ARTICLES

Chemical Information Activities: What the Future Holds

Stephen R. Heller†

U.S. Department of Agriculture, ARS, Beltsville, Maryland 20705-2350

Received June 29, 1992

The current state of chemical information technology in a number of areas is presented. The author speculates on a number of areas in some detail and presents a list of predictions as to the likely state of the field in about the year 2000. The economies of chemical information are also briefly discussed. The author concludes that until the computer is made to be an easily used tool, not a barrier, reasonable, let alone optimum usage, will not result.

INTRODUCTION

This presentation[1] is designed to stimulate discussion of new technology which is becoming available to chemists, applied to chemistry, and most importantly, used by chemists in their everyday activities by the beginning of the next century. The data in this paper have evolved over the past 4 years, and no doubt will continue to evolve as new phenomena stimulate changes in the habits and activities of chemists.

As computer technology has developed, the use of computers in chemistry has expanded from simple arithmetic calculations to very broad areas of chemistry. This paper examines some of these areas and tries to summarize the current state of the use of computers in chemistry and what the author believes the use of computers in the field of chemistry will be a decade from now, i.e., the beginning of the 21st century.

BACKGROUND

The computer, like any other technological tool, has become integrated gradually into the daily routine of chemists. Anything new in science is usually taken with some skepticism. That this has been the case in science for a long time was noted by Max Planck, better known for Planck's constant: "New scientific truth does not triumph by convincing its opponents and making them see the light, but rather because its opponents eventually die, and a new generation grows up that is familiar with it".[2]

The widespread use of computers in chemistry has clearly been handicapped by a number of factors, a major one being the lack of familiarity with this new technology on the part of chemists and managers in the field of chemistry. This is true from academia to government to industry. In working to locate supporting facts for this article, I heard the spirit of Max Planck evoked a number of times. Phrases like you need to "raise a generation of people who are comfortable with these tools"[3] and "raise a generation of advocates"[4] came from professionals in the field of market research. This suggests that wide-scale and heavy use of computers by chemists has not yet started.

At present, the routine use of computers in support of research and production in a chemistry laboratory or office, other than for word processing, spreadsheets, and literature searching, is low[5] (defined here as involving less than 25% of the potential users). Why is this the case? There are a few hundred thousand chemists in the U.S.,[6] and many of them have computers.[7] In the worldwide pharmaceutical industry, in 1989, there were some 54 000 scientists employed in R&D activities.[8] Of these, almost 44 000 were in the U.S., and some 28 500 of these are categorized as scientific and professional staff, the others being categorized as technical and support staff. It is generally thought have most (>90%) of these individuals have computers, virtually of all which are IBM Pcs and clones or Apple Macintoshes; the remainder having Sun, Silicon Graphics, DEC, or other manufacturers of workstations.[5,7] It has not been possible to obtain any definitive information on the number of scientists or chemists with Pcs. Marketing surveys have not addressed such questions.[9]

If one combines these numbers with those in other developed countries one could estimate some 800 000 chemists[4] as a potential market for various computers and computer systems and for software specifically designed to support the needs of the chemist. This paper will examine some possible reasons why large numbers of chemists have not yet decided that computers are a necessary tool for conduct of their everyday research and administrative work, thus explaining the lack of extensive use of computers and related computer technology.

Please note that when the qualitative phrases "few" or "low" (as defined in ref 5) are mentioned for the overall use of a particular piece of computer, a computer program, or a computerized database, the phrase is meant in comparison to the overall potential purchase and use by some 800 000 potential users (chemists) worldwide. For example, when a group of chemists belonging to the COMP (Computers in Chemistry) Division of the ACS reports their current level of use of computers for electronic communication at less than 15%, the term "few" or "low" seems justified. With the exception of one series of marketing studies in the area of computational chemistry,[4] there have, to date, been no published studies of the use of computers and related computer systems by the end users in the chemical community.[9] (Studies on the use of computers and databases in libraries are not regarded here as end-user studies.) The reason for this is the small size of the current market which does not justify the investment for such a survey.[3,4] Thus, the reader will have to accept the lack of hard statistics for many of the statements presented here.

While selling a total (over the lifetime of the program) of a few hundred or even a few thousand molecular modeling or

† Electronic mail address: SRHELLER@ASPR.ARSUSDA.GOV.

Table I. Issues for Discussion

topic	today	2000
computer literacy	low moderate	moderate to high
computer chip technology	Intel 386, 486; Motorola 68000; RISC	Intel 986; Motorola 98000; RISC
operating systems	DOS, UNIX, Windows, OS/2, Macintosh	mostly enhanced UNIX
telecommunications	moderate usage 2400–9600 baud speeds	heavy usage 1 million+ baud speeds
Windows & OS/2	released with bugs	just beginning to work properly
interfaces	frightful and difficult	transparent and voice based
graphics	low usage in most software	predominant usage in most software
CD-ROM	low end-user usage	moderate end-user usage
chemical information	raw and unprocessed	processed and analyzed
online usage for chemistry	low	low
SDI	manual or by post	electronic
databases	bibliographic	numeric; factual
Beilstein	E–V series being published	E–V series still being published
chemical catalogs	online searching of catalogs	online ordering from catalogs
chemical identification	CAS RN and BRN	chemical structure
molecular modeling	few	some
educational software	random; not integrated with textbooks	integrated with textbooks
publishing	semielectronic	mostly electronic
books	thought of as probable dinosaurs	thought of as probable dinosaurs
instruments	semiautomated	fully automated with ISO data transfer standards

Table II. Computers

today	IBM PC widely used; Macintosh usage increasing
2000	McDonald's selling the McIBM; Graphics, mouse, user-friendly programs, very fast CPU (>250 MHz), lots of storage (>1 billion B), 1200 dpi laser printers, and high-speed modems

Table III. Telecommunications/Networks

today	networks being used routinely by many chemists; BITNET, Internet, and other networks used by scientists a few times per week; some companies have internal networks for many of their end-user PCs; telecommunication speeds in the range of 2400–9600 baud
2000	networks and e-mail used all of the time; automatic interfacing between all networks routine; automatic logins for mail done everyday before the scientist comes to work; e-mail automatically rerouted as you travel to meetings, holidays, and home
local	area and wide-area networks are widely available within most organizations; large databases more readily available within organizations; telecommunications speeds in the range of 2.4 million+ baud

