BASICS FOR THE SERIOUS SEARCHER

In writing this book, I have made the assumption that the reader knows the internet basics—what it is, how to get connected, and so forth. The "basics" covered in this chapter involve background information that serious searchers need to know to be fully conversant with internet content and issues, as well as general ways of approaching internet resources to find just what you need. I go over some details already familiar to many readers, but I include this background material for two purposes: (1) to allow readers to understand more fully the characteristics, content, utility, and nuances of the internet in order to use it more effectively, and (2) to help those who find themselves teaching others how to use the internet, by providing answers to some of the more frequently asked questions.

As for general approaches to finding the right resources, this chapter provides an overview and comparison of the kinds of "finding tools" available and a set of strategies that can be applied. The coverage of strategies goes into some detail on topics (such as Boolean logic) that will also be encountered elsewhere in the book. Integral to all of this are some aspects and issues regarding the *content* that is found on the internet. These aspects include the questions of retrospective coverage, quality of content, and general accessibility of content, particularly the issue of the "Deep Web" (a.k.a., the "Invisible Web", the "Hidden Web"). Woven into this content fabric are issues, such as copyright, that affect how information found on the internet can be used. Although only lightly touched upon, it is important that every serious user have an awareness of these issues. Lastly, the chapter provides some useful resources for keeping up with the latest internet tools, content, and issues.

THE PIECES OF THE INTERNET

First, the *internet* and the *web* are not synonymous, although the terms are frequently used interchangeably. As late as the mid-1990s, the internet had

some clearly distinguishable parts, as defined by their functions. Much internet usage could be thought of as internet *sans* content. It was simply a communications channel that allowed easy transfer of information. Typically, a user at one university could use the internet to send or request a file from someone at another university using FTP (File Transfer Protocol). Sending email via the internet was becoming tremendously popular at that time. A user of a commercial search service such as Dialog or LexisNexis could harness the internet as an alternative to proprietary telecommunications networks, basically sending and receiving proprietary information. "Content" parts of the internet could indeed be found, such as Usenet newsgroups, where anyone with a connection could access a body of publicly available information. Gophers (menu-based directories allowing access to files, mainly at universities) were also beginning to provide access to content.

The world changed, and content was destined to become king, when Tim Berners-Lee at CERN (Conseil Européen pour la Recherche Nucléaire) in Geneva created the World Wide Web in 1991. The web provided an easy-to-use interface for both potential content providers and users, with a GUI (Graphical User Interface) incorporating hypertext point-and-click navigation of text, graphics, and sounds, and created what was for most of us at that time an unimaginable potential for access to information.

Within less than five years, the web had overtaken email and FTP in terms of internet traffic. By 2000, usage of the other parts of the internet was becoming fused into the web. Usenet newsgroups were being accessed through a web interface, and web-based email was becoming the main—or only—form of email for millions. FTP was typically being managed through a web interface. Gophers were replaced by web directories and search engines, and any gophers you find now are likely to be in your backyard.

A VERY BRIEF HISTORY

The following selection of historical highlights provides a perspective for better understanding the nature of the internet. It should be emphasized that the internet is the result of many technologies (computing, time-sharing of computers, packet-switching, etc.) and many visionaries and great technical thinkers coming together over a period of a few decades. In addition, what they were able to accomplish was dependent upon minds and technologies of preceding decades. This selection of highlights is merely a sampling and

leaves out many essential technical achievements and notable contributors. The points here are drawn primarily from the resources listed at the end of this timeline.

1957 The USSR launches *Sputnik*.

1958 Largely as a result of the *Sputnik* launch, ARPA (Advanced Research Projects Agency) is established to push the U.S. ahead in science and technology. High among its interests is computer technology.

1962 J. C. R. Licklider writes about his vision of a globally interconnected group of computers providing widespread access to data and programs; the RAND Corporation begins research on distributed communications networks for military purposes.

Early 1960s Packet-switching moves from theory to practice.

Mid- to Late-1960s ARPA develops ARPANET to promote the "cooperative networking of time-sharing computers" with four host computers connected by the end of 1969 (Stanford Research Institute, UCLA, UC Santa Barbara, and the University of Utah).

1965 The term "hypertext" is coined by Ted Nelson.

1968 The Tymnet nationwide time-sharing network is built.

1971 ARPANET grows to 23 hosts, including universities and government research centers.

1972 The International Network Working Group (INWG) is established to advance and set standards for networking technologies; the first chairman is Vinton (Vint) Cerf, who is later often referred to as the "Father of the Internet."

1972–1974 Commercial database services—Dialog, SDC Orbit, Lexis, The New York Times DataBank, and others—begin making their subscription services available through dial-up networks.

1973 ARPANET makes its first international connections at the University College of London (England) and the Royal Radar Establishment (Norway).

1974 "A Protocol for Packet Network Interconnection," which specifies the details of TCP (Transmission Control Protocol), is published by Vint Cerf and Bob Kahn.

1974 Bolt, Beranek & Newman, contractor for ARPANET, opens a commercial version of the ARPANET called Telenet, the first public packet-data service.

1977 There are 111 hosts on the internet.

1978 TCP is split into TCP and IP (Internet Protocol).

1979 The first Usenet discussion groups are created by Tom Truscott, Jim Ellis, and Steve Bellovin, graduate students at Duke University and the University of North Carolina, and Usenet quickly spreads worldwide.

The first emoticons (smileys) are suggested by Kevin McKenzie.

1980s The personal computer becomes a part of millions of people's lives.

There are 213 hosts on ARPANET.

BITNET (Because It's Time Network) is started, providing email, electronic mailing lists, and FTP service.

