The Web has grown so fast and is changing so rapidly that we’ve had to change the way we use it to find information. In the olden days—that is, just last year—you could get away with hotlisting or bookmarking the URLs for sites that interested you. As these lists grow into hundreds of entries they become unusable, and as Webs shut down and change locations, URLs (and bookmarks) become obsolete—as a recent NYT article put it “Home Pages Never Die; You Must Kill Them” (New York Times, January 2, 1996).

Today the best approach to working with this amorphous and chaotic mass of information is to mine it with a good Web search tool. Usually these are search engines that query large, descriptive indexes of Web pages. Each search engine typically searches its own index of page descriptions, not the Web itself. Some engines add thousands of entries a day to their databases.

Although most search tools encourage you to register your site with them, these vast indexes are primarily built by robots. Robots, also called bots, spiders, wanderers, worms, and crawlers, are programs that live on the search services’ host computers. Despite the names, spiders and such don’t actually traverse the Web. They stay on a host computer and retrieve information using the usual Web protocols such as http (hyper-text transfer protocol). The information retrieved is then used to build the indexes. Effective search engines revisit sites periodically to update their databases.

Meta-indexes

So—what search engine should you use to find your way through the Net? One good option is to try a meta-index such as MetaCrawler. Simultaneously querying eight search engines—Open Text, Lycos, WebCrawler, InfoSeek, Excite, Inktomi, Alta Vista, and Galaxy—MetaCrawler sends back an integrated set of results. Search options include phrase, all words (Boolean AND), or any words (Boolean OR), and searches can be limited by region (country, domain).

Consider the following:

- As of January 1996 there were nearly 10 million host computers connected to the Internet (http://www.nw.com/zone/WWW/report.html).
- Home use of the Web doubled in the last six months of 1995; 7.5 million households—8 percent of the U.S. total—have access to the Web. (WSJ, March 7, 1996).
- CompuServe began offering home page service at the end of November 1995 and within the first 10 days 10,000 people posted pages. More than 50,000 subscribers now have CompuServe-based Web pages (WSJ, March 4, 1996).
- The Library’s Web came up in March 1995 and a year later has well over 3,400 pages.
- The Web has “mass mediafied” the Internet. Product advertisements routinely cite URLs. Hotwired (the Web sibling of the magazine Wired) had an estimated advertising revenue of $720,000 in the fourth quarter of 1995 and Netscape had advertising revenues of $1.8 million.
may 1996 9

and by site type, for example, .edu or .com. No single search engine can give comprehensive coverage of Internet sites, nor are all search engines equally effective for all types of searches. With a meta-index you don’t need to learn a variety of interfaces or know the relative strengths of each index, and you’re closer to getting a comprehensive search.

Savvy Search is a new meta-index that reduces network traffic and claims to be sensitive to server loads and types of searches. This experimental system constructs a search plan for each query by ranking and grouping 19 search engines and directing the search to run simultaneously on the engines in the top-ranked group. The searcher can opt to continue the search through the next set of (second ranked) indexes. We found that Savvy Search’s top choices produced variable results—at times it was hard to believe that the best engines really had been selected.

But—at some point you’ll want to get to know some specific indexes. Here are a few things to consider when choosing a Web search engine:

• Check out whether it’s fast and accessible.

• Read the help pages describing advanced search options and suggestions for refining searches. (If you’re using Alta Vista for example, you’ll want to know that you need to search for a phrase by putting the words within quotes.) The relevance ranking on which these engines are based can give some pretty disconcerting results unless you use the power features that most services offer.

• Be aware of the scope of the service: Does it aim for comprehensiveness (Lycos, Alta Vista, Open Text), or is it selective and evaluative (Magellan)?

• Make sure you know what’s indexed—all the text in a document (WebCrawler, OpenText) or parts of documents (Lycos, World Wide Web Worm)—just Web documents (AliWeb) or FTP and Gopher sites also (Open Text, Lycos, WebCrawler)? Are news groups included (Alta Vista, Yahoo)?

• Look at the format used for displayed results. Does the display give you a good sense of what you’ve found?

• And be aware of useful features like Open Text’s indexing of every word, or World Wide Web Worm’s option to search on hyperlinks?

The Web indexing scene is developing so rapidly that any comparative review of engines is inevitably outdated before its published. Here we’ve selected four favorites—WebCrawler, Magellan, Lycos, and Alta Vista—to illustrate the differences and best uses of various engines. We suggest that you bookmark the Internet Search Tools page on the Library’s Web (see below) and revisit it now and then to check out new engines and new features. (continued on next page)
WebCrawler

Bigger isn’t necessarily better! If you’re in a hurry and you’re looking for a known item, e.g., the Daily Pennsylvania home page, the text of the Communications Decency Act, or Penn’s notable African Studies page, you may want to try a non-comprehensive, economical search engine. Fast and simple, WebCrawler was one of the first search engines on the Web.

Originally developed at the University of Washington, it’s been maintained by America Online since June 1995. The easy-to-use interface presents a limited number of options: all or any words, and a choice of 10, 25, or 100 results (with an option to continue with more). WebCrawler searches the full text of all pages and presents very brief results—the title of the Web page only. A useful relevance number at the left of the result line indicates the relative relevance ranking of each result. WebCrawler doesn’t waste your time retrieving all the sub-pages for a site—it generally takes you straight to the top level if that’s what you’re asking for.

