

Chimia 52 (1998) 658–663
© Neue Schweizerische Chemische Gesellschaft
ISSN 0009–4293

The ChemFinder WebServer: Indexing Chemical Data on the Internet

Jonathan S. Brecher*

Abstract. The Internet offers a wealth of data on all topics, but chemical information provides unique challenges for searching. General-purpose indexing techniques are summarized, with a discussion of why those techniques are unsuited for use with chemical data. The ChemFinder WebServer, a WWW-based chemical index, addresses some of the weaknesses in nonchemical search engines.

Introduction

The recent growth of the Internet has transformed what was once a small network devoted to scientific research into a ubiquitous part of modern society. Recently estimated [1] as containing 320 million public World-Wide Web pages, its size makes the Internet an uniquely valuable information resource. This sheer size and decentralized nature, however, also makes it difficult to find any particular piece of information on demand. Towards this end, public directories and search engines have been created that allow the location of information by topics, keywords, and related identifiers. We describe here the CS ChemFinder WebServer (<http://chemfinder.camsoft.com>), an Internet index specifically tailored to store and present information about individual chemical substances.

Architecture of Search Engines and Directories

Internet information indexes may be broadly categorized as *search engines* or as *directories*, which take different approaches to the same goal of making it easier to find information. Search engines function *via* automated traversal of the WWW using software agents known variously as spiders, (ro)bots, or worms [2].

These programs start at some page, which they read and parse into an internal database. Having parsed the starting page, the agent identifies all other pages referenced in the first one, and repeats the process with each of them in turn. Search engines excel at creating word- and phrase-searchable indexes. As such, they are often the best choice when looking for information that can be summarized by a single phrase that would not likely appear in contexts other than the one desired. Popular search engines include AltaVista (<http://altavista.digital.com>), HotBot (<http://www.hotbot.com>), Excite (<http://www.excite.com>), and Lycos (<http://www.lycos.com>).

Directories differ from search engines in that they do not attempt to index every word on every page, but rather try to classify pages into broad (and usually hierarchical) categories. Creating directories tends to be very labor-intensive, with the final decision on where and whether to reference any given page usually given to trained individuals. Because of this human involvement, directories are always smaller than search engines in absolute size, but are expected to be denser in terms of quality information content. The inherent categorization of directories makes them the indexes of choice when looking for general information about a topic. Yahoo! (<http://www.yahoo.com>) is the prototypical example of a directory, and is the most visited site on the WWW [3].

Locating chemical information provides challenges not encountered with other types of information. First among these is the variety of ways in which a single substance may be described. A given compound (Fig. 1) may reasonably be de-

scribed by name (benzotrile, phenyl cyanide, cyanobenzene, benzoic-acid nitrile), formula (C_6H_5CN , C_7H_5N), or any of several ID codes (100-47-0, DI2450000), as well as by the structural formula itself in any orientation and *Kekulé* form.

A search engine query for 'benzotrile' would locate any mention of that specific word, but would find no information about 'cyanobenzene' unless both words were present on the same page. Finding information about this compound would require many searches – ideally, searches for each of the terms listed above, plus many more. Such queries depend on some alphanumeric descriptor being available for the substance in question; if only a graphical structure is available, a name must be derived before any searching can begin. The broad categorization that is the goal of directories makes them also unsuitable for this type of search for information about a specific substance.

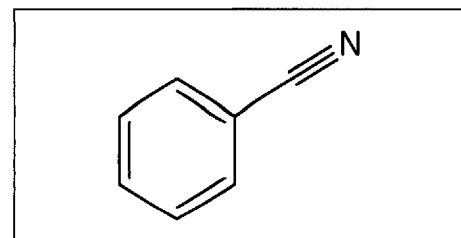


Fig. 1. Benzotrile

Methodology of the ChemFinder WebServer

Begun in 1995, the ChemFinder WebServer has been designed to address the deficiencies in general-purpose information indexes by providing a centralized means of searching for chemical information on the Internet. A comprehensive database has been dynamically assembled from numerous public sources and is continually being augmented. As new chemical information is located on the Internet, it is analyzed and merged into this master ChemFinder WebServer database.

