Chapter One
Fundamentals of Searching Digital Resources

Joseph Janes, PhD
School of Library and Information Science
University of Washington
jwj@u.washington.edu

For the introductory chapter of the first edition of this book, I wrote a little story about what people often went through in discovering the Internet: big media hype and overblown expectations, initial excitement of discovery of some of the really cool stuff that's there and eventual disillusionment about too much stuff, too much crap, and the difficulty of finding anything really good. I wrote that chapter not that very long ago in real time, but in Web time it's Paleozoic. It actually contains the sentence "Gophers aren't maintained."

That section would seem positively quaint now if it weren't so telling. A lot of what I wrote is pointless because the technology has moved so quickly, but a lot still holds up, in part because the technology hasn't moved that far. There is still quite a bit of good stuff on the Web, a lot of stuff that's not so good, and it can still be very difficult to find that good stuff.

The popular-media story has changed somewhat. It still seems as though we're being told that the entire human record is on the Internet (how many ads with URLs have you seen today?), and that if you just choose the correct service (search engine, portal, directory, etc.), you can find what you're looking for no muss no fuss. Somehow it also seems that 90 percent or so of the Net is porn and therefore we have to have filters or laws to protect us from our baser natures, but that's a matter for another book.

It's not true, of course, but it does pervade our thinking about the Internet. In this chapter, I want to give you an overview of networked information resources and their environment, give some words of wisdom about search strategy and technique, as well as some advice about how to look for stuff out there effectively.

Keep in mind, though, that by the time you read this, many things will have changed. I'll focus less on specific services or features and more on ideas and strategies. Who knows what will have happened by the time the third edition is written?

The Big Picture: Networked Information Resources

The range of information resources available in digital formats, either exclusively, primarily, or as companion versions of print resources, continues to grow daily. We are now accustomed to library catalogs in digital form, often incorporating access to other tools as well (journal indexes, for example). Similarly, access to databases via commercial vendors and CD-ROMs is commonplace.

New genres of resources have arisen, in the distributed networked environment, many of which have no print analogs or which share little with the more stable and familiar sources described above. It is important to understand something about the nature of these new sources and how they differ from their ancestors to be able to use them, search them, and decide when they might be most appropriate.

The Environment: distributed and dynamic

The two most important things to understand about the networked world are these:

- it is a distributed environment, and
- it is a dynamic environment.

These two different but related characteristics define this world and help you to know how to best use it and live in it.

By "distributed," we mean that it has no center, no overall authority, no tangible sense of coherence. There are thousands upon thousands of computers connected to the Internet, each of which has the capability to make information resources available, instantly and in most cases free of charge, to a global audience in the millions. Nobody can stop you, once you're connected, from putting up whatever you like. This unprecedented freedom to publish and communicate ideas clearly has enormous potential for intellectual exchange and the sharing of knowledge.

Furthermore, this environment is dynamic. Each of these resources can change by the second. New ones arise daily; other ones move or become unavailable for one reason or another. Most are not updated that rapidly (and indeed, many are never updated, raising altogether different problems), but the potential for nearly instantaneous responsiveness and creativity is also exciting.

It's also a pain in the neck. Since there are virtually no controls over who can put what out there, things change continually, and since there are no standards for what librarians think of as intellectual control (cataloging, indexing, organization, etc.), it's a mess. It can be fun and challenging and occasionally enlightening to simply wander around, surfing through the contours of the Net, coming across all sorts of new and different things. But there's also a lot of worthless, idiosyncratic garbage, and finding anything good (or indeed anything at all) on a particular topic can be difficult, at best. Most of the things

that we take for granted in the world of books, libraries, and commercial information products exist in crude, simplistic forms or not at all.

Net Resources v. Standard Commercial Resources

Many of the differences we can identify between Net and standard commercial sources stem from the lack of standards in the networked world. The freedom and flexibility offered by the Net have not yet given way or forced the development of standard information structures, search facilities, styles, and so on. There is a growing appreciation among many people who create and use Net resources that such things are important, and some tentative steps have been taken (largely in the realm of metadata, with such ideas as the Dublin Core, see Robert DeCandido, chapter 3), but little has yet become widely accepted or established.

