Therefore, some users in Spain want integrated information services in the form of one-stop shops. They are advocating the establishment of a national Spanish host which offers all these databases from just one source.[26] It is also difficult to search several files on the same host. Therefore, some online hosts, led by ESA[27] and later Dialog and STN International, have provided a multifile environment where the customer retrieves information units or pools and not single databases, thus extending the simple crossover capability. Linkage of the individual databases is also desired by users of the Beilstein[11] database, who are asking for cross references to other databases.[23] Since 1993 the Beilstein database contains such cross references[28] to chemical supplier catalogs, EINECS,[29] and the spectra database SPECINFO.[30] Other important integrating features are the "master indexes" (Dialog's Dialindex, Data-Star's CROS, etc.), and the Dialog Finder Files for journal and company names. Seals[18] considers CD-ROM, at least as it exists today, as "retrogressive in that it permits only one search at a time in only one database, and does not permit crossover and coordination of searches across multiple files". In the patent information field[3] the demand for a better integration of the many sources of information goes beyond this multifile searching capability. Users want the ability to receive information electronically, interpret it, annotate it, merge it with other internally generated sources and initiate action. The complex problems of chemistry also demand the simultaneous analysis of different information sources.[31] Barth[32] thinks that the major aspects of an integrated information system are (1) linkage of the individual databases; (2) integration of additional application software packages; (3) connection of various information systems; and (4) support by decentralized user software. STN International has begun to accommodate users' requests. They offer, for example, calculation programs in the spectra database SPECINFO.[30] Schreieck[2] desires a comprehensive collection of calculation programs and simulators which can cooperate and reflect the properties and behaviors of the actual objects.

2.6. Costs and Prices. Users showed great concern for the price level and especially for the online pricing metrics.

Retrieved information should be worth its cost. Provision must be made to provide "poor users" with information services at a lower price.[1] Pötzscher and Wilson[1] classified universities and lesser developed nations as such. East European users[33,34] are demanding that hosts should introduce some form of complementary pricing for their countries to stimulate their local information market. One user[33] is pleased to find STN International, already offering east European users an 80% discount on several databases, including the Chemical Abstracts file. Other hosts should introduce similar initiatives.[33] Some users are faced with travelling abroad to attend courses;[33–36] the costs for participants is a further disincentive for making use of online services. Besides local training, another user[34] states her desire for local help desks to alleviate the high costs of international calls. She adds: "Expense is not the only barrier to seeking help—there is also a psychological one. Not everybody can telephone a foreigner in a foreign city like London or Palo Alto with a problem, or a simple question." Basch[20] asks for a toll-free customer service line that should be staffed by a knowledgeable, informed, and resourceful person whenever the system itself is up, or supposed to be. This is necessary because many users do their search during the evening or on weekends and often are faced with problems that cannot wait until the next regular business day.

A number of users[37–40] voiced support for the policy of low or even no connect charges, coupled with a higher price for each document. One user reasons as follows: "Connect-time pricing has kept online searching out of thousands of libraries, created the tension-filled "running meter" syndrome where it is used, encouraged the growth of CD-ROM and locally-mounted tapes, and deterred vast numbers of end-users from accessing online information."[40] Searchers[14,19] feel strongly that costs should accrue only when information is actually being retrieved and manipulated; the online clock should stop when the user is browsing, thinking, constructing a strategy, looking at help screens *etc.* A partial solution was the introduction of commands such as Dialog's PAUSE, Data-Star's PARK etc. which hold your search and keep the user logged on while replacing your database's connect charge with a small parking charge.[42] Some users[20,40,43] prefer either free use or at least lower charges for browse formats, such as keywords and titles, that do not display crucial data elements (information that would enable the user to track down a reference with no further help from the database). These factors are essential in creating an efficient search strategy and for refining what has already been obtained. Low connect charges coupled with search-term pricing is not favored by users.[38,44,45] Charging for specifying a single data element, in a custom or user-defined format, in a way that triggers the same high display charge as the fixed format from which that element was drawn is also criticized.[20] An example is asking for a publication data only and being charged for the full bibliographic citation. An online display cost greater than the offline print charge and surcharges for higher baud rates and for multifile searching are also criticized.[20] A member of the Portuguese Online Users Group considered the discomfort users have felt in not knowing the cost of a search until it is complete as a great barrier to the use of online services.[46] In order to better control costs, an on-screen, dynamic cost display[19,20] was proposed, but this is only a partial solution. The complete solution to this problem seems to be the flat-rate pricing options which the major databanks offer or are developing now.[41] Flat rate refers to unlimited use of a database or set of databases for a fixed period and are individually negotiated. These flat rates are most suitable in high-volume-use cases and were welcomed enthusiastically by such customers,[41] but are impractical for less used databanks. Occasional users would like to eliminate monthly minimums, maintenance fees, and other barriers to casual use of a service.[14,37,39,40] Some occasional customers of Genios[47] suspended their contracts when the *Handelsblatt*-owned host introduced a monthly minimum charge.[37] In a survey conducted for the European Community,[48] many respondents expressed their desire for simplified invoicing (79 very important, 102 important, 87 not important).

