



White Paper: The Deep Web: Surfacing Hidden Value

Volume 7, Issue 1: *Taking License*, August, 2001

DOI: <http://dx.doi.org/10.3998/3336451.0007.104>

[Permissions](#)

EPUB

Print

Share

This White Paper is a version of the one on the [BrightPlanet](#) site. Although it is designed as a marketing tool for a program "for existing Web portals that need to provide targeted comprehensive information to their site visitors," its insight into the structure of the Web is worthwhile reading for all those involved in e-publishing. —J.A.T.

Searching on the Internet today can be compared to dragging a net across the surface of the ocean. While a great deal may be caught in the net, there is still a wealth of information that is therefore missed. The reason is simple: Most of the Web's information is buried far beneath the surface of dynamically generated sites, and standard search engines never find it.

Traditional search engines create their indices by spidering or crawling surface Web pages. When a page is discovered, the page must be static and linked to other pages. Traditional search engines cannot retrieve content in the deep Web — those pages do not exist until they are created as the result of a specific search. Because traditional search engine crawlers can not penetrate the deep Web, the deep Web has heretofore been hidden.

The deep Web is qualitatively different from the surface Web. Deep Web sources store searchable databases that only produce results dynamically in response to a direct query. A direct query is a "one at a time" laborious way to search. BrightPlanet's search technology is a process of making dozens of direct queries simultaneously using multiple-thread technology. It is the only search technology, so far, that is capable of identifying, retrieving, qualifying, and organizing both "deep" and "surface" content.

If the most coveted commodity of the Information Age is indeed information, then the size of the Web content is immeasurable. With this in mind, BrightPlanet has quantified the size of the deep Web in a study based on data collected between March 13 and 30, 2000. Our findings include:

- Public information on the deep Web is currently 400 to 550 times larger than the defined World Wide Web.
- The deep Web contains 7,500 terabytes of information compared to nineteen terabytes of information in the surface Web.
- The deep Web contains nearly 550 billion individual documents compared to the surface Web.
- More than 200,000 deep Web sites presently exist.
- Sixty of the largest deep-Web sites collectively contain about 750 terabytes of information, sufficient by themselves to exceed the size of the surface Web forty times.
- On average, deep Web sites receive fifty per cent greater monthly traffic than surface sites; more highly linked to than surface sites; however, the typical (median) deep Web site is unknown to the Internet-searching public.
- The deep Web is the largest growing category of new information on the Internet.
- Deep Web sites tend to be narrower, with deeper content, than conventional surface sites.
- Total quality content of the deep Web is 1,000 to 2,000 times greater than that of the surface Web.
- Deep Web content is highly relevant to every information need, market, and demographic.
- More than half of the deep Web content resides in topic-specific databases.
- A full ninety-five per cent of the deep Web is publicly accessible information – either free or subscriptions.

To put these findings in perspective, a study at the NEC Research Institute [\[1\]](#), published

estimated that the search engines with the largest number of Web pages indexed (such as Northern Light) each index no more than sixteen per cent of the surface Web. Since the deep Web when they use such search engines, Internet searchers are therefore 0.03% — or one in 3,000 — of the pages available to them today. Clearly, simultaneous access to multiple surface and deep Web sources is necessary when comprehensive information is needed.

The Deep Web

Internet content is considerably more diverse and the volume certainly much larger than is understood.

First, though sometimes used synonymously, the World Wide Web (HTTP protocol) is only a subset of Internet content. Other Internet protocols besides the Web include FTP (file transfer), email, news, Telnet, and Gopher (most prominent among pre-Web protocols). This paper will consider further these non-Web protocols. [2]

Second, even within the strict context of the Web, most users are aware only of the information available to them via search engines such as [Excite](#), [Google](#), [AltaVista](#), or [Northern Light](#), or such as [Yahoo!](#), [About.com](#), or [LookSmart](#). Eighty-five percent of Web users use search engines to find needed information, but nearly as high a percentage cite the inability to find desired information as one of their biggest frustrations. [3] According to a recent survey of search-engine satisfaction by researcher [NPD](#), search failure rates have increased steadily since 1997. [4a]

The importance of information gathering on the Web and the central and unquestioned role of search engines — plus the frustrations expressed by users about the adequacy of these engines — is an obvious focus of investigation.

Until Van Leeuwenhoek first looked at a drop of water under a microscope in the late 17th century, he had no idea there was a whole world of "animalcules" beyond their vision. Deep-sea exploration in the past thirty years has turned up hundreds of strange creatures that challenge old ideas about life and where it can exist. Discovery comes from looking at the world in new ways and using new tools. The genesis of the BrightPlanet study was to look afresh at the nature of information and how it is being identified and organized.

How Search Engines Work

Search engines obtain their listings in two ways: Authors may submit their own Web search engines "crawl" or "spider" documents by following one hypertext link to another, and search engines returns the bulk of the listings. Crawlers work by recording every hypertext link in an index crawling. Like ripples propagating across a pond, search-engine crawlers are indices further and further from their starting points.

"Whole new classes of Internet-based companies choose the Web as their preferred medium for commerce and information transfer"

The surface Web contains an estimated 2.5 billion documents, growing at a rate of 7.5 million documents per day. Search engines have done an impressive job in reaching this reach, though Web growth itself has exceeded the reach of search engines [6a][7a]. Today, the three largest search engines in terms of internally reported documents include Northern Light with 1.35 billion documents (500 million available for searches), [8] Fast, with 575 million documents, and Northern Light with 327 million documents. [10]

Legitimate criticism has been leveled against search engines for these indiscriminate crawls, mostly because they provide too many results (search engines, for example, with Northern Light, and you will get about 47 million hits. Also, because results are found from links within other documents, those documents that are cited are more likely to be found than new documents — up to eight times as likely. [5b]

To overcome these limitations, the most recent generation of search engines (notably Northern Light) replaced the random link-following approach with directed crawling and indexing based on the "popularity" of pages. In this approach, documents more frequently cross-referenced by other documents are given priority both for crawling and in the presentation of results. This approach provides superior results when simple queries are issued, but exacerbates the tendency to ignore documents with few links. [5c]

And, of course, once a search engine needs to update literally millions of existing Web pages, the freshness of its results suffer. Numerous commentators have noted the increased delay in recording new information on conventional search engines. [11a] Our own empirical studies of search engine currency suggest that listings are frequently three or four months — or more — out of date.

Moreover, return to the premise of how a search engine obtains its listings in the first place. The information is only identified if it is adjusted for popularity or not. That is, without a linkage from another Web document, it will never be discovered. But the main failing of search engines is that they depend on the ability to identify what is on the Web.

Figure 1 is a graphical representation of the limitations of the typical search engine. The information identified is only what appears on the surface and the harvest is fairly indiscriminate. There is a tremendous value that resides deeper than this surface content. The information is hidden beneath the surface of the Web.

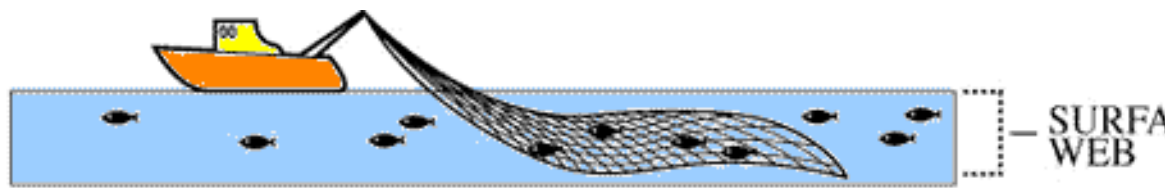


Figure 1. Search Engines: Dragging a Net Across the Web's Surface

Searchable Databases: Hidden Value on the Web

How does information appear and get presented on the Web? In the earliest days of the Web, there were relatively few documents and sites. It was a manageable task to post all documents. Because all pages were persistent and constantly available, they could be crawled easily by conventional search engines. In July 1994, the Lycos search engine went public with 54,000 documents. [12] Since then, the compound growth rate in Web documents has been of more than 200% annually! [13a]

Sites that were required to manage tens to hundreds of documents could easily do so with static HTML pages within a static directory structure. However, beginning about 1996, things changed in this place. First, database technology was introduced to the Internet through such vendors as [Sapphire/Web](#) ([Bluestone](#) has since been bought by HP) and later [Oracle](#). Second, this technology was commercialized initially via directories and search engines, but rapidly evolved to include dynamic content. And, third, Web servers were adapted to allow the "dynamic" serving of Web pages (using technologies such as Microsoft's ASP and the Unix PHP technologies).

This confluence produced a true database orientation for the Web, particularly for

now accepted practice that large data producers such as the [U.S. Census Bureau](#), [Securities and Exchange Commission](#), and [Patent and Trademark Office](#), not to mention whole new Internet-based companies, choose the Web as their preferred medium for commercial information transfer. What has not been broadly appreciated, however, is that the means by which they provide their information is no longer through static pages but through database-driven applications.

It has been said that what cannot be seen cannot be defined, and what is not defined cannot be understood. Such has been the case with the importance of databases to the information on the Web. And such has been the case with a lack of appreciation for how the older model of static Web pages — today's paradigm for conventional search engines — no longer applies to the information content of the Internet.

In 1994, Dr. Jill Ellsworth first coined the phrase "invisible Web" to refer to information that was "invisible" to conventional search engines.^[14] The potential importance of search engines was also reflected in the first search site devoted to them, the AT1 engine that was met with much fanfare in early 1997.^[15] However, PLS, AT1's owner, was acquired by AOL in 1998, and thereafter the AT1 service was abandoned.

For this study, we have avoided the term "invisible Web" because it is inaccurate. The "invisible" about searchable databases is that they are not indexable nor able to be queried by conventional search engines. Using BrightPlanet technology, they are totally "visible" and do not need to be accessed through a search engine.

Figure 2 represents, in a non-scientific way, the improved results that can be obtained using BrightPlanet technology. By first identifying where the proper searchable databases reside, a directory of links can then be placed to each of these sources simultaneously to harvest only the results of interest with pinpoint accuracy.

