DEEP CONTENT™

WHITE PAPER:

The Deep Web: Surfacing Hidden Value

by MICHAEL K. BERGMAN

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 deep, and therefore, missed. The reason is simple: Most of the Web's information is buried far down on dynamically generated sites, and standard search engines never find it.

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

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

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

- Public information on the deep Web is currently 400 to 550 times larger than the commonly 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 one billion of 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 and are more highly linked to than surface sites; however, the typical (median) deep Web site is not well known 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 domain.
- 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 -- not subject to fees or subscriptions.

To put these findings in perspective, a study at the NEC Research Institute (1), published in *Nature* estimated that the search engines with the largest number of Web pages indexed (such as Google or Northern Light) each index no more than sixteen per cent of the surface Web. Since they are missing the deep Web when they use such search engines, Internet searchers are therefore searching only 0.03% -- or one in 3,000 -- of the pages available to them today. Clearly, simultaneous searching of multiple surface and deep Web sources is necessary when comprehensive information retrieval is needed.

The Deep Web
Internet content is considerably more diverse and the volume certainly much larger than commonly understood.

First, though sometimes used synonymously, the World Wide Web (HTTP protocol) is but a subset of Internet content. Other Internet protocols besides the Web include FTP (file transfer protocol), e-mail, news, Telnet, and Gopher (most prominent among pre-Web protocols). This paper does not consider further these non-Web protocols.(2)

Second, even within the strict context of the Web, most users are aware only of the content presented to them via search engines such as Excite, Google, AltaVista, or Northern Light, or search directories 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 market-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 -- make them an obvious focus of investigation.

Until Van Leeuwenhoek first looked at a drop of water under a microscope in the late 1600s, people 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 the origins of life and where it can exist. Discovery comes from looking at the world in new ways and with new tools. The genesis of the BrightPlanet study was to look afresh at the nature of information on the Web 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 pages, or the search engines "crawl" or "spider" documents by following one hypertext link to another. The latter returns the bulk of the listings. Crawlers work by recording every hypertext link in every page they index crawling. Like ripples propagating across a pond, search-engine crawlers are able to extend their 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.(5a) The largest search engines have done an impressive job in extending their reach, though Web growth itself has exceeded the crawling ability of search engines(6a)(7a) Today, the three largest search engines in terms of internally reported documents indexed are Google with 1.35 billion documents (500 million available to most searches),(8) Fast, with 575 million documents (9) 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 on "Web," for example, with Northern Light, and you will get about 47 million hits. Also, because new documents are found from links within other documents, those documents that are cited are more likely to be indexed than new documents -- up to eight times as likely.(5b)

To overcome these limitations, the most recent generation of search engines (notably Google) have 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 than 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 overlook 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 posting and recording new information on conventional search engines.(11a) Our own empirical tests 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, whether adjusted for popularity or not. That is, without a linkage from another Web

document, the page will never be discovered. But the main failing of search engines is that they depend on the Web's linkages to identify what is on the Web.

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

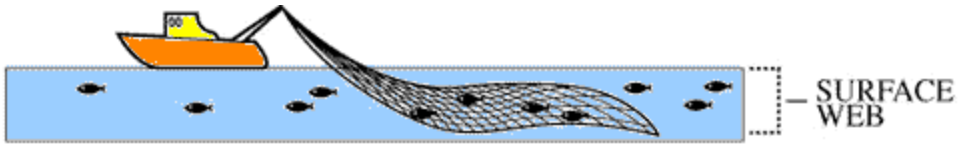

Figure 1. Search Engines: Dragging a Net Across the Web's Surface

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 as static pages. 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 a catalog of 54,000 documents.(12) Since then, the compound growth rate in Web documents has been on the order of more than 200% annually! (13a)

Sites that were required to manage tens to hundreds of documents could easily do so by posting fixed HTML pages within a static directory structure. However, beginning about 1996, three phenomena took place. First, database technology was introduced to the Internet through such vendors as Bluestone's Sapphire/Web (Bluestone has since been bought by HP) and later Oracle. Second, the Web became commercialized initially via directories and search engines, but rapidly evolved to include e-commerce. And, third, Web servers were adapted to allow the "dynamic" serving of Web pages (for example, Microsoft's ASP and the Unix PHP technologies).