CSNET (Computer Science Network) is created by computer scientists at Purdue University, University of Washington, RAND Corporation, and BBN, with National Science Foundation (NSF) support. It provides email and other networking services to researchers without access to ARPANET.

1982 The term "Internet" is first used.

TCP/IP is adopted as the universal protocol for the internet.

Name servers are developed, allowing a user to get to a computer without specifying the exact path.

There are 562 hosts on the internet.

France Telecom begins distributing Minitel terminals to subscribers free of charge, providing videotext access to the Teletel system. Initially providing telephone directory lookups, then chat and other services, Teletel is the first widespread home implementation of these types of network services.

1984 Orwell's vision, fortunately, is not fulfilled, but computers are soon to be in almost every home.

There are more than 1,000 hosts on the internet.

1985 The WELL (Whole Earth 'Lectronic Link) is started. Individual users, outside universities, can now easily participate on the internet.

There are more than 5,000 hosts on the internet.

1986 NSFNET (National Science Foundation Network) is created. The backbone speed is 56K. (Yes, as in the total transmission capability of a 56K dial-up modem.)

1987 There are more than 10,000 hosts on the internet.

1988 The NSFNET backbone is upgraded to a T1 at 1.544 Mbps (megabits per second).

1989 There are more than 100,000 hosts on the internet.
ARPANET fades away.
There are more than 300,000 hosts on the internet.

1991 Tim Berners-Lee at CERN (Conseil Européen pour la Recherche Nucléaire) in Geneva introduces the World Wide Web.
NSF removes the restriction on commercial use of the internet.
The University of Minnesota releases the first gopher, which allows point-and-click access to files on remote computers.
The NSFNET backbone is upgraded to a T3 (44.736 Mbps).

1992 There are more than 1,000,000 hosts on the internet.
Jean Armour Polly coins the phrase "surfing the internet."

1994 The first graphics-based browser, Mosaic, is released.
Internet talk radio begins.
WebCrawler, the first successful web search engine, is introduced.
A law firm introduces internet "spam."
Netscape Navigator, the commercial version of Mosaic, is shipped.

1995 NSFNET reverts to being a research network; internet infrastructure is now primarily provided by commercial firms.
RealAudio is introduced, meaning that you no longer have to wait for sound files to download completely before you begin hearing them, and allowing for continued ("streaming") downloads.
Consumer services such as CompuServe, America Online, and Prodigy begin to provide access through the internet instead of only through their private dial-up networks.

1996 There are more than 10,000,000 hosts on the internet.

1999 Microsoft's Internet Explorer overtakes Netscape as the most popular browser.

1999 Testing of the registration of domain names in Chinese, Japanese, and Korean languages begins, reflecting the internationalization of internet usage.

2001 Mysterious monolith does not emerge from the Earth and no evil computers take over any spaceships (as far as we know).

2002 Google is indexing more than 3 billion webpages.

2003 There are more than 200,000,000 IP hosts on the internet.

2004 Weblogs, which started in the mid-1990s, gain widespread popularity and attention.

2005 More than 50 percent of Americans who access the internet at home have a high-speed connection.

2006 Developmental focus is on a more interactive, personalized web, with collaboration, sharing, desktop-type programs, social networking, and use of APIs (Application Program Interfaces) to integrate data from multiple sources over the web. This shift is tagged "Web 2.0".

2009 Worldwide, there are over 1.5 billion internet users, with the largest number of users in Asia (over 650 million users).

Internet History Resources

Anyone interested in information on the history of the internet beyond this selective list is encouraged to consult the following resources.

A Brief History of the Internet, version 3.1

www.isoc.org/internet/history/brief.shtml

Compiled by Barry M. Leiner, Vinton G. Cerf, David D. Clark, Robert E. Kahn, Leonard Kleinrock, Daniel C. Lynch, Jon Postel, Larry G. Roberts, and Stephen Wolff, this site provides historical commentary from many of the people who were actually involved in the internet's creation.

Internet History and Growth

www.isoc.org/internet/history/2002_0918_Internet_History_and_Growth.ppt

This PowerPoint presentation by William F. Slater provides a good look at the internet's pioneers and provides an excellent collection of statistics on internet growth.

Hobbes' Internet Timeline

www.zakon.org/robert/internet/timeline

This detailed timeline emphasizes technical developments and who was behind them.

Internet World Stats
www.internetworldstats.com/stats.htm

This website provides a compilation of statistics, with graphs, for internet usage worldwide

The "New" Web: Web 2.0

By 2006, most heavy-duty internet users had begun to hear the term "Web 2.0" fairly frequently—a term coined (and trademarked) in conjunction with a series of web development conferences that began in 2004. Web 2.0 refers to a "second generation" of the web that provides a much greater focus on—and use of—desktop applications made available on the web, and on collaboration and sharing by users. Forerunners of this include wikis, weblogs, RSS, folksonomies (tagging), and podcasts. Though Web 2.0 has no precise definition, some people also define this new generation of the web in terms of the kinds of programs and techniques used, including APIs (Application Programming Interfaces), social software, and Ajax (Asynchronous JavaScript and XML). The glossary of this book has brief definitions of those terms. From one perspective, what the new web is really about is the *user*, with a focus on areas of user interaction such as participation, publication, social software, sharing, and "the web as platform."

Though individual websites are not usually labeled as Web 2.0, if you look closely, you will have seen these elements in more and more websites. You are seeing manifestations of it when you encounter sites that allow for user-applied "tags" (such as Flickr), in the way a search engine might "suggest" search phrases as you type in your terms, in the ability to zoom and drag maps, and in the instant windows that open on pages in response to moving your cursor or clicking (such as some of the tabs on Yahoo!'s main page). This flexible interactivity with webpages and with the web carries over into increased interactivity with others on the web and can also make web-based software (such as Google Docs) flow as smoothly as similar programs on your desktop.