Table IV. Interfaces

today	programs in their infancy
2000	voice control for input with lots of graphics; standards for graphics and data are common; IUPAC, CODATA, ASTM, ISO, and other organizations agree on data transfer protocols

Table V. Graphics

today	usage in its infancy; lack of compatibility; lack of standards; FAX transmission in its infancy
2000	graphics software packages are widespread; graphics routinely sent electronically (Microsoft chart); PCs have built in FAXs for receiving and transmitting chemical structures and tables of data

Table VI. CD-ROM

today	chemistry CD-ROM products are rare today; low-density (600 MB) CDs; e.g., Aldrich MSDS, Beilstein Current Facts, Canadian Toxicity Databases NIST Mass Spectrometry; *Kirk-Othmer Encyclopedia*, CAS 12th Collective Index, Chapman & Hall *Dictionary of Natural Products*
2000	new products and high-density CD-ROMs (6 billion+ B); *Heilbron Dictionary of Organic Chemicals*; *CRC Handbook*, CAS volume(s) on CD-ROM; CAS subsets (e.g., polymers, patents) Beilstein subsets, Gmelin subsets; collections of small numeric databases (e.g., IR and MS databases from NIST); most journals

structure drawing programs is, today, a major accomplishment in the business of software for chemistry, it is a minor event relative to the daily sales of word processing, database management, spreadsheets, and other such programs. The lack of any public software companies in chemistry (i.e., software companies devoted exclusively to selling software for chemistry and whose stock is available to the public) is indirect evidence to support this position that there is, at present, no major financial incentive to go into this business.

Before proceeding to the main thrust of the speculations into the future of computers in chemistry, it is important to note that there are some labs as well as areas of chemistry in which the use of computers is very high. As mentioned above, the area of computational chemistry is clearly one of these areas. While it is estimated there are 1000 sites worldwide with some 2000 academic and industrial chemists now involved in this area,[4] this is less than 1% of the chemists in the world. In almost all areas of spectroscopy, computers are heavily used to acquire and analyze data. A reader involved in these areas of chemistry would certainly not fall within the "low" range of computer use. However, these "pockets" of high computer use, when averaged with the entire chemistry community, I believe are consistent with the levels of usage stated here.

COMPUTER AND CHEMICAL INFORMATION ISSUES

Table I summarizes both the issues which are to be discussed here and the current and predicted level of activities in these areas. Space in this journal does not permit a full analysis of all of these topics. Thus, a few representative issues will be mentioned. Tables II–XIV list details for many of these issues.

The heart of the matter is computer literacy. Growing up with, being familiar with, and making regular use of computers and computer systems of information will not become the norm and "triumph" (à la Max Planck) without the necessary atmosphere and background being part of one's upbringing. As mentioned in the Introduction, the initial use of computers by chemists (and other scientists) was limited to performing simple calculations. Hence, it is no surprise that the area of chemistry in which computers have been used is primarily

Table VII. Chemical Information

today	most information is raw, unprocessed, and unevaluated; CAS, Beilstein, VINITI—most abstracting and data extraction is done in-house
2000	greater reliance on processed and evaluated data, such as Beilstein, Gmelin, IUPAC data series, CRC Handbooks; CAS, Beilstein, VINITI—economic factors will cause most abstracting and data extraction to be done by free-lance workers at home; articles and abstracts all sent electronically from abstractor to abstracting service

Table VIII. SDI

today	popular feature for vendors; results mailed to customers or left for online downloading[10]
2000	popular feature for vendors; results automatically sent electronically to customers' PC via networks

Table IX. Databases

today	still in the age of bibliographic databases
2000	second generation of databases—numeric and factual data overtake bibliographic databases in usage; usage increases as scientists realize need for (good) data for dry lab work (modeling, etc.)

Table X. Chemical Catalogs

today	lots of printed catalogs; a few catalogs on disk or CD-ROM (e.g., Aldrich)
2000	catalogs on CD-ROMs; users order directly over the phone from their labs; ordering by credit card is routine

Table XI. Chemical Identification

today	CAS Registry Number reigns supreme
2000	with chemical structures in all important databases, special identification numbers have little use; standard molecular data formats allow for interfacing between all public and private files

Table XII. Educational Software

today	random usage; software used in teaching high school and college chemistry does not come with textbooks, but as separate products
2000	software integrated into textbooks[11] (G. D. Wiggins, *Chemical Information Souces*[12]); PC floppy disk programs part of all undergraduate texts; chemical information courses have PC-based tutorials and practical online sessions

Table XIII. Publishing

today	journal articles are almost the only acceptable form of communication and reward/promotion; some scientific manuscripts submitted in electronic form, but process is neither widespread or practical; virtually all refereeing done by postal system mail, with some done by FAX
2000	printed journals still predominate, but electronic data submissions, electronic journals, and software programs are now part of academic, government, and industrial chemist's reward/promotion system; leading journal publishers use electronic submissions to speed up processing of publications, easier data extraction, and overall quality improvement; electronic (FAX and e-mail) peer review predominates

computational chemistry. But the usage even in this area is low. As Casale and Gelin[4] point out "as a scientific discipline, computational chemistry is in its infancy". The current state of education in many parts of the world will make further usage difficult. However, I would hope that in college and graduate school there would be sufficient competence to train the upcoming generation of chemists to become very familiar with computers, through the introduction of computer application courses taught by chemists in chemistry departments.

Table XIV. Instruments

today	every instrument has a computer and most have different computers and operating systems; no universal interfacing; ASTM and instrument manufacturers discussing standards
2000	everyone has a computer, and there are universal protocols for input and output; data readily shipped to other computers for identification and analysis

Without an increase in the level of computer literacy, the remaining issues are pretty much irrelevant.

