Chemistry on the Internet: The Library on Your Computer

Gary Wiggins

Indiana University, Chemistry Library, Bloomington, Indiana 47405

Received July 13, 1998

The Internet today has much to offer to chemists and other scientists. A review of the major uses of the Internet and a look at future developments is presented. Special attention is given to electronic journals and the impact that they will have on the scholarly communication process.

INTRODUCTION AND HISTORICAL REVIEW

In a sense, the Internet is a public library for the entire world. As with the service received at a well-staffed library reference desk, the Internet offers a place to ask a question and stand a good chance that it will be answered. Literature now abounds in various formats on the Internet, and the marvelous Internet search engines of recent years help us find it, much as an electronic catalog leads to materials in a traditional library. Though some have questioned the quality of the information obtained on the Internet, there can be no doubt that it is having a tremendous impact on the information-seeking behavior of people all over the world.

The history of the Internet from its beginnings in the late 1960s to the mid 1990s is well covered by Dusold and will not be repeated in this article.[1] Furthermore, Hobbes' Internet Timeline can be consulted for very up-to-date facts on the development of the Internet.[2] It includes information on the growth of Internet hosts, domains, and networks. For chemistry, Varveri provides a useful overview of the Internet's role in chemical information retrieval, broadly defined.[3]

Although chemists as a whole were relatively slow to embrace the Internet, there have been impressive strides in chemical Internet applications and widespread acceptance of the Internet by chemists since the early 1990s. A list of articles on chemistry and the Internet that is maintained by ChemConnect contains only one 1994 article, but 24 published in 1997.[4] Helping to popularize the Internet is the series of articles initiated by Heller in *TrAC; Trends in Analytical Chemistry*.[5] As of this writing, 25 *TrAC* articles had appeared between April 1995 and September 1997. Another source of information is the "Chemistry on the Internet" articles that are published in *The Chemical Educator*'s "Computers in Chemistry" section from 1996 onward.[6] Finally, the American Chemical Society book *The Internet: A Guide for Chemists* collects in one place much of the information needed to effectively utilize the Internet.[7]

CURRENT STATUS OF THE INTERNET

Guides to Internet Chemistry Resources. Commenting on the many guides to chemistry resources on the Internet, Rzepa has recently written, "Whilst 'collections' of bookmarks for chemistry materials on the Web continue to vary in quality, and many fail to really review, structure or categorize the materials, there are signs that more resources are being devoted to this important area".[8] There have been numerous attempts to compile the available chemical information sources on the Internet, some of them resulting in large databases or lists of resources. An example is the American Chemical Society's ChemCenter (http://www.ChemCenter.org), which claims to be a "one-stop Website for chemical professionals in industry, academe, and government worldwide", while providing "resources for educators, students and individuals who want reliable, accurate information about the chemistry-related sciences and the American Chemical Society (ACS)—the world's largest scientific society". Perhaps the most nearly comprehensive source is Winter's ChemDex, which had 3623 entries as of August 14, 1998.[9] A version called ChemDex Plus now is found among the offerings at another "one-stop shopping" place: ChemWeb (http://ChemWeb.com).

Many directories exist on the Web, but some are not of very high quality or not kept up to date. One of the more frustrating things about the use of the Internet today is the tendency of the resources to move or simply cease to exist, resulting in a "File (or Server) Not Found" message. The sheer number of chemistry sites on the Internet has led some to develop very selective compilations of the better Web sites. Huber's "Chemistry Resources on the Internet" has 42 entries,[10] while another list, targeted at special librarians, is the "Chemistry" section of *Internet Tools of the Profession*, with 79 links.[11] Even the venerable *Encyclopaedia Britannica* has produced an annotated collection of rated Internet resources, the Britannica Internet Guide.[12]

Finding properties and graphical depictions of chemical substances has been considerably enhanced by CambridgeSoft's ChemFinder.[13] Hundreds of sites can be searched by chemical name, CAS Number, molecular formula, or molecular weight. The result is presented both as handbook data and links to other Internet sources, as seen in Figure 1.

IP (Internet Protocol) and TCP/IP. The Internet runs today on v. 4 of the Internet Protocol (IP), the backbone of TCP/IP (Transmission Control Protocol/Internet Protocol) networking.[14] When computers talk to each other via TCP/IP, the connected computers may be running incompatible operating systems, but the TCP/IP suite enables the user to perform data transfer among dissimilar machines using a common procedure. IP version 4 has no built-in security and is running out of address space, due to its limited 32-bit

