Saturday, February 28, 2009 03:50:18 PM

To search, type and hit ente SEARCH

RSS

SUBSCRIBE TO OUR MAILING LIST

First Name:

Last Name:

Email Address:

SUBSCRIBE

Privacy Policy

- # Featured Archive

Artists of the 55th Carnegie International: Andro Wekua

July 3rd, 2008

- Archives
  - February 2009
  - January 2009
  - December 2008
  - November 2008
  - October 2008
  - September 2008
  - August 2008
  - July 2008
  - June 2008
  - May 2008
  - April 2008

Print
Email
Share
DIGG
STUMBLEUPON
MIXX
DEL.ICIO.US

The Knowns, the Unknowns, and the Unknown Knowns of the Web: The Deep Web

February 27th, 2009 by John Eastman

It is frightening to think of how dependent on the internet modern society has become. Perhaps the greatest communication tool in the world to date, the world wide web has a vast amount of servers, databases, and information pages, that serve up useful information in abundance on almost any subject. Its use is growing daily, the useful web2.0 tools that developers continue to add are constantly expanding that use in new ways. In 2008, Google reported that it has added the one trillionth address to the list of web pages it knows about, that is, pages of information that it can find using its search engines and web crawler-like technologies. As significant as that is, and as vast as informational sources appear to be to the general public, there are many unknowns (information that is accessible through varying user interfaces, but is unattainable via search engine results). The web pages that are currently locatable are only a fraction of what is available in the world's content of published knowledge.

Millions of databases containing trillions of pages of information do not surface using existing web search technology according to industry experts and researchers who are working on new search technologies to bring into the mainstream a broader set of information. While there are many industry-specific databases in existence used by industry insiders, their data is confined in a number of ways. Information regarding flight schedules, for instance, is available from airline companies to their

staff and travel agents, and is also available to organizations such as priceline.com, cheatickets.com, etc. But specific airline information– flight numbers, routes, and fee schedules– is largely unavailable to the general public in the manner that search engines search the web and present data to users. The public's gateway to these service companies who make it available is upon specific request. Booking a flight from New York to Chicago, for instance, through one of these websites forces you to pick your location, your destination, travel times and the like, but not to view all of the possible airlines, travel routes, times, and fees. Medical research of already published information, including medical records of an individual as treated by multiple doctors and medical facilities is largely unavailable. Financial information that large institutions house (banking for instance, a keen interest at this time) is not available unless you look for specifics by institution. Asking a question such as "what banking institution has the best rates on CD's in the US?" yields the user a host of information, but that information cannot be presented in an unbiased ranking. It is presented and manipulated by companies who pay for positioning on the web, and a direct answer is not presented. Make no mistake, in many instances this information is already in digital form, it is just not on the radar of web crawlers due to the design and how it is manipulated.

The realm of hidden internet data is referred to as the "Deep Web". Researchers from various institutions and companies are working on technology that searches database for much more vast and detailed content than the public is accustomed to seeing when conducting a search. If successful, these type of search results may well reshape how individuals use the web, as well as change the way that many firms who now offer products and services operate. An entire new strategy may need to be designed for conducting commerce on the web. Deep Web technologies will essentially extend the reach of all search engines on the web to locate and present trillions of pages of content that is now housed in databases in multiple languages across the world. These technologies will greatly improve the quality of the search for public users, academic researchers, those in the financial industry, and the like. Mostly apparent to the public will be the ability to see "pure" information presented in detail form, with a much broader manner in which to query by. The ability for a web search to cross reference multiple databases and return results from multiple queries will be at the root of a successful Deep Web based product.

As previously alluded to, search engines using web crawler or spider-like technology that gather informational pages published on the web by seeking out and subsequently following hyperlinks that tie or link the web together. Web crawler technology essentially locates information that is published at what may be considered the surface of the web. Web crawler technology is not efficient at conducting queries of specific user-requested information, the so-called "database style". A totally new method of search is required that entails a major shift of thinking and further development of technological tools to enable it to become a reality. Many experts indicate that much of the information contained in these databases is text, as opposed to graphic images, and a more structured way of searching and conducting queries is necessary.

Due to the huge volume of data present in the Deep Web, there has been a increasing interest to find a concrete approach that would allow users and computer applications to leverage this information.

Participants and startup firms in the industry who are working on Deep Web technology include www.kosmix.com and Google. A notable investor of Kosmix is Jeffrey P. 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.