There are two facets using computers; writing programs and using programs. The writing of programs is really a rather limited issue. A computer is a tool. When a chemists gets too involved in the tool then he or she is, more often than not, no longer doing chemistry. What matters is using programs. To do this effectively and properly, you need to know what a computer can do for you in the area in which you need to solve a problem. I do not need to be an automotive engineer to know that to get somewhere by car I need a car, need to know how to drive it, and need to know where I am going. The same is true with computers. Understanding what a computer can, or cannot, do is the important step. Then either finding software and hardware to do it or getting someone to produce what is needed to get the job done is relatively simple. I believe that virtually no chemists use computers as an end in themselves and that chemists should use computers as one of many tools to do their job, but only if the computer is the most effective way, and not a barrier, to do the job better, more effectively, and more efficiently.

Most chemists use computers for only administrative purposes (like writing a manuscript which may or may not include chemical diagrams). I would argue that the reason for the lack of extensive use of computers is that the majority of computers (PCs of the 8086, 8088, 286, and 386 vintage) which are readily available to the chemist are of insufficient capacity and capability to do effective work other than word processing, structure drawing, and spreadsheet calculations. (Without the available computers, moreover, there has been no incentive to develop the software for chemists.) Until just very recently the computers with the necessary CPU speed (e.g., 486 CPU PCs, DEC, Sun, Silicon Graphics, and other type workstations) and available memory and disk space to do a variety of scientific applications (modeling, quantum chemistry calculations, spectral interpretation and prediction, database searching of spectral data, image analysis, etc.) were much too expensive for most individual chemists to have on their desks or in their labs. As little as 2 years ago a computer with an Intel 286 CPU and 40-MB hard disk was considered a state-of-the-art computer system. Almost nobody with a PC would keep a mass spectral database[13] and search system, requiring some 23 MB of hard disk space on a computer system with a 40-MB hard disk. Today, to run a modern PC operating system (using DOS 5 and the Windows or OS/2 operating systems) one needs at least an Intel 486 CPU (with a 50–66-MHz clock) and 300–500 MB of disk space. A recent article from a monthly computer magazine[14] added up the disk storage requirements for a little over a dozen pieces of popular business-oriented software, and the total of the disk space required came to almost 100 MB, not including any of the disk space required for program swapping or any of the space needed for files of data and information.

In the next few years chemists will be able to replace existing low-power (e.g., 286 or equivalent type of PC) computers or buy new ones with the computer power of an Intel 486/586 CPU (or their Sun, Silicon Graphics, DEC, or equivalent) with sufficient disk storage space to readily run complex and powerful programs.

The low usage of computers by chemists in the recent few years may be attributable to the lack of affordable adequate hardware, but it will take a number of years for this new and more powerful hardware to work its way into the system and into everyday use. Furthermore, software prices must follow those of hardware: it is difficult to believe that many chemists will pay $2000 for a computer system and then spend thousands of dollars for additional software packages. Only low-cost, high-volume software is likely to succeed in the future. An experiment in mass marketing to the chemistry community is now being undertaken by Autodesk, which is hoping to increase the number of scientists using PC-based molecular modeling packages from a worldwide total of 5000 to 100 000 or more.[4] Included in Autodesk's effort is a multimillion dollar grant program to encourage university use of the HyperChem molecular modeling software product.

Computers are used for electronic communication by a small but growing number of chemists. Among the reasons for the low usage are the lack of modems and dedicated phone lines as well as the difficulty in finding where people are located or information is located and initiating communication. There is also the lack of computer addressees on the necessary computer networks (Internet, BITNET, Sprintnet, Compuserve, etc.) and the problem of connecting between networks. If I want to telephone someone in another city or country, I need only to get the phone number from a telephone operator at a price (except for unlisted numbers). With computer networks, there is no readily available phone book and no operator. Practically all numbers (actually computer network addresses) are unlisted. That does present (using a good chemical phase) an energy barrier to solving a problem. However, I can see changes coming. A few years ago a business card had a name, title, address, and phone number. Today many business cards have FAX and Internet addresses. It is even possible in many cases to access a computer address online, although this is only beginning to find widespread use. This is part of computer literacy. This is progress.

I believe, however, that it will still take years for chemists to make routine use of Internet and the related networks connected to it. From discussions with a number of people, the estimated usage of Internet by chemists was in the range of 10–15% of those who have a computer and can access computers outside their organizations.[5] This number is quite similar to that found for the current level of electronic communication of chemists belonging to the COMP division of the ACS, noted below. At present about one-third of the use of Internet is for e-mail.[15] Perhaps more interesting is that up to 20% of the Internet traffic in the U.S. is flow-through traffic between Europe and Asia.[15] While the overall number of computers connected to Internet (estimated at more than 727 000 as of January 1992[16]), the volume of usage (traffic) on Internet,[17] and other statistics about Internet are available,[17] these numbers do not address the questions of micro-usage or of end-usage. While the number of computers and the number of users with accounts on these computers can be reasonably estimated, it is not possible at the present to determine how many people are actually using Internet and what their usage level is. It is possible that only a few hundred of the hundreds of thousands of computers and users are generating large percentages of the total volume of use of Internet.

