Pure & Appl. Chem., Vol. 65, No. 2, pp. 251–256, 1993.
Printed in Great Britain.

Chemical information in the 21st Century

Stephen R. Heller

USDA, ARS, Beltsville, MD 20705-2350 USA

Abstract - The current state of the chemical information in a number of areas is presented. The author then discusses a number of areas in detail and predicts what is likely to be the state of the field by the year 2000. The economics of the chemical information are also briefly described.

INTRODUCTION

This presentation [1] is designed to stimulate discussion of what new technology will be available to chemists, applied to chemistry, and most importantly used by chemists in their everyday activities by the beginning of the next century. This paper has evolved over the past four years, and no doubt, will continue to evolve as new facts arise which produce changes in the habits and activities of chemists.

The use of computers in chemistry has gone from just calculations to cover a very broad area of chemistry. This paper delves into many of these areas and tries to give the current state of the use of computers and what the author believes the use of computers in the field of chemistry will be at the beginning of the 21st century.

BACKGROUND

Max Planck, more widely known in some places for Planck's constant, is, I feel not given proper credit for Planck's Law, which is: "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 wide-spread use of computers in chemistry has clearly been handicapped by a number of factors, a major one of which is the lack of familiarity with this new technology by chemists and managers in the field of chemistry. This is true in all job areas, from academia to government to industry. At present, the routine use of computers in a chemistry lab or office is low. Why is this the case? There are a few hundred thousand chemists in the USA. Combined with those in other developed countries one could estimate some 500,000 chemists as a potential market for computers and computer systems. This paper will examine some possible reasons for the lack of overwhelming use of computers and related computer technology.

Please note that when the qualitative phrases "few" or "low" 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 500,000 users. While selling, over the lifetime of the program, a total of a few hundred or even a few thousand molecular modeling or structure drawing programs is now a major accomplishment in the chemistry, it is a minor event relative to the daily sales word processing, database management, spreadsheets, and other such programs.

COMPUTER AND CHEMICAL INFORMATION ISSUES

Table 1 summarizes both the issues which are to be discussed here as well as the current and predicted level of activities in these areas. Space in this journal does not permit a proper and full analysis of all of these topics. Thus a few representative issues will be mentioned. Tables 2-8 list details for a few of these issues.

Table 1

Issues for Discussion

Topic	Today	2000	Topic	Today	2000
Computer Literacy	Low - Moderate	Moderate - High	Databases	Bibliographic	Numeric Factual
Computers	PC-DOS	Mac/Windows/OS-2	Beilstein	E-V Series being published	E-V Series still being published
Tele-Communications	Moderate Usage	Only the dead don't use INTERNET	CAS	Markush searching ready for testing	Markush searching ready for testing
Windows & OS/2	Released with bugs	Just beginning to meeting specs	Chemical Catalogs	Online Searching	Online Ordering
Interfaces & Difficult	Frightful & Voice Based	Transparent	Chemical Identification	CAS RN & BRN	Chemical Structure
Graphics	Usage in Infancy	Predominant Usage	Molecular Modeling	Few	Some
CD-ROM	Few	Many	Educational Software	Random Usage	Integrated with Textbooks
Chemical Information	Raw	Processed	Publishing	Semi-Electronic	Mostly Electronic
Online Usage for Chemistry	Low	Low	Books	Thought of as probable dinosaurs	Thought of as probable dinosaurs
SDI	Manual or by Post	Electronic	Instruments	Semi-automated	Full-Automated with ISO Data Transfer Standards

Table 2
CD-ROM

Today -

Chemistry CD-ROM products are rare today. Low density (600 MB) CD's.

e.g., Aldrich MSDS, Canadian Toxicity Databases NIST Mass Spectrometry, Kirk-Other Encyclopedia CAS 12th Collective Index

2000 -

New products and high density CD's (6 Billion + Bytes)

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., from NIST)
Most Journals

Table 3
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 4
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 5
Instruments

Today -

Everyone has a computer & everyone has a different computer. No universal interfacing.

2000 -

Everyone has a computer & there are universal protocols for input and output. Data readily shipped to other computers for identification and analysis.

Table 6
Chemical Information

Today -

Most is raw, unprocessed, and un-evaluated. CAS, Beilstein, VINITI - most abstracting done in-house

2000 -

Greater reliance on processed and evaluated data, such as Beilstein, Gmelin, IUPAC data series, CRC Handbooks.

CAS, Beilstein, VINITI - most abstracting done by free-lance workers at home. Articles and abstracts all sent electronically from abstractor to abstracting service.

Table 7
Tele-Communications

Today -

Networks being used routinely by many chemists. BITNET, INTERNET, and other networks used by scientists a few times per week.

2000 -

Networks and e-mail used as an absolute routine. Automatic interfacing between all networks routine. Automatic logins for mail done everyday before scientists comes to work. E-mail re-routed as you travel to meetings, holidays, and home.

Table 8
Publishing

Today -

Journal articles are almost the only socially acceptable form of communication and reward. Some scientific manuscripts submitted in electronic form, but process is neither widespread or practical. Virtually all refereeing done my mail.

2000 -

Printed journals still predominate, but electronic data submissions, electronic journals, software programs are now part of academic, government, and industry reward 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.

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" (according to Max Planck) without the necessary atmosphere and background being part of one's educational upbringing. The current state of education in many parts of the world will make this difficult. However one would hope that in college and graduate school there would be sufficient competence to train the upcoming generation of chemists to familiar with computers. Without an increase in the level of computer literacy the remaining issues are pretty much irrelevant.