SEARCHING THE INTERNET: WEB "FINDING TOOLS"

Whether your hobby or profession is cooking, carpentry, chemistry, or anything in between, the right tools can make all the difference. The same is true for searching the web. A variety of tools are available to help you find what you need, and each tool does things a little differently, sometimes with a different purpose or different emphasis, as well as different coverage and different search features.

To understand the variety of tools, it can be helpful to think of most finding tools as falling into one of three categories (although many tools will be hybrids): (1) general directories, (2) search engines, and (3) specialized directories. The third category could indeed be lumped in with the first because both are directories, but for a couple of reasons discussed later, it is worthwhile to treat them separately.

All three categories may also incorporate another function, that of a "portal," which is a website that provides a gateway not only to links, but also to a number of other information resources that go beyond just the searching or browsing function. These resources may include news headlines, weather, stock market information, alerts, yellow pages, and other kinds of handy information. A portal can be general, as in the case of My Yahoo! or iGoogle, or it can be specific for a particular discipline, region, or country.

Other finding tools provide identification of other kinds of internet content, such as discussion groups (forums), images, and audio. These tools may exist either on their own sites, or they may be incorporated into any of the three main categories of tools. These specialized tools will be covered in later chapters.

General Web Directories

The general web directories, such as the Yahoo! Directory and Open Directory, are websites that provide a large collection of links arranged in categories to enable browsing by subject area (see Figure 1.1). Interestingly, general directories, though once the major web "finding tool," are now almost an historical artifact, displaced very largely by search engines.

The advantages of general directories had been the categorization and the selectivity. The categories provide easy browsing of topics, and the selectivity provides a focus on sites that are generally highly regarded for

Figure 1.1

Open Directory main page

their content and usefulness. However, with the greatly improved relevance ranking provided by search engines (particularly with respect to the much greater role in ranking that "popularity" of sites plays), the selectivity provided by directories has become much less needed and much less used. The prominence of the Yahoo! directory on Yahoo!'s main page has rapidly diminished—as late as 2002, the directory was Yahoo!'s most prominent feature, whereas today it is not even a link on Yahoo!'s main page.

The Role of General Directories

General web directories can be a good starting place when you have a very general question (museums in Paris, dyslexia), or when you don't quite know where to go with a broad topic and would like to browse down through a category to get some guidance.

General web directories are discussed in detail in Chapter 2.

Web Search Engines

TIP:

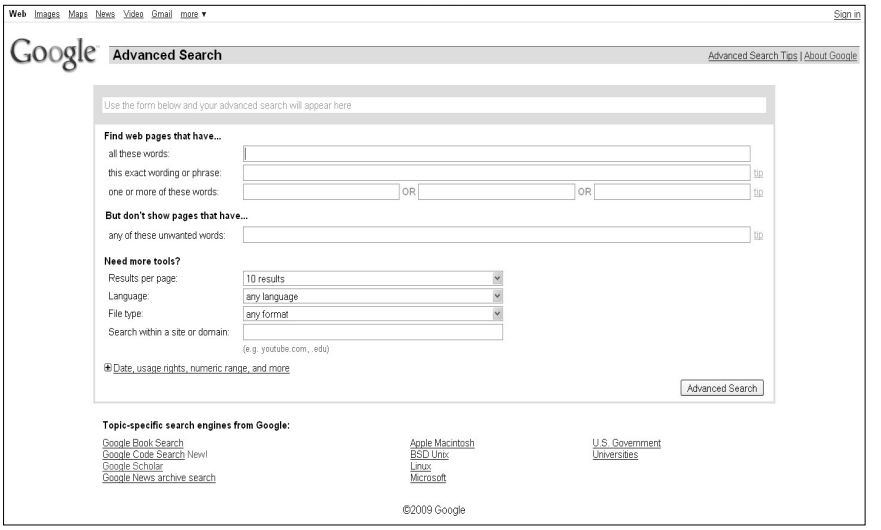
If your question contains just one or two concepts and you only want a small handful of results, you may want to consider using a directory. If your question contains three or more concepts, definitely start with a search engine.

Whereas a directory may be a good start when you want to be directed to just a few selected items on a fairly general topic, search engines are the place to go when you want something on a fairly specific topic (ethics of human cloning, Italian paintings of William Stanley Haseltine). Instead of searching brief descriptions of, at most, a few million websites, as with directories, search engine services allow you to search virtually every word from several billion webpages. In addition, web search engines allow you to use much more sophisticated techniques, so you can focus on your topic more effectively (Figure 1.2). The pages included in web search engines are not placed in categories (hence, you cannot browse a hierarchy), and no prior human selectivity was involved in determining what is included in the search engine's database. As the searcher, you provide the selectivity, by the search terms you choose and by the further narrowing techniques you apply.

The Role of Search Engines

If your topic is very specific or you expect that very little is written on it, a search engine will be a much better starting place than a directory. If your search needs to be exhaustive, use a search engine. If your topic is a combination of three or more concepts (e.g., *"Italian" "paintings" "Haseltine"*), use a search engine. Find out more about search engines in Chapters 3 and 4.

Figure 1.2



Google advanced search page

Specialized Directories (Resource Guides, Research Guides, and Metasites)

Specialized web directories are collections of selected internet resources (collections of links) on a particular topic. The topic could range from something as broad as medicine to something as specific as biomechanics. These sites go by a variety of names such as resource guides, research guides, metasites, cyberguides, and webliographies. Although their main function is to provide links to resources, they may also incorporate some additional portal features such as news headlines.