The needs and requirements stated above are at times contradictory and some are impracticable. Judging from the range and depth of concern of users, however, there seems to be a clear need to work toward improving the overall situation.

3. REQUIREMENTS AND IMPROVEMENTS FOR CHEMICAL INFORMATION

The following requirements and improvements (not necessarily in order of importance) have been proposed:

(a) After a CAS Registry Number has been assigned, it should be quoted in all later publications.[1] This number or other relevant numbering systems should be given by all services processing chemical information.[1] The authors[49] of a comparison of data sources for environmental chemicals noticed that searching by CAS Number is only possible in 58% of the CD-ROMs chosen, in around 40% of the online databases, and in less than 30% of the Manual Sources. There is a demand for the Registry file, the source of these CAS Registry Numbers, to be available on hosts other than STN.[50]

(b) The encoding of chemical structures in graphical form should be standardized unambiguously.[1,2] Stereochemistry must be included.[1] Biochemists, especially, need information about three-dimensional (3D) molecular structures in order to develop drugs efficiently.[51] The bond energy should be stored with these 3D structures.[31]

(c) Spectra and reactions should be standardized.[31] Reaction data bases should be supplemented with kinetic data.[31] Provisions for searching should be improved, so that consecutive reactions can be retrieved even if distributed throughout various literature sources.[1]

(d) A database with the results of quantum mechanical calculations should be built.[31]

(e) Safety, environmental behavior, toxicity, *etc.* should always be searchable, even if they are not the main objective of the primary source. The economic damage that may result from lack of information on such matters can be much greater than the cost of providing the information.[1]

(f) In the past, most emphasis was placed on purely chemical information. A representative of the chemical industry now observes an increasing demand on ecological and business information.[52] In a statistical study of on-line searches in a chemical company, Smith[53] detected a definite increase in the field of chemical business. Inquiries for chemistry and patents, though, remained fairly stable. Searches in the area of economic and trade data in the ICI (Imperial Chemical Industries PLC) are increasing more than searches for technical information.[54]

Some of the above points are discussed further in other paragraphs of this text.

4. SEARCH REQUIREMENTS

There has been remarkable progress in searching online files. STN has introduced sorting of retrieval answers. Because of this, presentation of search results is more flexible. Multifile searching and a capability for detection and elimination of duplicates have also been provided by most major online hosts. Users do still experience difficulties searching and integrating data from several hosts and in-house databases. One reason for this is the conspicuous lack of (and little obvious attempt at) standardization of commands and protocols among search systems.[55] Users[1,14] have therefore proposed that a single command language should be introduced for all publicly accessible hosts and for in-house systems. The International Organisation for Standardization (ISO) gave a proposal for a standard command language (ISO/DIS, 1988). This so-called Common Command Language (CCL) had already been introduced *e.g.* on STN.[16] Users of STN can now carry out their searches by CCL or by the retrieval language of the host, Messenger. CCL, however, is only the "greatest common factor" and does not cover all the features of the specific host languages. In a survey for the European Community,[48] 156 respondents considered a single command language to search several hosts on intelligent gateways as very important, 87 respondents considered it as important, and only 39 as not important. In the same survey[48] many respondents also expressed their desire for a single log-on procedure to access several hosts (101 very important, 110 important, 65 not important) and an automated selection of relevant hosts on intelligent gateways (85 very important, 73 important, 100 not important). Basch[14] suggests making selection of a database a transparent process, driven by the nature of the request at hand. The selection process might even cross system lines, proceeding by gateways to one database or another, depending on the criteria previously specified by the searcher. Users favor mnemonic filename abbreviations rather than numbers.[55] Some non-English mother tongue users[46,56,57] did not perceive the use of online services in English to be a problem. Nevertheless, English speakers and non-English speakers requested a dictionary in the database and descriptors as well as keywords in their mother tongue, thus making the database at least partially bilingual.[19,45] Documentation should be detailed, current, accurate, and reasonably priced.[20] Lack of adequate documentation for many databases was a common complaint.[45] In many countries[46,58] telecommunications are unreliable and/or expensive and therefore more and more information users are turning to CD-ROM.