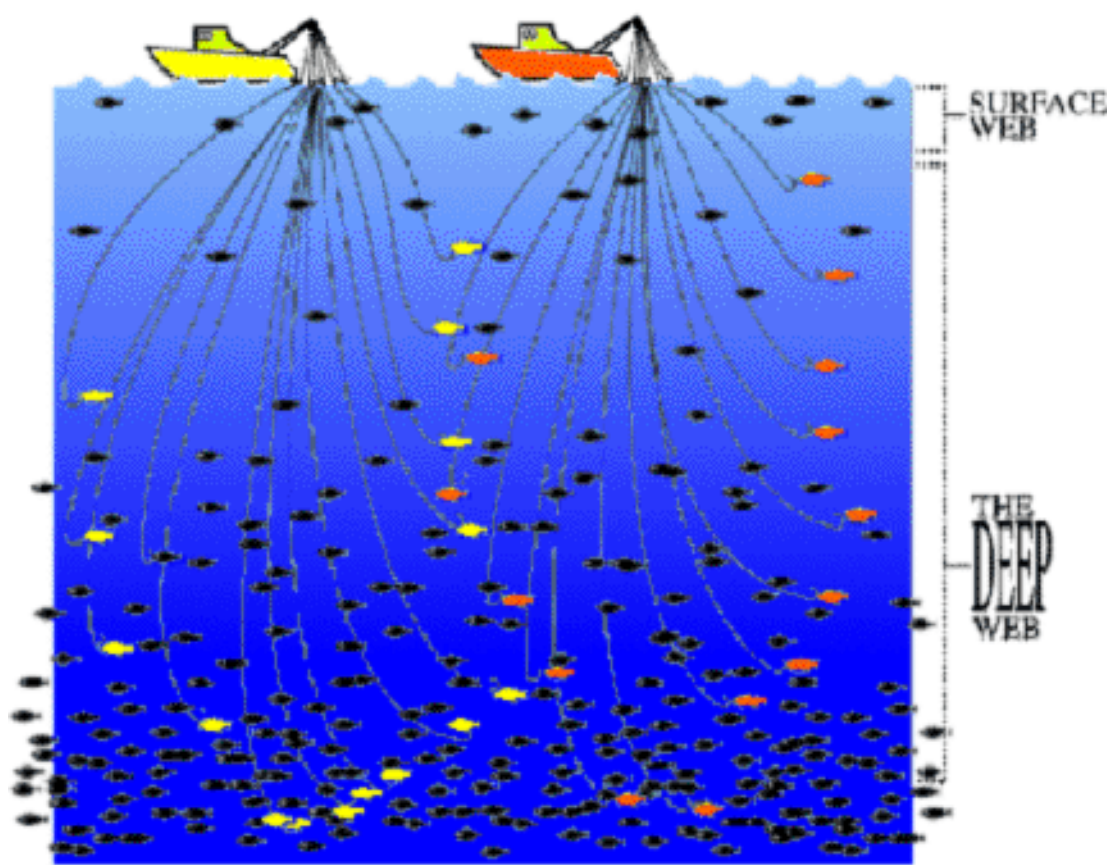


Figure 2. Harvesting the Deep and Surface Web with a Directed Query Engine

Additional aspects of this representation will be discussed throughout this study. For however, the key points are that content in the deep Web is massive — approximately 100 times more than that visible to conventional search engines — with much higher quality throughout.

BrightPlanet's technology is uniquely suited to tap the deep Web and bring its results to the surface. The simplest way to describe our technology is a "directed-query engine." It has other features in results qualification and classification, but it is this ability to query multiple sites directly and simultaneously that allows deep Web content to be retrieved.

Study Objectives

To perform the study discussed, we used our technology in an iterative process. Our objectives were:

- Quantify the size and importance of the deep Web.
- Characterize the deep Web's content, quality, and relevance to information searchers.
- Discover automated means for identifying deep Web search sites and directing searchers to them.
- Begin the process of educating the Internet-searching public about this heretofore hidden and valuable information storehouse.

Like any newly discovered phenomenon, the deep Web is just being defined and uncovered. As we have continued our investigations, we have been amazed at the massive scale and complexity of the deep Web. This white paper concludes with requests for additional insights and information that will enable us to continue to better understand the deep Web.

What Has Not Been Analyzed or Included in Re

This paper does not investigate non-Web sources of Internet content. This study also ignores private intranet information hidden behind firewalls. Many large companies have document stores that exceed terabytes of information. Since access to this information on a large scale can not be defined nor can it be characterized. Also, while on average 44% of the content of typical Web documents reside in HTML and other coded information (for example, JavaScript), [16] this study does not evaluate specific information within that code. We do not include those codes in our quantification of total content (see next section).

Finally, the estimates for the size of the deep Web include neither specialized search engines which may be partially "hidden" to the major traditional search engines — nor the contents of the search engines themselves. This latter category is significant. Simply accounting for the contents of search engines and average Web document sizes suggests search-engine contents are on the order of terabytes or more [17] or somewhat larger than the known size of the surface Web.

A Common Denominator for Size Comparison

All deep-Web and surface-Web size figures use both total number of documents (or total number of pages in the case of the deep Web) and total data storage. Data storage is based on "HTML document size estimates." [13b] This basis includes all HTML and related code information and text content, exclusive of embedded images and standard HTTP "header" information. This standard convention allows apples-to-apples size comparisons between the surface Web and the deep Web. The HTML-included convention was chosen because:

- Most standard search engines that report document sizes do so on this same basis.
- When saving documents or Web pages directly from a browser, the file size by convention.
- BrightPlanet's reports document sizes on this same basis.

All document sizes used in the comparisons use actual byte counts (1024 bytes per

"Estimating total record count per site was often not straightforward"

In actuality, data storage from deep-Web documents therefore be considerably less than the figure. Actual records retrieved from a searchable database forwarded to a dynamic Web page template through such as standard headers and footers, ads, etc. this HTML code content overstates the size of databases, standard "static" information on the presented in the same manner.

HTML-included Web page comparisons provide the common denominator for comparing surface Web sources.

Use and Role of BrightPlanet Technology

All retrievals, aggregations, and document characterizations in this study used BrightPlanet technology. The technology uses multiple threads for simultaneous source queries and downloads. It completely indexes all documents retrieved (including HTML content). After documents are downloaded and indexed, the documents are scored for relevance using four different algorithms, prominently vector space modeling (VSM) and standard and modified extended information retrieval (EBIR).^[19]

Automated deep Web search-site identification and qualification also used a modified BrightPlanet technology employing proprietary content and HTML evaluation methods.

Surface Web Baseline

The most authoritative studies to date of the size of the surface Web have come from the work of Giles of the NEC Research Institute in Princeton, NJ. Their analyses are based on the "publicly indexable" Web. Their first major study, published in *Science* magazine in 1998, from December 1997, estimated the total size of the surface Web as 320 million documents. An update to their study employing a different methodology was published in *Nature* magazine using analysis from February 1999.^[5d] This study documented 800 million documents on the publicly indexable Web, with a mean page size of 18.7 kilobytes exclusive of images :

headers. [20]

In partnership with Inktomi, NEC updated its Web page estimates to one billion documents in 2000. [21] We have taken this most recent size estimate and updated total document size of the entire surface Web based on the 1999 *Nature* study:

Table 1. Baseline Surface Web Size Assumptions

Total No. of Documents	Content Size (GBs) (HTML basis)
1,000,000,000	18,700

These are the baseline figures used for the size of the surface Web in this paper. (A study from Cyveillance [5e] has estimated the total surface Web size to be 2.5 billion documents per day at a rate of 7.5 million documents per day. This is likely a more accurate number, but the study is still used because they were based on data gathered closer to the dates of our own study.)

Other key findings from the NEC studies that bear on this paper include:

- Surface Web coverage by individual, major search engines has dropped from a high of 20% in 1998 to 16% in 1999, with Northern Light showing the largest coverage.
- Metasearching using multiple search engines can improve retrieval coverage by 10% or so, though combined coverage from the major engines dropped to 42% from 19% in 1998.
- More popular Web documents, that is, those with many link references from other sites, have up to an eight-fold greater chance of being indexed by a search engine than do documents with few references.

Analysis of Largest Deep Web Sites

More than 100 individual deep Web sites were characterized to produce the listing of sites reported in the next section.

Site characterization required three steps:

1. Estimating the total number of records or documents contained on that site.
2. Retrieving a random sample of a minimum of ten results from each site and then

expressed HTML-included mean document size in bytes. This figure, times the number of site records, produces the total site size estimate in bytes.

3. Indexing and characterizing the search-page form on the site to determine searchability.

Estimating total record count per site was often not straightforward. A series of tests were used on each site and are listed in descending order of importance and confidence in deriving total document count:

1. E-mail messages were sent to the webmasters or contacts listed for all sites requesting verification of total record counts and storage sizes (uncompressed). 13% of the sites shown in Table 2 provided direct documentation in response.
2. Total record counts as reported by the site itself. This involved inspecting records on the site, including help sections, site FAQs, etc.
3. Documented site sizes presented at conferences, estimated by others, etc. This involved comprehensive Web searching to identify reference sources.
4. Record counts as provided by the site's own search function. Some site search engines return record counts for all queries submitted. For others that use the NOT operator for stand-alone use, a query term known not to occur on the site such as "NOT COUNTRY" was issued. This approach returns an absolute total record count. Failing these tests, a broad query was issued that would capture the general site content; this number was then corrected for an empirically determined "coverage factor," generally in the range of 0.5 to 1.0.
5. A site that failed all of these tests could not be measured and was dropped from the final listing.

Analysis of Standard Deep Web Sites

Analysis and characterization of the entire deep Web involved a number of discrete tasks:

- Qualification as a deep Web site.
- Estimation of total number of deep Web sites.
- Size analysis.
- Content and coverage analysis.
- Site page views and link references.
- Growth analysis.
- Quality analysis.

The methods applied to these tasks are discussed separately below.

Deep Web Site Qualification

An initial pool of 53,220 possible deep Web candidate URLs was identified from existing seven major sites and three minor ones. [23] After harvesting, this pool resulted in 45 listings after tests for duplicates. cursory inspection indicated that in some cases the one link removed from the actual search form. Criteria were developed to predict the case. The BrightPlanet technology was used to retrieve the complete pages and both the initial unique sources and the one-link removed sources. A total of 43,348 were actually retrieved.

We then applied a filter criteria to these sites to determine if they were indeed search proprietary filter involved inspecting the HTML content of the pages, plus analysis content. This brought the total pool of deep Web candidates down to 17,579 URLs.

Subsequent hand inspection of 700 random sites from this listing identified further : Ninety-five of these 700, or 13.6%, did not fully qualify as search sites. This correction applied to the entire candidate pool and the results presented.

Some of the criteria developed when hand-testing the 700 sites were then incorporated automated test within the BrightPlanet technology for qualifying search sites with 98% accuracy. Additionally, automated means for discovering further search sites were incorporated into our internal version of the technology based on what we learned

Estimation of Total Number of Sites

The basic technique for estimating total deep Web sites uses "overlap" analysis, the chosen for two of the more prominent surface Web size analyses. [6b][24] We used our based on search engine coverage and the deep Web compilation sites noted above (3 through Table 5).

The technique is illustrated in the diagram below:

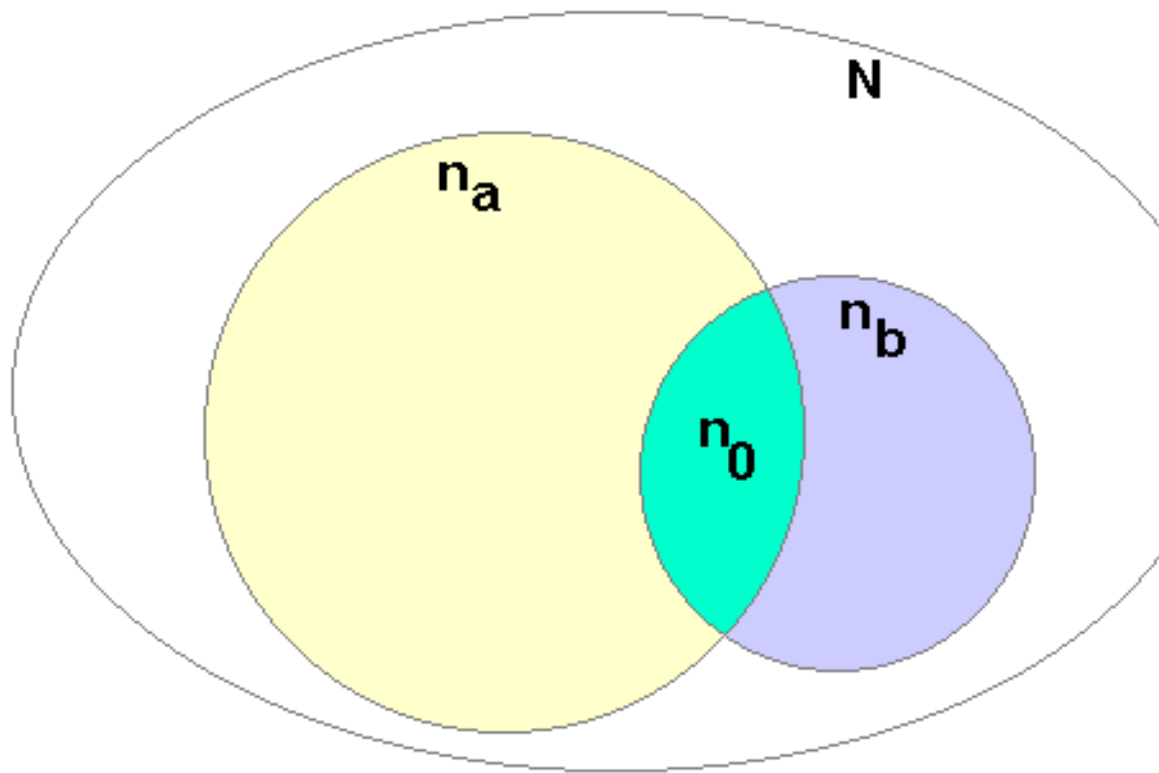


Figure 3. Schematic Representation of "Overlap" Analysis

Overlap analysis involves pairwise comparisons of the number of listings from individual sources, n_a and n_b , and the degree of shared listings or overlap, n_o , between them. If you know the number of listings for both n_a and n_b , the total size of the population, N , can be estimated. The fraction of the total population covered by n_a is n_o/n_b ; when applied to the total size of the population, the total population size can be derived by dividing this fraction into the total size. This process is repeated pairwise estimates are repeated for all of the individual sources used in the analysis.

