

CNET News

February 23, 2009 6:57 AM PST

Exploring a 'deep Web' that Google can't grasp

By Alex Wright

[The New York Times](#)
nytimes.com
 The New York Times

One day last summer, Google's search engine trundled quietly past a milestone. It added the one trillionth address to the list of Web pages it knows about.

But as impossibly big as that number may seem, it represents only a fraction of the entire Web.

Beyond those [trillion pages](#) lies an even vaster Web of hidden data: financial information, shopping catalogs, flight schedules, medical research, and all kinds of other material stored in databases that remain largely invisible to search engines.

THE LATEST BUSINESS NEWS FROM
nytimes.com
 The New York Times

- ➔ [U.S. Agrees to Raise Its Stake in Citigroup](#)
- ➔ [Economic Scene: A Bold Plan Sweeps Away Reagan Ideas](#)
- ➔ [Flowers' Bank Bets Not Always So Sweet](#)
- ➔ [Time Warner Cable Spin-Off to Finish Next Month](#)

Last Update: 08:20 AM ET

The challenges that the major search engines face in penetrating this so-called deep Web go a long way toward explaining why they still can't provide satisfying answers to questions like "What's the best fare from New York to London next Thursday?" The answers are readily available--if only the search engines knew how to find them.

Now a new breed of technologies is taking shape that will extend the reach of search engines into the Web's hidden corners. When that happens, it will do more than just improve the quality of search results--it may ultimately reshape the way many companies do business online.

Search engines rely on programs known as crawlers, or spiders, that gather information by following the trails of hyperlinks that tie the Web together. While that approach works well for the pages that make up the surface Web, these programs have a harder time penetrating databases that are set up to respond to typed queries.

- [Yahoo! Buzz](#)

"The crawlable Web is the tip of the iceberg," said Anand Rajaraman, co-founder of [Kosmix](#), a deep Web search start-up whose investors include Jeff Bezos, chief executive of Amazon.com.

Kosmix has developed software that matches searches with the databases most likely to yield relevant information, then returns an overview of the topic drawn from multiple sources.

"Most search engines try to help you find a needle in a haystack," Rajaraman said, "but what we're trying to do is help you explore the haystack."

Haystack of databases

That haystack is infinitely large. With millions of databases connected to the Web, and endless possible permutations of search terms, there is simply no way for any search engine--no matter how powerful--to sift through every possible combination of data on the fly.

To extract meaningful data from the deep Web, search engines have to analyze users' search terms and figure out how to broker those queries to particular databases. For example, if a user types in "Rembrandt," the search engine needs to know which databases are most likely to contain information about art (say, museum catalogs or auction houses), and what kinds of queries those databases will accept.

That approach may sound straightforward in theory. But in practice, the vast variety of database structures and possible search terms poses a thorny computational challenge.

"This is the most interesting data integration problem imaginable," says Alon Halevy, a former computer science professor at the University of Washington who is now leading a team at Google that is trying to solve the deep Web conundrum.

[Google's deep Web search strategy](#) involves sending out a program to analyze the contents of every database it encounters. For example, if the search engine finds a page with a form related to fine art, it starts guessing likely search terms--"Rembrandt," "Picasso," "Vermeer" and so on--until one of those terms returns a match. The search engine then analyzes the results and develops a predictive model of what the database contains.

Nixing the 'naive way'

In a similar vein, Professor Juliana Freire at the University of Utah is working on an

ambitious project called [DeepPeep](#) that eventually aims to crawl and index every database on the public Web. Extracting the contents of so many far-flung data sets requires a sophisticated kind of computational guessing game.

"The naive way would be to query all the words in the dictionary," Freire said. Instead, DeepPeep starts by posing a small number of sample queries, "so we can then use that to build up our understanding of the databases and choose which words to search."

Based on that analysis, the program then fires off automated search terms in an effort to dislodge as much data as possible. Freire claims that her approach retrieves better than 90 percent of the content stored in any given database. Freire's work has recently attracted overtures from one of the major search engine companies.

As the major search engines start to experiment with incorporating deep Web content into their search results, they must figure out how to present different kinds of data without overcomplicating their pages. This poses a particular quandary for Google, which has long resisted the temptation to make significant changes to its tried-and-true search results format.

"Google faces a real challenge," said Chris Sherman, executive editor of the Web site Search Engine Land. "They want to make the experience better, but they have to be super-cautious with making changes for fear of alienating their users."

Beyond the realm of consumer searches, deep Web technologies may eventually let businesses use data in new ways. For example, a health site could cross-reference data from pharmaceutical companies with the latest findings from medical researchers, or a local news site could extend its coverage by letting readers tap into public records stored in government databases.

This level of data integration could eventually point the way toward something like the [semantic Web](#), the much-promoted--but so far unrealized--vision of a Web of interconnected data. Deep Web technologies hold the promise of achieving similar benefits at a much lower cost, by automating the process of analyzing database structures and cross-referencing the results.

"The huge thing is the ability to connect disparate data sources," said Mike Bergman, a computer scientist and consultant who is credited with coining the term deep Web. Bergman said the long-term impact of deep Web search had more to do with transforming business than with satisfying the whims of Web surfers.

Entire contents, Copyright © 2009 The New York Times. All rights reserved.

See more CNET content tagged:

[haystack](#), [search engine](#), [Internet search](#), [query](#), [database](#)