According to Schreieck,[2] users are often forced to "shovel" the obtained data manually between several information islands. Improved postprocessing capabilities from major hosts, such as the ability to load results into a word-processing package without having to remove and break-up fields, line fields, and labels is desired.[36] Users are demanding further improvements of the retrieval software. They desire a more intelligent[4,31] or more "user-friendly"[3,23] software. This is an often abused term that has been taken to mean a combination of techniques like windowing, consistent menu design, good "help" information, transparent transfer between systems, *etc.*[3] Schreieck asks for a self-explanatory user interface.[2] Retrieval functions of the current online systems are based on (inverted) index files with Boolean logic. This technique is very useful for searches of simple text terms in bibliographic databases, but for more complex data, as in spectroscopic databases, Barth[32] suggests probabilistic retrieval functions or fuzzy logic approaches. He also discusses a new concept for information retrieval which is based on the idea of similarity and dissimilarity. Such retrieval may be applied to isomeric or tautomeric substances and also allows relating two terms which are not identical but which belong to the same family, *i.e.*, two terms whose features are closely related.[32] The problem of tautomerism is discussed further in the following paragraph.

4.1. Structure Search, Graphics Management, and Hypermedia. There has been remarkable progress in chemical structure search. Without having to pay for connection time, it is now possible, with front-end software such as STN Express, to draw structures on the PC and to load them into the database. Users are not completely satisfied with the available front-end software. Instead, they want to build their own front ends in a tool-kit approach using various modules all with a common look and feel.[59] For the draw module the user would choose his or her favorite structure drawing package. Further modules might be chemical name entry for conversion to chemical structures, and optical character recognition (OCR) of chemical structures, but the user still has to deal with problems of tautomerism and aromaticity during structure input. For this reason, Zass[4] concludes that these problems should be dealt with within the database by the host computer.

Structure search on host computers, however, is not ideal because it requires too much time. On STN, *e.g.*, sometimes simple queries can take nearly 15 min before the system finally admits that the answer set is incomplete.[60] Thus the costs are high and are not predictable in advance. The crossfire (XFIRE)[60] concept greatly reduces these problems by bringing the structure file in-house at a one-off leasing cost in a client–server architecture. All structure searching is then free of charge. Online sessions are only started when (and if) the searcher has targeted the hits, whereby further communication with the online host for property data is carried out as required with the cheap and fast use of the Beilstein Registry Number and a graphic window. The XFIRE system is not limited to the Beilstein file and is designed to be able to cope with other megafiles such as the CAS file.

Besides chemical structures there is also graphical information like spectra and patent drawings transmitted by STN. The graphical images that can be handled are rather small in size due to the speed of the telecommunication lines. A speed of at least 9600 baud is necessary for receiving graphical images within an acceptable time frame.[1,20,32] Ideally the spectroscopist would like a large, high-quality data set containing multidimensional spectral database (*i.e.*, one containing more than one type of spectrum, *e.g.*, NMR, mass spectra (MS) and IR); integration of his preferred chemical structure handling system; a facility to add his own data to the library and a link to his sample management software; networking of all the independent spectroscopic facilities in the company; and, ultimately, an expert system for structure prediction. These requirements are in approximately descending order of priority and were compiled by Warr.[61] She adds: "No one system yet meets the majority of the user needs". Prism and grating infrared instruments have a lower resolution than the modern Fourier transform infrared instruments. As Lias[62] has noted, many spectroscopists believe that such older data are not adequate for good reference databases. There are remarkable differences in the quality of MS data. The NIST scientists, *e.g.*, found[63] that almost 42% of the spectra unique to the Wiley database[64] have fewer than 10 peaks, whereas this is true in 3.5% of the NIST spectra.