Indeed, this category could have been lumped in with the general web directories, but it is kept separate for two main reasons. First, the large general directories, such as the Yahoo! Directory and Open Directory, have several things in common besides being general: They provide categories you can browse, they have a search feature, and when you get to know them, they tend to have the same look and feel in other ways as well. The second main reason for keeping the specialized directories as a separate category is that they deserve greater attention than they often get. More searchers need to tap into their extensive utility.

The Role of Specialized Directories

Use specialized directories when you need to get to know the web literature on a topic, in other words, when you need a general familiarity with the major resources for a particular discipline or area of study. These sites can be thought of as providing some immediate expertise in using web resources in the area of interest. When you are not sure of how to narrow your topic and would like to browse, these sites can also often be better starting places than a general directory because they reflect a greater expertise in the choice of resources for a particular area than would a general directory, and they often include more sites on the specific topic than are found in the corresponding section of a general directory.

Specialized directories are discussed in detail in Chapter 2.

General Strategies

For starters, there is no right or wrong way to search the internet. If you find what you need and find it quickly, your strategy is good. Keep in mind, though, that finding what you need involves other issues: Was it really the correct answer? Was it the best answer? Was it the complete answer?

At the broadest level, assuming that your question is one for which the internet is the best starting place, one approach to finding what you need on the internet is to start by answering the following three questions:

1. Exactly what is my question? (Identify what you need to know and how exhaustive or precise your answer needs to be.)
2. What is the most appropriate tool to start with? (See the previous sections on the categories of finding tools.)
3. What search strategy should I start with?

Answering these questions often takes place without much conscious effort and may take a matter of seconds. For instance, if you wanted to find out who General Carl Schurz was, you could go to your favorite search engine and type in those three words. The quick-and-easy, keep-it-simple approach is often the best.

Even with a more complicated question, it is often worthwhile to start with a very simple approach to get a sense of what is out there, then develop a more sophisticated strategy based on an analysis of your topic into concepts.

Organizing Your Search by Concepts

Thinking in terms of concepts is both a natural way of organizing the world around us and a way of organizing your thoughts about a search. Thinking in concepts is a central part of most searches. The concepts are the ideas that must be present in order for a resultant answer to be relevant, each concept corresponding to a required criterion. Sometimes a search is so specific that only a single concept may be involved, but most searches involve a combination of two, three, or four concepts. For instance, if our search is for *hotels in Albuquerque*, our two concepts are *hotels* and *Albuquerque*. If we are trying to identify webpages on this topic, any web page that includes both concepts possibly contains what we are looking for, and any page that is missing either of those concepts is not going to be relevant.

The experienced searcher knows that for any concept, there will often be more than one term (*cars* as well as *automobiles*) that may indicate the presence of the concept, and these alternate terms also need to be considered. Alternate terms may include, among other things, (1) grammatical variations (e.g., *electricity, electrical*), (2) synonyms, near-synonyms, or closely related terms (e.g., *culture, traditions*), and (3) a term and its narrower terms. For an exhaustive search on the concept *Baltic states*, you may also want to search

for *Latvia, Lithuania,* and *Estonia*. In an exhaustive search for information on the production of electricity in the Baltic states, you would not want to miss the webpage that dealt specifically with "Production of Electricity in Latvia."

When the idea of thinking in concepts is expanded further, it naturally leads to a discussion of Boolean logic, which will be covered in Chapter 4. In the meantime, the major point here is that, in preparing your search strategy, you need to think about what concepts are involved, and remember that, for most concepts, looking for alternate terms may be important.

A Basic Collection of Strategies

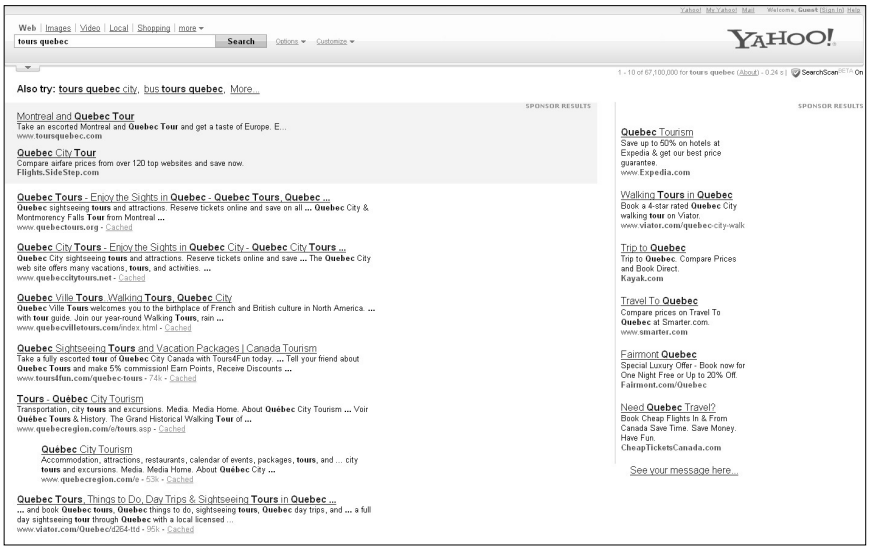
Just as there is no one right or wrong way to search the internet, there can be no list of definitive steps or one specific strategy to follow in preparing and performing every search. Rather, it is useful to think in terms of a toolbox of strategies and select whichever tool or combination of tools seems most appropriate for the search at hand. Among the more common strategies, strategic tools, or approaches for searching the internet are the following:

1. Identify your basic ideas (concepts) and *rely on the built-in relevance ranking* provided by search engines. When you enter terms in the major search engines and many other search sites, only those records (webpages) that contain all those terms will be retrieved, and the engine will automatically rank the order of output based on various criteria (Figure 1.3).
2. Use simple *narrowing techniques* if your results need narrowing:
   - Add another concept to narrow your search (instead of *hotels Albuquerque*, try *inexpensive hotels Albuquerque*).
   - Use quotation marks to indicate phrases when a phrase defines your concept(s) more exactly than if the words occur in different places on the page, for example, "*foreign policy*." Most websites that have a search function allow you to specify a phrase (a combination of two or more adjacent words in the order written) by the use of quotation marks.
   - Use a more specific term for one or more of your concepts (i.e., instead of *intelligence*, try *military intelligence*).
   - Narrow your results to include only those pages that contain your most important terms in the title of the page. (These kinds of techniques will be discussed in Chapter 4.)

