Computer Software Reviews

Beilstein Current Facts in Chemistry on CD-ROM. Version 1.0. Distributed by Springer-Verlag New York Inc.: P.O. Box 2485, Secaucus, NJ 07096-2491, (800)-777-4673. List price $490.00 for personal use, $1490.00 academic rate, and $2990.00 regular institutional rate. Academic and institutional rates are discounted with a *Beilstein Handbook* subscription.

The Beilstein Institute is developing a family of chemical information products called the Beilstein Information System (BIS), which will consist of the printed Handbooks, the online Beilstein database (now available on three online vendors), Current Facts on CD-ROM, and related tools and utilities. A major factor in the implementation of the BIS is a new computerized production system, in which all data from the literature will be indexed from the beginning of the production process, rather than with the multistep indexing procedures used previously. This new production system will allow Beilstein to publish current information, with an anticipated processing delay of 5 months, that is edited, sorted, and indexed. Beilstein's strategy will be to issue the most recent data on the BIS on a quarterly basis on the Current Facts CD-ROM. The data on the CD-ROM will later be integrated into the online database and eventually will be published in the printed Handbooks. Beilstein has decided to disseminate the latest data on CD-ROM rather than orchestrate frequent updates to its large online file. Also, the facility of the S4 structure search system to use a large CD-ROM file, and operate efficiently in a PC environment, makes this mode of delivery practical. When the BIS is fully implemented in 1995, the online Beilstein file, which presently only contains data up to 1979, will essentially be made current, less the most recent data on the Current Facts CD-ROM. It is anticipated the online file will be updated only annually or semiannually, so that Current Facts will have 6–12 months of exclusive information not available in the online database. In addition, the new electronic production system should accelerate the preparation of the printed handbooks.

Each quarterly update will contain cumulative data from the previous four quarters and will consist of about 300 000 records from more than 80 key journals received annually at the Beilstein Institute. Data from patents will not be included in Current Facts. The record format is similar to the online format. Each record contains a unique Beilstein Registry Number (BRN), a graphically searchable structure, and text and numeric searchable fields. This intimate association of the searchable structure and factual data is the most powerful feature of the Beilstein database. Unlike bibliographic files, such as Chemical Abstracts, locating a structure automatically provides instant access to a large hierarchy of sorted and edited information. Each fact also has a literature reference, and citations are also provided to the printed Handbooks.

Current Facts requires an IBM or compatible personal computer with MS-DOS 3.1 or higher, 640 K of RAM, a graphics card (all major types are supported), 10 MB of hard disk space, a CD-ROM drive with MSCEDEX 2.10 or higher, and a mouse. The program requires at least 500 K of free memory after the drivers are loaded. A variety of printers are supported, including HP LaserJet, DeskJet, and Plotter, as well as PostScript (registered trademark of Adobe Systems, Inc.) and dot matrix printers. The software is being upgraded and will be released in the Windows (registered trademark of Microsoft Corp.) environment in 1992, which should provide better integration of the various modules in Current Facts. A Macintosh (registered trademark of Apple Computer, Inc.) version is under development, but no release date has been given.

The documentation consists of two books and a short Quick Start Guide. The User Guide provides details on using the query masks and OptiSearch software, the Molkick structure editor, the S4 structure search system, and answer display modes. The Reference Guide contains a cogent explanation of Beilstein's philosophy, content, and indexing policies. A description of each field in the database consumes most of the book. All the documentation is clear, complete, and well illustrated.

Current Facts consists of essentially four modules at present: the display mode/main screen, the factual search mode using software from OptiSearch by Lasec of Berlin, the structure searching mode using S4 software produced by Softron of Munich, and the Molkick structure editor for generating structural queries. When fully implemented under Windows, the Beilstein Information System will be an integrated package consisting of the Current Facts CD-ROM database, Optisearch, S4, a structure editor, Autonom, Sandra, and a new communications package. Beilstein will probably release a new structure editor for Windows to replace Molkick. Autonom generates an IUPAC compatible chemical name from a structure; Autonom is currently available for DOS computers with the Molkick structure editor. Sandra generates System Numbers and Lawson Numbers (LN) from the structure editor. Lawson Numbers are an expanded form of the System Number for similarity searching in the electronic Beilstein databases. The similarity searching feature is an exciting tool designed to retrieve positional isomers which are not readily retrievable with a substructure search. The LN is a generally underutilized feature of Beilstein, because proper documentation has not been published, and because the updated Sandra, which will generate LN's, has not yet been released. The communications module will allow queries created on Current Facts to be uploaded to an online vendor for a search against the full online file. This will give the user the ability to search the chemical literature published from 1830 to the previous 5 months with a single query, albeit on two databases, the CD-ROM and the online file.