Finally, it should be noted that the current focus lies only on the display of graphics and not on the possibility for performing graphical searches similar to the 3-D searching of biomolecules.[32] This aspect of information retrieval is still a subject of research and an implementation in a production environment is not yet possible.[32] Because of this, biochemists especially need to be patient. Schreieck[2] visualizes electronic books and Barth is already noting a tendency to make large handbooks (Landot-Börnstein, Ullmann's *Encyclopedia of Industrial Chemistry*, *etc.*) available as online files. A page in such a handbook may contain all types of data, *e.g.* images or photos embedded in a full-text environment. In adition there may be tables of these different data entities. It is clear that the concept of a line-oriented dialogue mode which is still the standard in the online world is no longer adequate for supporting these complex information units.[32] A hypermedia environment is required to deal with these types of entities. Not a hypermedia, but a hypertext, facility has already been introduced by ESA-IRS, in order to help users browse the thesauri of major databases.[42] The online host OCLC[65] offers a hypertext user interface called GUIDON which provides easy and convenient access to their online databases.

4.2. Searching Bibliographic Databases. Searching bibliographic databases is essentially based on the indexing and searching of words and phrases using Boolean operators, proximities, and truncation symbols. If several search terms and their synonyms have to be combined by Boolean operators, the user is faced with many combinations and permutations of different depths and relevances. Therefore, a more intelligent retrieval system is demanded,[4] which is capable of conducting these combinations and permutations automatically. Automatic pluralization (with an "on-off" switch) and left-hand truncation are further requirements.[15]

The Boolean operators, proximities, and truncations are sufficient to handle most searches. Barth[32] indicates that in some textual databases the text terms do not correspond to the natural meaning of the words, *e.g.*, the codes in sequence databases, and that a formal comparison of characters does not meet the needs of the specialists. He indicates that there are also cases where a string-search capability is more appropriate than an index search as used in the online retrieval system. Basch[14] also points to alternatives to Boolean logic like the software DowQuest, which processes natural-language queries by means of a relevance-feedback algorithm, and the hierarchical text-retrieval software packages, like Persoft's IZE and Topic by Verity. As mentioned previously, Zass[4] calls for a hierarchical thesaurus. He explains that, in addition to the consistent use of search terms (*e.g.*, reduction), the user needs a means of finding subtopics (*e.g.*, "Birch reduction", "reduction, electrochemical"). Zass noticed that users of online databases often do not use printed thesauri, such as the index guide of Chemical Abstracts. He emphasizes, therefore, that the thesauri should be filed with the related database. Other users also favor online thesauri.[55,42] Zass indicates that the Chemical Abstracts implementation on Dialog already contains an important part of this guide. A superthesaurus has also been suggested[19] in order to search different databases of the same branch more easily.

If a user searches the CA file for a CAS Registry Number appended with a P, he should get references dealing with the preparation of a certain substance. Zass[4] indicates that in many references the substance is only formed in small quantities and a chemist would not call this a preparation. In addition, several hosts use a different "P-algorithm". Zass[4] calls on data base suppliers to develop a better algorithm jointly. He suggests that there should be a document analyst at CAS to decide which documents are preparative ones.

The CA file is available via several hosts (Dialog, STN, *etc.*). The abstracts for this file, however, are offered only via STN, so that only users of STN can take the advantage of this additional data source. CAS is reluctant to offer abstracts on other hosts. Dialog is determined to offer these abstracts to their users and has begun a lawsuit with CAS. Warr[50] writes on this subject: "First and foremost of the user wishes is an end to the costly lawsuit between CAS and Dialog, which benefits no-one but the lawyers". In the meantime a mutual agreement has been reached on this lawsuit—whatever that will mean for the availability of the abstracts. It is not only the CA file that needs abstracts. Given a choice between an index without abstracts and one with abstracts, Tenopir and Jacsó suspect most users would pick the abstract file every time.[66] However, the abstracts must be of adequate quality.

According to the American National Standards Institute (ANSI) an abstract is an abbreviated, accurate representation of the contents of a document.[66] ANSI also provides guidelines for writing style. Abstractors should as far as possible observe the ANSI standard in order to get readability and consistency of style. A database that relies only on author-written abstracts will lack consistency in length and style of abstracts.