3. *Examine your first results* and look for, and then use, relevant terms you might not have thought of at first.
4. *If you do not seem to be getting enough relevant items, use the Boolean OR* operation to allow for alternate terms; for example, *electrical OR electricity* would find all items that have either the term *electrical* or the term *electricity*. How you express the OR operation varies a bit with the finding tool, but in most cases it is the word OR, in capital letters.
5. *Use a combination of Boolean operations* (AND, OR, NOT, or their equivalents) to identify those pages that contain a specific combination of concepts and alternate terms for those concepts (for example, to get all pages that contain either the term *cloth* or the term *fabric* and also contain the words *flax* and *shrinkage*). As will be discussed later, Boolean is not necessarily complicated and is often implied without you doing anything; it can be as simple as choosing between "all of these words" or "any of these words" options.
6. *Look at what else the finding tools (particularly search engines) can do* to allow you to get as much as you need—and only what you need. Advanced search pages are probably the first place you should look.

Ask five different experienced searchers and you will get five different lists of strategies. The most important thing is to have an awareness of the

Figure 1.3



Ranked output from Yahoo! (note "Sponsor Results")

kinds of techniques that are available to you for getting everything you need and, at the same time, only what you need.

Content on the Internet

Not only the amount of information but also the kinds of information available and searchable on the internet continue to increase rapidly. In understanding what you are getting—and not getting—as a result of a search of the internet requires consideration of a number of factors, such as the time frames covered, quality of content, and a recognition that various kinds of material exist on the internet that are not readily accessible by search engines. In *using* the content found on the internet, other issues must also be considered, such as copyright.

TIP:

If you don't immediately see a link to get back to the home page of a site, try clicking on the site's logo. It usually works.

Assessing Quality of Content

A favorite complaint of those remaining people who are still a bit shy of the internet is that the quality of information they find is often low. The same could be said about information available from a lot of other resources. A newsstand may have both the *Economist* and the *National Enquirer* on its shelves. On television, you will find both The History Channel and infomercials. Experience has taught us how, in most cases, to make a quick determination of the relative quality of the information we encounter in our daily lives. In using the internet, many of the same criteria can be successfully applied, particularly those criteria we are accustomed to applying to traditional print resources, both popular and academic.

These traditional evaluation techniques and criteria that can be applied in the internet context include:

1. Consider the source.

From what organization does the content originate? Look for the name of the organization both on the webpage itself and in the URL. Is the content identified as coming from a known source such as a news organization, a government, an academic journal, a professional association, or a major investment firm? Just because the information does not come from such a source is certainly not cause enough to reject it outright. On the other hand, even if it does come from such a source, don't bet the farm on this criterion alone.

Look at the URL. Often you will immediately be able to identify the owner. Peel back the URL to the domain name. If that does not adequately identify its origins, you can check details of the domain ownership on sites that provide access to a Whois database, such as Network Solutions' WHOIS Search (www.networksolutions.com/whois) or DomainTools (www.domaintools.com). For most countries, whois-type sites are available. The Internet Assigned Numbers Authority provides a list of Whois sites by country (www.iana.org/domains/root/db).

Be aware that some look-alike domain names are intended to fool the reader as to the origin of the site. The top-level domain (.edu, .com, etc.) may provide some clues about the source of the information, but do not make too many assumptions here. An .edu or .ac domain does not necessarily assure scholarly content, given that students as well as faculty can often easily get a space on the university server.

A tilde [~] in a directory name is often an indication of a personal page. Again, don't reject something on such a criterion alone. There are some very valuable personal pages out there.

Is the actual author identified? Is there an indication of the author's credentials? The author's organization? Search for other things by the same author. Does she or he publish a lot on spontaneous human combustion or extraterrestrial origins of life on Earth? If you recognize an author's name and the work does not seem consistent with other works from the same author, question it. It is easy to impersonate someone on the internet.

2. Consider the motivation.

What seems to be the purpose of the site—academic, political, consumer protection, sales, entertainment (don't be taken in by a spoof!)? There is nothing inherently bad (or for that matter necessarily inherently good) in any of those purposes, of course, but identifying the motivation can be helpful in assessing the degree of objectivity. Is any advertising on the page clearly identified, or is advertising disguised as something else?

3. Look at the quality of the writing.

If there are spelling and grammatical errors, assume that the same level of attention to detail probably went into the gathering and reporting of the "facts" given on the site.

4. Look at the quality of the documentation of sources cited.

First, remember that even in academic circles, the number of footnotes is not a true measure of a work's quality. On the other hand, and more importantly, if facts are cited, does the page identify the origin of the facts? If a lot rests on the information you are gathering, check out a few of the cited sources to be sure they really do give the facts that were quoted.