Factual Search Mode: The OptiSearch module presents a "Query Input Mask" which provides a simple structured interface for entering search terms, data fields, and the Boolean operators AND, OR, and WITHOUT (usually called not). Two additional masks are provided: the "Identification Mask" for locating information on specific compounds, such as chemical name or molecular formula, and the "Citation Mask" for bibliographic searching. A powerful LIST command is provided for certain fields, such as journal title, author name, and use data. This command displays a listing of all the entries in the field. An entry may be selected with the mouse for inclusion in the search strategy. For the field names, the LIST command has a special mode that shows the hierarchy of searchable fields, which is very useful for visualizing the logical relationships among the fields. Optisearch allows range searching on numerical data, such as a range of melting points. Answer sets from prior factual and structural searches can be combined with each other and additional search terms in the Query Input Mask.

Structure Search Mode: In Current Facts, the S4 structure search system is fully implemented, with all search capabilities, including tautomeric, stereogenic, and extensive generic search facilities. For generic searching, a group of four predefined atom lists is available for the common substitutions of halogen, any atom, any atom except C or H, and any metal. Additionally, atom lists may be defined by the user to include any atom in the periodic table. Atom lists may also be negated, which is a unique feature to Current Facts. A hierarchy of generic groups, similar to the superatoms in Markush DARC, is provided with 14 groups in all. For example, there is an Acyclic generic group (ACY), and a subdivision of that is Carbacyclic (ABC), and a subdivision of that is Alkyl (ALK). Finally, there is the "Markush" search capability, in which a list of specific fragments is drawn that are allowed as substituents at a node in the main structure indicated by Gi, where i can be 1–99. The Markush feature is familiar to STN users, where it is called the "generic group" or Gk feature, although the implementation is more like DARC, with separate screens for each Markush group. Unlike STN, the Markush groups may be nested. There is no stated nesting limit.

Tautomers are retrieved by S4, and are not an artifact of an indexing policy rule. There is a special tautomeric search mode that must be switched on in the S4 options menu. A tautomeric structure may be indexed under one tautomeric form only, or as several different tautomeric forms, each with its own BRN. During a tautomeric search, the S4 software will retrieve all tautomeric forms that conform to the query. If the tautomeric search is switched off, only tautomers indexed exactly as drawn in the query will be retrieved. Because the search time is approximately doubled with tautomeric searching switched on, it is best to use the tautomeric search mode only when necessary.

One of the most sophisticated features of S4 is the stereo search capability. As with tautomeric searching, this is a special search mode that must be switched on, but unlike tautomeric searching, stereo searches are accomplished more quickly. Stereogenic centers are defined in the query with the use of wedge (stereo above), dashed (stereo below), and stereo double bond types, not by directly specifying *R,S* nomenclature. The system also distinguishes *E,Z* isomers. There are two levels to stereo searching. Level 1 retrieves records with the exact stereo structure drawn in the query, i.e., the same enantiomer only, while level 2 retrieves the exact structure and the mirror image, i.e., the same diastereomer, with both enantiomers plus the racemate. With stereo

searching switched off, all diastereomers will be retrieved.

Queries for structural searching are created with an adaptation of the Molkick structure editor. For Current Facts, Molkick has been enhanced with the features required for the advanced search capabilities of S4, such as stereogenic searching and generic searching. For users unfamiliar with Molkick, it is easy and intuitive to use with a short learning curve.

Search times vary considerably according to the nature of the query. Presumably, faster hardware will also accelerate the search time. Current Facts was tested for this review on a slower computer, an AT clone at 12 Mhz, and a typical search for an uncomplicated query with 0–20 hits took about 40 s. Large queries with several rings took 60–90 s. If the query was tautomeric, the search time roughly doubled. In the stereo search mode, the search time was about 5 s less. Use of the predefined or user defined atom lists added no time to the search, but use of the generic groups was a problem. A test of the heterocyclic generic group (CHC) took over 11 min, but this search found 16000 hits. It is not clear if the lengthy search was due to the generic search directly or the large number of hits that had to be posted. When several Markush structures were substituted for the CHC code, only 3 hits were retrieved and the search time dropped back to 40 s again.

Display Modes: After a search is complete, from either OptiSearch or S4, the answers are automatically passed to one of several display modes. The short display mode, designed for reviewing hits quickly, displays six structures on the screen simultaneously. There are two views in the short display mode. It is possible to view the structure only, or alternatively, the BRN, chemical name, and a short list of key fields available, such as preparation and spectral data. The full display mode shows the complete record for a compound including the structure. The structure graphic is superimposed over the identification information, but it is scalable with the mouse. There are several toggles for various features in the full display, including turning references off and turning the structure off. Printing the full display is time consuming (about 90 s per page when tested on an HP LaserJet II) but produces handsome output. The lengthy print times for graphics on LaserJet II printers is notorious, but it is reported that the LaserJet IIISi is much faster, requiring about 15 s per page.

It takes about 5 s for a record to be accessed from the CD-ROM itself. Thus, every short display screen takes 30 s to fill up and return control of the computer to the user, which seems like a very long time. After a record is displayed once, however, additional manipulation is fast, such as viewing the full display or going back to earlier records. Unfortunately, this is probably an artifact of the retrieval of data from the CD and is probably fairly constant, regardless of hardware. For some users, this sluggishness may be an obstacle to using Current Facts.