Several characteristics of networked resources are worth noting in more detail:

- **Dynamism:** Although we mentioned this above, it's worth restating. We are not used to books or articles or other print resources that change overnight. To be sure, databases are continually updated and new editions are common, but the nature of these changes are less dramatic than what is found on the Net. It is not unusual, on a typical day of working with networked resources, to find that one has changed its address (and, with luck, has left a link leading you to the new location), another has been updated and thus the interface has changed, a third is gone entirely because its creator has graduated from college and no longer maintains it, and a new one has come up with a great deal of

potentially interesting material. This is something of an exaggeration; not every resource changes every day, but the degree and speed of change on the network are often disorienting.

· **Quality, Review, Authority:** An article does not get published in the *New England Journal of Medicine* or any other scholarly journal without undergoing a rigorous process of peer review and approval. Books do not get published by Random House or major houses without editing both of content and style. Entries are not added to databases like Sociological Abstracts without being checked that they conform with indexing and other policies. These sorts of checks on style, grammar, authority and quality have not been widely implemented in the networked world. There are some scholarly journals that appear in electronic form and a few attempts have been made to enforce standards on other resources, but in general it is *caveat lector* — let the reader beware. Many people are aware of the need of such procedures to make the Net an attractive and worthwhile medium for serious communication and sharing of knowledge, but the lack of central authority and consensus have so far prevented anything from taking hold.

· **Currency:** Despite the fact that networked resources can be updated with ease, many aren't. It is easy to create a resource and even easier to leave it alone once it is up. Maintaining, updating, and developing networked resources is a continual challenge, and one which not all creators meet successfully. Again, a great many resources are up-to-date and current, and thus very useful, and for others continual updating is not necessary, but it can be a problem in some cases.

· **Functionality:** In addition, these resources often have features and functions that more traditional, print resources do not. Some are fairly pedestrian: the ability to search in

new ways (say, in quotation fields of the *Oxford English Dictionary* or in full text of the *Encyclopedia Britannica*) is really just a simple matter once a resource is digital. Others, however, take advantage of the hypertextual environment of the Web — the resource on Parliamentary and Presidential Elections Around the World (http://www.agora.stm.it/elections/election.htm) not only is a good repository of information but also links to Web sites from national parliaments, governments and even political parties and election coverage. Still others make use of the cooperative nature of the Web world; the Internet Movie Database (http://us.imdb.com) is maintained in part by the participation of hundreds of movie nuts around the world and is thus probably the finest resource on film today.

This, then, is the environment in which any searcher must operate in trying to identify potentially valuable information on the Net. It is not as completely hopeless as it may sound from this discussion. A number of tools have been developed which can make finding things easier.

Work the System: Technique and Strategic Tips

Broadly stated, there are three major categories of search tools on the Net: virtual libraries or catalogs (the Argus Clearinghouse, the Internet Public Library), Internet directories (Yahoo!), and search tools (AltaVista, HotBot, Excite, Metacrawler). The first two are primarily browsing tools, although many incorporate search features; the third is primarily for specific searching. The terminology is pretty fluid, and it would not be unusual to find the same resource called by any of these three names. They tend to differ

on the degree of consistency and coherence they provide, as well as, functionality, editorial control, and focus.

When to Use What

Simply put, it's usually best to search for specific words and phrases and known items if they have specific, unambiguous names, using search engines (things like AltaVista, Lycos, Excite, HotBot). It's what they do, and they do it reasonably well. You have to be careful sometimes, because you can get a lot of junk, especially with common or ambiguous words, search engines can often lack features you'd like, and it's hard to know what the databases are like that you're searching.

You're better off using directories such as Yahoo! when you want a number of things that are alike or on the same topic. I go to Yahoo! primarily for this — looking more for categories than sites, and I use their search engine rather than the now cumbersome and overloaded category system.

So, for example, if I were looking for the contents of the last meal before the execution of Ted Bundy, the serial murderer, I'd probably use a search engine, since his name is relatively unambiguous and the phrase "last meal" is pretty meaningful too.

On the other hand, if I were looking for good sites which talk about the television series *Babylon 5*, I'd go to Yahoo!, see if there is a category on the show (there is), and explore what's in that category. As it turns out, there are lots of sites, and Yahoo often provides little or no descriptive information and next to no evaluation, so that process can be tedious, but at least you have all the candidates in front of you.

