FEATURE ARTICLES

Evolution of Information Technology and Its Impacts on Chemical Information

RONALD L. WIGINGTON
Chemical Abstracts Service, Columbus, Ohio 43210

Received October 14, 1986

Chemistry has benefited more than any other discipline through the application of information technology over the past several decades. The information technology available for widespread application to chemical information problems is, for the most part, that which has become available for other purposes, and chemical information interests have had little influence on technology's rate of development. Although information technology has improved several orders of magnitude in performance/cost over the past 30–35 years and continues to improve rapidly, many fundamental problems remain in fully realizing information service visions.

INTRODUCTION

Information technology has been both a facilitator of and limitation on the ability to handle and use chemical information. This paper selects a few examples that illustrate key stages of development and provide insight into the development of today's state of the art of chemical information processing and use. As technology continues to provide improved and, especially, cheaper tools, perhaps the visions of the past for instantly and widely available comprehensive chemical information in the workplace useful for all types of endeavors will be achieved.

By "information technology", I mean the combination of techniques and processes that are, or can be, used to record, process, merge, transmit, search, display, and use information in any form. Today, and for the past two decades, attention has been dominated by various electronic technologies—computers, telecommunications, and associated input/output terminal equipment. Many habits are still deeply rooted in technologies of the past—printing on paper and ordinary photography. Other forms of information transfer, such as sound recordings and video programs, are playing an increasing role in education, an information-transfer-intensive activity. The cataloging and retrieval of such materials are serious topics for the library and information community, but are not normally thought of as a part of the central core of "chemical information", and they are not treated here.

THE VISION OF THE KNOWLEDGE WORKER'S WORKSTATION

So many of the aspirations that have driven the development of information support tools for individuals through the application of technology can be traced back to the "urge" expressed in the famous paper[1] by Vannevar Bush, called "As We May Think", published in July 1945. As science turned from the massive effort in support of World War II, the Editor, in introducing the paper, commented about Dr. Bush:

> He urges that men of science should then turn to the massive task of making more accessible our bewildering store of knowledge. For years inventions have extended man's physical powers rather than the powers of his mind. Trip hammers that multiply the fists, microscopes that sharpen the eye, and engines of destruction and detection are new results, but not the end results, of modern science. Now says Dr. Bush, instruments are at hand which, if properly developed, will give man access to and command over the inherited knowledge of the ages.

The need for applying technology to information handling was expressed then much as it might be today. Choosing just one of Bush's sentences to illustrate the point:

> The difficulty seems to be, not such much that we publish unduly in view of the extent and variety of present day interests, but rather that publication has been extended far beyond our present ability to make real use of the record.

Today, this comment can be extended to cover not only the formal and archival record of scientific communication but also the internal recording of knowledge and communications within industrial environments, the growing forms of informal exchanges of information that modern technology enables, and information well beyond core science as it was then known. Bush included in his imagined applications the needs of teachers and merchants, not just scientists.

Bush's version, named the "memex", was expressed in terms of advanced photography, thermionic and cathode ray tubes, and relays. He commented that "The world has arrived at an age of cheap complex devices of great reliability; and something is bound to come of it". This statement was made before the invention of the transistor, which after long and expensive development has finally led to the very cheap VLSI-based technology that today is taken for granted. What he described was a "mechanized private file and library...in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility". It was described as a desk with viewing screens and a keyboard, capable of being operated at a distance. Its memory was large enough that it was not a limit on how much information could be entered and thus would not be a limit on thinking. Means for selecting, indexing, annotation, and linking of related information through mechanical association were included. Even networking was anticipated by the capability to prepare customized material on one memex for insertion into another.

0095-2338/87/1627-0051$01.50/0

As old as it is, Bush's article still represents a very demanding "requirements document" for a modern knowledge worker's workstation, although we are only approaching the ability to build it. Where we remain in jeopardy of being seriously deficient is in assembling the comprehensive and accessible network of mankind's knowledge compatible with the envisioned use of such a workstation. It would be analogous to being parked in a Ferrari on a mountain road navigable only by a 4-wheel-drive Jeep and with no path to reach other roads.