Magellan

You can waste a lot if time retrieving pages only to find that they weren’t worth the wait! Search engines that provide useful descriptions can save you time. So a product like The McKinley Group’s Magellan is welcome. The Magellan result sets come with text written by real people! Readable and grammatical, they’re a nice change from the more commonly found computer-generated text. For example, a Magellan search for the text of the Maastricht Treaty presents the official European Union page as the top-ranked document with a useful accompanying blurb: “this site provides visitors with the full text of the Maastricht Treaty, the treaty which is the backbone of the European Union...” Good to know this isn’t a personal home page with a glancing mention of the Treaty.

Magellan bears watching. It’s pretty new, ambitious, and different. With a highly informed editorial staff it ranks 40,000 of its 2 million indexed sites with a four-star ranking for depth, accuracy and timeliness, organization and ease of access, and net appeal. Magellan itself gets a low ranking from us on timeliness—at press time it still listed the Penn Library Gopher as alive and well even though we killed it over a year ago!

Lycos and Alta Vista

But if you’re wondering if your best friend from first grade has a personal home page yet, or you need to find the text of an obscure poem or a quote from Oscar Wilde, you’ll want to try one of the search engines that doesn’t even consider selectivity useful! For comprehensiveness

references


URLs cited in the article are:

• ALIWEB (http://www.traveller.com/aliweb).

• Alta Vista (http://www.altavista.digital.com).

• Lycos (http://www.lycos.com).

• Magellan: McKinleys Internet Directory (http://www.mckinley.com).


• Open Text (http://www.opentext.com:8080).

• Savvy Search (http://savvy.cs.colostate.edu:2000).

• WebCrawler (http://webcrawler.com).


• Yahoo (http://www.yahoo.com).
it’s hard to beat Lycos, the engine that claims unabashedly to be the biggest catalog on the planet. Lycos’ most recent count (February/March 1996) indicates that it indexes 11.6 million Web pages (adding at a rate of 50,000 a day!).

Fighting it out for biggest and best is Digital Equipment’s Alta Vista, which also claims to give access to all 8 billion words found in over 16 million Web pages! Alta Vista is worth knowing about because it indexes both the Web and Usenet news groups—you choose which you want to search up front. But watch out for systems like this. If you’re looking for the DP home page, the Alta Vista simple search won’t do it—it finds 9000 matches. The advanced search works fine—but you need to read the instructions before trying it!

What’s ahead?

Unless there are some significant changes, chances are that it’ll be increasingly difficult to find exactly what you need on the Web. Sun Microsystems Java programming language for the Web is generating all kinds of commotion—could Java applets make it easier to find information on the Web? While it’s still much too early to tell, it’s certainly conceivable that Java programs could do automatic, unattended database-oriented searches across the Web and return the results to you. Clearly systems like Alta Vista, with the power to search on a URL, title, links, specific newsgroups, etc., will be increasingly valuable—but the effectiveness of Web search tools will still be limited by the format of the data they’re indexing.

The downside of the “everyone a publisher” environment of the Web is the lack of standards. In contrast to the highly controlled library model of MARC standards (machine readable cataloging), which have resulted in the consistency and effectiveness of online library catalogs, Web document creators follow few rules. The National Center for Supercomputing Applications (NCSA) and the Online Computer Library Center (OCLC) are sponsoring promising efforts to identify a set of metadata elements (The Dublin Core) that could become the required standards that Web page creators would use to describe their documents (subject, author, title, date, type of file, genre of work, etc.). Whether the free-wheeling democracy of the Web will respond to such an initiative is an open question, but it’s possible that serious information providers would conform.

Meantime, stay tuned! You can be sure that none of this will stay the same!

PATRICIA RENFRO is Director of Public Services for the University of Pennsylvania Libraries; WILLIAM GARRITY is Associate Director for Information Services at the Biomedical Library.

What if Dr. Seuss Did Technical Writing?
—Anonymous

Here’s an easy game to play.
Here’s an easy thing to say:
If a packet hits a pocket
on a socket on a port,
And the bus is interrupted
as a very last resort,
And the address of the memory
makes your floppy disk abort,
Then the socket packet pocket
has an error to report!

If your cursor finds a menu item
followed by a dash,
And the double-clicking icon puts your
window in the trash,
And your data is corrupted ’cause the
index doesn’t hash,
Then your situation’s hopeless, and your
system’s gonna crash!

You can’t say this? What a shame sir!
We’ll find you another game sir.

If the label on the cable on the table
at your house,
Says the network is connected to
the button on your mouse,
But your packets want to tunnel
on another protocol,
That’s repeatedly rejected
by the printer down the hall,

And your screen is all distorted
by the side effects of gauss
So your icons in the window are
as wavy as a souse,
Then you may as well reboot and
go out with a bang,
’Cause as sure as I’m a poet,
the sucker’s gonna hang!

When the copy of your floppy’s
getting sloppy on the disk,
And the microcode instructions cause
unnecessary risc,
Then you have to flash your memory and
you’ll want to RAM your ROM.
Quickly turn off the computer and
be sure to tell your mom!