Only core information about each substance is retained by the server: the name of the substance and the address of the page that contains information about it. When available, select additional common data are retained, including *Chemical Abstracts Service* Registry Numbers and basic physical property information such as boiling points and melting points. Structural information is rarely available on other sites. Even when available, it is generally provided only as a static image in .gif or .jpg format, and not in any format that preserves the chemical information in a way that can be interpreted automatical-

*Correspondence: J.S. Brecher
CambridgeSoft Corporation
100 Cambridge Park Drive
Cambridge, MA 02140, USA
E-Mail: jsb@camsoft.com

ly. Accordingly, structural information is not retained, but used as a reference when structures are redrawn.

During the indexing process, several methods are used to improve the quality of the data. Principal among these is normalization of the chemical names. During normalization, all spaces, parentheses, and other punctuation are removed, and the names are converted entirely to lower case. Regional spelling variations (aluminum *vs.* aluminium, sulfur *vs.* sulphur) are recognized and standardized. Names written in CAS-style inverted form are un-inverted (acetic acid, sodium salt *vs.* sodium acetate), and several dozen of the most common typographical errors are automatically corrected. The normalized name is stored along with the original form. In addition to name normalization, CAS Registry Numbers, if present, are checked for validity using their internal checksum [4].

The new data are compared to the data already present in the ChemFinder WebServer through comparisons of the normalized names and the CAS Registry Numbers. Any new substances that fail to match existing records are examined individually for typographical errors. Approximately 5% of all data indexed by the ChemFinder WebServer is recognized to be erroneous by this stage [5]. Where the intended data are obvious (*e.g.*, when the errors are typographical in nature), the errors are corrected; otherwise the datum is discarded.

During this premerge process, the data are also analyzed for transitive logic errors. For example, one site might reference 'Ritalin' and 'methylphenidate' as synonyms. Another might state that 'methylphenidate hydrochloride' is the generic name for Ritalin. From this, it could be concluded that 'methylphenidate' and 'methylphenidate hydrochloride' are interchangeable names that refer to the same compound. This is clearly an incorrect conclusion, and other reference sources must be consulted to resolve the discrepancy.

The remaining entries are then merged into the main database. During the merge process, any physical property information, if present, is compared to values already in the database. Numerical differences of one unit (generally, degree Celsius) or less are assumed to be insignificant and are discarded. Differences of greater than one unit are examined on an individual basis, and other reference works are consulted to identify the correct values. Chemical structures are added, with molecular weights and formulas calculated directly from the structures. Finally,

the completed database is made available for searching.

The ChemFinder WebServer is a descriptive index, not a peer-reviewed data source; many errors remain even after the verification process. These are addressed on a case-by-case basis as they are found, often by users of the ChemFinder WebServer service. In some cases, the 'error' is encountered more frequently in common use than is the correct data. For example, the name 'alanine' is chemically un-specific about the stereochemistry at C(2), and could refer to the D-form, the L-form, or any mixture of the two. In practice, the term almost always refers to the natural form. Where common usage contradicts official nomenclature rules, the error is not corrected. The ChemFinder WebServer lists 'alanine' and 'L-alanine' as synonymous, and does not list 'alanine' as a synonym for 'D-alanine'.

A site qualifies for inclusion only if it references individual chemical substances. In its current design, the ChemFinder WebServer does not index information about chemical conferences, chemical education, chemical software, or any of a host of other chemistry-related topics. Additionally, a site must provide at least one piece of information about the substance. A simple textual list of compound names – with no other information – would not qualify for indexing in the ChemFinder WebServer, since purely textual data is already well suited for indexing in existing text-based search engines.

Chemical Information Sources on the Internet

As an index, the ChemFinder WebServer would have little purpose if there were no data available to be indexed. Chemical data is fairly abundant on the Internet, but can be classified into one of only a small number of forms.

Health and safety data comprise one of the largest classes of chemical data encountered. Data of this type are typically oriented towards consumers or health-care professionals, rather than laboratory chemists. These data are found at sites focusing on the environmental impact of chemicals [6] and sites discussing specific diseases [7]. Extensive information is available about individual medications [8], particularly those involved in treating high-profile diseases [9].

Most chemical manufacturers and distributors have a presence on the Internet, usually on the WWW, but the information content of these sites is highly variable. A

small number of sites [10] provides extensive physical property data and full Materials Safety Data Sheets (MSDSs) for all products. More typically, chemical manufacturers will provide limited physical property or usage information for some of their products [11]. Many manufacturers provide only a textual list of their substances [12], or simply a list of classes of substances [13], neither of which would qualify for indexing in the ChemFinder WebServer for reasons discussed earlier.