5. Is the site and its contents as current as it should be?

If a site is reporting on current events, the need for currency and the answer to the question of currency will be apparent. If the content is something that should be up to date, look for indications of timeliness, such as a "last updated" date on the page or telling examples of outdated material. For example, if it is a site that recommends which search engines to use, and WebCrawler is still listed, don't trust the currency (or for that matter, accuracy) of other things on the page. What is the most recent material that is referred to? If you find a number of dead links, assume the author of the page is not giving it much attention.

6. For facts you are going to use, verify using multiple sources, or choose the most authoritative source.

Unfortunately, many "facts" given on webpages are simply wrong, whether from carelessness, exaggeration, guessing, or other reasons. Often facts are wrong because the person creating that page's content did not bother to check the facts. If you need a specific fact, such as the date of a historic event, look for more than one webpage that gives the date and see if they agree. Also remember that some websites are more authoritative than others. If you have a quotation in hand and want to find who said it, you might want to go to a source such as Bartleby.com (which includes very respected quotations sources), instead of taking the answer from a webpage of lesser-known origins.

For more details and other ideas about evaluating quality of information found on the internet, the following two resources will be useful.

The Virtual Chase: Evaluating the Quality of Information on the Internet
www.virtualchase.com/quality

Created and maintained by Genie Tyburski, this site provides an excellent overview of the factors and issues to consider when evaluating the quality of

information found on a website. She provides checklists and examples of sites that demonstrate both good and bad qualities.

Evaluating the Quality of WWW Resources
www.valpo.edu/library/user/evaluation.html

This site from Valparaiso University provides a detailed set of criteria and also about three dozen links to other sites that address the topic of evaluating web resources. Links to exercises and worksheets on the topic are also included.

Retrospective Coverage of Content

It is tempting to say that a major weakness of internet content is lack of retrospective coverage. This is certainly an issue for which the serious user should have a high level of awareness. It is also an issue that should be put into perspective. The importance and amount of relevant retrospective coverage available depends on the kind of information you are seeking at any particular moment and on your particular question. It is safe to say that no webpages on the internet were created before 1991.

Books, Ancient Writings, and Historical Documents

The lack of pre-1991 webpages does not mean that earlier *content* is not available. Indeed, if a published work is moderately well-known and was written before 1922 or so, you are at least as likely to find it on the internet as in a small local public library. Take a look at the list of works included in the Project Gutenberg site and The Online Books Page (see Chapter 6) where you will find works of Cicero, Balzac, Heine, Disraeli, Einstein, and thousands of other authors. Also look at some of the other websites discussed in Chapter 6 for sources of historical documents.

Scholarly and Technical Journals and Popular Magazines

If you are looking for full-text articles from journals or magazines written several years ago, you are not likely to find them free on the internet (and, for most journal articles, you are not even very likely to find the ones written this week, last month, or last year). This lack of content is more a function of copyright and requirements for paid subscriptions than a matter of the retrospective aspect. The distinction also needs to be made here between free material and "for fee" material on the internet. On a number of internet

sources (such as IngentaConnect and Google Scholar), you can find references to scholarly and other material going back several years. Most likely you will need to pay to see the full text, but fees tend to be very reasonable. Whatever source you use for serious research, whether it's the internet or other, examine the source to see how far back it goes.

Newspapers and Other News Sources

If, when you speak of news, you think of "new news," retrospective coverage is not an issue. But if you are looking for newspaper articles or other news reports dating back more than a few days, the time span of available content on any particular site is crucial. In 2000, many newspaper websites contained only the current day's stories, with a few having up to a year or two of stories. Fortunately, more and more newspaper and other news sites are now archiving their material, and you may find several years of content on the site. Look closely at the site to see exactly how far back the site goes.

Old Web Pages

A different aspect of the retrospective issue centers on the fact that many webpages change frequently and many simply disappear altogether. Pages that existed in the early 1990s are likely either to be gone or to have different content than they did then. This becomes a significant problem when trying to track down early content or citing early content. Fortunately, there are at least partial solutions to the problem. For very recent pages that may have disappeared or changed in the last few days or weeks, a search engine's "cache" option may help. For webpages in their databases, major search engines have stored a copy. If you find the reference to the page in search results, but when you try to go to it, either the page is completely gone or the content that you expected to find on the page is no longer there, click on the "cached" option and you will get to a copy of the page as it was when the search engine last indexed it. Even if you found the page elsewhere initially, search for it using a search engine, and if you find it there, try the cache.

For locating earlier pages and their content, try the Wayback Machine.

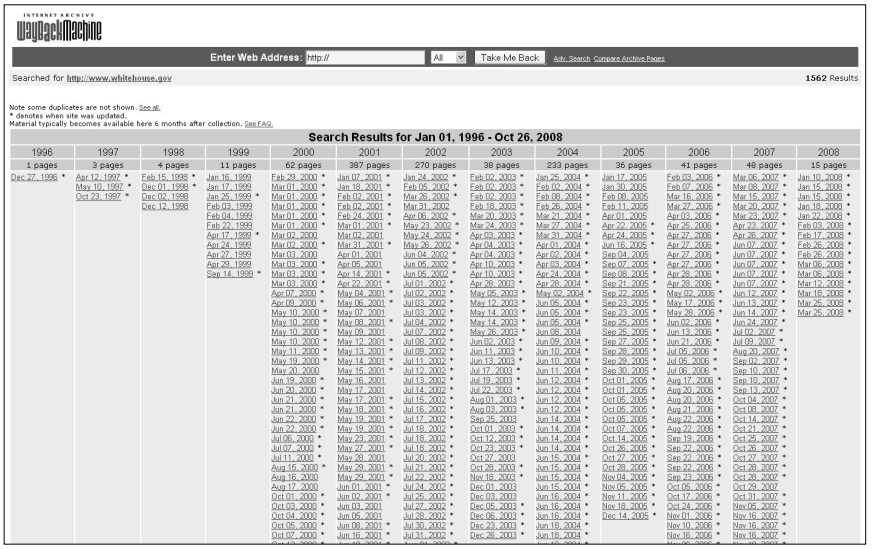
Wayback Machine—Internet Archive

www.archive.org

The Wayback Machine provides access to the Internet Archive, which has the purpose of "offering permanent access for researchers, historians, and scholars to historical collections that exist in digital format." It allows you to

search more than 150 billion pages and see what a particular page looked like at various periods in internet time. A search yields a list of what pages are available for what dates as far back as 1996 (Figure 1.4). As well as web-pages, it archives moving images, texts, music, and other audio. Its producers claim it is the largest database ever built.