10.1021/ci980125b CCC: $15.00
Published on Web 10/22/1998

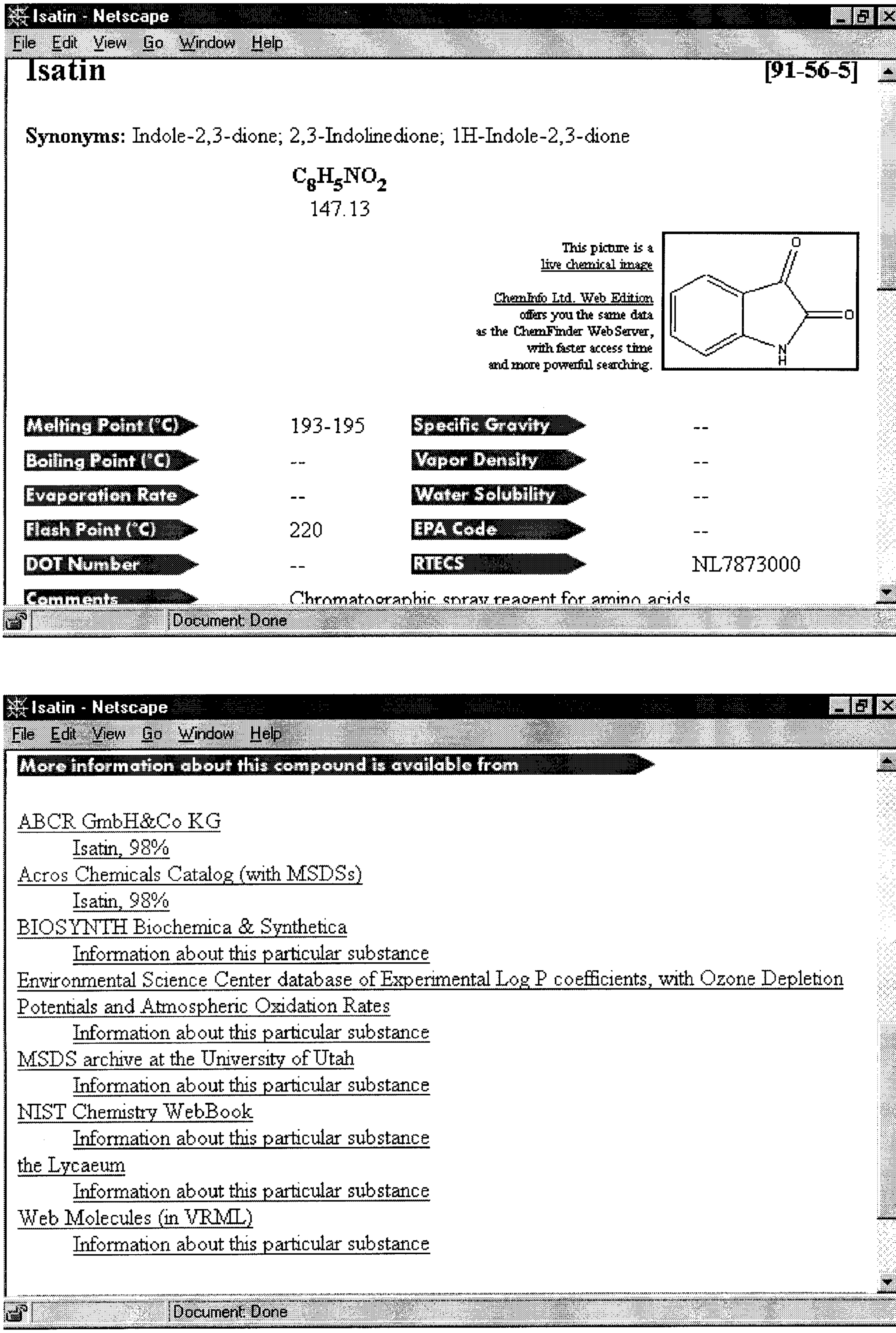

Figure 1. Entry for isatin found with CS ChemFinder.

address length. Hence, the next generation of the Internet will run on IPv6 and will use the Aggregatable Global Unicast Address Format (a 128-bit addressing scheme). The aggregatable address format is divided into three major sections: the public routing topology prefix (48 bits), the site topology (16 bits), and the interface identifier (64 bits).

Those who use personal computers in their places of employment usually have some experience with networks. Within an organization, networking helps to share resources or information from department to department, building to building, or from one geographic site of the organization to another. The Internet has expanded for the computer user the range of such sharing to the whole world, not just the boundaries established by the organization where a chemist may be employed.

FTP and Telnet. There are parallels between stand alone or local area network (LAN) uses of computers and the activities possible on the Internet. Whether a PC, Macintosh, or UNIX workstation is employed, certain functions are performed routinely by all who use computers. Copying of files and programs is often done on the personal computer or workstation, and that is possible on the Internet as well. However, on the Internet, files may be copied from thousands of repositories of textual or graphical information or computer software, often at no cost.

The program that enables copying on the Internet is FTP (File Transfer Protocol), part of a suite of programs that conform to the TCP/IP standard. There are a number of excellent free graphical FTP programs that greatly simplify the process of file transfer, and the main Internet browsers have the FTP capability built in.

TCP/IP also includes a facility for connecting to a remote computer and interacting in real time in the same manner as with a modem over telephone lines but at much higher speeds. The remote login software is called "telnet". Using only telnet, it is possible to access some free databases, for example, library catalogs. The telnet facility is still widely used to access library catalogs in some of the greatest libraries of the world, but more and more libraries are switching to Web interfaces or client/server architecture. The speed and relatively low cost of Internet access have driven most commercial vendors of online database searching to offer telnet access as an option in recent years. Likewise, most of the older online search services have provided at least an alternative Web access mechanism by which to search their databases, and some companies, for example, Ovid and SilverPlatter, now offer Web-based searching options.

Some free chemistry databases are searchable on the Web. Maizell, in a new book on chemical information sources, compares Web databases unfavorably to commercial proprietary online services and systems, which he characterizes as "... mature and well developed, yet still capable of new and innovative approaches".[15] Free specialized databases in chemistry are not very numerous on the Internet, compared to those in the related fields of molecular biology, biotechnology, and other biosciences. The 1998 database issue of *Nucleic Acids Research* (volume 26 number 1) is the fifth such annual compilation of factual biological databases. It contains descriptions of 102 databases, an increase of 27 over the 1997 database issue. Examples of Web databases with rich chemical content are the structure searchable versions of the National Library of Medicine's Hazardous Substances Databank and the National Cancer Institute's NCI-3D database, which are available at http://chem.sis.nlm.nih.gov. The NIST Chemistry WebBook, which includes the NIST Standard Reference Database Number 69 (March 1998 Release), was still free as of this writing at http://webbook.nist.gov/chemistry/. However, it is likely that this and other chemistry databases that debuted as free resources on the Internet will eventually have charges associated with them. For example, Beilstein Abstracts, originally known as the free database NetFire, is now available on Elsevier Science's ChemWeb (http://ChemWeb.com). A new company which quickly established a reputation via the Internet for its innovative property prediction databases and other software is Advanced Chemistry Development (http://www.acdlabs.com).

Electronic Mail, Mailing Lists, and News Groups. A use of the Internet that chemists have found extremely valuable is electronic mail. Whether corresponding directly with a colleague halfway around the world or accessing the postings of various newsgroups or discussion lists, TCP/IP's e-mail capability has enabled the rapid exchange of information on a multitude of topics.