Governmental and other regulatory data are a third category of information that is well represented on the Internet. Broad and unrestricted distribution of regulatory information is often mandated by the organization that compiles it, and the Internet provides an ideal way to distribute information to a wide audience. Regulatory information typically includes little more than a compound name and (sometimes) CAS Registry Number, but the fact that a substance is regulated is often an important datum in its own right. Given the bureaucratic division between governmental departments, specific pieces of regulatory data can be difficult to locate, but are available from many branches of the US federal and state governments [14], from the governments of many other countries [15], and from international organizations [16].

The molecular biology community has long had an active presence on the Internet, with data on substances large (The Human Genome Project [17]), and small (The Pherolist: List of Sex Pheromones of Lepidoptera and Related Attractants [18]), and in between (The Protein Data Bank [19]). Because of the large size of many of these substances, and because the molecular biology community has already developed excellent databases such as listed above for handling these data, we have made a conscious decision not to index biological macromolecules in the ChemFinder WebServer. Exceptions are made on a case-by-case basis, and generally only for 'small' biological molecules such as those listed in The Pherolist, and for 'popular' biological substances such as insulin and hemoglobin.

In recent years, there has been an increasing availability of commercial sources of chemical information on the WWW. Many journals are now available in electronic format, and the *Chemical Abstracts Service* provides Internet-based access to their data as well. With few exceptions, however, this information is available only by subscription. We have decided to limit the ChemFinder WebServer to indexing unrestricted data sources only.

Although most chemical information on the Internet can be considered as one of the types discussed, there is still much information that cannot be neatly categorized. Often these sites will be the work of one individual who has spent the time to document some substances of personal interest. University courses occasionally provide information about molecules discussed in class. Information about a chemical substance might appear in a context not specifically directed to chemists. All of these would qualify for indexing in the ChemFinder WebServer.

Additionally, four sites deserve special mention for their comprehensiveness or innovation. The Protein Data Bank mentioned earlier was one of the first public sources of chemical information on the Internet, and has been in operation for over 20 years [20]. As a doubly rare source of both peer-reviewed information and

crystallographic data, the Protein Data Bank is a particularly valuable resource, although, as a collection of biological macromolecules, it has not been indexed in the ChemFinder WebServer.

Klotho, the Biochemical Compounds Declarative Database, was one of the first small-molecule data sources on the WWW, becoming available in February, 1994 [21]. Klotho's collection of biological molecules demonstrated the potential of the WWW in presenting and distributing chemical information, and was a source of inspiration in creating the ChemFinder WebServer. The entries in Klotho have been indexed in the ChemFinder WebServer.

The WWW Chemical Structures Database was the first (and, apparently, only) attempt to utilize the techniques of a general purpose search engine in creating a chemistry-specific database [22]. Its search

program, the CACTVS System Chemistry Spider, traversed a portion of the WWW looking for chemical structures stored in any of several chemical formats. Over 2250 such files were identified, with some compounds represented by more than one file. A search of this type is necessarily hindered by the overall scarcity of chemical structures stored in chemical file formats publicly available on the Internet. The entries in the WWW Chemical Structures Database have also been indexed in the ChemFinder WebServer, with references to their original source.

Finally, special mention must be made of the NIST Chemistry WebBook [23], which since August, 1996, has provided one of the few large (over 30000 records currently) databases of substances. It is also one of the few databases to provide complete references for each physical property value listed. The entries in the NIST Chemistry WebBook have also been indexed in the ChemFinder WebServer.

Fig. 2. The simple query interface of the ChemFinder WebServer

Fig. 3. The advanced query interface of the ChemFinder WebServer, with diagram-based structure searching

Use of the ChemFinder WebServer

The ChemFinder WebServer is publicly available *via* two similar form-based WWW interfaces. Typical of other general-purpose search engines, the default search form for the ChemFinder WebServer accepts a single text phrase as query input (Fig. 2). This phrase is analyzed on the server, and a search is performed by chemical name, formula, molecular weight, or CAS Registry Number. These types of searches were selected both because they are the most common searches

Table 1. *Effects of Normalization.* All names shown will normalize to the same string, and will return the same records when used as a query.