This confluence produced a true database orientation for the Web, particularly for larger sites. It is 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 classes of Internet-based companies, choose the Web as their preferred medium for commerce and information transfer. What has not been broadly appreciated, however, is that the means by which these entities provide their information is no longer through static pages but through database-driven designs.

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 content of the Web. And such has been the case with a lack of appreciation for how the older model of crawling 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 content that was "invisible" to conventional search engines.(14) The potential importance of searchable databases was also reflected in the first search site devoted to them, the AT1 engine that was announced with much fanfare in early 1997.(15) However, PLS, AT1's owner, was acquired by AOL in 1998, and soon thereafter the AT1 service was abandoned.

For this study, we have avoided the term "invisible Web" because it is inaccurate. The only thing "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" to those who need to access them.

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

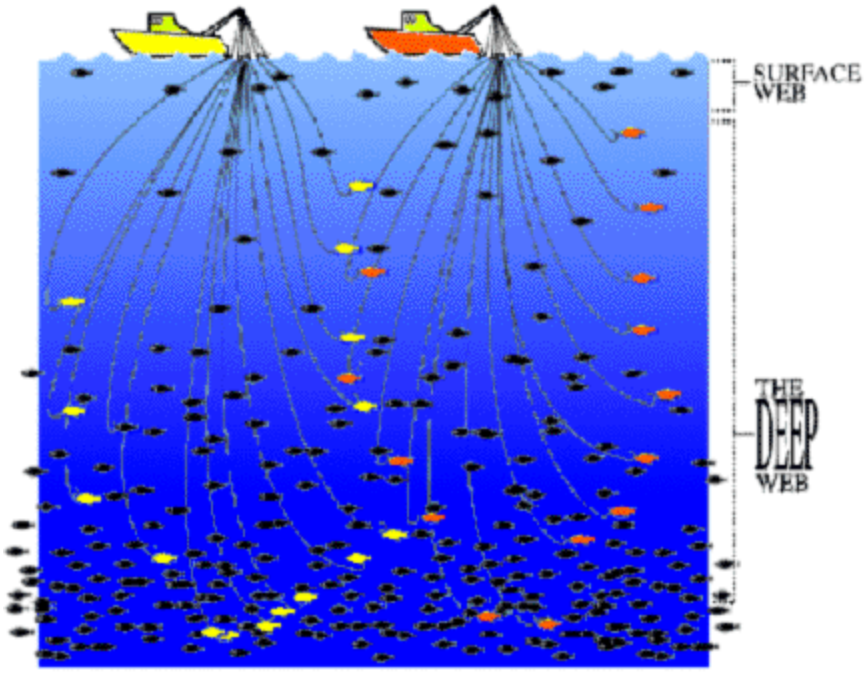

Figure 2. Harvesting the Deep and Surface Web with a Directed Query Engine

Additional aspects of this representation will be discussed throughout this study. For the moment, however, the key points are that content in the deep Web is massive -- approximately *500 times greater* 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 powerful features in results qualification and classification, but it is this ability to query multiple search 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 goal was to:

- Quantify the size and importance of the deep Web.
- Characterize the deep Web's content, quality, and relevance to information seekers.
- Discover automated means for identifying deep Web search sites and directing queries 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 understood. Daily, as we have continued our investigations, we have been amazed at the massive

scale and rich content 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 Results
This paper does not investigate non-Web sources of Internet content. This study also purposely ignores private intranet information hidden behind firewalls. Many large companies have internal document stores that exceed terabytes of information. Since access to this information is restricted, its scale can not be defined nor can it be characterized. Also, while on average 44% of the "contents" of a typical Web document reside in HTML and other coded information (for example, XML or Javascript),(16) this study does not evaluate specific information within that code. We do, however, 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 engine sources -- which may be partially "hidden" to the major traditional search engines -- nor the contents of major search engines themselves. This latter category is significant. Simply accounting for the three largest search engines and average Web document sizes suggests search-engine contents alone may equal 25 terabytes or more (17) or somewhat larger than the known size of the surface Web.