One of the more rewarding uses of electronic mail is to join a discussion group composed of colleagues who share a common interest. There are discussion groups on topics as diverse as computational chemistry (CCL), chemical structure indexing (CHEMIND-L), and corrosion (CORROS-L). All messages mailed to the list server are automatically sent to each subscriber. Such resources often post conference announcements, job opportunities, calls for help with an information need, etc. Furthermore, the archives of the mailing lists are rich searchable databases that can be mined for marketing information as well as factual data.

Van der Valk addresses the use of mailing lists and newsgroups, including how to subscribe to them, how to post messages and where to get specific information.[16] There is a large number of mailing lists or discussion groups of interest to chemists, over 60 of which are found in a compilation maintained by Boulez.[17] Cheng and Zhu have developed a tool to count the number of messages collectively posted to 60 chemistry groups in their Searchable Chemistry Mailing and Newsgroup Database.[18] Between May 21, 1996 and May 5, 1998, a total of 131 250 messages were exchanged. The greatest activity among the groups listed was on CHEMED-L (10 182) and sci.chem (34 174).

MIME. E-mail packages that are MIME-aware have an advantage over the text-only e-mail software still used by many. MIME stands for "Multipurpose Internet Mail Extensions", a standard that permits the transmission in an e-mail message of multipart, multimedia, and binary data. Thus, images, sound, wordprocessing files, and those with other data formats as well as plain text can be transmitted leaving it to the MIME-aware e-mail program on the receiving end to determine how to handle the various parts. A new standard, the MIME Encapsulation of Aggregate HTML Documents (MHTML), would allow complex Web pages or even entire Web sites to be attached to an e-mail message (http://www.ietf.org/ids.by.wg/mhtml.html). Another proposed standard that will protect Internet e-mail from eavesdropping and tampering is S/MIME, Secure Multipurpose Internet Mail Extensions, a data encryption standard.[19] Privacy, integrity of the message sent, and proof of origin are all guaranteed by S/MIME. It is expected to emerge as the standard for industry and other organizations, while PGP (Pretty Good Protection) will continue to be used for encoding messages by individuals.

Pine is a popular MIME-capable mailer that is freely available on the Internet, but virtually all proprietary e-mail systems, such as Eudora, now provide MIME translation. MIME in effect replaces the UUENCODE program that has been used for encoding binary files so that they can be transmitted over the Internet. MIME's encoding, known as "Base64", can withstand the types of message transformation caused by certain e-mail gateways.

Rzepa and others were quick to realize that MIME provides a way to transmit chemical objects over the Internet. They submitted a draft of a chemical MIME standard to the Internet Engineering Task Force (IETF), a body that has great influence on the standards adopted for the Internet. The IETF rejected the proposal to establish a chemical MIME standard. According to Rzepa, the IETF reasoned that if a MIME-encoded message of type "text/plain" sends text over the Internet and "image/gif" sends an image, then "chemical/*" ought to send a chemical substance! Nevertheless, the MIME file designation "chemical/*" has found widespread acceptance on the Internet, and Nelson's early test page for the various browser helper applications and plug-in programs that interpret chemical objects is still available at Lawrence Livermore National Laboratory.[20]

Early attempts were made to hold electronic conferences by e-mail. The first electronic conference in chemistry of real importance is thought to be INCINC '94, the first International Chemometrics InterNet Conference.[21] The organizers of INCINC '94 pointed out that the reolutionary browser Mosaic did not even exist when they began to plan the conference. This was quickly followed by others (see at: http://www.indiana.edu/~cheminfo/ca_cispd.html), and now we have the innovative series of Virtual Electronic Lectures found at ChemWeb (http://ChemWeb.com). These take full advantage of the chemical MIME and other novel chemical applications developed for the World Wide Web. Most recently ChemInt'98, the Chemistry & the Internet conference held in Irvine, CA on September 13th–15th, 1998, featured via ChemWeb a live virtual broadcast of a selection of the lectures presented.

THE WORLD WIDE WEB (WWW)

Almost all large and many small chemical companies now have WWW "home pages" on the net to promote their products and services. WWW's "forms" capability enables input by a remote user to everything from database search engines to order forms for a company's products. The development of the secure sockets layer (SSL) standard of the Internet transmission protocols is giving rise to a multitude of business transactions that can be conducted on the Web, and there is an inexorable move of many tasks toward the Web that in the past had to be performed using proprietary software and databases.

Markup Languages. In the past few years, the World Wide Web (WWW) and the browsers that make the Internet so much easier to use and more visually pleasing than before have really captivated Internet users. Microsoft's Internet Explorer and Netscape's Navigator are battling for the lead in the browser market. Netscape announced in the Spring of 1998 that it would release the source code for its Communicator 5.0 client program that includes the browser, e-mail, network news readers, and many other features. Nevertheless, there was still room in 1998 for the development of yet another browser, the lean, but capable Opera.[22,23] The key to success in the browser wars will be the implementation of features that enable secure electronic commerce, thereby encouraging companies to exchange business data over the Internet.

WWW employs links among various resources on the Internet by means of the hypertext transfer protocol (HTTP). The links, coded as anchors in Hypertext Mark-up Language (HTML), may connect to traditional sources such as FTP or telnet sites, but they may also lead to resources that are visually oriented, including even images or film clips, perhaps associated with sound.

Version 4.0 of HTML, the hypertext markup language that has enabled the riches of the Web to be developed, is now set to be adopted. Bremser includes a brief history of HTML from its beginnings in 1990.[24] He notes that version 4.0 of HTML will enable many of the features needed to move the Web much closer to a publishing medium, especially the new <OBJECT> element, which incorporates extensions such as Dynamic HTML, Cascading Style Sheets (CSS), and Extensible Markup Language (XML).