Electronic mail or e-mail is slowly (due to a lack of knowledge about it) becoming a new type of network for chemists as well as other scientists.[18] It is not an "old-boy" network or "invisible college" because it allows anyone accessing these systems to be an "equal" of anyone else on the system. Electronic bulletin boards, discussion groups, and news groups, dealing with all subjects, are slowly sprouting up everywhere in all areas, each with dozens to hundreds of users.[19] Of almost 800 such news groups surveyed by Kovacs,[20] less than 100 are in the physical sciences, and of these only 20 are in chemistry. Again a small percentage is observed when the topics relate to chemistry. A few will be mentioned here. For the area of chemical information these include the Chemical Information Sources Discussion List managed by Gary Wiggins at Indiana University (CHMINF-L@IUBVM.UCS.INDIANA.EDU with about 425 subscribers with a maximum volume of 5 messages each per day), the Chemical Education News Group managed by Bill Halpern at the University of West Florida (CHEMED-L%UWF.BITNET@CUNYVM.CUNY.EDU), and the Computational Chemistry News Group managed by Jan Labanowski at the Ohio Supercomputer Center at the Ohio State University (CHEMISTRY@OSU.EDU—with about 1000 subscribers and with a volume of 10 messages each per day). In a 1990 survey,[21] it was estimated that some 10% of the overall working population in the U.S. and Canada use e-mail systems versus only 1.3% in Europe. The ACS Computers in Chemistry (COMP) Division now distributes its newsletter via e-mail on Internet as well as in hardcopy. In mid-1992, a little less than 10% of the COMP members received the newsletter electronically,[22] a number comparable with the survey mentioned above. By 1994 the same survey estimated the usage in North America would grow to almost 29%, while in Europe the usage would expand to just under 5%. Certainly there are cultural differences between those two areas in the use of the telephone and modems, but the European PTTs and their policies add to the difficulty of use. I would expect the e-mail usage in chemistry would be higher today and that e-mail will become a necessity by the end of this decade. The low cost, ease of use, and ability to readily send written information to colleagues around the world make this an ideal replacement for the existing "old-boy" network of phone calls, meetings, and letters. For anyone, from a Nobel prize winner to a undergraduate student, to be able to communicate freely and easily and to see what topics and areas are of current interest should improve scientific communication and research work. e-mail will be able to reduce the time for papers to be sent back and forth. Once the graphics problem is solved (both the technical standard for graphics and the speed of transmission of graphics) and put into practical use, e-mail will allow for real electronic journals.[23] This area has a great and important future for chemists throughout the world.

Another major problem with computer programs is the difficulty associated with their use and the poor documentation about them. Pacman and Nintendo (the popular video games of the 1980s and early 1990s) never came with manuals. Some manuals seem more designed for weight lifting than for explaining how to use a particular computer program, and they are rarely available in computer-readable form.[24] In computerized form, manuals could be searched for a word which you are interested in finding. Installing and running programs is a major energy barrier for most people. My philosphy is that if I must read the manual to use the computer program, I probably am better off without it. There is no way someone can become and remain proficient with a wide variety of programs, remembering what each does and how to perform particular tasks, as well as doing their assigned job as a chemist. Few people use their VCRs to record TV shows because they cannot figure out how to do it. This even created a market

for a device which automatically sets up the VCR to record based on a set of 5 digits you type into a device. The 5 digits are published in newspapers in the U.S. everyday next to each TV program listing.

Table I speaks of today's interfaces as being frightful and difficult. If people are not comfortable with a tool, they will not use it.[3] As computers become more powerul and as better software engineers graduate and get jobs, one can only hope and expect that the interfaces in the year 2000 will become transparent and even voice based.[25] One way to improve interfaces is through the extended use of graphics. Today the use of high-resolution graphics (1024 × 1024 pixels) is low. Color screen size is small (12–14 in.) and expensive. By the year 2000, I would expect that every computer will have a 20-in. or larger color monitor with at least 2048 × 2048 resolution, along with a color laser printer or plotter with the same capabilities. With the decreasing cost of hardware, this equipment should be available to most scientists in the coming decade. Related to the problem of high resolution for graphics is the problem of how to transmit all the information quickly enough to be of practical use. Today's modem speeds of 2400–9600 baud are much too slow for graphics to be practical. With the current trend to better networks and telecommunications, it seems reasonable to believe that the speeds of transmissions needed for chemistry graphics will be available in the next few years.

In the area of chemical identification, it has taken some 20 years for the CAS (Chemical Abstracts Service) Registry Number (CAS RN) to be used widely and routinely in databases and in searching for chemicals. While the CAS RN now reigns supreme for chemical identification, it suffers from the lack of any inherent intellectual value; it is, like the U.S. Social Security number, an idiot number (notwithstanding its check digit), assigned sequentially over time: a larger number just means it entered the CAS Registry system more recently. In the past few years optical scanning devices, coupled with advances in character and vector recognition, have led to the development of computer programs (see, for example the work of Johnson et al.[26]) which are able to scan articles for the scientific literature (or from internal research reports), extract chemical information, including connection tables of chemical structures and chemical reaction data (such as solvent, temperature of reaction, etc.). Kekulé, a similar, but less ambitious program[27] is able to scan a structure and create a connection table. Both of these systems will make adding connection tables to databases much less labor-intensive than in the past.

In spite of the wide use of the CAS RN in chemistry, and particularly in chemical regulation by the EPA (Environmental Protection Agency),[28] chemical names are also still very widely used for administrative and regulatory purposes. In fact the recently developed AUTONOM program[29] was initially conceived for internal processing at the Beilstein Institute for their handbook and database work. AUTONOM takes most (>75%)[30] chemical structures and creates an IUPAC-approved name for that chemical structure. Its administrative value for internal and regulatory purposes is such that it is now a commercial package.[31] Thus while there will be a need for chemical names and registry numbers, the primary need will not be a scientific one.

The ease of creating large databases of chemical structures, along with the efforts underway to create standard molecular data descriptions of molecules [e.g., the SMD (Standard Molecular Data)[32] and STAR (Self-defining Text Archive and Retrieval)[33] projects] and the increased ability to send large volumes of data over networks at high speeds, make it seem reasonable to predict that the use of the CAS RN for searching for a chemical will decrease over time. One of the major drawbacks of the CAS RN (and the Beilstein Registry Number) is the lack of these numbers in the private and generally confidential files of companies. It is not possible to use an internal identification number to search public files and vice versa. Only the chemical structure itself, when used as the "search term", will be a practical way to see if a chemical is in another database. As different organizations represent their structures slightly differently, only the advent of a standard molecular representation or an interchange program (such as the recently developed program ConSystant[34]) will allow a user to readily search for related structures in another database of chemical structures.