Preparing to Search

As with any searches in the print or commercial digital domain, a search on the Net requires several steps of preparation: understanding the topic or topics of interest, extracting one or more concepts inherent in the question, identifying potentially useful terms that adequately represent those concepts, selecting possible resources and tools to use, and executing the search.

But of course, as we know, this doesn't always happen. Many searches are the quick-and-dirty, take-no-prisoners type and most Net search services facilitate that sort of thing. There's nothing wrong with them, and we all do them, but if you're serious, you really ought to think about the search for just a moment before diving in headfirst.

Again, it is important to acknowledge the environment in which the search will take place. For example, since there are few standards of indexing, classification, or vocabulary control in networked resources, it is almost always impossible to use any controlled vocabularies or thesauri. Familiar tools such as the Library of Congress Subject Headings or the Thesaurus of Psychological Index Terms are of little help. Rather, colloquial expressions, slang, metaphorical uses of words, and nonstandard writing are common on the Net, and may either obscure potentially useful resources or combine with content words and thus produce lots more hits. (Witness what's happened to the word "spam", which now not only means the meat-in-a-can we all know and love but also unwanted e-mail. Searching for Spam™ recipes is now a more difficult proposition.) Furthermore, since much of what exists on the Net revolves around the Net

itself and computing more generally, searching on words like "Internet", "Net", “computer”, “archive”, “software” and so on is typically useless.

This all means that decision-making on the part of the searcher becomes even more critical. In fact, many of these decisions are similar to those familiar to most librarians. What are the best resources to use in this circumstance, based on experience with the tools, coverage of various resources, features available for use, constraints of time and money, quality of the information, reputation of the source, and so on? The questions don’t really change all that much. This sort of professional assessment of the environment and how best to work within it, though, becomes even more important in such a chaotic and dynamic climate.

Search Features

There are a number of features available in most commercial or library systems which experienced (and not-so-experienced) searchers can use to refine or improve the quality of their searching. These features have evolved over the years as technologies have grown more sophisticated, and they require a substantial amount of preparation and work to implement. In this section, we’ll discuss the more important of these, and see whether they can be used in the networked environment.

In exploring these features, we’ll use the example of looking for information about millennial cults.

Truncation

Many commercial systems support truncation, which allows the searcher to specify that documents must contain at least a particular character string. In different systems, this is represented by various punctuation marks: the question mark, the asterisk, the pound sign, or the percent sign, so a search on, say, CULT? will produce all documents containing words that begin with CULT, such as CULT, CULTS, CULTIST, CULTISM, and so on. This broader search will probably gather most of the documents in the database on cults.

Truncation is a powerful tool, but also has some obvious side effects, the most important of which is the problem of over-truncation (this search will also get us CULTIVATE, CULTURE, and so on). Indeed, some systems have several truncation commands, and an experienced searcher would probably choose to search on CULT? ?,indicating a wish for only one additional character, rather than an arbitrary number.

Truncation is actually rather common in network-based tools, but is not always obvious, may be called other things, and might even be invisible. For example, some search engines automatically truncate. Typing CULT will get you all the variants listed above automatically. You would actually have to actively ask it not to truncate by adding a period to the ends of words.

It's common to see some simple search engines ask if you wish to search for your words as complete words or substrings. "Substring" is a term from computer science, and it means that the series of characters you ask for will occur somewhere in the word but not necessarily at the beginning. (Experienced searchers will recognize that this amounts to implicit left-hand and right-hand truncation.) So a substring search, for example, on CULT will produce all the other variations as well as AGRICULTURE,

HORTICULTURE, etc. This will produce even more false retrievals. Asking to search as complete words stops truncation altogether; there is often no middle ground.

Boolean searching

In our example, the searcher wants documents which have not only the word CULT in some form but also the word MILLENNIUM. This can easily be accomplished by using the Boolean operator AND: MILLENNIUM AND CULT? produces the set of documents containing both required words. The other two Boolean operators, OR (used to combine spelling variations, synonyms, and related words) and NOT (used to eliminate terms) are also available. (A note: some engines require AND NOT, and some require all Boolean operators to be capitalized, or they'll be searched as words.)

