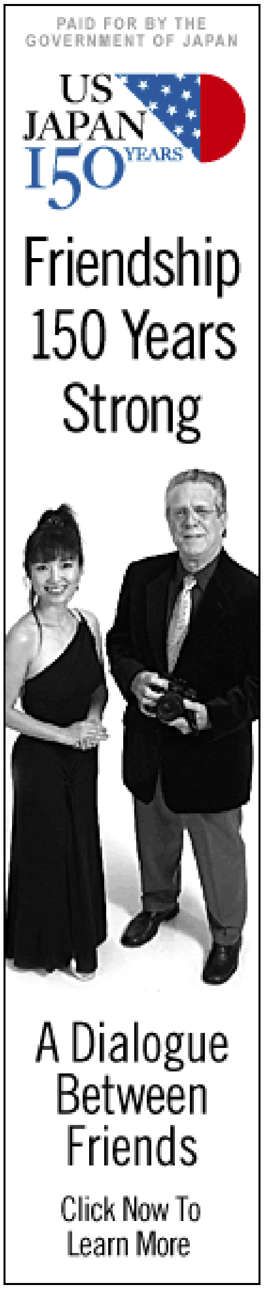
washingtonpost.com

Online Search Engines Help Lift Cover of Privacy

By Yuki Noguchi
Washington Post Staff Writer
Monday, February 9, 2004; Page A01


PAID FOR BY THE GOVERNMENT OF JAPAN
US JAPAN 150 YEARS
Friendship 150 Years Strong
A Dialogue Between Friends
Click Now To Learn More

Sitting at his laptop, Chris O'Ferrell types a few words into the Google search engine and up pops a link to what appears to be a military document listing suspected Taliban and al Qaeda members, date of birth, place of birth, passport numbers and national identification numbers.

Another search yields a spreadsheet of names and credit card numbers.

"All search engines will get you this," O'Ferrell said, pointing to files of spoils he has found on the Internet: Medical records, bank account numbers, students' grades, and the docking locations of 804 U.S. Navy ships, submarines and destroyers.

And it is all legal, using the world's most powerful Internet search engine.

Cybersecurity experts say an increasing number of private or putatively secret documents are online in out-of-the-way corners of computers all over the globe, leaving the government, individuals, and companies vulnerable to security breaches. At some Web sites and various message groups, techno-hobbyists are even offering instructions on how to find sensitive documents using a relatively simple search. Though it does not technically trespass, the practice is sometimes called "Google hacking."

"There's a whole subculture that's doing this," said O'Ferrell, a long-time hacking expert and chief technology officer of Herndon-based security consultancy Netsec Inc.

In the decade they have been around, search engines like Google have become more powerful. At the same time, the Web has become a richer source of information as more businesses and government agencies rely on the Internet to transmit and share information. All of it is stored on computers called servers, each one linked to the Internet.

For a variety of reasons -- improperly configured servers, holes in security systems, human error -- a wide assortment of material not intended to be viewed by the public is, in fact, publicly available. Once Google or another search engine finds it, it is nearly impossible to draw back into secrecy.

That is giving rise to more activity from "Googledorks," who troll the Internet for confidential goods, security engineers said.

"As far as the number of sites affected by this, it's in the tens of thousands," said Johnny Long, 32, a researcher and developer for Computer Sciences Corp. and veteran hacker who maintains a Web site that he says keeps him connected to the hacker community. He spoke about Google hacking at the Def Con hacker convention in Las Vegas last summer, which has led to more awareness of vulnerabilities, he said.

Google gets singled out for these searches because of its effectiveness.

"The reason Google's good is that they give you more information and they give you more tools to search," O'Ferrell said.

Its powerful computer "crawls" over every Web page on the Internet at least every couple weeks, which means surfing every public server on the globe, grabbing every page, and every link attached to every page. Those results are then catalogued using complex mathematical systems.

The most basic way to keep Google from reaching information in a Web server, security experts said, is to set up a digital gatekeeper in the form of an instruction sheet for the search-engine's crawler. That file, which is called robots.txt, defines what is open to the crawler and what is not. But if the robots.txt file is not properly configured , or is left off inadvertently, a hole is opened where Google gets in. And because Google's crawlers are legal, no alarms will go off.