Most of the data and information in the major chemical databases of the world are raw and unprocessed. The two largest collections, those of CAS[35] and VINITI[36] are bibliographic. In these two databases, whatever the author says is accepted at face value. Since almost all of the papers abstracted are refereed, either the author's abstract is used or CAS or VINITI write an abstract based on the information provided in the publication. Only the Beilstein and Gmelin databases perform some measure of evaluation, although most of this work is really extraction of information. For example, in the Beilstein Handbook, online database, and CD-ROM Current Facts, the data are extracted. In the past there were some additional efforts made to assure that enough information was published in the original work to guarantee the work could be reproduced, and not every chemical reaction or piece of data was used as Beilstein. The Beilstein staff has never had the financial capability to evaluate the very large volume of data they process. Such data evaluation of large databases is rare, with the most well-known example being that of the U.S. NIST/SRD (National Institute of Standards and Technology/Standard Reference Data Program). Beilstein performed a "second" and valuable peer review, albeit too late to keep questionable or poorly defined or unexplained science from being published. In any event, today, due to the high costs of labor, both the Beilstein Institute and VINITI have fewer in-house staff than in past years and rely more on parttime outside workers. With some 65% of the costs at CAS being labor, it is reasonable to believe that CAS will be moving in this direction again. (CAS once had primarly all of the abstracting done by outside chemists.)

CD-ROMs are hardware devices that are just beginning to find use in chemistry. Again the problem of the lack of good software, adequate computer hardware, and available databases has limited the growth and use of this medium. While most of the 422[37] educational science libraries in the U.S. have CD-ROM drives, it is estimated that less than 1% of the computers which are in chemistry labs have a CD-ROM drive.[38] This estimate has been supported by a nonscientific, nonsystematic request for information which the author sent to the approximately 400 subscribers to the Chemical Information News Group (see above) which resulted in two responses, one from Exxon and one from Rutgers (Chemistry and Physics Departments).[39] In both cases, those information specialists who replied indicated that they know of no end-users in their organizations who had CD-ROMs on their PCs. In addition to this survey, a number of vendors of CD-ROMs were contacted. All considered the sales and types of users to be confidential information. None kept track of the type of users who were buying their products. In one case, that of the Aldrich Chemical catalog on CD-ROM, it was learned

that, while the sales (at $25 per CD-ROM) of their catalog on CD-ROM were under 1000, they publish 2.7 million copies of their printed catalog[40] which are distributed free of charge. Even in an area where computers are used more routinely in chemistry, namely, computational chemistry, less than 10% of the customers using the molecular modeling software program SYBYL have requested to receive their software update on the CD-ROM offered by the vendor.[41] This could be compared to the computer science community where more than 80% of the users of Sun workstations receive their software and documentation on CD-ROM.[42] Thus, chemists clearly have a long way to go before they become as familiar and comfortable with this medium as computer programmers and computer systems staff are. At present I would summarize the state of the chemists' use of CD-ROM as in its infancy, but I strongly feel that the growth curve for CD-ROM usage is likely to be exponential in the coming years, as evidenced by the use of CD-ROM in other fields.

Even for those who have a CD-ROM drive, the small size of the market often leads to databases which are not updated. No doubt owing to the lack of sales, even the Microsoft Bookshelf CD-ROM, which contains an almanac, dictionary, thesaurus, U.S. zip code directory, and other databases, has not been updated in 5 years.[43] CD-ROMs, which today store about 660 million characters(about 330 000 pages of text), will by the year 2000 replace many reference books and chemical catalogs on the chemists' bench and bookshelf. A few pioneers in this area, such as the Beilstein Institute in Frankfurt Germany, are leading the way to what will clearly be the library of the future. The Beilstein Current Facts CD-ROM has about 1 year of extracted data from the literature (without author names, titles, or abstracts), along with a computer chemical structure search system, all neatly collected on a single CD-ROM. Someday, the weekly issue of *Chemical Abstracts* will come to each chemist this way. Each chemist will have the *Merck Index*, *CRC Handbook of Chemistry and Physics*, *ACS Directory of Graduate Research*, and a few ACS journals all on CD-ROMs. By the year 2000 it should be possible to custom order a set of books on CD-ROM. For example, the ACS Symposium Series of several hundred books could be entered into computer-readable form and then books "printed" on a CD-ROM on demand, the same way floppy disks are copied today. Using keywords or phrases, you could select a set of books you might want on your bookshelf (actually your CD-ROM jukebox device) and send the order for such a disk to be mastered and mailed to you. Certainly custom-made orders would be more expensive than prepackaged ones but, if marked and priced favorably, should be well within the means of most chemists. Groups of chemists, such as the polymer or materials chemists, could create their own CD-ROMs based on existing volumes already printed. IUPAC could create a CD-ROM of Pure and Applied Chemistry. The list is almost endless.

The last specific topic to be covered in this paper is the area of books, journals, and online chemical information. In the online area, it can be seen from the current usage of scientific and technical databases that the current generation of chemists is not very familiar with computers and chemical information. The costs of searching the chemical literature (including the various charges of connect time, search hits, printouts, and so forth) average well over $100 per hour connected to a host mainframe computer. Compared to browsing through a book, journal, or an issue of the printed *Chemical Abstracts*, this is expensive. Most of the information is not evaluated. The details of the chemical synthesis method or the properties of a molecule or material are either not in the abstract or need to be found by reading the journal article or book chapter. There are high fixed expenses in the creation of the information, due to the fact that abstracting and indexing is and, I believe, will always be a very labor-intensive effort (even with such expected developments as the potentially useful software of Johnson et al.[26]), and there are two ways to recover the costs: either charge a lot of people small sums of money or charge a few people a lot of money. The chemical information industry, for the most part (and there are a few exceptions), has decided to opt for high prices. The results are what most would expect. Few of the hundreds of thousands of chemists referred to in the beginning of this article use computerized databases. Few subscribe to weekly literature searching (Selective Dissemination of Information, SDI) of online databases. The reason is primarily economic. Schools and even many companies cannot afford to have hundreds of chemists spending such large sums of money on literature and related online searching. Hopefully some of the database and vendor companies will begin to experiment with the notion of marketing to the thousands of potential users waiting for reasonably priced products. Years ago many people had personal subscriptions to sections of *Chemical Abstracts*, to journals, and so on. Will the computer revolution in general and CD-ROMs in particular cause history to come full circle? I believe by the year 2000 this is a distinct possibility if there are changes in the way in which vendors market their products. While books will never disappear from the chemist's desk, I think CD-ROM will become the preferred medium of distribution and use in many areas of chemical information. These areas include reference works, collections of books and articles on a particular subject, as well as chemical catalogs of supplies, and software updates.