The specification for Cascading Style Sheets (CSS), developed under the auspices of the World Wide Web Consortium (W3C), defines elements of style and the layout of a document, thus providing precise control over the presentation of Web pages.[25] Up to now, the look of a Web page could vary considerably, depending on the Web browser used. CSS2, a part of Dynamic HTML (DHTML), offers enhanced page printing, support for absolute positioning and layered elements, and the possibility of downloadable fonts.[26]

Bremser feels that HTML 4.0 will standardize the use of tools as diverse as Java applets, ActiveX, and Shockwave movies.[27] Nevertheless, the implementations of HTML 4.0 by both Microsoft and Netscape have been criticized for inconsistent and incompatible style sheets. Furthermore, DHTML requires users to have Version 4.0 browsers, thus forcing those who want to take advantage of these techniques to upgrade. Hope is on the horizon, and the authors of "Weaving a Better Web" paint a picture of the Web at the end of the millennium that uses open standards with DHTML, style sheets, a document object model, and XML in conjunction with HTML 4.0 to solve the problems commonly experienced with HTML today.[28]

CML (Chemical Markup Language) and XML (Extensible Markup Language). For a period of time, there was a great deal of excitement in the chemical Web community about the visualization possibilities produced by suitable helper or plug-in applications in combination with browsers. Despite the failure of the Internet Engineering Task Force to adopt the chemical MIME type, many persisted in developing innovative ways of making chemistry visible on the Web using the chemical MIME designation. Unfortunately, some of those applications proved to be incompatible with the latest generation of Web browsers.

An innovation that should gain wide acceptance is the Chemical Markup Language (CML), developed by Peter Murray-Rust.[29] Murray-Rust has written CML in XML and even developed JUMBO, the Java Universal Molecular (or Markup) BrOwser.[30] JUMBO is a set of Java classes which act either as an application or applet for viewing XML files, especially CML. It reads numerous chemical/* MIME types.

Mackenzie notes that for XML to work, common-interest groups will have to agree on a shared vocabulary. CML is singled out as one such example because "CML enables the viewer's computer to meld information stored in separate databases into a seamlessly linked, interactive document, made to order".[31]

XML, or Extensible Markup Language, is actually a specification for the implementation of a syntax for specific subjects.[32] It has been characterized as a thinner version of SGML, the Standard Generalized Markup Language that has been in use in the publishing industry for a number of years. Publishers have been unable to utilize SGML in a very effective manner to serve various roles, for example, simultaneous delivery of journal information in multiple formats that will facilitate the ability of users to quickly locate information in the journals. XML promises to do just that.[33] The basic concept of XML is that content and presentation must remain separate. The stylesheets will interpret XML tags using eXtensible Style Language, whereas the Xlink feature of XML provides a way to link to particular sections of a document.[34,35] It is likely that each discipline will provide its own specific tools to manage domain-specific information.[36]

Search Engines. The many free pieces of information, programs, and databases on the Internet constitute a resource comparable in some ways to the world's finest libraries. However, as anyone who has used the Internet extensively will attest, it is a library sorely in need of cataloging. Tools have been introduced to assist the user in finding needed information or software, but at this point in time, they lack the sophisticated search techniques found in a well-developed library catalog or commercial database of information. The early search tools such as ARCHIE (for searching FTP archives) and VERONICA (for searching menus at gopher sites) made the use of the Internet considerably easier, but those have largely been displaced by global search engines.

One of the most important developments in the history of the Internet was the appearance of the powerful search engines capable of indexing millions of pages of full-text information. Among them are AltaVista and HotBot (the two with the largest coverage) but also Lycos, Excite, Infoseek, and the newest to appear, Northern Light (which is currently the third largest).[37] Using robots to prowl the Internet and bring back for indexing huge volumes of information, the Internet search engines have given users both enormous delight at the wealth of information available and great frustration at their relative lack of ability to fine-tune a search. Lawrence and Giles found that none of the search engines covered more than about 1/3 of the available Web pages. In fact, the range of pages searched by the various engines is only 3–34%.[38]

A good way to keep up with the various changes and new features of the search engines is to consult Mecklermedia's Search Engine Watch.[39] It includes "A Webmaster's Guide To Search Engines", which explains how search engines find and rank Web pages, with an emphasis on what webmasters can do to improve how search engines list their Web sites. Another up-to-date comparison of the features of 10 of the Internet search engines is maintained in the "Quick Reference Guide to Search Engine Syntax".[40] Notess maintains a list that compares how many pages are covered by the search engines (http://imt.net/~notess/search/).

Digital Equipment Corporation claims that its AltaVista, which first appeared in 1995, is able to find any word in any document on the Web.[41] All pages manually submitted to AltaVista as well as pages brought back that day by Digital's robot, Scooter, are indexed each day and added to the database. Examples of recent features incorporated in AltaVista are

- language tags—25 choices for limiting a search
- multiple network search sites—5 choices
- search refinement capabilities
- date range searching.

One impressive feature of the Alta Vista system is the capability to translate the search results from English into Spanish, French, Portuguese, German, and Italian, an instantaneous free feature added in 1998.

Despite the capabilities of the search engines, they have come under criticism for a number of reasons. *Scientific American* devoted a large portion of the March 1997 issue to a special report, "The Internet: Fulfilling the Promise".[42] Many good recommendations on making Internet information easier to locate, more nearly comprehensive, more secure, and universally accessible are found in the articles. Although 1997 saw some improvement in Boolean and adjacency searching, truncation capabilities commonly found in commercial online services are still rudimentary in most Internet search engines. In early 1998, it was not possible to sort the results (by site or by date, for example), nor was it possible to combine the results of a previous search with a new one. Notess sees the need for a number of improvements on the Web search engines, among them

- improved Web interfaces to online databanks
- larger, more comprehensive Web page databases
- more search and sort features on the Internet search engines
- ability to track and combine search sets
- more quality subject-specific search engines and subject directory combinations (including classified subject directories and descriptive and evaluative reviews of contents).[43]