"The scariest thing is that this could be happening to the government and they may never know it was happening," Long said. "If there's a chink in the armor, [the hackers] will find it."

Google and other search-engine officials said they are sensitive to the problem, but are not in a position to control it.

With a vast system of more than 10,000 computer systems constantly collecting new information on more than 3 billion Web sites, the company cannot and does not want to police or censor what goes on the Web, said Craig Silverstein, Google's chief technology officer.

"I think Web masters have to be careful," he said. "The basic problem is that with 3 billion [Web sites], there's a lot of information out there." It offers a tool on its own Web site, "Webmaster guidelines," on how to remove Web sites from Google's system, including Google's vast store of cached pages that may no longer be available online, Silverstein said.

For hacking experts, Google-hacking has a kind of populist allure: any one with Internet access can do it if they know the right way to search.

"It's the easiest point-and-click hacking -- it's fun, it's new, quirky, and yet you can achieve powerful results," said Edward Skoudis, a security consultant for INS Inc., which helps government and business clients monitor what is visible from the Web. "This concept of using a search engine for hacking has been around for a while, but it's taken off in the last few months," probably because of a new-found enthusiasm in the underground hacking community, he said.

Search strings including "xls," or "cc," or "ssn" often brings up spread sheets, credit card numbers, and Social Security numbers linked to a customer list. Adding the word "total" in searches often pulls up financial spreadsheets totaling dollar figures. A hacker with enough time and experience recognizing sensitive content can find an alarming amount of supposedly private information.

"On a [client's] bank site, I found an Excel spread sheet with 10,000 Social Security and credit card numbers," said Skoudis, of one of his successful treasure hunts.

The bank's Web server had been properly configured to keep such documents private, but someone had mistakenly put the information on the wrong side of the fence, he said. "Google found the open door and crawled in."

Skoudis confronted the "red-faced executives" with his findings, he said, and was told: "Just fix it, damn it."

Google and other search-engine operators are unable to gauge how frequently private documents are accessed using their sites, or how many are removed for security reasons.

"The challenge is that as the search-engine tool evolved, people got more lax about what they put on a publicly available Web server," said Tom Wilde, vice president and general manager of Terra Lycos's 19 search engines. "It would be impossible to monitor" the tens of millions of searches that take place every day, Wilde said, adding that he has never been notified of a security breach on his sites.

Government officials said they were familiar with Google hacking, and were working with government agencies and businesses to secure sensitive documents on Web servers.

"It's an issue we're aware of and tracking," said Amit Yoran, director of the cybersecurity division of the Homeland Security Department. By law, each agency is responsible for its own security, and although hacking or security breaches are reported to Homeland Security, the cybersecurity division does not monitor the content of the Web, he said.

It is unclear who is at fault when someone digs up a confidential document.

"I don't know what law's been violated just for searching" on a publicly available search engine, said Paul Bresson, a spokesman for the FBI, noting the bureau has not yet taken actions against individuals who have found secure documents by using search engines. "If they use it for some sinister purpose, that's another issue."

The availability of private information contributes to rising incidence of identity theft, which for the last four years has been the No. 1 consumer problem for the Federal Trade Commission. Last year the FTC received nearly 215,000 complaints about identity theft, up from about 152,000 in 2002.

Since 2001, the FTC has settled cases with Eli Lilly & Co., Microsoft Corp. and clothing maker Guess Inc. for not taking "reasonable" measures to keep medical or financial information secure, said Jessica Rich, assistant director of the commission's bureau of consumer protection. Letting customer information reside on an unsecure server can open up a business to such liability.

"There are unique vulnerabilities because of databases that are accessible through the Web," Rich said, adding that the FTC anticipates bringing more security-related cases in the future.

Once confidential pages are found, it is not easy to get them back under wraps.

Even after a document has been pulled off of a Web server, as was the case when MTV removed from its Web site a pre-Super Bowl press release promising "shocking moments" at the halftime show, documents often remain cached, or stored, in other search engines' computers so they can still be accessed.

"Once it is placed online, it's very hard to get the digital horse back in the electronic barn," said Marc Rotenberg, executive director of the Electronic Privacy Information Center. "It's close to impossible to get it back."