3-chlorobenzoic acid, sodium salt
3-chloro-benzoic acid, sodium
sodium 3-(chloro)-benzoate
m-chloro benzoate, sodium salt
m-chlorbenzoate, sodium

Table 2. *Disallowed Field Combinations for Searching*

Molecular formula with molecular weight
Exact name with any other fields
CAS Registry Number with any other fields
Structure with molecular formula or molecular weight

executed by users of the ChemFinder WebServer, and because a single text phrase can be identified as one of the four input types mentioned above with a high degree of certainty. Numerical values that conform to the CAS Registry Number checksum algorithm are assumed to be CAS Registry Numbers. Other numerical values are identified as molecular weights. Alphanumeric phrases that can be parsed as molecular formulas (ignoring case) are treated as formulas, while all other queries are searched as names. Since a chemical formula could also be interpreted as a chemical name, a formula search that fails to match any entries in the database is automatically re-queried as a search by chemical name. This simple interface is popular, accounting for 72% of all searches.

As with general-purpose search engines, substances can be searched by full or partial chemical name; however, all name searches are normalized according to the same rules used in the database merging process (Table 1). Since the ChemFinder WebServer has a very diverse user base – both geographically and in educational level – normalization is very important in addressing common typographical and translation errors. This normalization greatly increases the percentage of successful searches.

An advanced search form (Fig. 3) offers greater control of the search process, allowing chemical names, formulas, molecular weights, and CAS Registry Numbers to be searched explicitly. Queries are also supported for boiling and melting points. Using this form (or its available historic sibling, Fig. 4), chemically related data can be searched in chemically meaningful ways. Molecular weights and numerical physical properties can be searched by range as well as by exact values, with automatic recognition of significant figures. Formulas can be searched by element ranges as well as by exact formula with any ordering of elements. Chemical structures can be drawn directly in the browser window when using the CS ChemDraw Netscape Plug-In (or entered in the textual SMILES [24] format in the older form). Searches such as these are commonplace when looking for chemical information, but rarely or never used otherwise, and are not supported in general-purpose search engines and directories.

The advanced search forms allow concurrent searching over multiple fields. This feature was intended to permit the creation of more specific searches, but an analysis of failed searches over several months revealed that many searches were failing

Fig. 4. The historic query interface of the ChemFinder WebServer, with SMILES-based structure searching

Poor Query

If you specify a Molecular Formula, it is never useful to specify a Molecular Weight. At best, any Molecular Weight you enter will be redundant with the Molecular Formula. At worst, they will be contradictory, and prevent you from getting any useful hits.

Press the button below to continue your query, searching for all fields except Molecular Weight, or press the Back button on your browser to modify your query.

Search for all fields except Molecular Weight

Fig. 5. An example of the message returned when entering redundant search terms

Fig. 6. Results of a search that matches more than one substance

due to terms that were inherently contradictory. A search of this type might have looked for a substance whose name was 'benzene' and whose molecular weight was 76. Since the molecular weight of benzene truly is 78, this search would fail even though the database contained both

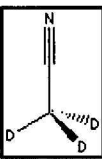
benzene and many other substances with a molecular weight of 76. The server has since been modified to disallow certain field combinations (Table 2). If a search is attempted with one of these combinations, the query is not executed, but the user is presented with an explanation of why the

Acetonitrile-d3 [2206-26-0]

Synonyms: Deuterated acetonitrile

C_2D_3N
44.071

This picture is a live chemical image
The ChemDraw Plugin lets you search by drawing structures in your web browser. Have you downloaded it yet?



Melting Point (°C)	--	Specific Gravity	0.844
Boiling Point (°C)	80.7	Vapor Density	--
Evaporation Rate	--	Water Solubility	--
Flash Point (°C)	2	EPA Code	--
DOT Number	--	RTECS	--
Comments	LACHRYMATOR/HYGROSCOPIC.		