In the Net world, there is another set of commands which achieve similar, but not exactly the same, aims. You can often use the plus sign (+) to specify that a word or phrase must be in documents which are retrieved, which is sort of an AND, and the minus sign (-) to specify that a word or phrase must not be in retrievals, a sort of NOT. Thus, you might try

+millennium +cults -carter

to get documents about these cults, but which don't mention Chris Carter, the creator of the television series *Millennium*.

Adjacency, Ranking and Phrases

Experienced online searchers will know that they could do better than MILLENNIUM AND CULT? for this search. Using proximity or adjacency operators, we can specify that these two words must appear next to each other, in this order, for a document to be retrieved. Asking for MILLENNIUM(W)CULT? would not retrieve documents which mention, for example, horticulture in one paragraph and the millennium bug in the next, since AND alone only requires that the two words be in the same document, whether they have anything to do with each other or not.

Many Internet search tools appear as though they provide this feature. They allow searchers to type words and phrases, and magically produce retrievals. However, they really employ algorithms which look for the target words in documents and then calculate scores for each based on several factors: for example, how often the words occur, whether they occur early or late in documents or in titles, whether they occur close together, their overall frequency, and so on. Some of these formulas are quite complex, and they are almost always hidden from the user. So typing MILLENNIUM CULTS, for example, produces a ranked list of documents, largely on topic, but also ranging from the television show *Millennium*, several software sites, and *National Geographic*. These systems are quite powerful and often produce good results, but it can be a bit disconcerting, especially to those experienced with commercial systems, not to be able to have more control over the retrievals they produce.

Although it provides fewer options than proximity operators, it is now common in Net searching to be able to specify that a search be conducted on a phrase (as opposed to ranking individual words). So you could type “millennium cults” —using the quotation

marks, as they are the usual way of specifying phrase searching —and usually only retrieve documents with that complete intact phrase.

At least one Net engine (AltaVista) has permitted weak proximity searching of a sort with the NEAR operator, which will retrieve documents where words are phrases are within five words of each other. A beginning, but there's still quite a way to go for really precise searching. On the other hand, some kinds of searching on the Web could not be done any other way, like searching for graphics by type.

Fields & Context

In commercial systems, there are a great many other commands at the searcher's disposal. We'll discuss only two more: the ability to search for words in particular fields of the document (the title, the abstract, the index terms) and the ability to focus a search based on the context of documents (say, only retrieving documents from particular years or in particular languages). One might see a search such as this:

(MILLENNI?(“N)CULT? ?)/TI AND PY>=1998

which would look for variants of the phrase “millennium cults “ or “millennial cults” in the title field of documents published in 1998 or later.

These features are only now emerging on the Internet. It is likely that something like this will arise as more time and effort are spent on developing search facilities, but at the moment, only a few rudimentary features of this sort are available. They often take the

form of check boxes or fill-in fields, removing the need for users to remember a large or cumbersome set of commands.

The reason these are only appearing slowly if at all is due to the amount of work required to make them available. Not only do the search engines need to be designed, but database indexes must be sophisticated enough to support searching of words together, in specific fields. Documents must include tags about dates, languages, authorship, and other contextual information. Then there is the whole question of indexing, classification, and name authority, so common in familiar library systems, which is only now dawning on the Net world. There is clearly a long way to go to make these systems as reliable and powerful as those found in the commercial realm.

Advice for the Searchlorn

Finally, let me offer a few pieces of advice which I think will be of assistance. Again, they're not specific to any particular service, but I find them to be helpful.

Use the most direct approach

Search engines and directories are fine, and indeed useful in many situations. But you should always try to figure out what is the most direct line of search. In particular, think if you need to search at all. It may be possible to go directly to a Web site without searching. For example, if you were looking for information about the American Library Association, you might go to http://www.ala.org. Web sites like

http://www.whitehouse.gov, http://www.biography.com, and http://www.nytimes.com are good examples of this phenomenon.