4.3. Searching Full-text Files. Searches in full-text databases are often less precise than in bibliographic databases. In the opinion of Barth[32] this is due to the presence of free-text vocabulary. Barth[32] states that a context-specific retrieval with the possibility for a semantic interpretation would be of greater help for the user. As a first step in this direction, he says, would be a thesaurus which allows the user to improve his search query by automatically including corresponding terms, *e.g.* synonyms and related terms. But there is also another opinion:[67] the problem is the absence of descriptors/standardized vocabulary, not the presence of free text (by volume the dominating text element in the CA file and other secondary sources). That implies that these descriptors have to be supplemented.

4.4. Searching Factual Databases. The results of a search of a factual database can now presented in a more flexible fashion than in the past. Capability of automatic conversion of physical units is provided by STN, which has also developed other features for the particular support of physical property data. These features include numeric-range overlap detection, table management, and an interface for the access to numeric tools.[32]

Chemists need a lot of properties for every compound. For example, in the database Beilstein[11] there are about 400 numerical and text data fields. Ideally, for every compound all data fields are filled. But what is the reality? The data matrix is almost empty. The molecular weight is known for every compound, but beyond that the knowledge is extremely fragmentary. The "data-holes" can be filled by experimental measurements, which, however, often necessitate great pains and are expensive. Gasteiger,[31] therefore, recommends that such missing data should be predicted by induction using statistical models or new learning methods like neural networks or machine learning. He emphasizes the importance of neural networks in chemistry by referring to the rapidly growing number of publications about the application in this field. Most users,[67] however, do not wish to add nonexperimental data to factual databases. There are several reasons for this argument, *e.g.*, model dependency and impreciseness of the values. However, this wish is only understandable if the user cannot exclude or include certain data types (*e.g.* experimental or nonexperimental; see section 2.4) because corresponding information was not stored. For example, the polymer material data base POLYMAT PC[68] provides experimental data as well as estimated values. This is to avoid the exclusion of a relevant product in a search only because measured data was not available. In POLYMAT PC, as a matter of fact, searches can be conducted only on the basis of measured values.

5. SUMMARY AND DISCUSSION

The achievements of modern science and technology have greatly improved our daily lives and have enabled us to collaborate without concern for the temporal and social barriers of the world. Today, all human activities, including scientific and technical information activities, are rapidly moving toward globalization. The user needs outlined above share in this drive toward globalization. Users want to monitor information as *widely* and as *quickly* as possible in order to obtain relevant accurate and comprehensive information. Their preferred method of charging is either by a flat rate or a cost per document known in advance. Necessary prerequisites for meeting user needs are reliable and user-friendly telecommunication systems.

Potential users are often barred from participation in these information activities for financial reasons. Such restriction is frequent in developing countries, eastern Europe, and many universities. Comprehensive searches involve the use of several databases and perhaps several hosts, though the bases available on STN satisfy the needs of most chemists. Hosts tend to ignore national borders and to expect users to understand instructions and output in English; the latter problem would be eased by the development of a single log-on procedure and a single command structure.

Retrieval of accurate data quickly and comprehensively requires as a minimum comparable, consistent, correct, and adequately cross-referenced information; online thesauri; registry numbers; standardized databases; and telecommunication lines with a speed of at least 9600 baud. Copies of the original documents should be obtainable rapidly and at reasonable cost; preferably it should be possible to place an order while the database information is still on-screen.

Many of the above user needs can be satisfied only through the cooperation of several institutions—the database producers, the hosts, *etc.*—and progress will necessarily be slow. On the other hand, there are problems that could be rectified quickly by a single institution. For example, the word "rearrangement" is misspelled nearly 500 times in the CA file on STN. One might think that this is negligible, since it affects less than 1% of the occurrences of the word, but every user searching the literature about rearrangement has to deal with more than 60 different spellings. This is labor-intensive and time-consuming, and every user has to pay for it in connect time and other charges. This needless work and cost should be prevented; the database producer should correct these and other spelling mistakes and thus aid retrieval. It is an anachronism that a database should have many spelling mistakes in an age when word processors routinely incorporate spelling checkers. Other mistakes, inconsistencies, blank fields, *etc.* are easily detected and rectified during input; this is far better than "bolt-on" packages designed to improve the output from imperfect databases; MEDWORD makes some attempt at this by accounting automatically for differences between UK and US spelling. In summary it may be said that all easily computer-detectable mistakes and inconsistencies should be given priority treatment.