Lynch notes that the Web indexing programs do not really have access to the content of the databases that underlie some of the most important Web sites, such as library catalogs or commercial electronic journals.[44] Thus, the search engines do not see most of the fulltext databases, the bibliographic records, and bit-mapped images that have been created over the last 3–4 decades and are accessible through library Web sites or commercial vendors of databases or serial aggregators. The same can be said of the many smaller legacy databases that are starting to appear on the Internet through the use of various development software, such as Cold Fusion.[45] These allow the searching of databases created with tools such as Microsoft Access to be searched on the Web. A list of such tools can be found in the June 1998 issue of *BYTE*. Web search capabilities are also beginning to appear from the personal bibliographic software producers. For instance, Research Information Systems' Reference Web Poster allows the posting and searching of databases created with Reference Manager, ProCite, and EndNote. (For information, see http://www.risinc.com.)

```
Date: Mon, 4 May 1998 11:06:47 -0600 (MDT)
From: uncover@csi.carl.org
To: wiggins@indiana.edu
Subject: UnCover Reveal - Journal of chemical information and computer sci

Article availability and price:
Service charge:    $   10.00
Copyright Fee:     $   14.25
------------------------------
Total Article Cost: $   24.25

 You may order articles from UnCover Reveal by faxing the citation
(including the UnCover Order Number) and your Profile Number
to 303-758-7547.  Your profile must have complete payment information.

You may also order articles by sending a REPLY message using your e-mail
editor.  Type the word 'ORDER' anywhere on the line that displays the
'UnCover # (the bottom line of each entry in the table of contents).
The article will be faxed to you, usually within 24 hours.
Charges will be made against the account number stored in your UnCover
profile.
JT Journal of chemical information and computer sci
DA MAR 01 1998 v 38 n 2
PG 85
AU Schultz, John L.
AU Wilks, Edward S.
TI Dendritic and Star Polymers: Classification, Nomenclature, Structure
  Representation, and Registration in the DuPont SCION Database.
SI 0095-2338(19980301)38:2L.85:DSPC;1-
>>     Profile #: 1107859        UnCover #:  251,094,094,143

JT Journal of chemical information and computer sci
DA MAR 01 1998 v 38 n 2
PG 100
AU Carabedian, Michel
AU Dubois, Jacques-Emile
TI Large Virtual Enhancement of a 13C NMR Database. A Structural Crowning
  Extrapolation Method Enabling Spectral Data Transfer.
SI 0095-2338(19980301)38:2L.100:LVE1;1-
>>     Profile #: 1107859        UnCover #:  251,094,094,153

[Contents deleted from this point]
```

Figure 2. Selected output from an UnCover reveal e-mail delivery of the contents of a recent *JCICS* issue.

Despite the criticisms, the capabilities of the Internet search engines are already quite impressive and will undoubtedly rival the commercial counterparts of the online search vendors in the coming years, as more attention is given to increasing the relevance (precision) of the searches.

ELECTRONIC JOURNALS

Getting access to information about the contents of journal issues published in the decade of the 1990s is easy now that the CARL UnCover service offers free access on the Internet to their UnCover database.[46] For a small fee, it is possible to subscribe to UnCover's REVEAL service that brings customized SDI current awareness service to the masses. The breadth of the subject coverage in UnCover can be judged from the REVEAL printout in Figure 2. There are many avenues besides UnCover for obtaining copies of journal tables of contents on the Web. Some of the newer possibilities are from publishers or serial vendors (aggregators), among them

- SwetScan Service
- Elsevier Contents Direct
- Springer-Verlag Link-Alert Service
- Institute of Physics Current Awareness Service
- Academic Press IDEAL-Alert Service.

As seen in Figure 2, it is possible to order a copy of the document from CARL. However, the delivery mechanism is either by postal mail or telefacsimile. The Web offers an opportunity to move from the traditionally slow document delivery service offered by the photocopy machine or fax to the almost instantaneous retrieval of a copy of an article on the personal computer. A large number of journals that have for many years existed in print form only are now available on the Internet, and some new ones that were designed for the Internet are beginning to appear. Those take advantage of features found only on the Web.

The imaging aspects of the World Wide Web have caused great excitement among chemists. Not only words or numeric data can be moved across the Internet but also representations universally understood by chemists: 3-D chemical structural diagrams and depictions. Of course, it is not the pretty colored pictures themselves that are being transmitted but renderings of molecules or other chemical objects in computer code. The World Wide Web has made it possible to share such data on a truly global scale by using appropriate viewing software, and chemists have been very generous in providing free Internet software tools to help visualize chemical information concepts, especially the depiction of molecules.[47] With these developments, the introduction of Internet electronic chemistry journals was enabled, and the first commercial Internet chemistry journal, the *Journal of Molecular Modeling*, made its debut on January 1, 1995.[48]

Even the process of bringing the traditional printed journal article to publication-ready status has felt the impact of the Internet. Manuscripts are often transmitted by e-mail, and the time between submission of the manuscript and the completion of the peer review process has been dramatically shortened. In the case of the American Chemical Society journals, the texts of new articles on the Internet are available before their appearance in the printed journals under the ACS ASAP (As Soon As Publishable) initiative. Springer-Verlag has a similar program with their Online-First project, and other publishers are sure to follow as the competition for manuscripts in the digital age becomes more intense. The review process itself is experiencing the beginnings of a revolution, as witnessed by the open review policy of the *Journal of Chemical Education* Internet edition. The *JCE* review guide issues this invitation to all who read it on the Internet, “You are welcome, even encouraged, to submit a review [of articles under consideration for publication].”[49]

Maizell notes that Chemical Abstracts Service began to cover Internet publications in April 1995, but that during the first year, only 197 papers from 14 online journals were included in *Chemical Abstracts*.[50] As of August 1998, Chemical Abstracts Service monitors 31 online-only journals.[51] There are many interesting chemistry resources on the Internet, but it is the publication of true scholarly journals that will inevitably tip the balance of the publishing activity and use of the Internet by chemists toward this new medium. A flowering of many electronic journals of interest to chemists has been seen in 1998. By January 1998, over one-half of the journals taken in paper format in the Indiana University Chemistry Library had full-text electronic counterparts on the Web. A printout from these products is almost indistinguishable from the printed counterparts, and there were a number of value-added features (for example, fulltext searching capabilities, hyperlinks among footnotes and figures within the articles, and links to bibliographic databases for abstracts of articles in the bibliographies) that made the Internet versions of the journals quite attractive. On the other hand, for truly innovative content options in electronic journal publishing, one must look at independent journals, such as the *Internet Journal of Chemistry*.[52] The user has a degree of control over the layout and features in the *IJC* that is unmatched among the journals of the larger commercial publishers, with everything from the background colors to position of the references being selectable by the reader.