Sometimes it takes a bit of guesswork (http://www.amnesty.org for Amnesty International, http://www.nwa.com for Northwest Airlines, http://www.state.mi.us for the State of Michigan), and you can go badly wrong (http://www.whitehouse.com —which you should know is an explicit sexual site), but this can often be an effective (and quick) technique.

Let somebody else do the work

Why search if somebody else has done the hard work for you? Use a resource like the Argus Clearinghouse (http://www.clearinghouse.net) or the Internet Public Library (http://www.ipl.org), which selects and organizes resources in categories and on particular topics.

Also, when searching in an area I know little about but in which there is a lot of information, I often try to find a site compiled by an expert (or at least somebody with a lot of interest in the area and time on their hands). Sites such as this not only provide access to a lot of information but also provide a point of view, an organizational structure, background, additional information, and so on.

A variant of this strategy is to find a Usenet group, listserve or Web ring in the area, look for an FAQ (frequently asked questions list), and, if you don't find the answer there, post a question to the list and see if anyone answers.

Take advantage of full text searching and ranking

As professional searchers know, full text searching is a double-edged sword. The ability to search the entire text of documents is powerful, but sometimes too powerful, and it opens the door to many problems of ambiguity, synonymy, metaphor and just plain weird uses of words (How many senses of the word "pitch" can you think of right now? Exactly.).

But you might as well take advantage of it. If you can think of a way to say something, there's always a chance that somebody else has thought of it too and more to the point has put up a Web site on it. If you're looking for information about numerical values assigned to the Greek alphabet, you could do worse than search on that phrase and see what happens.

Find a favorite, but don't get too attached to it

It seems that most experienced Web searchers I know have a favorite search engine or service. They're never really happy with it all the time, but it makes life easier to have a single place to start and not have to think about it really hard.

On the other hand, part of the trick here is to know when not to start with your favorite. Knowing special features of other search engines or when to use a metasearcher or directory can save time. Furthermore, be prepared to abandon your favorite when you feel it's not helping any more. Perhaps its interface has become too difficult to use, your results are suffering, or the database isn't being refreshed as often. When it's over, it's over, and you should just move on.

Don't be afraid to ask for help, but don't expect miracles

All of these sites have help screens, which vary in terms of actual help provided. None of them will tell you how a search engine works, because the methods they use are their competitive advantage, but at the very least they'll tell you what kinds of commands are available, suggestions on how to use them and so on. Be sure to notice the differences between "regular" and "advanced" searching —they can be substantial and surprising. To use features such as field searching, Boolean operators, and so on, you may have to use "advanced" searching screens, but sometimes you give up functionality to do that, so be sure you know what you're doing.

In addition, there are sites —my favorite is Search Engine Watch (http://www.searchenginewatch.com) — that provide information on all search mechanisms, comparison tests, how to use them effectively, how to write your documents to get high scores, even gossip (see Sara Ryan's chapter for more detail). These can be very useful, especially in getting the big picture and knowing how things are changing.

What will this chapter in the third edition of this book look like?

Good question. Much as I hate predictions, I think it's likely that as we proceed into an information environment which increasingly incorporates networked resources (but not to the exclusion of print and commercial digital stuff), we will see:

- more sophisticated technique available for searchers. I doubt it will ever be of the Dialog, Lexis/Nexis type technique, but it will get better.

- It will be easy to use (because it has to be), which probably means lots of clicky boxes and menus and options, hidden to beginning searchers but available for experts.
- more value-added services like categories of human-selected good sites that we now see, maybe context-sensitive automatic help or even human help (perhaps for a fee, but it's still hard to get small amounts of money over the Web -- a technological but primarily psychological hurdle).
- specialized search engines and services by subject area, population, language, etc., to make it easier to find good stuff. Of course, finding those services will now become a challenge.
- greater presence of libraries, librarians and our perspective on information. Call this one a hopeful scenario for the future. What librarians bring to the information world is of such value; I just hope we can find a way to incorporate it and help people to realize its importance.

I reserve the right to be completely wrong about these, and if I write the introductory chapter to the third edition, to not even mention them at all.

This is Chapter One of The Internet Searcher's Handbook, by Peter Morville, Louis Rosenfeld and Joseph Janes; revised by GraceAnne A. DeCandido (New York: Neal-Schuman, 1999), available from Amazon and Barnes & Noble.