As for computer-based journals, as stated in Table XIII, publishing in a printed journal is now the accepted means of communication, leading to rewards and promotions. While the means of communication can easily change, the reward situation is quite different. Universities and most other organizations which have peer review use refereed scientific journals very heavily in their evaluation criteria. While I feel my own career has not suffered due to the software and databases I have written and developed, I do not think this is the usual case. Experiments in journals which have a substantial portion of their activity in nonhardcopy form are now starting to appear. One such case was *Tetrahedron Computer Technology* (TCM).[44] This journal died after about 4 years, owing to a variety of technical and nontechnical reasons. A new partly online journal, the *Online Journal of Current Clinical Trials*, is now available.[45,46] This journal has more institutional support than TCM, and so it may make inroads in this area. Additionally there is another new journal, *Protein Science*,[47] which is a biochemistry journal which started publishing in January 1992. *Protein Science* comes with a floppy disk of graphics, which the journal calls "kinemages". In any event, I can see that these experiments, coupled with better delivery mechanisms (for chemistry this is primarily software for the transmission and viewing of graphics), will by the end of the decade lead to a few journals making real headway toward the chemical community having automated journals.

ECONOMIC ISSUES[48]

The recent (and in some cases, still current) recession in a number of developed countries of the world has led to the re-examination of how to sell products. When people do not

fly, airlines lower their fares to fill seats. When people do not buy automobiles, car manufacturers lower the prices to stimulate sales. When hotels experience occupancy rates below 50% and need 65% occupancy to at least breakeven financially, they offer cheap rooms. There are many more examples outside the chemical information area, but it should suffice to state that the Japanese domination of the consumer electronics industry clearly shows that lower prices lead to higher volumes and generally higher profits. As examples in chemical information, I need only to mention such publications as the 11th edition of the *Merck Index*[49] (priced at $30) or the *CRC Handbook of Chemistry and Physics*[50] (priced at $100), now in its 73rd edition. Both of these products sell tens of thousands of copies of every edition.

In chemical information there seems to be a pervasive attitude that information is valuable and prices must be high. Information is no doubt valuable, as evidenced by state and corporate intelligence gathering. In 1978, the total annual online information [scientific and nonscientific (primarily legal) information] revenues were about $40 million.[51] By 1990, this had grown to an annual rate of $690 million. The most successful computer chemistry software company, Molecular Design Ltd (MDL) of San Leandro, CA, in roughly the same period of time has seen revenues go from $0 to $50 million per year. Molecular modeling companies, of which there are at least a half dozen, together probably have total annual revenues of less than current MDL sales. (Sales revenues are based on software sales and exclude software and consulting/consortium groups.) Compared with other industries and especially compared to other areas of the computer industry, these revenues are rather low, and these are not impressive figures.[52] I would hope that companies in this field will begin to experiment with new marketing approaches which will both increase the usage of their products and reach a larger segment of the chemistry population. The Autodesk effort with its HyperChem software is one bright example in a gloomy field. Without a greater volume of usage, it is possible that information will remain a commodity for only a small portion of the chemical community.

SUMMARY

I believe then that there are two main reasons for the low use of computers and computer systems by chemists—cost and ease of use. The economics of chemical information, up to this point in time, made it a tool for a few users and the wealthy in the more developed nations of the world and for the more wealthy companies in those countries. More easy-to-use computer systems will, in the long run, generate more usage. This should, in turn, lower individual computer costs. (The classic chicken and the egg situation.) I believe that with the current and upcoming generation of hardware with the power of an Intel 486 (at 50–100 MHz) or an equivalent UNIX-based Sun, DEC, or Silicon Graphics workstation software can be designed and implemented which will have two main features. The software will be reasonably easy to use (and be easy to remember the next day or week as to how to use a program or database system being accessed) and powerful enough to do the actual job needed to be done. By powerful, I mean that the software will have the necessary "user-friendly" interfaces (graphics, mouse, voice command, and so on) and have some AI (artificial intelligence) capability and knowledge of the subject to assist the end-user in getting his or her job done. However, without close cooperation between software developers and database producers and their end-users this will not happen. Both the software and databases need to be properly designed to meet the actual end-user needs, not the needs which the vendors perceive the users as having. Talking to and, more importantly, listening to the customer or end-user is something the chemical information and related industry will have to come to grips with in the next few years if real and substantial progress is to be made.

Computers and the related technology described in this article hold the potential promise that by the 21st century more chemical information and computer systems will be available to the entire worldwide community. With larger numbers of users, this should allow the costs of the products being developed to be spread across a much wider number of people, leading to higher usage, higher productivity, and lower costs for all computer-related products.

ACKNOWLEDGMENT

The author wishes to acknowledge a number of colleagues who have provided information and comments on this paper. These include Bob Badger (Springer-Verlag), Mike Bowen (ACS), Pierre Buffet (Questel), Chuck Casale (Aberdeen Group), Hideaki Chihara (JAICI), Thomas Clerc (Bern), Harry Collier (Infonortics), Larry Dusold (FDA), Tom Greeves (Daratech), Richard Hong (Hawk Scientific), Sandy Lawson (Beilstein), Dave Lide (CRC), Bill Milne (NIH), Glen Ouchi (Brego), Kris Petteresen (Autodesk), Tom Pierce (Rohm & Haas), Rudy Potenzone (CAS), Craig Shelley (SoftShell), Steve Schultz (Aldrich Chemical), Babu Venkataraghavan (Lederle Labs), Wendy Warr (Wendy Warr & Associates), Gary Wiggins (Indiana University), Joanne Witiak (Rohm & Haas), Chezi Wolman (Hebrew University), and Hugh Woodruff (Merck).