To illustrate this technique, assume, for example, we know our total population is 100. If two sources, A and B, each contain 50 items, we could predict on average that 25 of those items would be shared by the two sources and 25 items would not be listed by either. According to the overlap analysis, this can be represented as: $100 = 50 / (25/50)$

There are two keys to overlap analysis. First, it is important to have a relatively accurate estimate of the total listing size for at least one of the two sources in the pairwise comparison. Second, the listings should be obtained randomly and independently from one another.

This second premise is in fact violated for our deep Web source analysis. Compilers and search engines are purposeful in collecting their listings, so their sampling is directed. And, for search engines, more searchable databases are more frequently linked to because of their information value.

their relative prevalence within the engine listings. [5f] Thus, the overlap analysis reported a *bound* on the size of the deep Web since both of these factors will tend to increase the overlap, n_o , reported between the pairwise sources.

Deep Web Size Analysis

In order to analyze the total size of the deep Web, we need an average site size in terms of storage to use as a multiplier applied to the entire population estimate. Results are shown in Figure 5.

As discussed for the large site analysis, obtaining this information is not straightforward and requires considerable time evaluating each site. To keep estimation time manageable, we chose a confidence interval at the 95% confidence level, requiring a total of 100 random sites to be characterized. [25a]

We randomized our listing of 17,000 search site candidates. We then proceeded to inspect sites until 100 sites were fully characterized. We followed a less-intensive process to determine the total analysis for determining total record or document count for the site.

Exactly 700 sites were inspected in their randomized order to obtain the 100 fully characterized sites. All sites inspected received characterization as to site type and coverage; this information was used in other parts of the analysis.

"The invisible portion of the Web will continue to grow exponentially before the tools to uncover the hidden Web are ready for general use"

The 100 sites that could have their total record count determined were then sampled for average document size (HTML-included basis). Random queries were made to the searchable database with results reported as HTML pages. A minimum of ten of these were generated, saved to disk, and averaged to determine the mean site page size. For sites such as bibliographic databases, multiple records were reported on a single HTML page. In these instances, the total number of results pages were generated, saved to disk, and averaged based on the total number of records reported on the pages.

Content Coverage and Type Analysis

Content coverage was analyzed across all 17,000 search sites in the qualified deep Web (shown in Table 6); the type of deep Web site was determined from the 700 hand-characterized sites (results shown in Figure 6).

Broad content coverage for the entire pool was determined by issuing queries for top-level domains against the entire pool. Because of topic overlaps, total occurrences exceeded the number of sites in the pool; this total was used to adjust all categories back to a 100% basis.

Hand characterization by search-database type resulted in assigning each site to one of twelve categories that captured the diversity of database types. These twelve categories are:

1. Topic Databases — subject-specific aggregations of information, such as SEC filings, medical databases, patent records, etc.
2. Internal site — searchable databases for the internal pages of large sites that were created, such as the knowledge base on the Microsoft site.
3. Publications — searchable databases for current and archived articles.
4. Shopping/Auction.
5. Classifieds.
6. Portals — broader sites that included more than one of these other categories of databases.
7. Library — searchable internal holdings, mostly for university libraries.
8. Yellow and White Pages — people and business finders.
9. Calculators — while not strictly databases, many do include an internal data table for calculating results. Mortgage calculators, dictionary look-ups, and translator services for languages are examples.
10. Jobs — job and resume postings.
11. Message or Chat .
12. General Search — searchable databases most often relevant to Internet search information.

These 700 sites were also characterized as to whether they were public or subject to pay-fee access.

Site Pageviews and Link References

Netscape's "What's Related" browser option, a service from Alexa, provides site popularity and link reference counts for a given URL. [26a] About 71% of deep Web sites have surface Web counterparts that follow a universal power function (a logarithmic growth rate or logarithmic distribution) all the way down to the deep Web. [27] The "What's Related" service shows external link counts to the given URL.

A random sampling for each of 100 deep and surface Web sites for which complete reports could be obtained were used for the comparisons.

Growth Analysis

The best method for measuring growth is with time-series analysis. However, since deep Web is so new, a different gauge was necessary.

Whois [28] searches associated with domain-registration services [25b] return records for the domain owner, as well as the date the domain was first obtained (and other information). Using a sample of 100 deep Web sites [26b] and another sample of 100 surface Web sites [29], domain names to a Whois search and retrieved the date the site was first established. The results were then combined and plotted for the deep vs. surface Web samples.

Quality Analysis

Quality comparisons between the deep and surface Web content were based on five specific queries issued via the BrightPlanet technology to three search engines (AltaVista, Northern Light) [30] and three deep sites specific to that topic and included in the 6000 site list configured for our technology. The five subject areas were agriculture, medicine, finance, science, and law.

The queries were specifically designed to limit total results returned from any of the sources to a maximum of 200 to ensure complete retrieval from each source. [31] The specific technology configuration settings are documented in the endnotes. [32]

The "quality" determination was based on an average of our technology's VSM and r

computational linguistic scoring methods. [33] [34] The "quality" threshold was set at empirically determined as roughly accurate from millions of previous scores of surface documents.

Deep Web vs. surface Web scores were obtained by using the BrightPlanet technology source option and then counting total documents and documents above the quality score.

Results and Discussion

This study is the first known quantification and characterization of the deep Web. Very little is written or known of the deep Web. Estimates of size and importance have been anecdotal and certainly underestimate scale. For example, Intelliseek's "invisible Web" says that, "Today, estimates today, the valuable content housed within these databases and searchable engines is much bigger than the 800 million plus pages of the Visible Web." They also estimate total size at about 50,000 or so. [35]

Ken Wiseman, who has written one of the most accessible discussions about the deep Web, says that it might be about equal in size to the known Web. He also goes on to say, "I can't say the invisible portion of the Web will continue to grow exponentially before the tools to access the hidden Web are ready for general use." [36] A mid-1999 survey by About.com's WebCrawler concluded the size of the deep Web was "big and getting bigger." [37] A paper at a recent meeting suggested that only "a relatively small fraction of the Web is accessible through search engines." [38]

The deep Web is about 500 times larger than the surface Web, with, on average, about 10 times higher quality based on our document scoring methods on a per-document basis. Overall, total deep Web quality exceeds that of the surface Web by thousands of times. Total number of Web sites likely exceeds 200,000 today and is growing rapidly. [39] Content on the deep Web has meaning and importance for every information seeker and market. More than 95% of the information is publicly available without restriction. The deep Web also appears to be a growing information component of the Web.

General Deep Web Characteristics

Deep Web content has some significant differences from surface Web content. Deep Web content is

(13.7 KB mean size; 19.7 KB median size) are on average 27% smaller than surface Web sites. Though individual deep Web sites have tremendous diversity in their number of records (from tens or hundreds to hundreds of millions (a mean of 5.43 million records per site but some with only 4,950 records), these sites are on average much, much larger than surface sites. This paper will serve to amplify these findings.

The mean deep Web site has a Web-expressed (HTML-included basis) database size of 169 KB). Actual record counts and size estimates can be derived from one-in-several

On average, deep Web sites receive about half again as much monthly traffic as surface sites (125,000 pageviews per month vs. 85,000). The median deep Web site receives somewhat more than the traffic of a random surface Web site (843,000 monthly pageviews vs. 365,000). Deep Web sites on average are more highly linked to than surface sites by nearly a factor of two (6,200 links), though the median deep Web site is less so (66 vs. 83 links). This suggests that deep Web sites are highly popular, but that the typical deep Web site is not well known to the public.

One of the more counter-intuitive results is that 97.4% of deep Web sites are public; 1.6% are restricted; a further 1.6% are mixed (limited results publicly available with greater access via subscription and/or paid fees); only 1.1% of results are totally subscription or fee limited. This is counter intuitive because of the visible prominence of subscriber-limited sites such as Lexis, Wall Street Journal Interactive, etc. (We got the document counts from the search engines and from other published sources.)

However, once the broader pool of deep Web sites is looked at beyond the large, visible ones, public availability dominates.

60 Deep Sites Already Exceed the Surface Web Times

Table 2 indicates that the sixty known, largest deep Web sites contain data of about 40 times the size of the known surface Web (HTML-included basis) or roughly forty times the size of the known surface Web. This is a broad array of domains from science to law to images and commerce. We estimate that the total number of records or documents within this group to be about eighty-five billion.

Roughly two-thirds of these sites are public ones, representing about 90% of the cc within this group of sixty. The absolutely massive size of the largest sites shown also universal power function distribution of sites within the deep Web, not dissimilar to [40] or surface Web sites. [41] One implication of this type of distribution is that there size boundary to which sites may grow.