Rather than be too proud of what technology has been able to bring to information handling, as important as it is, we should ask "Why has it taken so long and we have done so little to achieve the full capabilities of and, especially, the environment for, the memex?"

ESSENCE OF CHEMICAL INFORMATION

As stated by Michael Lynch and others,[2] "Chemistry has long enjoyed the reputation of being the best documented branch of science, since much of it is directly concerned with individual chemical molecules". Much of chemical information's conceptual development has been the definition and representation of the language used to describe, as completely and accurately as known by science, the structure and properties of substances. The three-dimensional molecular formula is the language of chemistry. The fact that these formulas can be handled mathematically made chemical information especially suitable for computerized processing. The applications of technology for chemical information transfer have thus been dominated by the need for representing, inputing, transforming, searching, and displaying structures.

In addition to structure handling as such, other areas of emphasis in applying technology to chemical information have been the mechanics of publication, which evolved into what is now called database publishing, and the "full interactive graphics" capabilities for molecular modeling. Even the ordinary text of chemistry has character-set requirements beyond those normally needed for printing nonscientific texts, and special development was necessary.

Mathematically, one may look at structural representations as being highly precise and dense partitioning parameters for chemical knowledge. When questions can be stated in terms of specific substances or unambiguous sets of specific substances, retrievals can have both very high relevance and very high recall. Chemical information that can only be stated in ordinary text terms is as "fuzzy" to deal with as the natural language terms for expressing any other concept.

EARLY MECHANIZATION

In 1946 G. Malcolm Dyson[3] described a cipher intended as a substitute for chemical nomenclature and as a representation of chemical structure that could be handled on punched cards. He presented the ideas for using punched cards to IBM. The result, in 1949, was the construction with Peter Luhn (of IBM) of a machine that would sort free field code cards.[4]

Another early demonstration of mechanizing structure handling was given by William J. Wiswesser[5] at an Industrial Health Conference in Cincinnati in 1952. A deck of 1800 cards, each containing the notation of a single compound and an auxiliary searching code, was processed through an IBM card sorter.

In June 1955, E. J. Crane, then Director of Chemical Abstracts Service (CAS), wrote about research[6] started in 1954 and continuing at that time. Striving for economy of time, effort, and money, CAS was developing several innovations, Crane reported, including new equipment for the following: sorting of index cards, recording and transcribing of index entries by dictation, photocopying for speeding abstracting workflow, microfilming to protect records, monthly index production using a "step" camera, and printing by card-operated machines.

All of these activities extending through the early 1950s were well before the computer age as we now know it. I was first introduced to computers when, in 1955, a program was prepared for me and run on a prototype of the Univac I to calculate certain phenomena for ultra-high-speed electronic signals. If I had not actually seen that machine, it would be difficult for me to appreciate the many orders of magnitude of change since that time. Now, one of the cars I drive probably has more computing power than did the computers of that day.

In September of 1959, G. Malcolm Dyson[7] gave another snapshot of early mechanization of handling chemical information as he described "Research Expansion at Chemical Abstracts Service". He spoke of building a million-record compound file of structural characteristics and of a proposal to produce a permuted title index by using an IBM 704 computer. Ultimately, in 1961, this resulted in the production, using an IBM 1401 computer, of *Chemical Titles* (CT), the first computer-produced journal. CT was followed shortly by the first computer-produced abstract files, *Chemical/Biological Activities* (CBAC) and *Polymer Science and Technology* (POST).

In mid to late 1961 approximately 30 industrial groups were interviewed during a study about chemical notation systems by the National Academy of Sciences.[8] Although one company spoke of forms of mechanical aids dating back to 1949, the typical industrial environment for handling chemical information in 1961 could be characterized as using IBM card-oriented systems, mostly involving card sorters and collators. Some systems used various edge-notched or peek-a-boo cards. Various fragment and indexing codes were entered on the cards. Some organizations spoke of looking forward to obtaining computers. A few already had computers, such as the IBM 704, 705, 1401, CDC 160A, and early Burroughs machines. Files were typically small, although one had already grown to 50 000 compounds, considered huge at that time.