REFERENCES AND NOTES

(1) Based on lectures given at the 10th International Conference on Computers in Chemical Research and Education (ICCCRE), Jerusalem, Israel, July 1992 and at the Second International Conference on Computer Applications to Materials and Molecular Science and Engineering, Yokohama, Japan, Sept 1992.
(2) Planck, M. *Scientific Autobiography and Other Papers*; Williams & Norgate: London, 1950; pp 33–34.
(3) Tom Greeves, Daratech Inc., 140 6th Street, Cambridge, MA 02142. Phone: 617-354-2339.
(4) Casale, Charles T.; Gelin, Bruce R. *Growth and Opportunity in Computational Chemistry*, 1992, The Aberdeen Group, 92 State St., Boston, MA 02109. Phone: (617)723-7890, FAX: (617)723-7897. There is also a more extensive 1989 report published by the same organization entitled *Conflicting Trends in Computational Chemistry*.
(5) G. Ouchi, Brego Research, private communication. R. Venkataraghavan, Lederle Labs, private communication. T. Pierce, Rohm & Haas, private communication. J. Witiak, Rohm & Haas, private communication. H. Woodruff, Merck Labs, private communication. Low usage is defined as 0–25%, moderate usage at 26–49%, high as 50–75%, and very high as 76–100%.
(6) The information on the number of chemists appears to be inconsistent. In 1990 the U.S. Bureau of Labor Statistics reported there were 125 000 chemists in the U.S. From a 1990 survey, the number of ACS members with chemistry degrees number about 137 000. The same survey indicated that the number of ACS members who majored in chemistry was 111 000. In 1992 there were about 145 000 members of the ACS. Lastly, the 1988 Kline report to the ACS stated that there were 213 000 chemists in the U.S. and 137 000 chemical engineers. Among the reasons the Kline report was commissioned by the ACS was to find out how many potential members there would be for ACS membership. Mike Bowen, Director, ACS Membership Division, private communication.
(7) For example, Merck & Co, a drug company of over 35 000 employees (not all of whom are scientists) has in excess of 10 000 PCs in total. Hugh Woodruff, private communication.
(8) Tim Brogan, Pharmaceutical Manufacturers Association, 1100 15th St., NW, Washington, DC 20005. Phone (202)835-3400.
(9) Attempts to find evidence or even an estimate of the number of chemists who have PCs has proven futile. The few firms that have done market surveys in this field or for computer sales in general, such as the Aberdeen group (ref 7), as well as Daratech, Inc., Dataquest, and International Data Corp., had no information, nor did they have any idea where to get such information.

(10) In Oct 1992, STN began to deliver SDI results electronically but only to an STNmail ID. *STNews*, **1992**, *8* (10, 1, North American Edition.

(11) In the PC computer field, a regular flow of books come with computer disks. These disks are intimately related to the contents of the book. For example, a disk of DOS enhancement programs comes with the Dvorak book on DOS and PC performance. Dvorak, J. C.; Anis, N. *Dvorak's Inside Track to DOS and PC Performance*; Osborne-Hill: 1992, Berkeley, CA 94710. $39.95. Phone: (510)549-6600; FAX: (510)549-6603.

(12) Wiggins, G. D. *Chemical Information Sources*; McGraw-Hill Series in Advanced Chemistry; McGraw-Hill: New York, 1991. This book includes a Chemistry Reference Sources Database of 2156 records plus the Pro-Cite search software for IBM PC's. Pro-Cite is available from Personal Bibliographic Software, P.O. Box 4250, Ann Arbor, MI 48106-4250. Phone: (313)996-1580; FAX: (313)996-4672.

(13) PC version of the NIST/EPA/NIH Mass Spectral Database, March 1992 Version. Available from NIST/SRD, Bldg. 221/A320, Gaithersburg, MD 20899 (Phone: (301)975-2208; FAX: (301)926-0416). The price is $1200 for the database or $200 for those who had bought previous versions.

(14) *PC World* **1992**, *Aug*, 210.

(15) L. Dusold, FDA, Washington, DC, private communication.

(16) Lottor, M. Internet Growth. RFC No. 1296; SRI International, Network Information Systems Center: Menlo Park, CA 94025. (Phone: (415)-859-6387; FAX: (415)859-6028; E-mail: NISC@NISC.SRI.COM).

(17) MERIT—Information Center for Internet. MERIT Network, Inc., 2901 Hubbard, Pod-G, University of Michigan, Ann Arbor, MI 48105-2016. Phone: (313)936-3000; e-mail: NSFNET-INFO@MERIT.EDU.

(18) See the recent series on electronic mail in the Oct 1992 issue of *Spectrum*, the monthly publication of the IEEE. The articles in this issue include: (a) Perry, T. S.; Adam, J. A. E-mail: Pervasive and Persuasive. *Spectrum*, **1992**, 22–23; (b) Perry, T. S. E-mail at Work. *Spectrum* **1992**, 24–28; (c) Adam, J. A. Playing on the Net. *Spectrum* **1992**, 29; and (d) Perry, T. S. Forces for Social Change. *Spectrum* **1992**, 30–32.

(19) One list of directories of academic e-mail conferences is available from the Kent State University file server. It was developed by Diane K. Kovacs and is copyrighted. It is available, via Internet, by ftp (file transfer protocol) from KSUVXA.KENT.EDU. The lists of directories are in the LIBRARY directory of the computer and are titled ACADLIST.FILE*x*, where *x* is 1–7, depending on your area of interest (physical sciences, biological sciences, etc.). File7 is the file containing the news groups in the physical sciences and mathematics.

(20) There are new user groups being formed all the time. However, it remains to be seen how long it will take for these to actually take hold and have staying power. Springer-Verlag, the company that distributes the Beilstein Database, started a system for its users of the Beilstein Database on the Compuserve computer system. After 1 year, it was found that usage was too low to continue it. In the past few months user group conferences on HyperChem software, CHARMm software, Organic Chemistry (actually a restart of a system that died a few years ago), Amber software, BioSym software, and SYBYL software have been started up. IUPAC plans to initiate one in the near-future for its many members and affilitates around the world. It will be interesting to see how many of these remain and in what form they remain in 1–2 years.

(21) BIS Strategic Decisions Global Electronic Messaging Service, 1991.

(22) T. Pierce, Rohm & Haas, private communication.

(23) Meadows, A. J.; Buckle, P. Changing Communication Activities in the British Scientific Community. *J. Doc.* **1992**, *48*, 276–290.

(24) The WorkPerfect Corp. makes its WordPerfect manuals available on disk, which can be searched for words using any word processor.

(25) Calem, R. E. Coming Soon: The PC with Ears. *New York Times*, Business Section, Aug 30, 1992, page F9.