Using computers consists of two parts. Writing programs and using programs. Writing programs is really a rather limited issue. A computer is a tool. When a chemist 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 one needs to know what a computer can do for you in the area in which you need to solve a problem. I don't need to be an automotive engineer to know that to get somewhere I need a car to drive there and how to drive a car. The same is true with computers. Understanding what a computer can 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. Chemists don't use computers as an end in themselves. Chemists should use computers a one of many tools to do their job.

Today the use of computers is relatively low. Most chemists use computers for administrative purposes (like writing this manuscript). Using computers for electronic communication is only done by a small, but growing number of chemists. Like any problem, there are reasons for this. Among the reasons are the lack of modems and related dedicated phone lines for user to have. The second problem is the lack of computer accounts on the necessary computer networks (INTERNET, BITNET, Compuserve, etc.). Along with this are the problems in connecting between networks. If I want to telephone someone in another city or country I need only get the phone number from a telephone operator (except for unlisted numbers). With computer networks, there is no phone book, no operator. All numbers (actually computer network addresses) are unlisted. That does present (using a good chemical phrase) a minor energy barrier to solve a problem. However one 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. This is part of computer literacy. This is progress.

Another major problem with computer programs is the downright difficulty in using them. Pacman and Nintendo (the popular video games of the 1980's and early 1990's) never came with manuals. Some manuals seem more designed for weight lifting than explaining how to use a particular computer program. Installing and running programs is a major energy barrier for most people. My preferred philosophy is that if I must read the manual to use the computer program, I probably am better off without it. There is no way a person can become proficient in using a wide variety of programs and remembering what each does and how to perform particular tasks, as well as doing their assigned job as a chemist. As we all know there are few people using their VCR's to record TV shows because they can't figure out how to do it. This even created a market for a device which auomatically 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 USA everyday next to each TV program listing.

Table 1 speaks of today's interfaces as being frightful and difficult. One can only hope and expect that as computers become more powerful and better software engineers graduate and get a job, that the interfaces in the year 2000 will become transparent and even voiced based. One way to accomplish this is through the extended use of graphics in computers. Today the use of high resolution graphics (1024 x 1024 pixels) is low. Color screen size is small (12 - 14 inches) and expensive. By the year 2000 I would expect that every computer will have a 20 inch color monitor with at least 2048 x 2048 resolution, along with a color laser printer or plotter with the same capabilities.

CD-ROM's are another hardware device that is 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. CD-ROM's, which today store about 660 million characters (about 330,000 pages of text), will, by the year 2000, replace many reference books on the chemists'

bench and bookshelf. A few pioneers in this area, such at the Beilstein Institute in Frankfurt Germany, under the leadership of Clemens Jochum, are leading the way to what will clearly be the library of the future. The Beilstein Current Facts CD-ROM has one year of extracted data from the literature, along with a computer chemical structure search system, all neatly tied together. 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-ROM's. 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 one 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 (sorry about that, the US Postal Service will still be in business in the year 2000). Certainly custom made orders would be more expensive than pre-packaged ones, but well within the means of most chemists. Groups of chemists, such as the polymer or materials chemists could create their own CD-ROM's 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 and online chemical information. As can be seen from the current usage of scientific and technical databases, the current generation of chemists are not very familiar with computers and chemical information. The costs of searching the chemical literature are high (an average of $100 or more per online hour of being connected to a host main-frame 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 processed. The details of the chemical synthesis method or the properties of a molecule or material is either not in the abstract or needs to found by reading the journal article or book chapter. With high fixed expenses in the creation of the information 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. Hopefully some 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-ROM's in particular cause history some 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 chemists desk by the 21st century, and Beilstein will, no doubt, still be publishing the 5th series of the *Beilstein Handbook*, 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.

ECONOMIC ISSUES [3]

The recent (and perhaps still current) recession in a number of developed countries of the world has led to the re-invention of how to sell products. When people don't fly, airlines have lowered their fares to fill seats. When people don't buy automobiles, General Motors, Ford, and Chrysler, along with foreign car companies lower the prices to stimulate sales. When hotels have occupancy rates below 50% and need 65% occupancy to at least break-even financially, hotels offer cheap rooms. There are many more examples outside of the chemical information area, but it should suffice to state that the Japanese predominance of the consumer electronics industry clearly shows lower prices lead to higher volumes and generally higher profits. As for examples in chemical information, one need only to mention such publications as the *Merck Index* (priced at $30) or the *CRC Handbook of Chemistry and Physics* (priced at $100), now in its 72nd edition and currently edited by David Lide, a leading authority in scientific databases. Both these products sell in the tens of thousands of copies.

In chemical information there seems to be a pervasive attitude that information is valuable and prices must be high. Information is no doubt valuable. In 1978 the total annual online information (scientific and non-scientific (primarily legal information) revenues were about $40 million [4]. 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, California, in roughly the same period of time has seen revenues go from $0 to about $50 million per year. Molecular modeling companies, of which are at least a half dozen, probably have total annual revenues of less than current MDL sales. Compared with other industries and especially compared to other areas of the computer industry, these revenues are rather low and are not impressive financial numbers. One 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. Without a greater volume of usage it is possible that information will remain a commodity for only a small portion of the chemical community.

CONCLUSION

The economics of chemical information, up to this point in time, have made it a tool for the wealthy in the more developed nations of the world. 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 world-wide community. These additional numbers of users 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.

REFERENCES

[1] Based on a lecture given at the 10th ICCCRE, Jerusalem, Israel, July 1992.

[2] M. Planck, "Scientific Autobiography and Other Papers", Williams & Norgate, London (1950), pages 33-34.

[3] S. Heller, "Proceedings of the 15th International Online Information Meeting, London, December 1991, pages 47 - 50.

[4] M. Williams, "Proceedings of the National Online Meeting, New York, May 1992, pages 1-4.