ACKNOWLEDGMENT

The authors thank the members of the International Council for Scientific and Technical Information (ICSTI), Chemical Abstracts Service (CAS), Columbus, OH, Institut de l'Information Scientifique et Technique (INIST), Nancy, France, and Japan Information Center of Science and Technology (JICST), Tokyo, Japan, for information on this topic derived from their databases as well as the members and particularly the chairman Dr. Weiske of the ICSTI Group in Chemistry for their worthwhile contributions.

REFERENCES AND NOTES

(1) Pötzscher, G.; Wilson, A. J. C. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 169–173.

(2) Schreieck, A. M. *Mitteilungsbl.-Ges. Dtsch. Chem., Fachgruppe Chem.-Inf.-Comput.* **1991**, No. 19, 47–55.

(3) Wales, J. L. *Chem. Inf.* **1990**, *2*, 81–8.

(4) Zass, E. *Mitteilungsbl.-Ges. Dtsch. Chem., Fachgruppe Chem.-Inf.-Comput.* **1993**, No. 27, 26–43.

(5) BIOSIS is a bibliographic database which covers the worldwide literature on all life science topics, including toxicology, public health, and ecology. The database is available on STN International and several other hosts. Producer: BIOSIS, 2100 Arch Street, Philadelphia, PA 19103-1399.

(6) Zass, E. *10 Jahr FIZ CHEM-Grussworte Vorträge* **1992**, 65–80.

(7) Jacsó, P. *Database* **1993**, *16*, 38–49.

(8) Jochum, C.; Moricz, P. *Database* **1987**, *10*(4), 41–46.

(9) ChemInform RX is available on STN International. Producer: FIZ CHEMIE.

(10) ChemInform RX and CASREACT selection criteria.

(11) BEILSTEIN is a major structure and factual database in organic chemistry. The database is available on STN International and Dialog. Producer: Beilstein Institute für Organische Chemie, Varrentrappstr. 40–42, 60486 Frankfurt/Main, F. R. Germany.

(12) Dolan, D. R. *Online* **1992**, *16*(2), 30–35.

(13) Data-Star Marketing Ltd.; Plaza Suite, 114 Jermyn Street, London SW1Y 6HJ, U.K.

(14) Basch, R. *Int. Online Inf. Meet. Proc. 14th* **1990**, 251–259.

(15) Mintz, A. P. *Online* **1990**, *14*(6), 15–23.

(16) Barth, A. *Datenbanken Naturwiss.* **1992**, 65–69.

(17) Westbrook, J. H. *Mater. Inf. Eur. Communities* **1990**, 25–31.

(18) Seals, J. V. *Mitteilungsbl.-Ges. Dtsch. Chem., Fachgruppe Chem.-Inf.-Comput.* **1991**, No. 18, 9–19.

(19) Bernhart, I. *Qualität von Informationsdiensten* **1993**, 135–146.

(20) Basch, R. *Online* **1992**, *16*(4), 22–25.

(21) Isler, N. *Eur. Inf. Users Pt. 1 Germany and Central Europe* **1993**, 42–43.

(22) Norman, S. *Inf. World Rev.* **1991**, No. 63, 21.

(23) Bucher, R. *Cogito* **1993**, No. 1, 8–15.

(24) Zhakarova, L. A. *Nauch. Tekh. Inf.* **1992**, Ser. 1, No. 2, 17–18.

(25) HYTELNET is a software for accessing the Internet which is a network of computer networks. Internet is not only a source for library information, but also a new source for secondary information, *e.g.*, the Buckyball Database (University of Arizona), Material Safety Data Sheets, *etc.* The following literature informs about HYTELNET and Internet: Scott, P. *Electronic Networking: Research, Applications and Policy* **1992**, *2*(1), 38–44. Krol, E. *The Whole Internet: Users Guide & Catalogue*; O'Reilly & Associates, Inc.: Sebastopol, CA, 1992.

(26) Baiget, T. *Inf. World Rev.* **1992**, No. 76, 34.

(27) ESA-IRS, European Space Agency Information Retrieval Service, Via Galileo Galilei, I-00044 Frascati, Italy.

(28) Weber, C. *Chem. Ind.* **1993**, *10*, 36–37.

(29) European Inventory of Existing Commercial Chemical Substances.