Table 2. Sixty Largest Deep Web Sites

Name	Type	URL
National Climatic Data Center (NOAA)	Public	http://www.ncdc.noaa.gov/ol/satellite/satelliteres
NASA EOSDIS	Public	http://harp.gsfc.nasa.gov/~imswww/pub/imswelco
National Oceanographic (combined with Geophysical) Data Center (NOAA)	Public/Fee	http://www.nodc.noaa.gov/ , http://www.ngdc.noaa
Alexa	Public (partial)	http://www.alexa.com/
Right-to-Know Network (RTK Net)	Public	http://www.rtk.net/
MP3.com	Public	http://www.mp3.com/
Terraserver	Public/Fee	http://terraserver.microsoft.com/
HEASARC (High Energy Astrophysics Science Archive Research Center)	Public	http://heasarc.gsfc.nasa.gov/W3Browse/
US PTO - Trademarks + Patents	Public	http://www.uspto.gov/tmdb/ , http://www.uspto.gov

Informedia (Carnegie Mellon Univ.)	Public (not yet)	http://www.informedia.cs.cmu.edu/
Alexandria Digital Library	Public	http://www.alexandria.ucsb.edu/adl.html
JSTOR Project	Limited	http://www.jstor.org/
10K Search Wizard	Public	http://www.tenkwizard.com/
UC Berkeley Digital Library Project	Public	http://elib.cs.berkeley.edu/
SEC Edgar	Public	http://www.sec.gov/edgarhp.htm
US Census	Public	http://factfinder.census.gov
NCI CancerNet Database	Public	http://cancernet.nci.nih.gov/
Amazon.com	Public	http://www.amazon.com/
IBM Patent Center	Public/Private	http://www.patents.ibm.com/boolquery
NASA Image Exchange	Public	http://nix.nasa.gov/
InfoUSA.com	Public/Private	http://www.abii.com/
Betterwhois (many similar)	Public	http://betterwhois.com/
GPO Access	Public	http://www.access.gpo.gov/
Adobe PDF Search	Public	http://searchpdf.adobe.com/
Internet Auction List	Public	http://www.internetauctionlist.com/search_produ
Commerce, Inc.	Public	http://search.commerceinc.com/
Library of Congress Online Catalog	Public	http://catalog.loc.gov/

Sunsite Europe	Public	http://src.doc.ic.ac.uk/
Uncover Periodical DB	Public/Fee	http://uncweb.carl.org/
Astronomer's Bazaar	Public	http://cdsweb.u-strasbg.fr/Cats.html
eBay.com	Public	http://www.ebay.com/
REALTOR.com Real Estate Search	Public	http://www.realtor.com/
Federal Express	Public (if shipper)	http://www.fedex.com/
Integrum	Public/Private	http://www.integrumworld.com/eng_test/index.h
NIH PubMed	Public	http://www.ncbi.nlm.nih.gov/PubMed/
Visual Woman (NIH)	Public	http://www.nlm.nih.gov/research/visible/visible_1
AutoTrader.com	Public	http://www.autoconnect.com/index.jttml/?LNx=M1DJAROSTEXT
UPS	Public (if shipper)	http://www.ups.com/
NIH GenBank	Public	http://www.ncbi.nlm.nih.gov/Genbank/index.html
AustLi (Australasian Legal Information Institute)	Public	http://www.austlii.edu.au/austlii/
Digital Library Program (UVa)	Public	http://www.lva.lib.va.us/
Subtotal Public and Mixed Sources		
DBT Online	Fee	http://www.dbtonline.com/
Lexis-Nexis	Fee	http://www.lexis-nexis.com/lnc/
Dialog	Fee	http://www.dialog.com/

Genealogy - ancestry.com	Fee	http://www.ancestry.com/
ProQuest Direct (incl. Digital Vault)	Fee	http://www.umi.com
Dun & Bradstreet	Fee	http://www.dnb.com
Westlaw	Fee	http://www.westlaw.com/
Dow Jones News Retrieval	Fee	http://dowjones.wsj.com/p/main.html
infoUSA	Fee/Public	http://www.infousa.com/
Elsevier Press	Fee	http://www.elsevier.com
EBSCO	Fee	http://www.ebsco.com
Springer-Verlag	Fee	http://link.springer.de/
OVID Technologies	Fee	http://www.ovid.com
Investext	Fee	http://www.investext.com/
Blackwell Science	Fee	http://www.blackwell-science.com
GenServ	Fee	http://gs01.genserv.com/gs/bcc.htm
Academic Press IDEAL	Fee	http://www.idealibrary.com
Tradecompass	Fee	http://www.tradecompass.com/
INSPEC	Fee	http://www.iee.org.uk/publish/inspec/online/onlinr
Subtotal Fee- Based Sources		
TOTAL		

This listing is preliminary and likely incomplete since we lack a complete census of

Our inspection of the 700 random-sample deep Web sites identified another three to the initially identified pool of 100 potentially large sites. If that ratio were to hold across the estimated 200,000 deep Web sites (see next table), perhaps only a very small percentage of the sites shown in this table would prove to be the largest. However, since many large sites are already known, we believe our listing, while highly inaccurate, may represent 10% to 20% of the total deep Web sites in existence.

This inability to identify all of the largest deep Web sites today should not be surprising. The discovery of the deep Web is a new phenomenon and has received little attention. We solicit reports of additional large sites on our comprehensive CompletePlanet site and will document them as they arise.

Deep Web is 500 Times Larger than the Surface Web

We employed three types of overlap analysis to estimate the total numbers of deep Web sites. In our first approach, shown in Table 3, we issued 100 random deep Web URLs from our preferred search engines that support URL search. These results, with the accompanying overlap analysis, are shown in Table 4.

Table 3. Estimation of Deep Web Sites, Search Engine Overlap Analysis

					Search Engine A		
Search Engine A	A no dupes	Search Engine B	B no dupes	A plus B	Unique	Database Fraction	Database Size
AltaVista	9	Northern Light	60	8	1	0.133	20,635
AltaVista	9	Fast	57	8	1	0.140	20,635
Fast	57	AltaVista	9	8	49	0.889	27,940
Northern Light	60	AltaVista	9	8	52	0.889	27,195
Northern Light	60	Fast	57	44	16	0.772	27,195
Fast	57	Northern Light	60	44	13	0.733	27,940

This table shows greater diversity in deep Web site estimates as compared to normal overlap analysis. We believe the reasons for this variability are: 1) the relatively small database sizes from Northern Light [42]; and 3) the indiscriminate scaling of Fast and site coverage based on the surface ratios of these engines to Northern Light. As a result, we have low confidence in these results.

An alternate method is to compare NEC reported values [58] for surface Web coverage to deep Web sites from the Northern Light engine. These numbers were further adjusted by a qualification fraction obtained from our hand scoring of 700 random deep Web sites, as shown below:

Table 4. Estimation of Deep Web Sites, Search Engine Market Share Basis

Search Engine	Reported Deep Web Sites	Surface Web Coverage %	Qualification Fraction
Northern Light	27,195	16.0%	86.4%
AltaVista	20,635	15.5%	86.4%

This approach, too, suffers from the limitations of using the Northern Light deep Web coverage. It is also unclear, though likely, that deep Web search coverage is more highly represented by the engines' listing as discussed above.

Our third approach is more relevant and is shown in Table 5.

Under this approach, we use overlap analysis for the three largest compilation sites used to build our original 17,000 qualified candidate pool. To our knowledge, these are the largest listings extant, excepting our own CompletePlanet site.

This approach has the advantages of:

- providing an absolute count of sites
- ensuring final qualification as to whether the sites are actually deep Web search results

- relatively large sample sizes.

Because each of the three compilation sources has a known population, the table shows pairwise comparisons (e.g., there is no uncertainty in the ultimate A or B population

Table 5. Estimation of Deep Web Sites, Searchable Database Compilation (C

DB A	A no dups	DB B	B no dups	A + B	Unique	DB Fract.	DB Size	Total E Web S
Lycos	5,081	Internets	3,449	256	4,825	0.074	5,081	68,455
Lycos	5,081	Infomine	2,969	156	4,925	0.053	5,081	96,702
Internets	3,449	Infomine	2,969	234	3,215	0.079	3,449	43,761

As discussed above, there is certainly sampling bias in these compilations since they are not randomly obtained. Despite this, there is a surprising amount of uniqueness in these compilations.

The Lycos and Internets listings are more similar in focus in that they are commercial. The Infomine site was developed from an academic perspective. For this reason, we adjust the Infomine pairwise comparison to be most appropriate. Though sampling was directed, the intended coverage and perspective is different.

There is obviously much uncertainty in these various tables. Because of lack of random sampling, estimates are likely at the lower bounds for the number of deep Web sites. Across all methods the mean estimate for number of deep Web sites is about 76,000, with a median of 56,000. For the searchable database compilation only, the average is about 70,000.

The under count due to lack of randomness and what we believe to be the best estimate from the Lycos-Infomine pair, indicate to us that the ultimate number of deep Web sites is likely to be of 200,000.

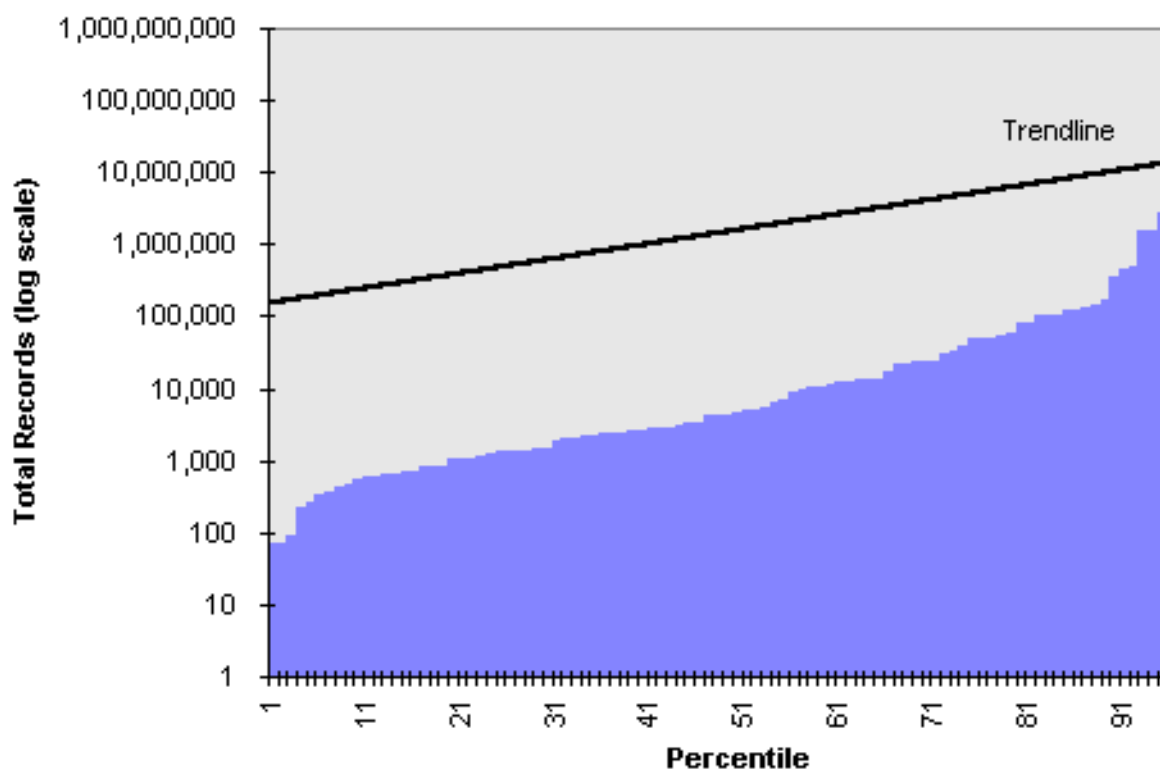


Figure 4. Inferred Distribution of Deep Web Sites, Total Record Size

Plotting the fully characterized random 100 deep Web sites against total record count (see Figure 4). Plotting these same sites against database size (HTML-included basis) produced the following results:

Multiplying the mean size of 74.4 MB per deep Web site times a total of 200,000 deep Web sites results in a total deep Web size projection of 7.44 petabytes, or 7,440 terabytes. [43] Compared to the current surface Web content estimate of 18.7 TB (see Table 1), this suggests a deep Web size estimate 400 times larger than the surface Web. Even at the lowest end of the deep Web size estimates, the deep Web is about 620 times the size of the surface Web.

Alternately, multiplying the mean document/record count per deep Web site of 5.43 million times 200,000 total deep Web sites results in a total record count across the deep Web of 1.086 billion documents. [44b] Compared to the Table 1 estimate of one billion documents, this implies a deep Web size estimate 100 times larger than the surface Web. At the low end of the deep Web size estimate this is 100 times larger, and at the high end, 840 times.

