Categorically NOT

Don't believe everything you read about search engines

For the last several years, it's seemed as if you couldn't pick up a newspaper or listen to news radio without reading, or hearing someone say, something like, "The word 'spin doctor' was used in 1,343 newspaper articles last year, according to a search on the widely used Nexis database."

Lexis-Nexis, formerly known as Mead Data Service, is a database of all news stories, in full, from a variety of publications that license their material for retrieval from the database —generally for a hefty fee per retrieval. (That's the Nexis part; Lexis is a legal information service, like Westlaw.)

In fact, prior to the Web, Lexis-Nexis, or one of its competitors, like Knight-Ridder Information Services (formerly Dialog Information Services), provided the only broad access to electronic versions of newspaper and magazine articles.

Reporters short on news would sometimes do searches, paid for by their publishers, to find out how many times particular terms, like "search doctors," "bully pulpit," or "digerati," were used. The total matches in the Nexis database would be weighed against each other or discussed individually.

The danger in this kind of reportage on searches is that the conclusions are generally ridiculous. The reporters making the searches might have used the search tools well or poorly; sometimes multiple versions of the same story are filed from successive editions of a newpaper, increasing the number of matches; finally, it's not real reporting to draw conclusions about trends merely from the number of occurrences of a given word—there's no context for that conclusion.

William Safire, on the other hand, has often used Nexis in his language

column in the *New York Times Magazine* as a method of showing the increasing popularity in use of a given word or phrase. Dictionary editors often use the public press as method of determing whether a word should be added to the lexicon, so the match is perfect for Safire's use of the database to discuss words and phrases in current circulation. The number of times given terms are used is a side-effect, or epiphenemon, of the collection of massive amounts of reportage.

All this is preamble—honest—to my real point, which is that the Web has become one giant Nexis engine for folks who aren't doing the legwork (or mousework, as the case may be).

Why so nasty?

I can't recall the number of times I've seen, in both mainstream and Net- or computer-oriented publications, the phrase, "A search on Yahoo yielded a list of 43 organizations selling fungicides targeted at common foot mold." Not that exact phrase, obviously. But its ilk are seemingly everywhere. (Even a

recent feature here at *adobe.mag* recently used that construction.)

Unfortunately, although Yahoo certainly does everything it can to imply that it's exhaustive, it's not—and the majority of Net users wouldn't know this unless told otherwise. What Yahoo *attempts* to do is catalog new resources as they become available on the Net; but, by its nature, it catalogs more self-listed entries—listings from people who submit their own Web sites—than a demographically or randomly sampled array of available information. Any attempt to use Yahoo to show either empirically arrived-at numbers or weighted percentages is therefore futile.

Under some circumstances it may be interesting to know if *any* companies, or whatever it is you're looking for, are listed by Yahoo or another Web search engine. But some may contend that it's useful to know if dozens, say, or hundreds are there, because (they suppose) if hundreds are listed, thousands may well exist. This supposition is wrong.

Carl Sagan uses logic like this, to a bit more effect, to speculate that it's pretty darn likely that life on other planets analogous to ours, and with tech-

nology relatively similar to ours, exist throughout the universe. The logic goes something like this: Given a sufficiently vast number of planets overall, a smaller but still vast number of planets could support life like ours, and therefore some smaller subset of those planets would develop consciousness and eventually use radio signals.

On the other hand, the number of Web sites is not in the trillions, so we can't presuppose large numbers of sites about any particular subject. Empiricism is *everything*. If you took the logic, say, that since Yahoo has cataloged 5 percent of all Web sites and lists 3 sites devoted to Jane Austen, that therefore there must be at least 60 such sites on the Web, it's time for you to remove that three-pronged device from its complementary receptable and go to bed.

On the other hand, if you used the three Austen sites to examine links to further sites, and then went on to use various search engines to track down more references, and finally

summarized the results, you'd be doing good research and could make some claims about the results. Where the Sagan argument is concerned, for example, now that astronomers have more or less proven that several planets besides those in our own solar system exist, they're a bit more confident in extrapolating the results out to the rest of space.

Web empiricism

Finding out what's really out there is, of course, a process of hard work. If you're trying to compile numbers on how many sites exist on a given subject or in a given industry, you have to start with the notion that nobody has everything: not Yahoo, not Alta Vista, not any of the sites attempting complete listings in particular categories.

The Web is constantly in flux, as I discussed in my June 15, 1996, WebSpy column. Thus, by definition, only constant, omnipotent, empirical observation of the medium could yield a definitive listing. Since neither I nor anyone I know is omnipotent, we can only approach the goal of total knowledge

without ever reaching it.

I say "approach" because, in the time you've taken to read this article, hundreds of sites have come up and dozens have probably disappeared; many others have changed their URLs. Given the numbers involved, it's happening that fast. The only way to find out if something is still there or if something new has arrived is to check and get a contemporary—and only temporary—answer.

Spiders like Alta Vista, Lycos, and excite constantly forage over the Web, retrieving millions of pages a week, indexing their contents, and updating the databases that visitors to these services use to search for terms (how search engines work will be the subject of a future column). However, the engines are always behind.

Solutions

Several simultaneous technological developments may change the picture slightly, though the basic conclusions above are still true.

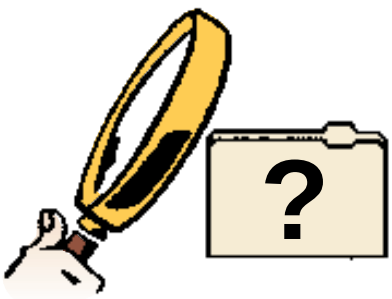
Brute force. New spiders, like the announced-but-not-yet-public Ultraseek (from Infoseek), plan to churn through information much more rapidly, possibly checking on a given page once a week to see if it's changed. This is the brute-force method. Alta Vista visits fairly often, but spot checks by this reporter indicate that they're clearly weeks and months out of sync on some sites.

Metasearches. The experimental and soon-to-be commercial Metacrawler *(http://metacrawler.cs.washington.edu)*, still running as a project at the University of Washington, allows a user to enter a single search request. Metacrawler rewrites that request into the language of each of ten Net search engines and indexes, then retrieves the results, removes overlaps, and compiles a nice, final product.

Distributed information architecture. The system called Domain Naming Service (DNS) allows any machine on the Internet to find the "real," unique address of any other machine connected to the Internet. The machine looking for this number doesn't rely on a central database with all information; it uses

a set of pointers that tell it where to go to find this out. This distribution of information means that each location maintains a small, always-accurate set of information, while the organization required to retrieve it must simply maintain an updated set of pointers to find it.

Put simply, information is local, maps are global. It's like having a map of Texas that shows you how to get to a Chamber of Commerce in Austin, where they'll tell you about all the different hotels and restaurants you can go to in the city. There's been a lot of discussion about developing this kind of distributed scheme for indexing information on the Web, but nothing has captured the wireheads' hearts and minds yet.



When in doubt, look it up

The moral of this story is that it's always easier to use an arbitrary fact as a starting point than to do the real research to back it up. Okay, that's not the moral, that's the snide commentary. The real moral is that you shouldn't accept any

source, whether it's a news story or a search index, that claims to have the absolute, exhaustive answer to what's out there. You have to take your own journey to know for sure.