Standards. What is largely missing from the conglomeration of electronic journal offerings on the Web today is standardization. There are efforts underway to standardize some of the basic publishing tools needed for widespread interoperability among Web publishers. For example, the STIX project, set up by the STIPUB group of publishers and spearheaded by Elsevier, aims to formulate a collection of characters used in scientific and technical publishing. The codes will be presented to the Unicode consortium for the next revision of the Unicode character set, and a complete set of fonts will be commissioned.

The publishing industry is now working on the concept of digital object identifiers.[53] The DOI has been equated by some publishers to the PII (Publisher Item Identifier). The PII is a 17-character string that consists of

- one character to indicate source publication type
- the identification code (ISSN or ISBN) of the publication type (serial or book) to which the publication item is primarily assigned
- (in the case of serials only) the calendar year (final two digits) of the date of assignment (this is not necessarily identical to the cover date)
- a number unique to the publication item within the publication type
- a check digit.[54]

Also known as the International Standard Digital Identifier, the DOI is already being used by a number of publishers to identify not only the journal article but, in some cases, the abstract and other parts of the article. Furthermore, it can provide a link to the full text of the articles from abstracting and indexing databases such as Chemical Abstracts or the National Library of Medicine databases. The DOI is ultimately a tool of commerce since it facilitates the linking of users of the materials to the rights holders. Even an individual who wants to obtain a DOI for objects placed on the Web will be able to do so.

DOIs are expected to help with the problem of the impermanence of URLs and to identify any portion of a work that is desired. Publishers speak of *granularity* in defining how specific the identifier is to be. For example, it may be sufficient for most purposes to assign a code at the article or book level, but, in other cases, it may be desirable to have multiple codes for individual chapters in books or other works. A code that defines chapters in books might ultimately allow the compilation on demand of a course pack of selected chapters from various books which are used in teaching a particular course.

Other identifiers in common use in the publishing trade are the ISSN (International Standard Serials Number) that is a component of the SICI code (Serial Item and Contribution Identifier) and the ISBN (International Standard Book Number), with a corresponding BICI. Examples of the SICI can be seen in the second reference in Figure 2, which is reproduced with explanatory notes in Figure 3.

Up one level from the ISSN and the ISBN is the International Standard Work Code (ISWC). Attached to that could be an Interested Party (IP), perhaps an author, who may not own the work at the point of publication but could

```
        Item              | Contri-   | Control
        Segment           | bution    | Segment
                          | Segment   |
 0095-2338(19980301)38:2L.100:LVE1;1-  ←Version No. &
    |          |        |  |      |      Check Character
   ISSN   Chronology    |  |    Contribution
                        | Location Abbreviation
                   Enumeration
```

Figure 3. Explanation of the SICI code.

still be associated with it. Collectively, all of these acronyms are known as International Standard Information Identifiers (ISII).

Metadata. Metadata is structured information about information and is coded at the beginning of an HTML document.[55] It bears some similarities to descriptive and subject cataloging in the library world, but it is not restricted to bibliographic data. Metadata may tell a great deal about an object on the Web—what it is, whether it exists, and if so, how it may be obtained. Such information may be picked up and indexed by the search engines, thus making retrieval of information much more reliable.

The Dublin Core of Metadata Elements has gained popularity in the past 3 years. The Dublin Core originally had 13 elements: subject, author, language, date, title, format, publisher, type, rights, identifiers, coverage, description, and other contributors. Author was later labeled creator, and source and subject were added, bringing the total to 15 elements.[56] Rzepa has proposed a prototype Dublin Core for chemistry.[57]

FUTURE DIRECTIONS

TCP/IP has no quality-of-service (QoS) attributes, thus making it incapable of reserving bandwidth for applications that require high-quality data transmission. Every bit of information sent on the Internet today competes with every other bit as if each were equally important. Dern and Mace point out that the Internet really needs network scalability (128-bit addressing), multicasting (the transmission of only one copy of information out to the network which is then copied as needed), security, and, especially, quality of service guarantees.[58]

In a recent book, Rosenfeld and Morville note that the Web suffers from problems of not being able to find the information, poor graphic design and layout, gratuitous use of bells and whistles, inappropriate tone, designer centeredness, and lack of attention to detail.[59] Furthermore, they note that all too often a site is "under construction" and never seems to be finished. The authors point to an emerging profession, the "information architect", whose function is to

- organize the patterns inherent in data, making the complex clear
- create the structure or maps of information which allow others to find their personal paths to knowledge
- focus upon clarity, human understanding and the science of the organization of information.

A general criticism of the Web is that it is necessary to use the latest version of the Web browsers in order to access the more sophisticated features of some sites. Those features place more demand on an already inadequate Internet infrastructure. Network bandwidth demands of the Internet are said to be growing much faster than the capacity of the Internet backbones, and traffic is now doubling every 3−6 months.

There is a widespread belief that the next generation of the Internet will alleviate many of the problems that are felt in the current environment by providing much greater access speeds. Gigabit Ethernet is likely to displace competing network technologies, such as Asynchronous Transfer Mode (ATM) and FDDI (Fiber Distributed Data Interface) in the corporate environment, as asymmetrical digital subscriber line (ASDL) and cable modems vie for the home market. Nevertheless, Van Houweling thinks that "The growth in demand for Internet-based service will continue to outstrip the Internet industry's ability to provide high quality service".[60] He describes the current initiatives in the United States that are designed to at least test the feasibility of improving the situation. Among those are Internet2 (http://www.internet2.edu), the Next Generation Internet (NGI) program (http://www.ngi.gov), and the very high-speed backbone network (vBNS) (http://www.vbns.net).