As has always been true, and probably will remain true, various prognosticators expect too much from information technology, especially computers, and are at the same time shortsighted about possibilities. For example, in 1959, Dyson was strongly interested in correlative searching of structure and concept information.[9] Having solved, in principle, how to deal with the structure of compounds, it was felt necessary "to provide an equally detailed record of their properties and behavior". The approach suggested was to code fields of interest, e.g., physiological activity, physical properties, etc., and to search on coincidence of codes for both structural and nonstructural aspects. Dyson felt that, once the principles were established, extension to all fields of interest would be "a matter of routine", a typical understatement resulting from research enthusiasm. The problem, of course, was that knowledge and the way people wanted to access it were not so neatly codable as initial ideas suggested and as was indeed possible for chemical structures. Of course, even for structures, there are "fuzzy" questions.

Dyson, on another occasion, said "One complete notation *cannot* be converted by a computer into another complete notation because there is no intermediate language available which the computer can handle. The only possibility for such an intermediate would be the structural formula or the name, but the computer cannot handle either of the two". As it has turned out, computers can handle both, at least when the

notations are truly complete and unambiguous in the conceptual description of the substance.

At the same time that thought was being given to representing structural notations in such a way as to facilitate computer searching, there was growing interest among publishers of scientific and technical abstracting and indexing journals to standardize the bibliographic information in published citations. The uniform bibliographic systems that resulted also—like molecular structures—lent themselves to computerized storage and retrieval systems.

FIRST WAVE OF COMPUTERIZATION

Based on the experimentation and results of the early 1960s, chemical information handling by computer started two major thrusts which lasted into the early 1970s.

One was the development of "registration", a means to uniquely describe substances and, through assignment of an identifier, to link information from various sources related to a specific substance. The activity of that type with which I am most familiar is the establishment of the CAS Registry system in 1965 which matured throughout that period as the key part of the system supporting the indexing for *Chemical Abstracts*. Subsequently, that system produced a database for searching as an access route to various kinds of information in the literature and in government and industrial files.

The second thrust was the computerization of the publication process, leading to "database publishing" in which information was prepared for delivery via computer-controlled composition and/or computer-readable files.

The information technology of the first wave of chemical information computerization was crude by today's standards, but much was done with it. The computers—from the mid 1960s on—were, typically, IBM 360s of various sizes with limited memories, although computers from other manufacturers were used by some organizations.[10] The file medium was magnetic tape. Disks on mainframes were of the order of 10 Mbytes and were used only for system programs and temporary work files. Output was electromechanical, with limited character sets. In the early part of this period, computer operating systems to handle more than one job at a time were introduced. The computer files, produced as a byproduct of the computer-controlled composition systems, were searched serially on magnetic tape and were used in establishing information centers to provide computer-based information retrieval services.

As the computerization of the publication process proceeded, electromechanical printers gave way to photocomposers which could produce extensive character sets of print quality. Although the use of photocomposers had actually been suggested as early as 1958,[11] it took several years to get them engineered into large-volume production systems.

Computer graphics were developed throughout the 1960s as a result of work related to development of the air defense system, to engineering design support, and to electronic circuit design automation. Particularly notable work was done by Ivan Sutherland of MIT at Lincoln Laboratories (whose thesis on "sketchpad" was the only thesis I have seen referred to in *Fortune* magazine![12]). These graphics techniques were applied to chemical structure depiction and manipulation by Wipke[13] and others and typically ran (during that period) on fairly expensive middle- to large-sized systems.

Chemical information input in the early part of that period used "chemical typewriters" for structures and special conventions on ordinary keyboards for chemical text. Development improved the recording and transfer medium from paper tape and cards, through incrementally recorded magnetic tape, to "key-to-disk" units—a cluster of keyboards supported by a minicomputer.

There was optimism during this period also. In January 1967 Fred Tate[14] made the comment: "By 1969 the entire body of information handled by CAS will go into computer-manipulable form...". Actually, it took until 1975 to get it all into the machine, and even now all of it has been available for searching only the past couple of years.

EMERGENCE OF ONLINE SYSTEMS

Throughout the 1970s, online systems developed, ultimately to become dominant for both database building and information retrieval.

The technological improvements that made this possible were larger internal computer memories, up to several megabytes; faster computers, by more than an order of magnitude; large direct-access storage systems, up to several billion bytes; and computer operating systems which could handle many jobs simultaneously. Searching of both text and structures moved from the serial batch systems to the disk-based, direct-access systems.

