Today's Chemist at Work
February 1999
Today's Chemist at Work, **1999**, 8*(2)*, 45-46, 49-50.

The History of the NIST/EPA/NIH Mass Spectral Database

Stephen R. Heller

Developed in the 1970s, this system has become the backbone of today's GC/MS identification methods.

Today, seemingly every analytical instrument comes with an attached personal computer (PC) to collect and process data. This occurrence may seem to be such a given that some people think it was always so. However, those of us who came of age in the 1960s know it was not. Today, we think little of extracting a wastewater or sludge environmental sample or a blood, urine, or other biological sample; injecting the solvent in a gas chromatograph/mass spectrometer or liquid chromatograph/mass spectrometer; and having an attached data system tell us what analytes are present in the extract and the degree to which we can trust the identification. Pretty printouts come without bidding from laboratory printers, and data can be transferred to a central laboratory information data system. Computers are everywhere in the modern laboratory, and with regard to the mass spectrometer - arguably the most information-rich source of analytical data - laboratories could not produce the information needed for environmental and biological analyses without computers. It was not always this way.

Although MS has been around for nearly a century, before 1970 such instruments generated data of only the most rudimentary type. Most laboratories used mass spectrometers as one part of a complex identification process to determine the chemical structure of natural products, such as alkaloid drug candidates, or to determine whether an organic synthesis had been carried out properly to generate a

compound of the correct molecular weight. MS data often were combined with NMR, IR, and UV spectroscopic data to give insight into the overall structure of a complex molecule. There were no data systems, and, actually, none was needed. For structural analysis, it was necessary only to introduce the sample directly to the mass spectrometer and generate a single representative spectrum to be manually deciphered. First, one had to identify the *m/z* value of each peak by manual counting (28 for nitrogen, 32 for oxygen, 149 for a phthalate ester impurity coming from plastic tubing, and so on). After the *m/z* values for each peak had been labeled, the analyst - typically, a Ph.D. chemist - then identified the molecular ion and each daughter ion. Knowledge of how the chemical bonds rearranged on ionization hopefully allowed the original structure to be written out.

However, when a gas chromatograph was combined with a mass spectrometer (Figure 1), analytical chemists were faced with exponentially increasing amounts of data. It was cumbersome, if not impossible, to generate a representative spectrum from each chromatographic peak, much less manually count each *m/z* peak to identify the chemical species. Fortunately, the introduction of laboratory computers effectively coincided with the development of GC/MS. Computerized GC/MS quickly revolutionized the field of trace organic analysis and made very significant contributions to research in medicine, biochemistry, flavors, fragrances, and organic geochemistry. Thirty years later, GC/MS is still the major analytical tool for environmental analysis. However, it is not only the automation of individual *m/z* mass spectral peak identification that makes the computer so necessary today; it is also the computer's ability to routinely identify compounds on the basis of their mass spectrum.

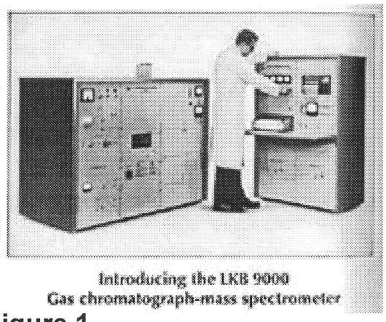

Figure 1
A 1966 advertisement from Analytical Chemistry showing a gas chromatograph/mass spectrometer from LKB Instruments (Rockville, MD; Stockholm, Sweden).

This ability to decipher identity depends on the existence of a database mass spectra from a wide range of compounds; the most universal database is that maintained by the National Institute of Standards and Technology (NIST). Since the development of this database in the 1970s, under the National Institutes of Health (NIH) and the U.S. Environmental Protection Agency (EPA), the NIST mass spectral database has become an essential tool in environmental monitoring for regulatory compliance.

The main advantage of the analytical technique and the database is that they have substantially increased the capacity of staff to handle many environmental samples and to accurately identify specific organic compounds without the need for a Ph.D.'s training in mass spectral structure elucidation. Back in the 1960s, few chemists knew MS. When the demand for GC/MS analysis arose, many chemists with other backgrounds - such as chromatography - were pressed into service to do environmental analysis. The mass spectral database, along with various search algorithms, became a crucial tool in meeting the demand for trained staff, which was much greater than the supply. In a sense, one could say that the mass spectral

database and search system has taken the chemist out of analytical chemistry.