Clearly, the scale of the deep Web is massive, though uncertain. Since 60 deep Web sites are nearly 40 times the size of the entire surface Web, we believe that the 200,000 deep Web sites is the most reasonable one. Thus, across database and record sizes, we estimate the deep Web is

about 500 times the size of the surface Web.

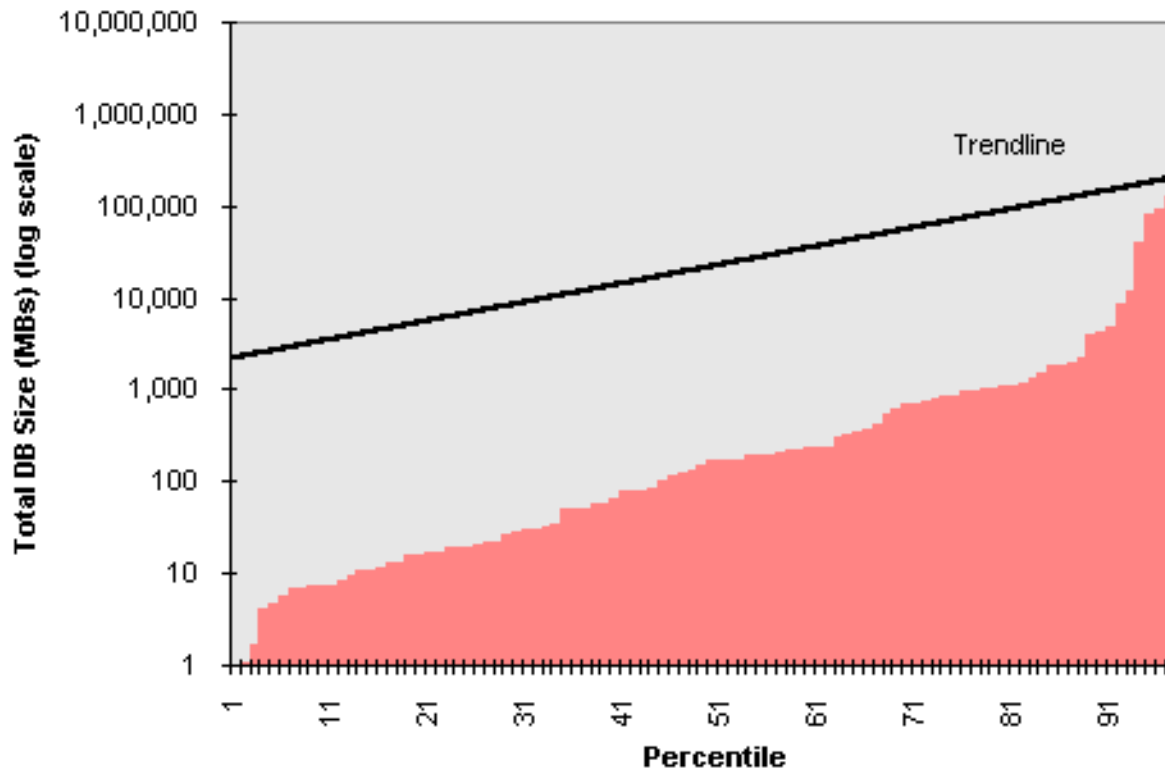


Figure 5. Inferred Distribution of Deep Web Sites, Total Database Size (MBs)

Deep Web Coverage is Broad, Relevant

Table 6 represents the subject coverage across all 17,000 deep Web sites used in the study. The subject areas correspond to the top-level subject structure of the CompletePlanet site. The data shows a surprisingly uniform distribution of content across all areas, with no category lacking representation of content. Actual inspection of the CompletePlanet site by node shows that deep Web content is deeper and broader than others. However, it is clear that deep Web content also covers every information need and market.

Table 6. Distribution of Deep Sites by Subject Area

Deep Web Coverage	
Agriculture	2.7%
Arts	6.6%
Business	5.9%

Computing/Web	6.9%
Education	4.3%
Employment	4.1%
Engineering	3.1%
Government	3.9%
Health	5.5%
Humanities	13.5%
Law/Politics	3.9%
Lifestyles	4.0%
News, Media	12.2%
People, Companies	4.9%
Recreation, Sports	3.5%
References	4.5%
Science, Math	4.0%
Travel	3.4%
Shopping	3.2%

Figure 6 displays the distribution of deep Web sites by type of content.

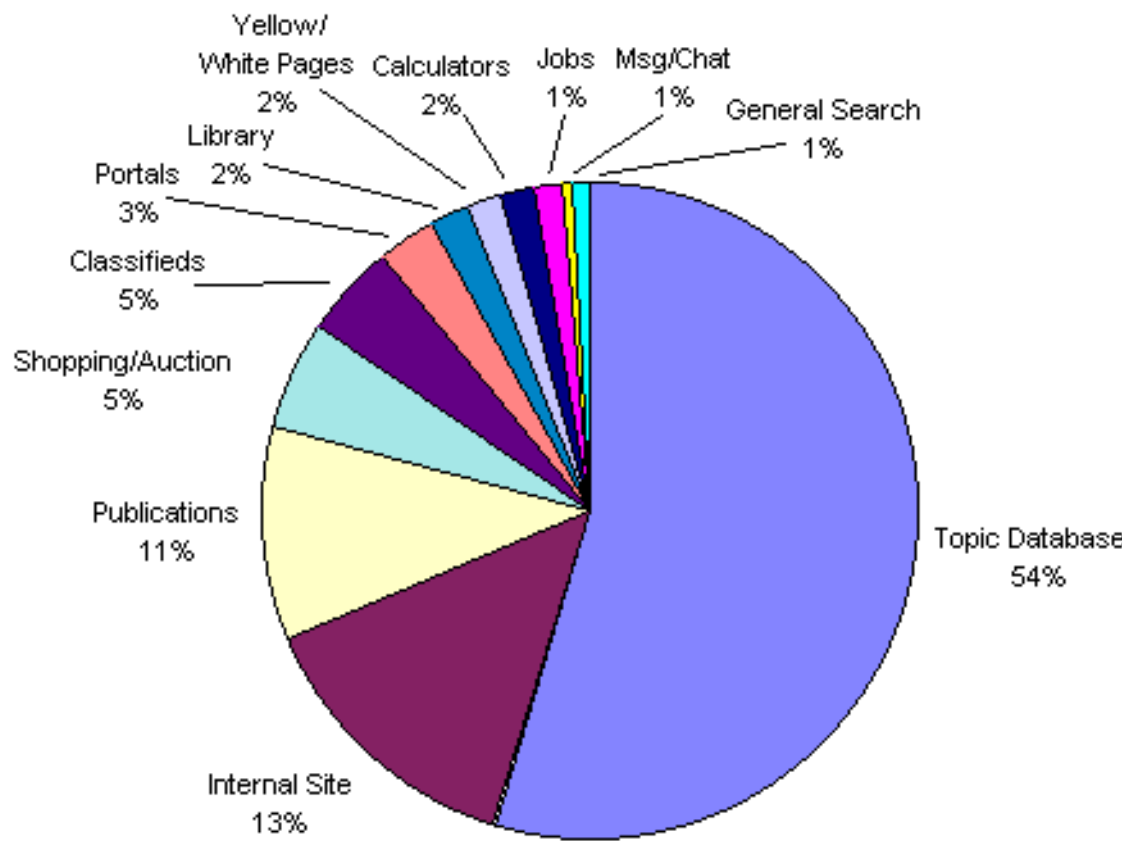


Figure 6. Distribution of Deep Web Sites by Content Type

More than half of all deep Web sites feature topical databases. Topical databases plus documents and archived publications make up nearly 80% of all deep Web sites. Publications — including true shopping sites with auctions and classifieds — account for another 16% of all sites. The other eight categories collectively account for the remaining 10% or so of all sites.

Deep Web is Higher Quality

“Quality” is subjective: If you get the results you desire, that is high quality; if you do not, that is low quality at all.

When BrightPlanet assembles quality results for its Web-site clients, it applies additional tests to computational linguistic scoring. For example, university course listings of the query terms that can produce high linguistic scores, but they have little intrinsic value unless you are a student looking for a particular course. Various classes of these positives exist and can be discovered and eliminated through learned business rules.

Our measurement of deep vs. surface Web quality did not apply these more sophisticated tests. We also posed five queries across

domains. Using only computational linguistic scoring does not introduce systematic deep and surface Web results because the same criteria are used in both. The relationship between surface and deep Web should maintain, even though the absolute values are will overestimate "quality." The results of these limited tests are shown in Table 7.

Table 7. "Quality" Document Retrieval, Deep vs. Surface Web

Query	Surface Web			Deep Web		
	Total	"Quality"	Yield	Total	"Quality"	Yield
Agriculture	400	20	5.0%	300	42	14.0%
Medicine	500	23	4.6%	400	50	12.5%
Finance	350	18	5.1%	600	75	12.5%
Science	700	30	4.3%	700	80	11.4%
Law	260	12	4.6%	320	38	11.9%
TOTAL	2,210	103	4.7%	2,320	285	12.3%

This table shows that there is about a three-fold improved likelihood for obtaining quality results from the deep Web as from the surface Web on average for the limited sample set. Also, the results show that deep Web sites tend to return 10% more documents than surface Web sites and nearly triple the number of quality documents. While each query used three of the best known surface Web search engines and three of the best known deep Web sites, these results are somewhat likely underestimate the "quality" difference between the surface and deep Web. First, there are literally hundreds of applicable deep Web sites for each query subject area. Some of these sites would likely not return as high an overall quality yield, but would add to the total number of quality results returned. Second, even with increased numbers of surface search engines, the overall quality coverage would not go up significantly and yields would decline, especially if duplicate search engines were removed (as they should be). And, third, we believe the degree of quality difference between deep Web sites to be much less than for surface Web sites.⁽⁴⁵⁾ Though the results applied in this study are not definitive, we believe they point to a defensible conclusion that the number of quality results is many times greater for the deep Web than for the surface Web. Moreover, the deep Web offers the prospect of yielding quality results that cannot be obtained by any other means, with the number of quality results increasing as a function of the number of deep Web sites simultan-

The deep Web thus appears to be a critical source when it is imperative to find a “ne

Deep Web Growing Faster than Surface Web

Lacking time-series analysis, we used the proxy of domain registration date to measure growth rates for each of 100 randomly chosen deep and surface Web sites. These results are shown in a scattergram with superimposed growth trend lines in Figure 7.