There were many uses for minicomputers: as input/output support processors, as communication handlers, and in small, free-standing systems.

Database-management software for large systems matured over this period in the software industry, but it did not contribute in most cases to improvements in chemical information systems. CAS wrote its own database-management system, because no commercially available software was adequate for handling large-volume chemical information. Simple things, such as the inability to handle adequate character sets for scientific text, were impediments. More complex requirements for variability of data elements could not be handled by the commercial software. Information searching systems were specialized, engineered for efficiency in the searching process, severely compromising flexibility to handle complex information structures.

Teletypes and other typewriters began to give way to full screen softcopy terminals. Laser printers became commercially available, at first in the form of large, expensive, high-volume units.

During this period, scientific information searching for industrial, governmental, and many academic purposes shifted away from local centers and centers associated with universities to large centralized online facilities. It is interesting to remember that before this happened, computer-readable files were distributed to users who arranged for their own searching facilities, especially in industry. The shift to the centralized online service centers was caused by the economies of scale of the systems of the time, and the homogenation of information from many sources was a service to the users, bringing some order out of chaos.

TOWARD NETWORKING

The 1980s could be characterized by networking. Although there has long been communications networking applied to computer systems—pioneered by ARPANET for geographically dispersed systems—networking at the information service level is still in the very early stages. The OSI (Open Systems Interconnection) protocols are gradually being sorted out at the lower end of its seven levels, but agreements on the upper, user-oriented levels are nowhere near being reached.

There are two different approaches now being used to bring unity to the user view of networked systems and to help users cope without long and detailed training and experience on many systems.

One might be termed "homogeneous networking", in which the same software is used on all service nodes to create a homogeneous user environment. This is the model being followed by STN International, the network founded by the

American Chemical Society and its German and Japanese associates. While this cannot be universal for all information for all purposes, it is aimed at satisfying the core needs of its users. Extensions to peripheral purposes will ultimately have to use techniques similar to the other approach, described below.

The second approach is application of computing power to aid users in coping with dissimilar systems for which they may not know the detailed protocols and conventions. An example of this approach is CSIN (Chemical Substances Information Network), which has produced a Micro-CSIN workstation.[15] It has been successfully interfaced with several online search systems, presenting a consistent, user-friendly interface to the searcher. The computing power necessary to do this task is a 512K RAM IBM-PC/XT or AT with a 10 MB hard disk—equivalent to the power of typical mainframes of the late 1960s and 1970s. This "intelligent assistant" is very useful and can be implemented either at a remote terminal, such as Micro-CSIN, or at other levels of the computer hierarchy in the service network. At the present stage of development, it converts among equivalent protocols and executes prearranged scripts. This is useful, but it cannot interpret semantic structures and bridge among systems that are fundamentally different in information structure and search strategy. Perhaps advanced artificial intelligence techniques implemented on even more powerful microcomputers will do that some day. But that is straying beyond "history".

The information technology now being applied to information networking can be characterized as "more at less cost". Advanced technology mainframes with many million of bytes of internal memory operate so fast that a significant portion of the operation time is spent in the signal delay in the wiring between components—even on VLSI chips! Large direct-access storage complexes of hundreds of billions of bytes are common. Laser printers are becoming cheaper and more ubiquitous and, together with bit-mapped graphic terminals, provide the hardware potential to distribute useful graphics widely.

In a similar way, minicomputers are as powerful as large mainframes of only a few years ago, and new storage technology is producing cheap, small-size but large-capacity storage. This is leading to a variety of local systems for functions once carried out on large central systems.

The other big potential impact is due to the now ubiquitous personal computer, which itself has become very powerful. Since *Time* magazine crowned "the PC" as the "Man of the Year" a few years ago, a fundamental change has occurred in the computer industry, making it more like the consumer appliance industry. That change has had an impact on what is developed and how it is marketed.

Illustrative of what is now widely available, a survey among STN International users found that the typical workstation can be characterized as having IBM PC and compatible microcomputers; at least 256K bytes of memory, sometimes 512K or more; one or two floppy-disk drives, sometimes augmented by a hard disk; a printer, often an Epson or Thinkjet; no graphics card; no mouse or other direct positioning device.