Figure 1.4



Wayback Machine search results showing pages available for whitehouse.gov

CONTENT—THE DEEP WEB

No matter how good you are at using web search engines, there are valuable resources on the web that search engines will not find for you. You can *get to* most of them if you know the URL, but a search engine search will probably not find them for you. These resources, often referred to as the Deep Web, the Hidden Web, or the Invisible Web, contain a variety of content, including—and most importantly—databases of articles, data, statistics, and government documents. The term *invisible* here refers to "invisible to search engines." There is nothing mysterious or mystical involved.

Knowing about the Deep Web is important because it contains a lot of tremendously useful information—and it is large. Various estimates put the size of the Deep Web at from 200 to 500 times the content of the visible Web. Before that number sinks in and alarms you, keep in mind the following:

1. The Deep Web contains very important material.
2. For the information there that you are likely to have a need for and the right to access, there are ways of finding out about it and getting to it.
3. While the sheer volume seems overwhelming, most of the material may be meaningless except to those who already know about it, or to the producer's immediate relatives. Much of the material that can't be found is probably not worth finding.

To adequately understand what the Deep Web is all about, one must know why certain kinds of content are not visible to search engine searches. Note the use of the word *content* instead of the word *sites*. The main page of a Deep Web site is usually easy to find and is covered by search engines. It is the rest of the site (webpages and other content within the site) that may be hidden. Search engines do not index certain web content mainly for the following reasons:

1. The search engine *does not know about the page*. No one has submitted the URL to the search engine and no pages currently covered by the search engine have linked to it. (This falls in the category, "Hardly anyone cares about this page, you probably don't need to either.")
2. The search engines have *decided not to index* the content because it is too deep in the site (and probably less useful), the page changes so frequently that indexing the content would be somewhat meaningless (as, for example, in the case of some news pages), or the page is generated dynamically and likewise is not amenable to indexing. (Think in terms of "Even if you searched and found the page, the content you searched for would probably be gone.")
3. The search engine has been *asked not to index* the content by the presence of a robots.txt file on the site that asks engines not to index the site, or not to index specific pages or particular parts of the site. (A lot of this content could be placed in the "It's nobody else's business" category.)
4. The search engine *does not have or use the technology required* to index non-HTML content. This applies to files such as images and a few other file types. Until 2001, this category included file types such as PDF (Portable Document Format) files, Excel files, Word files, and

others that began to be indexed by the major search engines in 2001 and 2002. Audio and video content, such as "flash" movies, have been difficult to index, but with an increased amount of readable data attached to such files, the files are much more searchable and retrievable than they were just a few years ago. Because of this increased coverage, the Deep Web may actually be shrinking in proportion to the size of the total web.

5. The search engine cannot get to the pages to index them because *it encounters a request for a password or the site has a search box* that must be filled out in order to get to the content.

It is the last part of the last category that holds the most interest for searchers—sites that hold their information in databases. Prime examples of such sites would be phone directories, literature databases (such as Medline), newspaper sites, and patents databases. As you can see, if you can find out that the site exists, then you can search its contents (without going through a search engine). This leads to the obvious question of where one finds out about sites that contain unindexed (Deep Web) content.

The best way to find out about these sites is to find a good specialized directory (resource guide) that covers your area of interest. In such a directory, you will find reference to the major websites in that subject area, including websites that contain databases (see Chapter 2 for the discussion of specialized directories).

In the past, there were multiple sites that contained collections of links to major Deep Web websites. Some of the best known have now been discontinued or have not been updated because of the difficulty of adequately keeping up. The following site, however, is a directory of searchable databases that provides another way of finding Deep Web websites for a broad variety of subject areas. For more information on what the Deep Web is, why things are invisible to search engines, etc., you may also want to check out the excellent (though now somewhat dated) book by Chris Sherman and Gary Price, *The Invisible Web: Uncovering Information Sources Search Engines Can't See* (CyberAge Books, Medford, NJ, 2001).

CompletePlanet
completeplanet.com

The site claims to cover "70,000 searchable databases and specialty search engines," but a significant number of the sites are such things as company

website searches, university catalogs, and art gallery catalogs, and many are not necessarily "invisible." It does list a lot of useful resources, but the content on the CompletePlanet site also brings home the point of how trivial much Deep Web material can be.

TIP:

On virtually every site, look for a site index and a search box. They are often more useful for navigating a site than the graphics and links on its home page.

COPYRIGHT

Because of the serious implications of this topic, this section could extend for thousands of words. Because this chapter is about basics, however, a few general points will be made here, and the reader is encouraged to go for more detail to the sources listed next, which are much more authoritative and extensive on the copyright issue. For those in large organizations, particularly an educational institution, you may want to check your organization's website for local guidelines regarding copyright.