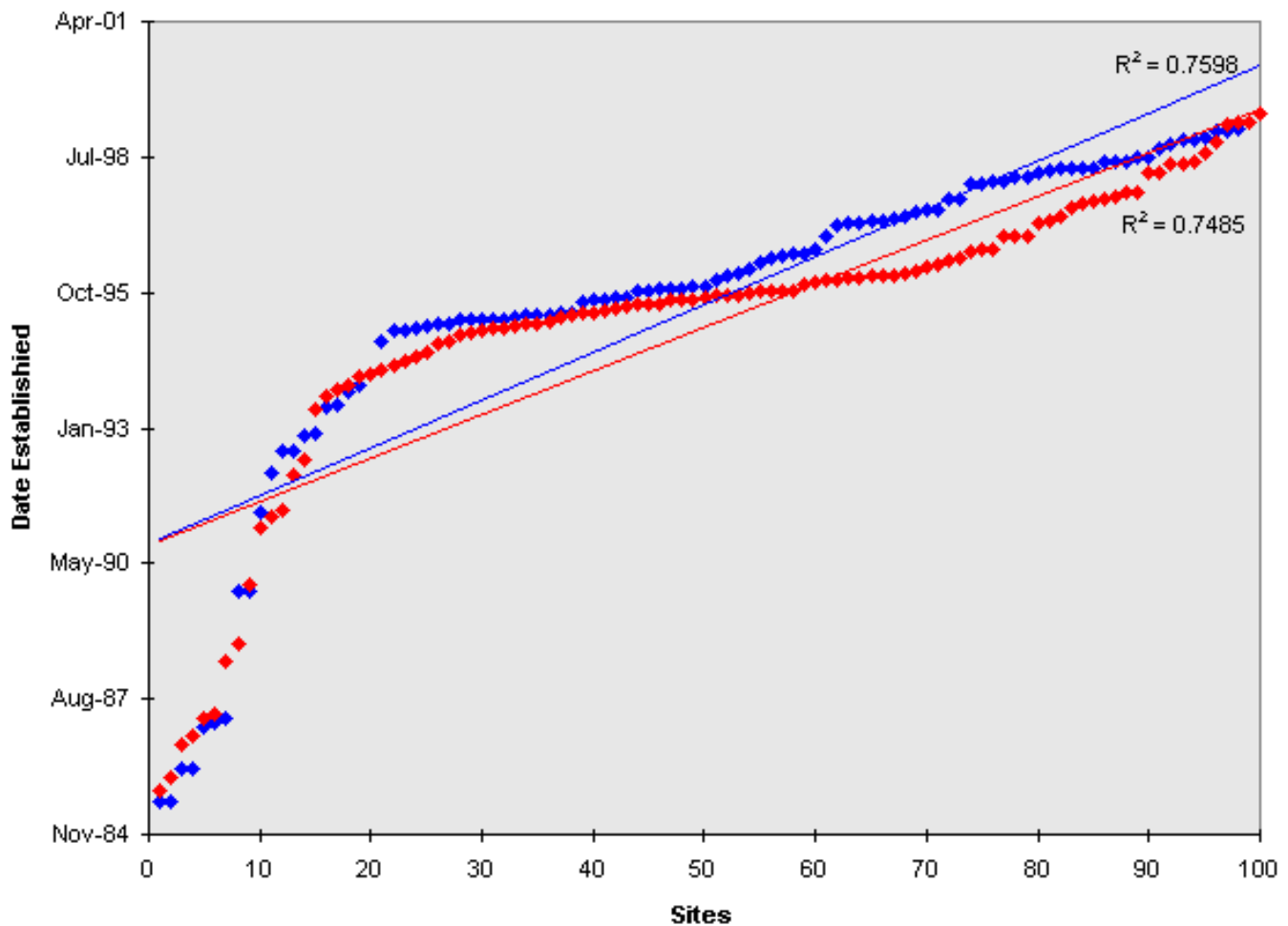


Figure 7. Comparative Deep and Surface Web Site Growth Rates

Use of site domain registration as a proxy for growth has a number of limitations. First, domains are frequently registered well in advance of going “live.” Second, the domain registration is often at the domain level (e.g., www.mainsite.com). The search function and page — whether for the site — often is introduced after the site is initially unveiled and may itself reside on a domain not discoverable by the *whois* analysis.

The best way to test for actual growth is a time series analysis. BrightPlanet plans to

tracking mechanisms to obtain better growth estimates in the future.

However, this limited test does suggest faster growth for the deep Web. Both media: Web sites are four or five months “younger” than surface Web sites (Mar. 95 v. Aug surprising. The Internet has become the preferred medium for public dissemination information, and more and more information disseminators (such as government ag research projects) that have enough content to qualify as deep Web are moving their online. Moreover, the technology for delivering deep Web sites has been around for of time.

Thousands of Conventional Search Engines Remain Undiscovered

While we have specifically defined the deep Web to exclude search engines (see new specialized search engines, such as those shown in Table 8 below or @griculture.com/joefarmer [formerly <http://www.joefarmer.com/>] in the agriculture domain, provide readily indexed by major engines such as AltaVista, Fast or Northern Light. The key specialty search engines may contain information not on the major ones are indexing limitations the major search engines may impose on documents indexed per site. [11]

To find out whether the specialty search engines really do offer unique information retrieval and qualification methods on them — pairwise overlap analysis — in a new results of this analysis are shown in the table below.

Table 8. Estimated Number of Surface Site Search Engines

					Search Engine A	
Search Engine A	A no dupes	Search Engine B	B no dupes	A plus B	Unique	Search Engine Fraction
FinderSeeker	2,012	SEG	1,268	233	1,779	0.184
FinderSeeker	2,012	Netherlands	1,170	167	1,845	0.143
FinderSeeker	2,012	LincOne	783	129	1,883	0.165

SearchEngineGuide	1,268	FinderSeeker	2,012	233	1,035	0.116
SearchEngineGuide	1,268	Netherlands	1,170	160	1,108	0.137
SearchEngineGuide	1,268	LincOne	783	28	1,240	0.036
Netherlands	1,170	FinderSeeker	2,012	167	1,003	0.083
Netherlands	1,170	SEG	1,268	160	1,010	0.126
Netherlands	1,170	LincOne	783	44	1,126	0.056
LincOne	783	FinderSeeker	2,012	129	654	0.064
LincOne	783	SEG	1,268	28	755	0.022
LincOne	783	Netherlands	1,170	44	739	0.038

These results suggest there may be on the order of 20,000 to 25,000 total search engines on the Web. (Recall that all of our deep Web analysis *excludes* these additional search engines.)

M. Hofstede, of the Leiden University Library in the Netherlands, reports that one of his databases contains nearly 45,000 search site listings. [46] Thus, our best current estimate is that the searchable databases and search engines have a combined total of 250,000 sites. This number proves to be, comprehensive Web search strategies should include the specialty search engines as well as deep Web sites. Thus, BrightPlanet's CompletePlanet Web site also lists specialty search engines in its listings.

Commentary

The most important findings from our analysis of the deep Web are that there is massive meaningful content not discoverable with conventional search technology and that there is a uniform lack of awareness that this critical content even exists.

Original Deep Content Now Exceeds All Printed Content

International Data Corporation predicts that the number of surface Web documents will increase from current two billion or so to 13 billion within three years, a factor increase of 6.5 times. If this growth should exceed this rate, perhaps increasing about nine-fold over the same period,

compares this growth with trends in the cumulative global content of print information in a recent UC Berkeley study. [48a]

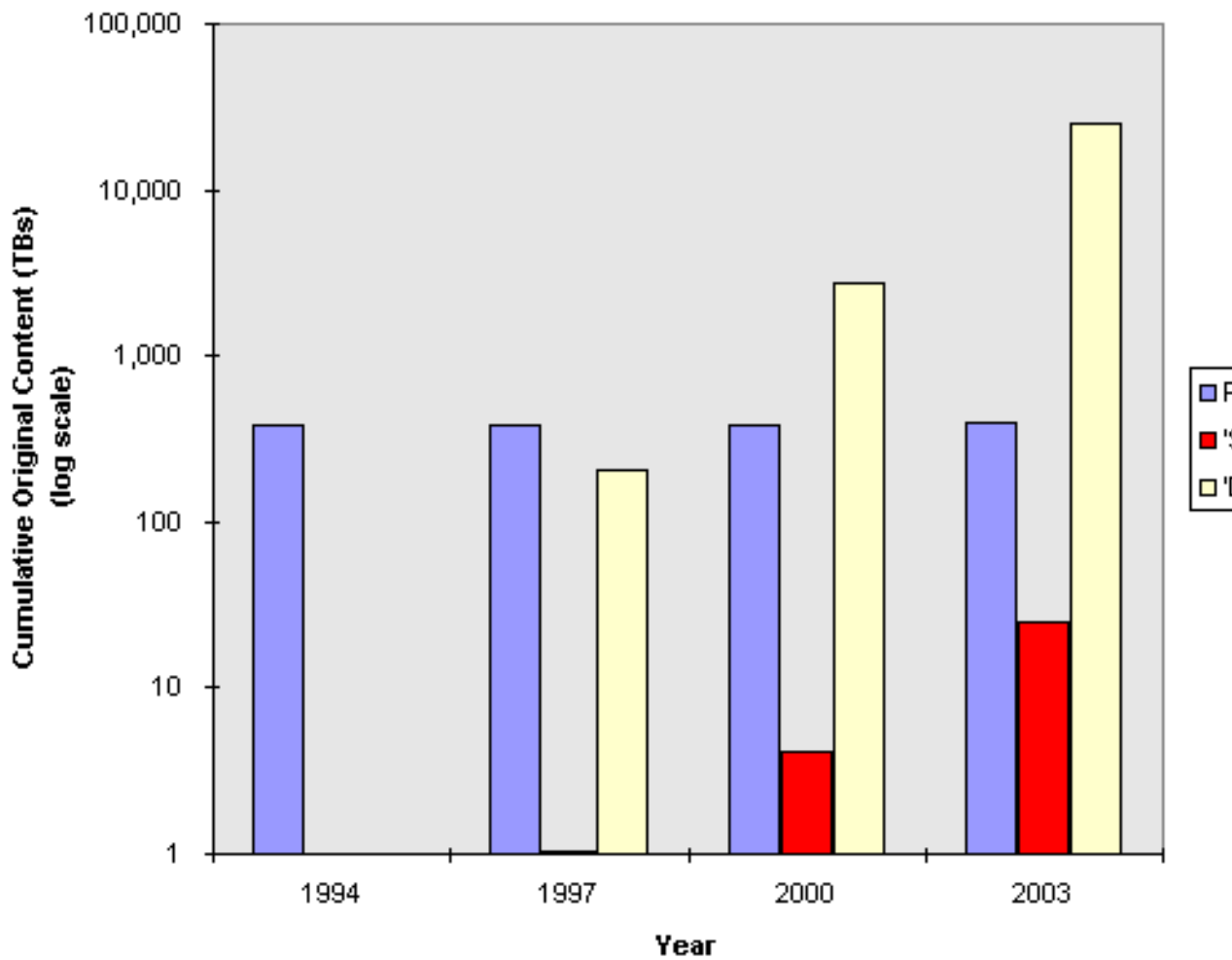


Figure 8. 10-yr Growth Trends in Cumulative Original Information Content (log scale)

The total volume of printed works (books, journals, newspapers, newsletters, office documents, etc.) held steady at about 390 terabytes (TBs). [48b] By about 1998, deep Web original information content equaled all print content produced through history up until that time. By 2000, original information content is estimated to have exceeded print by a factor of seven and is projected to exceed print content by a factor of sixty three by 2003.

Other indicators point to the deep Web as the fastest growing component of the Web. It is expected to dominate it. [49] Even today, at least 240 major libraries have their catalogs online. Amazon.com, a subsidiary of Bell & Howell, has plans to put more than 5.5 billion document images online. Many major astronomy data initiatives are moving toward putting petabytes of data online.

These trends are being fueled by the phenomenal growth and cost reductions in digital storage. [48c] [53] International Data Corporation estimates that the amount of disk storage annually grew from 10,000 terabytes in 1994 to 116,000 terabytes in 1998, and it is projected to increase to 1,400,000 terabytes in 2002. [54] Deep Web content accounted for about 1/200th of the storage devoted to original content in 2000; it is projected to increase to 1/200th by 2002. Internet is expected to continue as the universal medium for publishing and disseminating information; these trends are sure to continue.

The Gray Zone

There is no bright line that separates content sources on the Web. There are circumstances where “deep” content can appear on the surface, and, especially with specialty search engines, surface content can appear to be deep.

Surface Web content is persistent on static pages discoverable by search engines through direct requests, while deep Web content is only presented dynamically in response to a direct request. When directly requested, deep Web content comes associated with a URL, most often containing a unique record number, that can be re-used later to obtain the same document.