Unfortunately, for interactions with networked systems, all this computer power is usually used to emulate an ASCII or other "dumb" terminal, and on-site handling of information is crude capture of the search session and simple selection and replay or printout. Crude graphics, when used, are usually through a software package emulating the Tektronix 4010. This illustrates that what is possible and demonstrated in the laboratory or other advanced environments is not widely available until much later—as has been the case throughout the history of computers.

Information technology deficiencies limiting progress so far in this networking era include network-management software; software of various types for full scientific text, low-cost graphics, and user-comfortable interfaces (although the latter are improving); and high telecommunications costs and low data-transmission rates (only up to 9600 baud, but most commonly 1200 or 2400). And, while there is much potential for creating effective distributed systems to serve large groups of people, most dispersed computing power is used in decentralized, independent activities doing incompatible things and with much of the support overhead, once handled in computer centers, being dumped on the users. The lack of, or weaknesses in, widely available graphics systems is a special problem for chemical information users who depend on chemical structures so much.

LESSONS FROM HISTORY

In addition to nostalgia, one of the reasons to look at the past is to gain insight for application to the present and the future. In closing, I will choose for comment some of the many conclusions that could be formed and which affect chemical information handling.

(1) Chemical information has placed very demanding requirements on information technology, and, at the same time, chemistry has benefited more than any other discipline through the application of information technology.

(2) Although information technology has improved several orders of magnitude in performance/cost over the past 30–35 years, and continues to improve rapidly, many fundamental problems remain in fully satisfying information service visions.

(3) Major shifts in relative economics among basic facets of technology, e.g., computation power, memory, and communications, cause design strategies to be unstable. The high obsolescence rate is expensive.

(4) There are fundamental limits to computational speed, such as the speed of light, heat from the energy change required to change from one information state to another, and the limits on minimum size of constructable components, but these limits have not been reached yet. More limiting has been the specification and management of complexity and change which are the heart of the software problem.

(5) Despite what can be demonstrated in the laboratory, the information technology available for widespread application to chemical information problems is that which has become widely available for other purposes. Except for some specialized application software, chemical information interests have little influence on its rate of development.

(6) There has been a long sequence of "mass storage devices", such as data cells, the photodigital store, the 3850 Mass Store, various forms of automated tape libraries, and more recently juke boxes of optical disks, each considered as huge storage capacity in its day. A fundamental deficiency has been the limited number of access ports into and out of such devices and the low bandwidth of access paths. So far none of them has found widespread use. That experience suggests caution in evaluating the effectiveness of CD-ROMs and an expanding variety of digital optical storage media.

(7) For chemical information, hardcopy and softcopy graphics are essential, but are still not widely available except in crude forms.

(8) Intelligent gateways have been shown to be possible at a useful level of capability. However, the design and implementation of them has turned out to be far more difficult than expected.

(9) Effective distributed systems have not yet been created. The most misunderstood aspect of distributed systems is the requirement for system-wide discipline, which is necessary to enable dispersed resources to operate coherently.

(10) We are still far from producing the storehouse of mankind's knowledge, as envisioned by Vannevar Bush, in a form which will be usable in a system of memexes.

(11) The limits on information systems really are what people can and will do and are not due to technology.

Lest these lessons seem too negative and give the wrong impression, please remember that my assignment was to interpret history. If it had been to forecast the future potential for impact of technology on chemical information, I would have been as "bullish" as anyone on what opportunities exist and what could be achieved. The final lesson from history is that reality is achieved much more slowly and with more difficulty than the "vision" initially discloses, but without the hope and expectations and optimism, the real progress would not be attempted or achieved at all. Sometimes too much experience can be a burden.

Therefore, despite history, the vision of the memex is still worth achieving, and giant strides have been made in technology toward it. The challenge is to use that technology effectively. That leaves plenty to work on.