(30) The database SPECINFO contains NMR, IR, and mass spectra and is produced and supplied by Chemical Concepts GmbH, P.O. Box 100202, 69442 Weinheim, F. R. Germany.

(31) Gasteiger, J. *Mitteilungsbl.-Ges. Dtsch. Chem., Fachgruppe Chem.-Inf.-Comput.* **1993**, No. 27, 4–25.

(32) Barth, A. *Squaring Inf. Circle, ICSTI Symp. Proc.* **1991**, 39–51.

(33) Dologova, M. *Eur. Inf. Users Pt. 1 Germany and Central Europe* **1993**, 17–18.

(34) Sarkasian, A. *Eur. Inf. Users Pt. 1 Germany and Central Europe* **1993**, 19–20.

(35) Schmeikal, B. *Eur. Inf. Users Pt. 1 Germany and Central Europe* **1993**, 8–9.

(36) Cazan, C. *Eur. Inf. Users Pt. 1 Germany and Central Europe* **1993**, 6–7.

(37) Behrensen *Eur. Inf. Users Pt. 1 Germany and Central Europe* **1993**, 24–25.

(38) Hauffe, H. *Eur. Inf. Users Pt. 1 Germany and Central Europe* **1993**, 12–13.

(39) Puchmueller. *Eur. Inf. Users Pt. 1 Germany and Central Europe* **1993**, 28–29.

(40) Behrensen, J. *Password* **1993**, No. 7, 4–5.

(41) O'Leary, M. *Online* **1993**, *17*(1), 34–38.

(42) Basch, R. *Online* **1991**, *15*(6), 42–47.

(43) Personal message of an user to the author (P.L.).

(44) Kaschnitz, R. *Eur. Inf. Users Pt. 1 Germany and Central Europe* **1993**, 10–11.

(45) Dueltgen, R. R. *Database* **1990**, *13*(6), 103–104.

(46) Lopes da Silva, G. *Inf. World Rev.* **1992**, No. 76, 35.

(47) A German business host. Address: GENIOS Wirtschaftsdatenbanken, Kasernenstr. 67, 40213 Düsseldorf, FRG.

(48) BIS Mackintosh Limited. *Electronic Information Users in Europe Fourth Survey Report* **1989**, 18–19.

(49) Voigt, K.; Benz, J.; Pepping, T. *Int. Online Inf. Meet. Proc. 14th* **1990**, 261–268.

(50) Warr, W. A. *Inf. World Rev.* **1992**, No. 74, 17–18.

(51) Borman, S. *Chem. Eng. News* **1992**, *70*(32), 18–26.

(52) Donner, W. T. *Nachr. Chem. Techn. Lab.* **1993**, *41* (1), 21–24.

(53) Smith, E. D. *J. Inf. Sci.* **1990**, *17*, 119–125.

(54) Weeks, C. *J. Inf. Sci.* **1986**, *12*, 283–287.

(55) Shuman, B. A. *Online* **1992**, *16*(2), 54–59.

(56) Flaminia Ramos, M. *Inf. World Rev.* **1992**, No. 76, 36.

(57) Bassit, A. A. *Inf. World Rev.* **1990**, *No.* 54, 8–9.

(58) de la Viesca, R. *Inf. World Rev.* **1992**, *No*, 76, 34–37.

(59) Warr, W. A.; Wilkins, M. P. *Online* **1992**, *16*(1), 48–55.

(60) Lawson, A. J.; Swienty-Busch, J. *Int. Online Inf. Meet. Proc. 17th* **1993**, 187–194.

(61) Warr, W. A. *Chemometrics and Intelligent Laboratory Systems* **1991**, *10*, 279–292.

(62) Lias, S. G. *J. Res. Natl. Inst. Stand. Technol.* **1989**, *94*(1), 25–35.

(63) Heller, St. R. *Chem. Int.* **1991**, *13*, 235–238.

(64) John Wiley & Sons, 605 Third Avenue, New York, NY 10158.

(65) OCLC Online Computer Library Center, Inc., 6565 Frantz Road, Dublin, OH 43017.

(66) Tenopir, C.; Jacsó, P. *Online* **1993**, *17*(3), 44–55.

(67) Personal message of a reviewer of this article to the authors.

(68) The in-house database POLYMAT PC is available from FIZ CHEMIE. The in-house version is an excerpt from the extensive online database POLYMAT (STN International).