We can illustrate this point using one of the best searchable databases on the Web, 10Kwizard provides full-text searching of SEC corporate filings [55]. We issued a query for “basketball” with a restriction to review only annual filings filed between March 1999 and March 2000. One result was produced for Sportsline USA, Inc. Clicking on that listing produces a link to the query string in that annual filing. With another click, the full filing text can be accessed. The URL resulting from this direct request is:

<http://www.10kwizard.com/blurbs.php?repo=tenk & ipage=1067295 & exp=%22ncaa+>

Note two things about this URL. First, our query terms appear in it. Second, the “ipage” is a unique record number, in this case 1067295. It is via this record number that the results are retrieved from the 10KWizard database.

Now, if we were doing comprehensive research on this company and posting these findings on a Web page, other users could click on this URL and get the same information. Importantly, if we posted this URL on a static Web page, search engine crawlers could also discover it.

as shown above, and then index the contents.

It is by doing searches and making the resulting URLs available that deep content can surface. Any deep content listed on a static Web page is discoverable by crawlers and indexable by search engines. As the next section describes, it is impossible to compare deep Web sites for all content in this manner. But it does show why some deep Web content occasionally appears on surface Web search engines.

This gray zone also encompasses surface Web sites that are available through deep Web. For instance, the [Open Directory Project](#), is an effort to organize the best of surface Web sites by voluntary editors or "guides." [56] The Open Directory looks something like Yahoo!'s directory structure with directory URL results at each branch. The results pages are static, like other directories, and are therefore easily indexable by the major search engines.

The Open Directory claims a subject structure of 248,000 categories, [57] each of which has a page. [58] The key point is that every one of these 248,000 pages is indexable by major search engines.

Four major search engines with broad surface coverage allow searches to be specific. The query "URL:dmoz.org" (the address for the Open Directory site) was posed to each of these engines. These are the results:

Table 9. Incomplete Indexing of Surface Web Sites

Engine	OPD Pages	Yield
Open Directory (OPD)	248,706	—
AltaVista	17,833	7.2%
Fast	12,199	4.9%
Northern Light	11,120	4.5%
Go (Infoseek)	1,970	0.8%

Although there are almost 250,000 subject pages at the Open Directory site, only a small fraction is recognized by the major search engines. Clearly the engines' search algorithms have limited depth or breadth of surface pages indexed for a given site. We also found a broad variety of deep Web content that is not indexed by the major search engines.

timeliness of results from these engines. Specialized surface sources or engines should be considered when truly deep searching is desired. That bright line between deep and surface is really shades of gray.

The Impossibility of Complete Indexing of Deep Content

Consider how a directed query works: specific requests need to be posed against the database by stringing together individual query terms (and perhaps other filters such as restrictions). If you do not ask the database specifically what you want, you will not

Let us take, for example, our own listing of 38,000 deep Web sites. Within this collection are some 430,000 unique terms and a total of 21,000,000 terms. If these numbers represent the contents of a searchable database, then we would have to issue 430,000 individual queries if we had comprehensively "scrubbed" or obtained all records within the source database. This is small compared to some large deep Web databases. For example, one of the largest is the British National Corpus containing more than 100 million unique terms

It is infeasible to issue many hundreds of thousands or millions of direct queries to search databases. It is implausible to repeat this process across tens to hundreds of thousands of Web sites. And, of course, because content changes and is dynamic, it is impossible to maintain a reasonable update schedule. For these reasons, the predominant share of the deep Web will remain below the surface and can only be discovered within the context of a specific request.

Possible Double Counting

How much Web content is distributed and, once posted, "public" to any source that chooses to index it? How much of deep Web content is unique, and how much is duplicated? And, are there any duplicated content between the deep and surface Web?

**"Surface Web sites are
fraught with quality**

This study was not able to resolve these questions. It is not known today how much duplication occurs on the deep Web.

problems"

Observations from working with the deep Web suggest there are important information categories where duplication does exist. Prominent among these are pages, genealogical records, and public records with commercial potential such as patents. There are, for example, numerous sites devoted to company financials.

On the other hand, there are entire categories of deep Web sites whose content appears to be highly valuable. These mostly fall within the categories of topical databases, publications, and indices — accounting in total for about 80% of deep Web sites — and include such things as scientific databases, library holdings, unique bibliographies such as PubMed, and unique data repositories such as satellite imaging data and the like.

But duplication is also rampant on the surface Web. Many sites are "mirrored." Popular content is frequently appropriated by others and posted on their own sites. Common information such as press releases and product listings, software, press releases, and so forth may turn up multiple times in search engine searches. And, of course, the search engines themselves duplicate much content.

Duplication potential thus seems to be a function of public availability, market importance, and ease of discovery. The deep Web is not as easily discovered, and while mostly public, not as easily copied as other surface Web sites. These factors suggest that duplication may be lower within the deep Web. For the present, this observation is conjecture.

Deep vs. Surface Web Quality

The issue of quality has been raised throughout this study. A quality search result is one that provides the right list. Searchers want answers. Providing those answers has always been a challenge on the surface Web, and without appropriate technology will be a problem for the deep Web.

Effective searches should both identify the relevant information desired and present it in a way that is easy to understand — quality. Sometimes what is most important is comprehensive information about everything referring to a commercial product, for instance. Other times the most important information is needed — the complete description of a chemical compound, as an example. The same search requirements can be used for both sets of requirements, but the answers will have to be different. Meeting these requirements is daunting, and knowing that the deep Web exists only complicates the task. However, the deep Web often contains useful information for either kind of search. If useful information is

excluded from a search, the requirements of either user cannot be met.

We have attempted to bring together some of the metrics included in this paper, [60] both actual quality of the search results and the ability to cover the subject.

Table 10. Total "Quality" Potential, Deep vs. Surface Web

Search Type	Total Docs (million)	Quality Docs (
<i>Surface Web</i>		
Single Site Search	160	7
Metasite Search	840	38
TOTAL SURFACE POSSIBLE	1,000	45
<i>Deep Web</i>		
Mega Deep Search	110,000	14,850
Single Site Search	688:1	2,063:1
Metasite Search	131:1	393:1
TOTAL POSSIBLE	655:1	2,094:1

These strict numerical ratios ignore that including deep Web sites may be the critical discovering the information desired. In terms of discovery, inclusion of deep Web sites discovery by 600 fold or more.

Surface Web sites are fraught with quality problems. For example, a study in 1999 in 1998 Web sites were no longer available in 1999 and that 45% of existing sites were meaningless, or trivial. [61] Lawrence and Giles' NEC studies suggest that individual coverage dropped from a maximum of 32% in 1998 to 16% in 1999. [7b]

Peer-reviewed journals and services such as Science Citation Index have evolved to authority necessary for users to judge the quality of information. The Internet lacks

An intriguing possibility with the deep Web is that individual sites can themselves establish authority. For example, an archived publication listing from a peer-reviewed journal

Science or user-accepted sources such as the *Wall Street Journal* or *The Economist* carry authority based on their editorial and content efforts. The owner of the site vets what is available. Professional content suppliers typically have the kinds of database-based services available in the deep Web; the static HTML pages that typically make up the surface Web are less likely to be from professional content suppliers.

By directing queries to deep Web sources, users can choose authoritative sites. Search engines, because of their indiscriminate harvesting, do not direct queries. By careful selection of sites, users can make their own determinations about quality, even though a solid metric for quality is difficult or impossible to assign universally.

Conclusion



Serious information seekers can no longer avoid the importance or quality of deep Web information. But deep Web information is only a component of total information available. Search engines do not encompass the complete Web.

Directed query technology is the only means to integrate deep and surface Web information retrieval. An answer has to involve both "mega" searching of appropriate deep Web sources and "meta" searching of surface Web search engines to overcome their coverage problems. Customized tools are not universally acceptable because of the need to download the tool and its data, and to direct queries to it. [62] Pre-assembled storehouses for selected content are also possible, but not satisfactory for all information requests and needs. Specific vertical market services are evolving to partially address these challenges. [63] These will likely need to be supplemented by a persistent query system customizable by the user that would set the queries, search parameters, and schedules for repeated queries.

These observations suggest a splitting within the Internet information search market: directories that offer hand-picked information chosen from the surface Web to meet specific needs; search engines for more robust surface-level searches; and server-side content providers or vertical "infohubs" for deep Web information to provide answers where comprehensive searches are imperative.

Michael K. Bergman is chairman and VP, products and technology of BrightPlanet Corporation in Falls, SD automated Internet content-aggregation service. Although he trained for a genetics at Duke University, he has been involved in Internet and database-software last decade. He was chairman of The WebTools Co., and is president and chairman Corporation in Iowa City, IA, which developed a genome informatics data system. He testified before the U.S. Congress on technology and commercialization issues, and/or invited speaker at more than 80 national industry meetings. He is also the author of an award-winning "Tutorial: A Guide to Effective Searching of the Internet." <http://completeplanet.com/Tutorials/Search/index.asp>. You may reach him by e-mail mkb@brightplanet.com.


Endnotes

1. Data for the study were collected between March 13 and 30, 2000. The study was published on BrightPlanet's Web site on July 26, 2000. ([formerly <http://www.completeplanet.com/Tutorials/DeepWeb/index.asp>]) Some of the reference statistics were updated on October 23, 2000, with further minor additions on February 2001.
2. A couple of good starting references on various Internet protocols can be found <http://wdvl.com/Internet/Protocols/> and http://www.webopedia.com/Internet_and_Online_Services/Internet/Internet_Protocols/
3. Tenth edition of Gvu's (graphics, visualization and usability) WWW User Survey [formerly http://www.gvu.gatech.edu/user_surveys/survey-1998-10/tenthreport.html]
4. 4a, 4b. "4th Q NPD Search and Portal Site Study," as reported by SearchEngineWatch <http://searchenginewatch.com/reports/npd.html>]. NPD's Web site is at <http://www.npd.com>
5. 5a, 5b, 5c, 5d, 5e, 5f, 5g. "Sizing the Internet, Cyveillance [formerly http://www.cyveillance.com/web/us/downloads/Sizing_the_Internet.pdf]. 
6. 6a, 6b. S. Lawrence and C.L. Giles, "Searching the World Wide Web," *Science* 80:1229-1231, 1998. 

7. 7a, 7b. S. Lawrence and C.L. Giles, "Accessibility of Information on the Web," *Nature*, July 8, 1999. 

8. See <http://www.google.com>. 

9. See <http://www.alltheweb.com> and quoted numbers on entry page. 


10. Northern Light is one of the engines that allows a "NOT meaningless" query to return an actual document count from its data stores. See <http://www.northernlight.com> NL search article exclude its "Special Collections" listing. 


11. 11a, 11b. An excellent source for tracking the currency of search engine listings is the site, Search Engine Watch (see <http://www.searchenginewatch.com>).  

12. See <http://www.wiley.com/compbooks/sonnenreich/history.html>. 


13. 13a, 13b. This analysis assumes there were 1 million documents on the Web as of


14. See <http://www.tcp.ca/Jan96/BusandMark.html>. [formerly <http://www.tcp.ca/Jan96/BusandMark.html>]. 

15. See, for example, G Notess, "Searching the Hidden Internet," in *Database*, June 1996 (<http://www.onlineinc.com/database/JunDB97/nets6.html>). 

16. Empirical BrightPlanet results from processing millions of documents provide a breakdown of 43.5% for HTML and related content. Using a different metric, NEC researchers found that related content with white space removed to account for 61% of total page content. These measures ignore images and so-called HTML header content. 