Internet2 is a project that uses advanced networks in a mixed environment of university, corporate, and government users. It aims to provide end-to-end (workstation-to-workstation) application performance, quality of service guarantees, multicasting, and authentication and security features. Applications that require very high speed computer-to-computer communication (such as data mining and large scale multisite computation) or those that involve human/computer interaction in real time (for example, tele-immersion) are prime candidates for Internet2 experimentation. Internet2 was founded in October 1996 and now has over 135 universities among its members. The contract for the Abilene Network, as it is now called, has recently been awarded to Indiana University, with implementation expected in early 1999.[61]

The NGI is a U.S. federal government program announced within days after the Internet2 consortium was founded. Its goal is to provide national security agencies, federal laboratories, and research centers with a secure network that will operate at 100 to 1000 times the speed of the current Internet. Phase one is expected to begin operations in late 1998.

The vBNS is a National Science Foundation initiative provided by MCI for a 622 Mb per second, very high performance backbone network service. System speeds will eventually increase on the vBNS to 45 Gb per second. When speeds of this magnitude are reached, applications such as broadcast-quality videoconferencing with stereo sound will be possible. Even by the time the speeds envisioned in the NGI are attained, virtual reality applications become feasible. However, for most people and for many years to come, it is the original Internet that will carry signals to their computers. Tennant predicts that only a very small fraction of the millions of print items currently held by the world's libraries will ever be in digital form.[62] He sees digital library collections and services as complements to existing libraries and predicts that the staff of the traditional libraries will grow, not disappear.

Increasing costs of chemistry journals have led more and more libraries to cancel journals. With such attempts to cope with increases in the prices of traditional printed journals, the means to purchase even more subscriptions to electronic versions of those journals have been difficult to find. Instead, academic libraries are seeking alternative solutions to the expensive existing communication system of science. With print journals, ownership and access were linked. However, with changes in technology, questions of ownership of electronic data are clouded by issues such as the need for indefinite access after subscriptions have been canceled and which organizations should maintain the archival records of science.

To provide an alternative to the increasing concentration of scientific publishing in the hands of a few mega-corporations, the scholarly community has recently established SPARC, the Scholarly Publishing and Academic Resource Coalition. SPARC will attempt to put together partnerships between SPARC members and vendors ("publishers") to increase competition in the marketplace. SPARC is eschewing print on paper as the product to be supported. One of the first SPARC partners is the American Chemical Society. ACS has agreed to publish at least one new scholarly scientific journal each year for the next three years, the first being *Organic Letters*, a rapid communications journal set to appear in July 1999. To be sure, there have been other attempts by academic library consortia to collectively have an impact on the economics of scholarly publishing, almost without exception with no visible results. However, an effort that has already forged a partnership with such a major publisher as the American Chemical Society is assured of success.

CONCLUSION

Despite the drawbacks to the Internet in its current incarnation, there is ample evidence that the Internet is already a powerful force in information retrieval. It is likely to become an even more important source of information for chemists as the next generation of the Internet becomes a reality. A dramatic impact on primary scientific journal publishing is sure to occur in the next few years, as more and more libraries turn to electronic journals, inevitably to the exclusion of paper versions. Once copyright and ownership issues are agreed upon, the electronic journal will be the major source of primary information for chemists throughout the world.[63]

REFERENCES AND NOTES

(1) Dusold, L. R. History of the Internet. In Bachrach, S. M., Ed.; *The Internet: A Guide for Chemists*; American Chemical Society: Washington, DC, 1996; Chapter 1, pp 3−24.

(2) Zakon, R. H. Hobbes' Internet Timeline v. 3.3 (http://info.isoc.org/guest/zakon/Internet/History/HIT.html).

(3) Varveri, F. S. Information Retrieval in Chemistry Across the Internet. In *Using Computers in Chemistry and Chemical Education*; Zielinski, T. J., Swift, M. L., Eds.; American Chemical Society: Washington, DC, 1997; Chapter 6, pp 93−107.

(4) Published Articles About Chemistry and the Internet (http://www.chemconnect.com/library/articles.html).

(5) *TrAC* Internet Column (Originally edited by Stephen R. Heller, the column is now under the editorship of Michael Guilhaus.) (http://www.elsevier.nl:80/inca/homepage/saa/trac/intntcol.htm).

(6) *The Chemical Educator* (http://journals.springer-ny.com/chedr/).

(7) *The Internet: A Guide for Chemists*; Bachrach, S. M., Ed.; American Chemical Society: Washington, DC, 1996.

(8) Rzepa, H. Chemweb: A List for Chemical Applications of the Internet 13 April 1998 (http://www.lists.ic.ac.uk/hypermail/chemweb/).

(9) Winter, M. ChemDex (http://www.shef.ac.uk/~chem/chemdex/).

(10) Huber, C. F. Chemistry Resources on the Internet. *Issues in Science and Technology Librarianship* Winter 1998 (http://www.library.ucsb.edu/istl/98-winter/interne1.html).

(11) Wiggins, G.; Beckman, R.; Schreiber, K. Chemistry. In *Internet Tools of the Profession; A Guide for Information Professionals*, 2nd ed.; Tillman, H. N., Ed.; Special Libraries Association: Washington, DC, 1997; Chapter 5, pp 77−100 (http://www.sla.org/pubs/itotp/c5.htm).

(12) Britannica Internet Guide (http://www.ebig.com).

(13) CS ChemFinder (http://chemfinder.camsoft.com/).

(14) Stallings, W. The New and Improved Internet Protocol. *BYTE* **1996**, *21(9)*, 55−56.