A World-Class Product

It would be nice to look back now and see how a careful study was done on what needed to be done and was followed by a well-implemented plan. Certainly, the success of the database and the many problems it has solved would support such a story. Unfortunately, the true history is not so clean and neat. However, it is a nice story about many people in different organizations who worked together to produce a world-class product used today in thousands of analytical chemistry labs around the world.

In the 1960s, Scotty Pratt, the NIH Division of Computer Research and Technology (DCRT) director, was looking for projects to prove to NIH scientists that computers were more than calculators. He realized early on that for any such project to be a success, it needed a computer group willing to provide the necessary hardware and software support, an NIH lab willing to explore something new, and a scientist willing to work full-time as a computer chemist. Having tired of synthesizing nitrogen and sulfur compounds, I was fortunate to have the opportunity to work at DCRT as a senior staff fellow (the U.S. government title for a postdoc at that time) and to work with Henry Fales at NIH's National Heart, Lung, and Blood Institute. Fales was (and is) an eminent organic chemist and mass spectroscopist, and he realized the potential benefit of automating the mass spectral data analysis process. He was willing to test any computer-based analysis tool developed. He asked one of his lab chemists, Bill Milne (now editor of the American Chemical Society's Journal of Chemical Information and Computer Sciences), to work with us in testing the system.

The first task was to find a database. Fales convinced Klaus Biemann at the Massachusetts Institute of Technology to let NIH use his tape-based database of some 8000 spectra that he had been collecting as part of his research during the past decade. Biemann and students (Harry Hertz, now at NIST, and Ron Hites, now at Indiana University) had been developing batch-search software as a mass spec data analysis tool. The user manually input the *m/z* values, and the closely matching compounds were printed out.

In 1970-71, with the assistance of Richard Feldmann, a very talented DCRT programmer, we wrote a time-sharing program in the FORTRAN programming language on a Digital Equipment Corporation PDP-10 minicomputer to search the database provided by Biemann. The first of several papers to describe the database and search system was published in Analytical Chemistry in 1972, but that was only the beginning.

When Fales traveled, lectured, or welcomed chemists from all over the world to NIH, he would take a few minutes during his lecture or lab tour to show off the software, called the NIH Mass Spectral Search System (MSSS). One day, Fales called to tell me that a few of his scientific friends and colleagues (more like a few dozen) liked the MSSS. He asked whether it was possible to find a way to let them use it from their home laboratories. Accomplishing such a task was not as easy then as it is now. There was no Internet, no computer

networks - only 110-baud modems connecting an ASR teletype printer to a single mainframe computer. Nonetheless, the MSSS was opened to the public as an experimental project via a time-share arrangement with the NIH computer center. Originally, I wrote a simple manual, but as users of the search system and database asked for more features - which eventually grew to some two dozen options - I wrote a formal 59-page manual that NIH printed in November 1972.

Graphics on Microfiche

In the early 1970s, disk space was expensive, and few computer graphics were available. Thus, among the options added to the MSSS search program was one that linked search results to a computer-driven microfiche reader that contained the identified compound's full spectrum, with all the peaks nicely plotted. Figure 2 shows the author with an old teletype, a telephone 110-baud acoustic coupler, and a microfiche device.

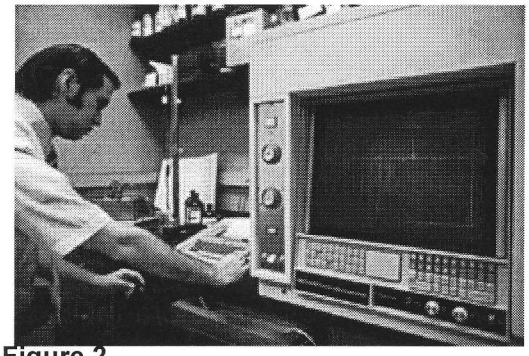

Figure 2
The author with an old teletype, a telephone 110-baud acoustic coupler, and a microfiche device.