Google launched "Google Scholar" in 2004. It is a search engine with an index able to search most peer-reviewed online journals of the world's major scientific publishers. Similar search engines commonly are only available to those who pay a subscription fee. Scholar also enables users to search for digital or physical copies of articles.

Using its "group of" feature, it shows the various available links to journal articles. As of December 2006, it can access both published versions of articles and major open access repositories, but it does still not cover individual university pages. Access to self-archived non-subscription versions is now provided by a link to Google, where one can find such open access articles.

A similar application, DeepPeep was started as a project at the University of Utah and is overseen by Juliana Freire, a professor at the university's School of Computing WebDB group. DeepPeep has the daunting goal of crawling and indexing every database on the public web. Unlike traditional search engines, which crawl easy-to-view existing webpages, DeepPeep aims to allow access to the Deep Web. A specific milestone is to make 90% of all world wide web content accessible, according to Freire. The project runs a beta search engine and is sponsored by the University of Utah and a $243,000 grant from the National Science Foundation.

While these firms are developing product and search technologies that will return detailed searches across a vast amount of databases, it is thought by many experts in the industry that a single search engine technology will not be the likely solution. This is due the fact that building a search engine for the public's use that scans every possible public database with billions of pages to sift through, and using the possibility of thousands of queries, all in seconds may be beyond the reach of any one individual search product. The data integration that challenges these type of searches is monumental.

Google's approach towards a Deep Web search strategy product involves dispatching an application to database content and analyzing the contents of every database it encounters. Some Deep Web strategies involve analyzing users' search terms and then making deductions as to how to use them, and predict a model to effectively search available databases. Google, as the major search engine firm, begin to explore and deploy their initial efforts for deep search, the risk is that they could complicate existing search results for end users. Furthermore, a giant company like Google has mastered a tried and true search format that they continue to enhance as well, in web crawler style format.

And the "unknown knowns", the information that specific databases hold, represents information that may become available in the future, that the web's current technology does not even know about… yet. ∎

Pages:

1 response so far.

- **Matthew Theobald** - Feb 28, 2009 at 3:18 am

  For you consdieration.

An emerging deep web standard called the Internet Search Environment Number.
http://www.isen.org

Feel free to ask open questions about our solution(s).

Matt

Leave a Comment

Name Email

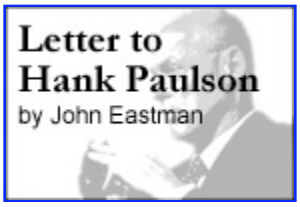
Submit Comment

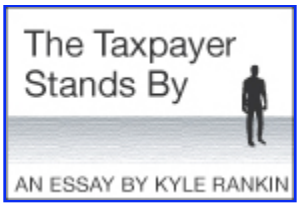
- Sponsors

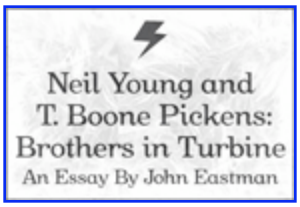
- Essays


U.S. Treasury Funds:
Hidden Intentions?
An Essay By John Eastman


We Are
the Dollars
and Sense
An Essay By
John Eastman

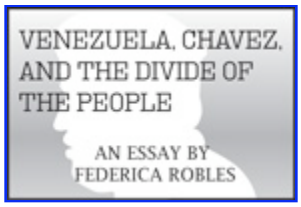

Letter to
Hank Paulson
by John Eastman


The Taxpayer
Stands By
AN ESSAY BY KYLE RANKIN

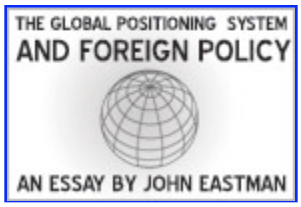

Neil Young and
T. Boone Pickens:
Brothers in Turbine
An Essay By John Eastman


The iPhone
and the
Killer App
An Essay By John Eastman

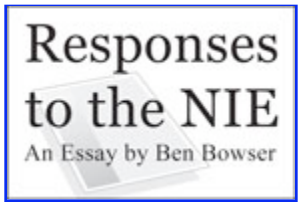

Global Ideas
Outside the Box
AN ESSAY BY JOHN EASTMAN



Voices from India



sovereign wealth funds



Take part in something, become part of the results. Participate in a Black and White public interest poll.