Special Uses: Current Facts may be useful as a current awareness tool. Currently, there is no real way to do a current awareness search based on a structure query. Current awareness searches of the CAS Registry file are possible but are of very limited value because CAS puts structures in the Registry file much faster than the corresponding citations appear in the CA file, so Registry file current awareness searches contain very few actual references. A structure search on Current Facts will return a citation for each structure located. However, a structure retrieved on Current Facts is not necessarily novel; it only has to be reported with new data. Unfortunately, there is no way to distinguish new and novel structures from known structures with new data. This would be a nice feature for certain current awareness searches. The inability to distinguish which update a record belongs to is another limitation of Current Facts as a current awareness tool. There are four updates on each disc, and a regular current awareness search should only retrieve records for the most recent update.

Another area with great potential, unrealized in the CD-ROM database, is reaction searching. In the online file, it is possible to search for a particular substance or group of substances as starting materials, reagents, catalysts, or products and obtain chemical reaction information. At present, Current Facts does not have the requisite linking capabilities in the OptiSearch module to perform this type of search, but this is a promised enhancement.

The Big Picture: Ultimately, the software in the Current Facts package will be the cornerstone of a search on the Beilstein system. Current Facts will be used to generate queries, including creation of structural and factual queries, creating chemical names for queries with Autonom, and generating LN's for similarity searching with Sandra. The most recent literature data will be on the Current Facts CD-ROM, with the online services readily available for a complete search of the literature from 1830. Current Facts and the online files will also provide direct references to the printed Handbook, for easier browsing and for users who prefer to use the Handbooks. The electronic databases do not obviate the desirability of a subscription to the printed Handbooks. The organization of the Handbook gives it great browsing potential, much more so than any electronic database can provide.

At the moment, in the absence of a more complete Beilstein Information System, Current Facts as a stand alone product is limited. There is probably little interest in just a 1-year slice of literature. However, Current Facts is a well designed and executed package, and a very significant chemical information product. Beilstein is attempting to reassert its historical role as the most significant source of chemical information for organic chemistry, which is a role played by Chemical Abstracts in recent years. Although parts of the Beilstein's traditional product, the printed Handbook, are 30 years behind in their coverage, Beilstein is catching up fast with its electronic databases. According to Beilstein's schedule, they should be as current as Chemical Abstracts by 1995, but of course the data from a Beilstein search is much more than just a list of references; Beilstein provides edited and sorted data. Current Facts will provide high-quality current information, as well as the essential user interface for using Beilstein in the future.

Andrew H. Berks, *American Cyanamid Company*

*Book Reviews**

Molecular Thermodynamics: A Statistical Approach. By James W. Whalen (The University of Texas at El Paso). John Wiley & Sons, Inc.: New York. 1991. x + 381 pp. $49.95. ISBN 0-471-51478-0.

As one who has for many years taught courses in macroscopic thermodynamics and in statistical mechanics, I looked forward to this book with great interest. What one expects from a book with this title is the introduction of molecular and statistical insight into the formalism and precision of classical thermodynamics. Unfortunately, I found it very disappointing. It reads like an undigested amalgam of a chemical engineering thermodynamics book with a decades-out-of-date book on statistical mechanics. There are many details about thermodynamic calculations and some useful tables, but these can be found in any good book; what is missing is real molecular insight.

The author seems somewhat uncomfortable with statistical mechanics. He repeats the standard mathematical derivations, indeed somewhat repetitively, but with little insight as to their significance. For example, I could find no mention of fluctuations, a discussion of which is the essential link between molecular ideas and macroscopic thermodynamics. The reader needs to understand why fluctuations of macroscopic properties are essentially unmeasurable, so macroscopic thermodynamics is "exact". Chapter 1 perpetuates the common but incorrect derivation of the molecular partition function by applying Stirling's formula to $N_i!$ when most of the N_i's are zero. Had not quantum statistics and the grand partition function been relegated to brief mention in appendices, these important concepts could have been used for an impeccable derivation. The author even manages to take the *difference* (rather than the ratio) of two partition functions (p 229).

The book introduces the equation $E = \sum p_i \epsilon_i$, differentiates it, and identifies the two differential terms as reversible heat and work, but then fails to make use of them, as, for example, in discussing adiabatic reversible (isentropic) expansion in terms of $\sum \epsilon_i \, dp_i = 0$. The discussion of the Third Law is inadequate, and, except for a very brief mention of CO, there is no discussion of the "exceptions", the molecular explanations of which are models of molecular thinking. If there is any discussion of nuclear spin (*o*- and *p*-H_2) or of isotope effects, I could not find it.

The discussion of interacting particle systems (Chapters 7 and 8) is conventional and inadequate. The virial equation of state and the virial coefficients are briefly discussed, but only in an appendix. The section on solid state discusses heat capacities and relates the Debye and Einstein θ's incorrectly in Figure 7.4 (but correctly in Figure 7.5). The electronic heat capacity of metals is included, but there is no mention of magnetism. The section on the liquid state is similarly inadequate; some discredited

*Unsigned book reviews are by the Book Review Editor.