As the popularity of the system increased and more people dialed into the DCRT PDP-10, NIH administrators decided that the DCRT could not continue to provide MSSS computer support. Instead, a collaboration was established with the Mass Spectrometry Data Centre (MSDC) in Aldermaston, England. The MSDC also had a mass spectral database, and the two systems were combined. Subsequently, the MSDC transferred its system to a commercial (General Electric) time-sharing system that was available throughout Europe and the United States. Soon, hundreds of chemists worldwide were using the database and the search system. The MSSS later moved to the Automatic Data Processing (ADP)-Cyphernet system and became part of a larger NIH-EPA Chemical Information Systems (CIS) database.

In 1973, I joined EPA at its headquarters in Washington, DC. Shortly before then, EPA had been given a congressional mandate to develop routine monitoring methods for environmental analysis. Working primarily with EPA scientists Bill Budde in Cincinnati, OH, and John McGuire in Athens, GA, we developed the MSSS and its attendant mass spectral database as a mainstay of EPA's environmental analysis program, much of which was based on GC/MS analysis.

Because EPA had a large and practical need for mass spectral search and analysis, considerable funds were provided to increase the size and quality of the database. The primary EPA database and software work was contracted to Fein-Marquart Associates. Their chief technologist, Dave Martinsen, began evaluating the quality of the spectra, and added Chemical Abstract Service (CAS) Registry Numbers to each chemical in the MS database.

In addition, Cornell University professor Fred McLafferty, under contract to EPA, developed a Quality Index computer program for the database, thereby permitting an analyst to judge how well unknown spectra matched those in the MSSS database. Through the work of Fein-Marquart, McLafferty, and others, by 1975 more than 30,000 spectra were in the database. In 1978, the National Bureau of Standards (NBS) agreed to print a five-volume collection of some 25,500 different mass spectra, along with the chemical names, synonyms, and chemical structures. More than 1500 copies of these books were sold.

NIST Given Responsibility

Although the driving force behind the rapid expansion and use of the database was EPA, the information was finding other research uses, including assisting in organic chemistry structure elucidation problems and toxicology research. For this reason, in 1980, EPA transferred responsibility for database maintenance to NBS (which later became NIST). NBS had recently been authorized, through a unique congressional authority to copyright databases, to disseminate information such as that in the MSSS. Under the leadership of Dave Lide and Lew Gevantman of NIST's Office of Standard Reference Data (OSRD), the database was expanded.

By the late 1980s, EPA's Budde had turned over all further environmental database development activities to Sharon Lias and Steve Stein at NIST. Using income from the lease of the database (more than 25,000 copies have been distributed since 1988), NIST enlarged the database and improved its quality by manually examining each mass spectrum. In addition, Stein developed a PC-based search algorithm. Contracts to obtain complete mass spectra for each compound, as opposed to spectra from the scientific literature which often contained only a few peaks because of journal space limitations were undertaken by NIST. Many spectra of "useful" compounds, such as those commercially available and those of medicinal chemical interest, were added to the database. These compounds came from such sources as the Toxic Substances Control Act (TSCA) Inventory, EPA Monitoring Methods Index (EMMI), the U.S. Pharmacopoeia/United States Adopted Names (USP/U.S.A.N.), the Chemical Rubber Company's Handbook of Data of Organic Compounds (HODOC), and the European Index of Industrial Chemical Substances (EINECS).

In the area of database searching, the original Biemann search algorithm using the two largest peaks in every 14 *m/z* interval also has evolved. Incos, a 1970s software company, had developed a search system that was incorporated into the Finnigan data systems at the beginning of that decade. The 1990s Stein/NIST system has taken the Incos approach and expanded it. In the middle to late 1970s, Cornell's McLafferty developed the Probability Based Matching (PBM) program, which improved on searching for compounds in mixtures. PBM is still used today by several mass spectrometer manufacturers. McLafferty also developed the Self-Training and Interpretation System (STIRS) program, which attempted to elucidate chemical structures for compounds not found in the NIST database. Today, the Stein software does substructure analysis and is able to perform a highly evolved library search algorithm with sophisticated probability calculations to show that a

given library identification is correct. An example of the Stein/NIST mass spectrum interpreter program is presented in Figure 3.