(15) Maizell, R. *How to Find Chemical Information*, 3rd ed.; Wiley: New York, 1998; p 220. Section 10.15, pp 220−235, is devoted to Internet Issues and Tools.

(16) Van der Valk, P. M. Mailing Lists and Newsgroups: The Cornerstones of the Internet. *Chemical Web Marketing & Technology* **1997**, *1(3)* (http://www.chemconnect.com/cwmt/cwmt03.html#mailing).

(17) Boulez, K. Overview of Chemical Mailing Lists (http://bionmr1.rug.ac.be/chemistry/overview.html).

(18) Cheng, G.; Zhu, Y. Searchable Chemistry Mailing and Newsgroup Database (http://asknpac3.npac.syr.edu/chem/).

(19) Stallings, W. S/MIME: E-Mail Gets Secure. *BYTE* **1998**, *23(7)*, 41−42.

(20) Nelson, S. D. Chemical Object Test Page (Lawrence Livermore National Laboratory) (http://www-dsed.llnl.gov/documents/tests/chem-.html).

(21) INCINC '94, the first Internetional Chemometrics InterNet Conference (http://www.emsl.pnl.gov:2080/docs/incinc/homepage.html).

(22) Price, G. Opera—The Little Browser That Can. *Online* **1998**, *22(3)*, 34, 36.

(23) Opera Software (http://www.operasoftware.com/index.html).

(24) Bremser, W. The Big 4.0. *Internet World* **1998**, *9(1)*, 64−65, 68−70.

(25) Cascading Style Sheets (http://www.w3.org/Style/css).

(26) Cascading Style Sheets 2 (http://www.w3.org/TR/PR-CSS2.

(27) Bremser. Op. Cit., pp 64−65.

(28) Mace, S.; Flohr, U.; Dobson, R.; Graham, T. Weaving a Better Web. *BYTE* **1998**, *23(3)*, 58−66, 68.

(29) Murray-Rust, P. Welcome to Chemical Markup Language − CML (http://www.venus.co.uk/omf/index.html).

(30) Murray-Rust, P. XML and Chemical Markup Language (http://www.vsms.nottingham.ac.uk/vsms/talks/chemweb.html).

(31) Mackenzie, D. New Language Could Meld the Web Into a Seamless Database. *Science* **1998**, *280*, 1840−1841.

(32) Extensible Markup Language (XML) (http://www.w3.org/XML).

(33) The SGML/XML Web Page (http://www.sil.org/sgml/sgml.html).

(34) eXtensible Style Language (http://www.w3c.org/Style/XSL/).

(35) XML Linking Language (Xlink) (http://www.w3c.org/TR/1998/WD-xlink-19980303).

(36) Moody, G. A New Dawn. *New Scientist Planet Science* 30 May 1998 (http://www.newscientist.com/ns/980530/xml.html).

(37) Notess, G. Northern Light: New Search Engine for the Web and Full-Text Articles. *Database* **1998**, *21(1)*, 32−37.

(38) Lawrence, S.; Giles, C. L. Searching the World Wide Web. *Science* **1998** (3 April), *280*, 98−100.

(39) Search Engine Watch (http://searchenginewatch.com).

(40) Quick Reference Guide to Search Engine Syntax (http://www.albany.edu/library/internet/syntax.html).

(41) About AltaVista (http://www.altavista.digital.com/av/content/about_our_strengths.html).

(42) The Internet: Fulfilling the Promise. *Scientific American* **1997**, *276-(3)*, 49ff.

(43) Notess, G. R. On the Net: Looking Ahead to 1998 on the Net. *Online* **1998**, *22(1)*, 74−76.

(44) Lynch, C. A. Networked Information: Finding What is Out There. *Educom Rev.* **1997**, *32(6)*, 42−50.

(45) Web-Based Application-Development Tools. *BYTE* **1998**, *23(6)*, 112L.

(46) CARL UnCover (http://www.carl.org).

(47) Molecular Visualization Tools and Sites (http://www.indiana.edu/~cheminfo/mvts.html).

(48) *Journal of Molecular Modeling* (http://www.ccc.uni-erlangen.de/jmolmod/index.html).

(49) JCE Internet Articles in Review (http://jchemed.chem.wisc.edu/JCEWWW/InReview/index.html).

(50) Maizell. Op. cit., p 222.

(51) Online-Only Journals Monitored by Chemical Abstracts, August 1998 (http://www.cas.org/EO/ejourn2.html).

(52) *Internet Journal of Chemistry* (http://www.ijc.com).

(53) The International DOI Foundation. Digital Object Identifier System (http://www.doi.org).

(54) Publisher Item Identifier Adopted (http://www.aip.org/epub/piipr.html).

(55) Griffin, S.; Wason, T. The Year of Metadata. *Educom Rev.* **1997**, *32-(6)*, 56−58.

(56) Description of Dublin Core Elements (http://purl.org/metadata/dublin_core_elements).
(57) Rzepa, H. Chemical Meta-data (http://www.ch.ic.ac.uk/talks/wm4/9.html).
(58) Dern, D. P.; Mace, S. The Internet Reinvented. *BYTE* **1998**, *23(2)*, 89−90, 92, 94, 96.
(59) Rosenfeld, L.; Morville, P. *Information Architecture for the World Wide Web*; Cambridge; Sebastopol, CA, O'Reilly, 1998.
(60) Van Houweling, D. Building the Internet's Future: Internet2, ICAID (sic) and NGI. *Educom Rev.* **1998**, *33(3)*, 16−20. p 17.
(61) University Consortium for Advanced Internet Development. The Abilene Project (http://www.ucaid.edu/abilene).
(62) Tennant, R. Digital Potential and Pitfalls. *Library Journal* November 15, **1997**, *122*, 21−22.
(63) An interesting perspective on copyright can be found in Bachrach, S.; Berry, R. S.; Blume, M.; et al. Who Should Own Scientific Papers? *Science* **1998**, *281*, 1459−1460 (http://www.sciencemag.org/cgi/content/full/281/5382/1459).

CI980125B