Copyright—Some Basic Points

Here are some basic points about copyright:

1. For the U.S., "Copyright is a form of protection provided by the laws of the United States (title 17, U.S. Code) to the authors of 'original works of authorship,' including literary, dramatic, musical, artistic, and certain other intellectual works" (www.copyright.gov/circs/circ1.pdf). As stated on the official U.K. Intellectual Property site, "Copyright gives the creators of a wide range of material, such as literature, art, music, sound recordings, films and broadcasts, economic rights enabling them to control use of their material in a number of ways, such as by making copies, issuing copies to the public, performing in public, broadcasting and use online. It also gives moral rights to be identified as the creator of certain kinds of material, and to object to distortion or mutilation of it" (www.ipo.gov.uk/types/copy/c-about/c-about-faq/c-about-faq-whatis.htm). Other countries will have similar definitions and descriptions according to their own legal definition of copyright. Regardless of the country, copyright (and any failure to acknowledge it appropriately) has legal, moral, and economic implications and repercussions.

2. Assume that what you find on a website is copyrighted, unless the site states otherwise or you know otherwise, based, for example, on the age of the item. See the site for the copyright office in your own country for details about the time frames for copyrights. (In the U.S., of considerable use for webpage creators is the fact that "Works by the U.S. Government are not eligible for U.S. copyright protection" (www.copyright.gov/circs/circ1.pdf). You should still identify the source when quoting something from a site, even if the material is not under copyright.
3. The same basic rules that apply to using printed material apply to using material you get from the internet, the most important being: For any work you write for someone else to read, cite the sources you use.

For more information on copyright and the internet, see the following sources.

U.S. Copyright Office
www.copyright.gov

The official U.S. Copyright Office site has copyright information (for the U.S.) directly from the horse's mouth.

The U.K. Intellectual Property Office—Copyright
www.ipo.gov.uk/copy

The copyright section of the U.K. Patent Office site describes in detail, but also in a very readable fashion, what both the creators and users of copyrighted material need to know

Canadian Intellectual Property Office—A Guide to Copyrights
www.cipo.ic.gc.ca/eic/site/cipointernet-internetopic.nsf/eng/wr00037.html

This is, as the site says, a "guide", not a legal document. Look particularly at the "Twenty Common Questions About Copyright" section. (For other countries, do a search for analogous sites.)

Copyright Website
www.benedict.com

This site is particularly good for addressing, in laypersons' language, the issues involved in the copyright of digital materials. It also provides background and discussion on some well-known legal cases on the topic.

Copyright and Fair Use in the Classroom, on the Internet, and the World Wide Web
www.umuc.edu/library/copy.shtml

This page, from the University of Maryland, is an example of an institutional site that provides practical guidelines—in this case, in the educational context—for use of copyrighted material on websites and elsewhere.

Citing Internet Resources

The biggest problem with citing a source you find on the internet is identifying the author, the publication date, and so forth. In many cases, the information just isn't there, or you have to really dig to find it. Basically, when citing internet sources, you need to give as much of the typical citation information as you would for a printed source (author, title, publication, date, etc.), add the URL, and include a comment such as "Retrieved from the World Wide Web, October 15, 2009" or "Internet, accessed October 15, 2009." If your reader isn't particularly picky, you can just give the information about who wrote it, the title (of the webpage), a date of publication if you can find it, the URL, and when you found the material on the internet. If you are submitting a paper to a journal for publication, to a professor, or including it in a book, you need to be more careful and follow whatever style guide is recommended. Since the details of exactly how you will write the latter kind of citation will vary both with the particular style (MLA, APA, Chicago, etc.) and with the type of publication (articles, books, newsletters, standalone website page, etc.), it is not feasible to provide examples here. Fortunately, many style guides are available online. The following two sites provide links to popular style guides online.

Journalism Resources—Guide to Citation Style Guides
bailiwick.lib.uiowa.edu/journalism/cite.html

Karla Tonella provides links to more than a dozen online style guides.

Citation Styles, Style Guides, and Avoiding Plagiarism: Citing Your Sources
www.lib.berkeley.edu/instruct/guides/citations.html

This site provides a compilation of guidelines based on the following well-known style guides: MLA, APA, Chicago, and Turabian.

KEEPING UP-TO-DATE ON INTERNET RESOURCES AND TOOLS

For those who want to be alerted to the more valuable resources that become available online, the following sites will be useful. Also, numerous specialized sites that cover specific areas (such as science) or tools (such as search engines) will be mentioned throughout the following chapters. All the sites listed here provide free email alert services and also provide archives of past content.

ResourceShelf
www.resourceshelf.com

This site, compiled and edited by Gary Price and Shirl Kennedy, and updated daily, provides extensive updates on new resources. The site also provides a blog newsletter that is extremely useful for being alerted to new sites, particularly those in the Deep Web.

FreePint
www.freepint.com

This U.K.-based site, created by William Hann, provides:

- A free email newsletter with tips on internet searching and reviews of websites
- FreePint Bar: Forums where subscribers can post internet-related research questions and comments
- Resources including book reviews and event listings, and the Free Pint Portal that brings together current and archived Free Pint articles and book reviews, etc.

ResearchBuzz
www.researchbuzz.org/wp

This site, maintained by Tara Calishain, covers news on a broad spectrum of internet research tools and provides articles, archives, and a weekly newsletter.

Internet Resources Newsletter
www.hw.ac.uk/libwww/irn

Produced by the Heriot-Watt University Library, "the free, monthly newsletter for academics, students, engineers, scientists and social scientists" contains descriptions and reviews of new, useful websites, and other internet-related news, reviews, press releases, etc.

The Internet Scout Project
scout.wisc.edu

The Internet Scout Project produces the Scout Report, published since 1994, which provides well-annotated reviews of new sites, with a weekly report on websites for research, education, general interest, and network tools.