More information about this compound is available from

ABC R GmbH & Co KG
[Acetonitrile-d3, 99.8%](#)
[Acetonitrile-d3, isotopic purity, 99%](#)

Fig. 7. Detailed information about an individual substance

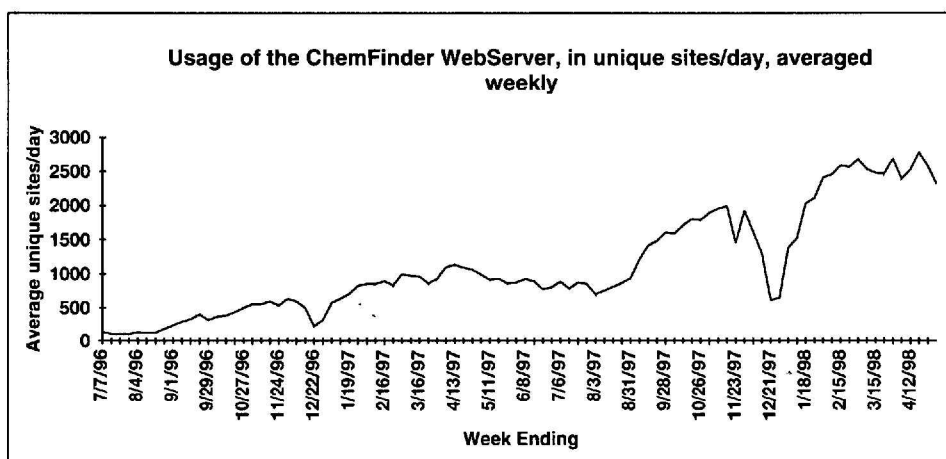


Fig. 8. ChemFinder WebServer daily usage. Several strong seasonal trends are readily apparent, including a slight decrease in traffic during the (northern hemisphere) summer, followed by a sharp increase in traffic in September and October. The year-end holidays are particularly obvious, while the lesser decrease at the end of November corresponds to the American Thanksgiving holiday.

search terms should not be used in combination (Fig. 5). This has resulted in a general increase in search quality.

Queries resulting in more than one matching substance produce a list of names of those substances (Fig. 6), with links to detailed information about each one. Queries that match a single substance display the detailed information page (Fig. 7) directly. Fulfilling its role as a search engine

for chemical information on the Internet, the ChemFinder WebServer provides a list of links at the bottom of this page, indicating other locations that provide additional information about the substance shown. Substance-specific information (usually limited to substance name and synonyms, structure, formula, and molecular weight) is shown at the top of the page. This information is not intended to

Table 3. Top 20 Most-Common Searches at the ChemFinder WebServer (data for April, 1998). Note that several of the most-common searches were for classes of compounds and not for specific compounds themselves; these searches would not have returned any data.

(DHQD)2PHAL
 ethanol
 KMnO4
 hexamethylenediamine
 benzene
 4,5-dihydroxy-1,3-benzenedisulfonic acid
 phenol
 acetone
 benzaldehyde
 benzoic acid
 methanol
 sodium hydroxide
 sulfuric acid
 absolute ethanol
 acid
 acetic acid
 NMO
 CO
 ester
 sulfanilic acid monohydrate

supplant that available at the other sites listed, but is often sufficient to answer basic identity questions ('What is the name of the following structure?').

The ChemFinder WebServer is not a single entity, but is composed of a suite of server-side applications. Queries are transmitted from a WWW browser to the WebSite WWW server software, published by O'Reilly and Associates. WebSite parses the initial query from HTTP format into its distinct fields, and transfers it via the Common Gateway Interface (CGI) [25] to a Visual Basic application. The CGI confirms the validity of the query as discussed above, then transmits the data to CS ChemFinder Pro (CambridgeSoft Corporation) via OLE Automation. CS ChemFinder is a chemical database designed specifically to search and store chemical data, and functions as the core of the ChemFinder WebServer. It performs the actual query, and retrieves the textual, numeric, and structural data, which it returns to the CGI. The CGI then formats the data and returns it to WebSite, which finally returns it to the user.

Usage of the ChemFinder WebServer is currently showing an approximately 200% yearly growth rate, with strong seasonal variations (Fig. 8). As of May, 1998,

over 75 000 individual substances were indexed from some 300 sites. Each substance had an average of 2.4 links to other sites on the Internet containing additional information about the substance. Although we have been unable to quantify the ways that most people are using the ChemFinder WebServer, an examination of the most popular searches suggests that much of its use is in locating general information about common substances.

Future Prospects

The ChemFinder WebServer represents a new approach to providing chemical data, directed not at the trained information specialist but at the lay users that now represent the majority of traffic on the Internet. Further developments of the ChemFinder WebServer will likely result in larger data sets, but will probably not reach the comprehensiveness achieved by the *Chemical Abstracts Service Registry File* (17 million entries) or the *Beilstein* database (6 million entries).