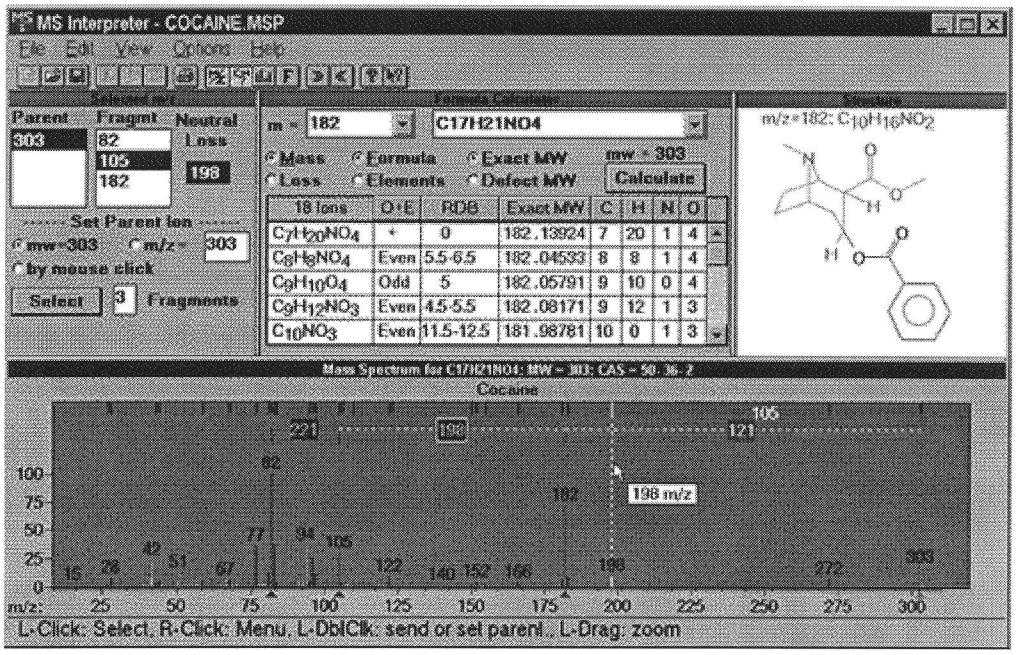

Figure 3
An example of the Stein/NIST mass spectrum interpreter program

The main effort in the database work in the 1990s has been to build a large high-quality database. Some areas of current and future work on the database and associated search software being undertaken at NIST include acquiring additional spectra of useful compounds, adding retention index information, developing fragmentation software to identify peaks that are consistent with fragmentation rules, improving chemical nomenclature, adding CAS Registry Numbers to database entries that currently lack this information, and improving the internal administrative evaluation and editing program.

The latest release of the database, NIST98, stands at 107,886 compounds - almost 100,000 more than the original database provided by Biemann almost 30 years ago. Today's NIST software allows a user to display and print not only a compound's mass spectra but also its chemical structure. The effort to create this database took a team of mass spectrometrists about 10 years to evaluate the entire database by hand. Including additional spectra for some compounds, there are a total of 129,136 spectra (out of a master database of 175,510 spectra, which include some duplicates and some low-quality spectra), all of which have been critically evaluated. Information about how to obtain the current version of this database is available at the NIST home page (www.nist.gov/srd/).

An Ongoing Endeavor

In summary, throughout three decades, the NIST-EPA-NIH mass spectral database has been a very useful tool for problem solving, almost anywhere mass spectrometers are used, and it will become even more so in the future. The Bieman database has become an ongoing, not a static, endeavor. It gives the ordinary chemist, untrained in spectral problem solving, the ability to solve today's

questions - whether they involve environmental analysis or the analysis of compounds from combinatorial chemistry studies.

Stephen R. Heller is past chair of the IUPAC Committee on Chemical Databases and the software review editor for ACS's Journal of Chemical Information and Computer Sciences. He is a consultant at Pool, Heller & Milne and can be reached by e-mail at steve@phm.com