REFERENCES AND NOTES

(1) Bush, V. "As We May Think". *Atlantic Monthly* **1945**, *76*(1), 101-108.
(2) Lynch, M. F.; Harrison, J. M.; Town, W. G.; Ash, J. E. "Computer Handling of Chemical Structure Information". Elsevier: New York, 1971; p 1.
(3) Dyson, G. M. "A New Notation for Organic Chemistry". Royal Institute of Chemistry lecture published jointly with The Chemical Society and the Society of Chemical Industry: London, 1946.
(4) National Academy of Sciences, National Research Council "Survey of Chemical Notation Systems—A Report of the Committee on Modern Methods of Handling Chemical Information". National Academy of Sciences: Washington, DC, 1964; p 180.
(5) *Ibid.*, p 441.
(6) Crane, E. J. "The Chemical Abstracts Service—Good Buy or Good-by". *Chem. Eng. News* **1955**, *33*(26), 2753-2754.
(7) Dyson, G. M. "Research Expansion at Chemical Abstracts Service". *Chem. Eng. News* **1959**, *37*(36), 128-131.
(8) National Academy of Sciences, National Research Council "Survey of Chemical Notation Systems—A Report of the Committee on Modern Methods of Handling Chemical Information". National Academy of Sciences: Washington, DC, 1964; pp 156-157.
(9) Dyson, G. M. "Research Expansion at Chemical Abstracts Service." *Chem. Eng. News* **1959**, *37*(36), 129.
(10) National Academy of Sciences, National Research Council "Survey of Chemical Notation Systems—A Report of the Committee on Modern Methods of Handling Chemical Information". National Academy of Sciences: Washington, DC, 1964; p 189.
(11) Kuney, J. H.; Belknap, R. H.; Lazorchak, B. G. "Progress in Photocomposition". *J. Chem. Doc.* **1961**, *1*, 44-45.
(12) Pfeiffer, J. "Machines That Man Can Talk With". *Fortune* **1964**, May, 153-156, 194-198.
(13) Corey, E. G.; Wipke, W. T. "Computer-Assisted Design of Complex Organic Syntheses". *Science (Washington, D.C.)* **1969**, *166*, 178-192.
(14) Tate, F. A. "Progress toward a Computer-Based Chemical Information System". *Chem. Eng. News* **1967**, *45*(4), 78-90.
(15) Page-Castell, J. A.; Hollister, C. "The Chemical Substance Information Network: User Service Office Evaluation and Feedback". *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 359-364.

ARTICLES

Chemical Information Flow across International Borders: Problems and Solutions†

DALE B. BAKER

Chemical Abstracts Service, Columbus, Ohio 43210

Received August 29, 1986

While we are deeply enmeshed in many of the changes needed to bring the basic functions and parts of information delivery into a highly coordinated and integrated international system, barriers to the uninhibited international flow of scientific and technical information continue to increase. Recently, there have been some new initiatives and approaches to information access and the issues of information flow across national borders. Networks are emerging as alternatives to bureaucratic hierarchies as ways to get things done and as the basic building block for a "New International Chemical Information Order".

INTRODUCTION

During the past four decades, I have seen many changes at Chemical Abstracts Service (CAS). We have evolved from the manual processing and production of print on paper to almost wholly electronic methods of processing and delivering chemical information. The American Chemical Society (ACS) and CAS have, indeed, over the past 2.5 decades, been on the cutting edge of these technological changes. We have been leaders in developing many of them. Being a futurist, and knowing that technological innovation and trends usually snowball, I have become very concerned about how inadequately we are prepared for moving into what I call the "New International Chemical Information Order". This is the critical issue for us to address today. We are, at this time, deeply enmeshed in the changes desired and needed for a New International Chemical Information Order, and I strongly believe that there has not been sufficient attention to, and discussion of, these many challenges.

† Herman Skolnik Award Address, presented at the Symposium on Challenges in Moving toward a New International Chemical Information Order, Division of Chemical Information, 191st National Meeting of the American Chemical Society, New York, NY, April 16, 1986.

THE NEW INFORMATION ORDER

I am most certain that the next millennium will bring a New Information Order. It will have, at least, the following components:

- A continued, rapidly evolutionary movement from the print-on-paper era into an era of electronic information delivery;
- Integration (or blurring) of, as well as reduction of, the discrete, value-added steps in the infor-

0095-2338/87/1627-0055$01.50/0