Additional effort will also likely be placed on further enhancements to the user interface, search engine, and normalization routines. All work will be directed to meeting what is the goal of all search engines and directories: the rapid, easy, and accurate location of information.

Received: September 11, 1998

- [1] S. Lawrence, C.L. Giles, *Science* **1998**, *280*, 98.
- [2] M. Kloster, 'The Web Robots Pages' (<http://info.webcrawler.com/mak/projects/robots/robots.html>).
- [3] *RelevantKnowledge, Inc.*, 'RelevantKnowledge Releases March's Top 25 Web Properties and Domains by Unique Visitors' (http://www.relevantknowledge.com/Press/release/04_08_98.html).
- [4] *Chemical Abstracts Service*, 'Check Digit Verification of CAS Registry Numbers' (<http://www.cas.org/EO/checkdig.html>).
- [5] J.S. Brecher, 'Chemical errors found on WWW sites' (<http://chemfinder.camsoft.com/errorsfound.shtm>).
- [6] See the Internet HazDat Site Contaminant Query at <http://atsdr1.atsdr.cdc.gov:8080/gsql/sitecontam.script> and The Coalition Opposed to PCB Ash in Monroe County, Indiana, at <http://copa.org>, for example.
- [7] See the Berkeley Carcinogenic Potency Project at <http://potency.berkeley.edu/cpdb.html> and the Contact Dermatitis Home Page at <http://www.mc.vanderbilt.edu/vumcdept/derm/contact/index.html>, for example.
- [8] See the Index to Allergy/Asthma Medications at <http://www.cmh.edu/Allergy/Drugs/> and the Australian Guide to Medications at <http://www.avm.com.au/agtm/agtm.html>, for example.
- [9] See the AIDS Drug Assistance Program at <http://www.aidsinfonyc.org/network/aces/index.html>, for example.
- [10] See the *Fisher Scientific* site at <http://www.fisher1.com>, for example.
- [11] See the *Callery Chemical Company* at <http://www.callery.com> (physical properties), *Spectrum Laboratories* at <http://www.spec-lab.com> (usage, analytical methods and environmental fate), and *Penn Bioorganics* at <http://www.pennbio.com> (structures, melting and boiling points) as representative examples.
- [12] See The *Coral Group* site at <http://www.coralchem.com>, for example.
- [13] See the *Hoechst* site at <http://www.hoechst.com> and the *Discovery Chemical, Inc.* site at <http://www.inc.com/users/kkeller.html>, for example.
- [14] See the EPA Restricted Use Products report at <http://www.epa.gov/docs/RestProd/> and the Alaskan Administrative Code at <http://www.touchngo.com/lglcntr/akstats/AAC.htm>, for example.
- [15] See a list of Australian approved food additives at <http://www.hawkesbury.uws.edu.au/~skurray/code.txt>, for example.
- [16] See the Helsinki Convention of 1992 at <http://www.helcom.fi>, for example.
- [17] <http://www.nhgri.nih.gov/HGP/>
- [18] <http://nysaes.cornell.edu/pheronet/>
- [19] <http://www.pdb.bnl.gov>
- [20] F.C. Bernstein, T.F. Koetzle, G.J.B. Williams, E.F. Meyer, Jr., M.D. Brice, J.R. Rodgers, O. Kennard, T. Shimanouchi, M. Tasumi, *J. Mol. Biol.* **1977**, *112*, 535.
- [21] 'What's New on Klotho' (see http://www.abc.wustl.edu/moirai/klotho/whats_new.html).
- [22] 'The WWW Chemical Structures Database' (see <http://www2.ccc.uni-erlangen.de/services/webmol.html>).
- [23] W.G. Mallard, P.J. Lindstrom, Eds., 'NIST Chemistry WebBook, NIST Standard Reference Database Number 69', March 1998, National Institute of Standards and Technology, Gaithersburg MD, 20899 (<http://webbook.nist.gov>).
- [24] D. Weininger, *J. Chem. Inf. Comput. Sci.* **1998**, *28*, 31.
- [25] National Center for Supercomputing Applications, 'The Common Gateway Interface' (<http://hoohoo.ncsa.uiuc.edu/cgi/>) and R.B. Denny, 'Windows CGI 1.3a Interface' (<http://www.dc3.com/wsdocs/32demo/windows-cgi.html>).