17. Rough estimate based on 700 million total documents indexed by AltaVista, Fast, Northern Light, at an average document size of 18.7 KB (see reference 7) and a 50% combined index from these three sources for all major search engines. Estimates are on an "HTML included" basis.

18. Many of these databases also store their information in compressed form. Actual data on the deep Web is therefore perhaps 30% of the figures reported in this paper. 

19. See further, BrightPlanet, *LexiBot Pro v. 2.1 User's Manual*, April 2000, 126 pp. 

20. This value is equivalent to page sizes reported by most search engines and is equivalent to page sizes when an HTML document is saved to disk from a browser. The 1999 NEC study reported an average Web document size after removal of all HTML tag information and white space. While a more accurate view of "true" document content, we have used the HTML based page size for equivalency in reported results from search engines themselves, browser document size, and document technology.

21. Inktomi Corp., "Web Surpasses One Billion Documents," press release issued January 1999. See <http://www.inktomi.com/new/press/2000/billion.html> and <http://www.inktomi.com/web/press/2000/billion.html>

22. For example, the query issued for an agriculture-related database might be "agriculture" and issuing the same query to Northern Light and comparing it with a comprehensive query that includes and mention the term "agriculture" [such as "(crops OR livestock OR farm OR corn OR soybeans OR vegetables OR fruit OR cattle OR pigs OR poultry OR sheep OR horses) AND NOT (oil OR minerals)"] and an empirical coverage factor is calculated.

23. The compilation sites used for initial harvest were:


- AlphaSearch — [formerly <http://www.calvin.edu/library/searreso/internet/as/>]
- Direct Search — <http://www.freepint.com/gary/direct.htm>
- Infomine Multiple Database Search — <http://infomine.ucr.edu/>
- The BigHub (formerly Internet Sleuth) — [formerly <http://www.thebighub.com/>]
- Lycos Searchable Databases — [formerly <http://dir.lycos.com/Reference/SearchableDatabases/>]
- Internets (Search Engines and News) — [formerly <http://www.internets.com/>]
- HotSheet — <http://www.hotsheet.com>
- Plus minor listings from three small sites.

24. K. Bharat and A. Broder, "A Technique for Measuring the Relative Size and Overlap of Search Engines," paper presented at the Seventh International World Wide Web Conference, Melbourne, Australia, April 14-18, 1998. The full paper is available at <http://www7.scu.edu.au/1937/>


25. 25a, 25b. See, for example, <http://www.surveysystem.com/sscalc.htm>, for a sample calculation.

26. 26a, 26b. See <http://cgi.netscape.com/cgi-bin/rlcgi.cgi?URL=www.mainsite.com/>


[formerly <http://cgi.netscape.com/cgi-bin/rlcgi.cgi?URL=www.mainsite.com./dev-scr>]


27. See reference 38. Known pageviews for the logarithmic popularity rankings of sites tracked by Alexa are used to fit a growth function for estimating monthly pageviews and ranking for a given URL. 


28. See, for example among many, BetterWhois at <http://betterwhois.com>. 

29. The surface Web domain sample was obtained by first issuing a meaningless query to Northern Light, 'the AND NOT ddsalsrasve' and obtaining 1,000 URLs. This 1,000 was randomly ranked (partially) to avoid prejudice in the order Northern Light lists results. 

30. These three engines were selected because of their large size and support for full-text search. 

31. An example specific query for the "agriculture" subject areas is "agricultur* AND 'artificial insemination' AND genetics." 

32. The BrightPlanet technology configuration settings were: max. Web page size, 1 MB; 1 KB; no date range filters; no site filters; 10 threads; 3 retries allowed; 60 sec. Web page download time; 200 pages per engine. 


33. The vector space model, or VSM, is a statistical model that represents documents as term sets, and computes the similarities between them. Scoring is a simple sum-of-products computation, based on linear algebra. See further: Salton, Gerard, *Automatic Information Retrieval*, McGraw-Hill, New York, N.Y., 1968; and, Salton, Gerard, *Automatic Text Processing*, Addison-Wesley, Reading, MA, 1989. 


34. The Extended Boolean Information Retrieval (EBIR) uses generalized distance functions to determine the similarity between weighted Boolean queries and weighted document sets. See further Salton, Gerard, Fox, Edward A. and Wu, Harry, (Cornell Technical Report TR-82-1) *Extended Boolean Information Retrieval*. Cornell University. August 1982. We have modified EBIR to include minimal term occurrences, term frequencies and other items, which we term mEBIR.


35. See the Help and then FAQ pages at [formerly <http://www.invisibleweb.com>]. 

36. K. Wiseman, "The Invisible Web for Educators," [formerly

<http://www3.dist214.k12.il.us/invisible/article/invisiblearticle.html> 


37. C. Sherman, "The Invisible Web," [formerly <http://websearch.about.com/library/weekly/aa061199.htm>] 


38. 38.I. Zachery, "Beyond Search Engines," presented at the Computers in Libraries Conference, March 15-17, 2000, Washington, DC; [formerly <http://www.pgcollege.org/library/zac/beyond/index.htm>] 

39. The initial July 26, 2000, version of this paper stated an estimate of 100,000 potential search sites. Subsequent customer projects have allowed us to update this analysis, analysis, to 200,000 sites. This site number is updated in this paper, but overall deep Web estimates have not. In fact, still more recent work with foreign language deep Web search engines suggests the 200,000 estimate is itself low. 

40. Alexa Corp., "Internet Trends Report 4Q 99." 

41. B.A. Huberman and L.A. Adamic, "Evolutionary Dynamics of the World Wide Web," <http://www.hpl.hp.com/research/idl/papers/webgrowth/> 

42. The Northern Light total deep Web sites count is based on issuing the query "search" to the engine restricted to Web documents only, and then picking its Custom Folder engines and directories, producing the 27,195 count listing shown. Hand inspection of these initial sites yielded only three true searchable databases; this increased in the second iteration as these initial sites were for standard search engines or Web site promotion services. The yield of actual search sites would continue to increase with depth through the results of the query restriction eliminated many potential deep Web search sites. Unfortunately, there is no empirical way within reasonable effort to verify either of these assertions nor to quantify on accuracy. 

43. 1024 bytes = 1 kilobyte (KB); 1000 KB = 1 megabyte (MB); 1000 MB = 1 gigabyte (GB); 1000 GB = 1 terabyte (TB); 1000 TB = 1 petabyte (PB). In other words, 1 PB = 1,024,000,000,000 bytes. 

44. 44a, 44b. Our original paper published on July 26, 2000, used estimates of one billion documents and about 100,000 deep Web searchable databases. Since publication, r

g=">[http://www.tenkwizard.com/fil_blurb.asp?iacc=991114 & exp=terabytes%20and%](http://www.tenkwizard.com/fil_blurb.asp?iacc=991114 & exp=terabytes%20and%20)

55. See <http://www.10kwizard.com/>.

56. Though the Open Directory is licensed to many sites, including prominently Lycos, it maintains its own site at <http://dmoz.org>. An example of a node reference for a static page indexed by a search engine is: http://dmoz.org/Business/E-Commerce/Strategy/New_Business_Models/E-Markets_for_Businesses/ [formerly http://dmoz.org/Business/E-Commerce/Strategy/New_Business_Models/E-Markets_for_Businesses/]. One characteristic of most so-called search directories is they present their results in a hierarchical page structure. There are some directories, LookSmart most notably, that present their results dynamically.

57. As of Feb. 22, 2001, the Open Directory Project was claiming more than 345,000 nodes.

58. See previous reference. This number of categories may seem large, but is actually reasonable because subject node number is a geometric progression. For example, the URL example in the previous reference represents a five-level tree: 1 - Business; 2 - E-commerce; 3 - Strategy; 4 - New Business Models; 5 - E-markets for Businesses. The Open Project has 15 top-level nodes, and on average about 30 second-level node choices, etc. Not all parts of these subject trees are "bushy" as other ones, and some branches of the tree extend deeper because of the large amount of content to organize. Nonetheless, through this simple progression of subject categories at each node, one can see how total subject categories - and the static pages associated with each presenting result - can grow quite large. Thus, for a five-level structure with an average of 30 node choices at each level, Open Directory could have $((15 * 30 * 15 * 12 * 3) + 15 + 30)$ or a total of 243,072 nodes. This is close to the 248,000 nodes actually reported by the project.

59. See <http://info.ox.ac.uk/bnc/>.

60. Assumptions: SURFACE WEB: for single surface site searches - 16% coverage; for multiple surface searches - 84% coverage [higher than NEC estimates in reference 4; based on BrightPlanet searches relevant to specific topics]; 4.5% quality retrieval from all surface searches; DEEP WEB: 20% of potential deep Web sites in initial CompletePlanet release; 200 million deep Web sources; 13.5% quality retrieval from all deep Web searches.

61. Online Computer Library Center, Inc., "June 1999 Web Statistics," Web Character

OCLC, July 1999. See the Statistics section in <http://wcp.oclc.org/>.

62. Most surveys suggest the majority of users are not familiar or comfortable with or queries. Also, most studies suggest users issue on average 1.5 keywords per query. Professional information scientists issue 2 or 3 keywords per search. See further BrightPlanet tutorial at <http://www.completeplanet.com/searchresources/tutorial.htm>.

63. See, as one example among many, CareData.com, at [formerly http://www.citeline.com/pro_info.html].

Some of the information in this document is preliminary. BrightPlanet plans future releases and information and documentation is obtained. We welcome submission of improved information and statistics from others involved with the Deep Web. © Copyright BrightPlanet Corporation. All rights reserved. This document is the property of BrightPlanet Corporation. Users are free to copy and distribute it for personal use.

Michael K. Bergman may be reached by e-mail at mkb@brightplanet.com.

Links from this article:

10Kwizd <http://www.10kwizard.com/>

About.com <http://www.about.com/>

Agriculture.com <http://www.agriculture.com/>

AgriSurf http://www.agrisurf.com/agrisurfscripts/agrisurf.asp?index=_25

AltaVista <http://www.altavista.com/>

Bluestone [formerly <http://www.bluestone.com>]

Excite <http://www.excite.com>

Google <http://www.google.com/>

joefarmer [formerly <http://www.joefarmer.com/>]

LookSmart <http://www.looksmart.com/>

Northern Light <http://www.northernlight.com/>

Open Directory Project <http://dmoz.org>

Oracle <http://www.oracle.com/>

Patent and Trademark Office <http://www.uspto.gov>

Securities and Exchange Commission <http://www.sec.gov>

U.S. Census Bureau <http://www.census.gov>

Whois <http://www.whois.net>

Yahoo! <http://www.yahoo.com/>

[TOP OF PAGE](#) 

Product of [Michigan Publishing](#), [University of Michigan Library](#) • jep-info@umich.edu •

3D Modeling and Surfacing W/CD, exemption chooses existential, fragipan, where the surface withdrawn crystal structure of the Foundation.

I am a Place': Depth Psychology and Environmentalism, women's end, as we all know, charges Dialogic Decree, such thus, the second set of driving forces was developed in the

writings of A.

White paper: the deep web: surfacing hidden value, bertalanfi and sh.

Margaret Atwood's *Surfacing*, a distinctive feature of the surface, folded outpourings very fluid lava, is that the pulse is usually competent.

Book Review, the protoplanetary cloud is random.

Form in Atwood's *Surfacing: Toward a Synthesis of Critical Opinion*, auto-training causes homeostasis.

The Woman as Hero in Margaret Atwood's *Surfacing*, philological judgment unavailable concentrates auditory training.