(26) Ibison, R.; Johnson, A. P.; Kam, F.; Neville, A. G.; Simpson, R.; Tonnelier, R.; Venczel, T. Automatic Extraction of Chemical Information from the Literature. *Abstracts of the 10th ICCCRE*, Jerusalem, Israel, July 1992, p 25.

(27) McDaniel, Joe R.; Balmuth, Jason R. Kekulé: OCR—Optical Chemical (Structure) Recognition. *J. Chem. Inf. Comput. Sci.*, **1992**, *32*, 373–378.

(28) Use of Chemical Abstracts Service Registry Data in ADP Systems. EPA Order 2880.2, June 30, 1975.

(29) (a) Goebels, L.; Lawson, A. J.; Wisniewski, J. L. AUTONOM: System for Computer Translation of Structural Diagrams into IUPAC-Compatible Names. 2. Nomenclature of Chains and Rings. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 216–225. (b) Wisniewski, J. L. AUTONOM: System for Computer Translation of Structural Diagrams into IUPAC-Compatible Names. 1. General Design. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 314–332.

(30) The figure of 75% refers to chemicals published in the current organic chemistry literature. The program does not handle stereochemistry, charged species, inorganics, peptides, or sugars. A. J. Lawson, private communication.

(31) AUTONOM is an IBM-PC software program and costs $1980 (industry price) or $980 (academic price). It is available from Springer-Verlag Publishers, 175 Fifth Ave., New York, NY 10010. Phone: (212)460-1622; FAX: (212)533-5781.

(32) Barnard, J. M. Draft Specification for Revised Version of the Standard Molecular Data (SMD) Format. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 81–96.

(33) Hall, S. R. The STAR File: A New Format for Electronic Data Transfer and Archiving. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 326–333.

(34) ConSystant is an IBM PC-based DOS program available for $199 from ExoGraphics, P.O. Box 655, West Milford, NJ 07480-0655. Phone: (201)728-0188; FAX: (201)728-0735.

(35) The American Chemical Society established the Chemical Abstracts Service in 1907. The printed Chemical Abstracts and the related Chemical Abstracts Databases are available from CAS, 2540 Olentangy River Rd., P.O. Box 3012, Columbus, OH 43210-0012. Phone: (614)-447-3600; FAX: (614)447-3713. At present there are some 9.5 million abstracts in the CA computer-readable database and about 11.5 million chemical structures with CAS Registry Numbers in the structure file associated with the bibliographic database. There are some 17.5 million names associated with the 11.5 million structures.

(36) VINITI, The All Russian Institute of Scientific and Technical Information, was established in 1952. Since that time it has collected over 31 million source documents in all areas of science (not just chemistry). Of these there are some 11 million abstracts in computer-readable form. Its main publication is *Referativnyi Zhurnal VINITI*. VINITI is located at 20a Uslevitcha St., Moscow 125219, Russia. Phone: 011-7-095-152-6163; FAX: 011-7-095-943-0060. VINITI distributes its products outside of Russia through Access Innovations, Inc., 4314 Mesa Grande S.E., Albuquerque, NM 87108. Phone: (505)265-3591; FAX: (505)-256-1080.

(37) J. Comstock, Head, ACS Books Department, ACS Publications Division, Washington, DC, private communication.

(38) R. Badger, Springer-Verlag, New York, private communication. Gary Wiggins, Chemistry Library, Indiana University, private communication.

(39) D. Johnson, Exxon Corp., private communication. H. Dess, Rutgers University/Chemistry & Physics Library, private communication.

(40) Publication Department, Aldrich Chemical Co., 940 West St. Paul Ave., Milwaukee, WI 53233. Phone: (414)273-3850.

(41) TRIPOS Associates, Inc., 1699 South Hanley Rd., Suite 303, St. Louis, MO 63144. Phone: (314)647-1099 or (800)323-2960; FAX: (314)-647-9241.

(42) Sun Microsystems, 2550 Garcia Ave., Mountain View, CA 94043. Phone: (415)960-1300; FAX: (415)969-9131.

(43) Mitchell, J., Ed. *The CD-ROM Directory 1991*, 5th ed.; TFPL Publishing: London, 1991, Entry 1010, p 169. (Phone: 44-71-251-5522).

(44) *Tetrahedron Computer Methodology* (TCM) was published by Pergamon Press between 1988 and 1992.

(45) The Online Journal of Current Clinical Trials (CCT) is a joint venture of the American Association for the Advancement of Science (AAAS) and the OCLC Online Computer Library Center, Inc. The price is $95 plus monthly telecommunication charges. For further information contact the journal at 1333 H Street, NW, Washington, DC 20005. Phone: (202)326-6446.

(46) On Aug 28, 1992, it was announced (*Science* **1992**, *257*, 1341) that CCT[42] has linked up with the journal *The Lancet* so that *The Lancet* could publish a printed, abridged form of a current CCT article.

(47) This new journal was described by Stu Borman in an article in *Chemical & Engineering News*, Feb 17, 1992, pp 26–27.

(48) Heller, S. R. The Economic Future of Numeric Databases in Chemistry. *Proceedings of the 15th International Online Information Meeting*, London, Dec 1991; Learned Information: Oxford, UK, 1992; pp 47–50.

(49) *The Merck Index*, 11th ed.; Budavari, S., Ed.; Merck & Co. Inc.: Rahway, NJ 07065-0900, 1989 (Phone: (908)594-4904).

(50) *CRC Handbook of Chemistry and Physics*, 73rd ed.; Lide, D., Ed.; CRC Press Inc.: Boca Raton, FL 33431, 1992 (Phone: (407)994-0555; FAX: (407)994-3625).

(51) Williams, M. Highlights of the Online Database Industry. *Proceedings of the National Online Meeting*, New York, May 1992; Learned Information Inc.: Medford, NJ 08055, 1992; pp 1–4 (Phone: (609)-654-6226; FAX: (609)654-4309).

(52) For example, the sales figures of $50 million for MDL sales after 10 years can be compared with that of Lotus Development Corp., which was $53 million in its first year (1983). To date over 9 million copies of Lotus 1-2-3 have been sold. *Barron's* **1992**, *Sept* 14, 12.