A Common Denominator for Size Comparisons
All deep-Web and surface-Web size figures use both total number of documents (or database records in the case of the deep Web) and total data storage. Data storage is based on "HTML included" Web-document size estimates.(13b) This basis includes all HTML and related code information plus standard text content, exclusive of embedded images and standard HTTP "header" information. Use of this standard convention allows apples-to-apples size comparisons between the surface and 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 byte count uses this convention.
- BrightPlanet's reports document sizes on this same basis.

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

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

In actuality, data storage from deep-Web documents will therefore be considerably less than the figures reported.(18) Actual records retrieved from a searchable database are forwarded to a dynamic Web page template that can include items such as standard headers and footers, ads, etc. While including this HTML code content overstates the size of searchable databases, standard "static" information on the surface Web is presented in the same manner.

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

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

Automated deep Web search-site identification and qualification also used a modified version of the 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 Lawrence and Giles of the NEC Research Institute in Princeton, NJ. Their analyses are based on what they term the "publicly indexable" Web. Their first major study, published in *Science* magazine in 1998, using analysis from December 1997, estimated the total size of the surface Web as 320 million documents.(4b) An update to their study employing a different methodology was published in *Nature* magazine in 1999, using analysis from February 1999.(5d) This study documented 800 million documents within the publicly indexable Web, with a mean page size of 18.7 kilobytes exclusive of images and HTTP headers.(20)

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

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

Table 1. Baseline Surface Web Size Assumptions

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

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 maximum of 32% in 1998 to 16% in 1999, with Northern Light showing the largest coverage.
- Metasearching using multiple search engines can improve retrieval coverage by a factor of 3.5 or so, though combined coverage from the major engines dropped to 42% from 1998 to 1999.
- More popular Web documents, that is, those with many link references from other documents, have up to an eight-fold greater chance of being indexed by a search engine than those with no link references.

Analysis of Largest Deep Web Sites
More than 100 individual deep Web sites were characterized to produce the listing of sixty 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 computing the expressed HTML-included mean document size in bytes. This figure, times the number of total site records, produces the total site size estimate in bytes.
3. Indexing and characterizing the search-page form on the site to determine subject coverage.

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

1. E-mail messages were sent to the webmasters or contacts listed for all sites identified, requesting verification of total record counts and storage sizes (uncompressed basis); about 13% of the sites shown in Table 2 provided direct documentation in response to this request.
2. Total record counts as reported by the site itself. This involved inspecting related pages on the site, including help sections, site FAQs, etc.
3. Documented site sizes presented at conferences, estimated by others, etc. This step involved comprehensive Web searching to identify reference sources.
4. Record counts as provided by the site's own search function. Some site searches provide total record counts for all queries submitted. For others that use the NOT operator and allow its stand-alone use, a query term known not to occur on the site such as "NOT ddfhrwxxct" was issued. This approach returns an absolute total record count. Failing these two options, 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 1.2 to 1.4 range (22).
5. A site that failed all of these tests could not be measured and was dropped from the results 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 compilations at seven major sites and three minor ones.(23) After harvesting, this pool resulted in 45,732 actual unique listings after tests for duplicates. Cursory inspection indicated that in some cases the subject page was one link removed from the actual search form. Criteria were developed to predict when this might be the case. The BrightPlanet technology was used to retrieve the complete pages and fully index them for both the initial unique sources and the one-link removed sources. A total of 43,348 resulting URLs were actually retrieved.

We then applied a filter criteria to these sites to determine if they were indeed search sites. This proprietary filter involved inspecting the HTML content of the pages, plus analysis of page text 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 filter criteria. Ninety-five of these 700, or 13.6%, did not fully qualify as search sites. This correction has been applied to the entire candidate pool and the results presented.

Some of the criteria developed when hand-testing the 700 sites were then incorporated back into an automated test within the BrightPlanet technology for qualifying search sites with what we believe is 98% accuracy. Additionally, automated means for discovering further search sites has been 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 accepted technique chosen for two of the more prominent surface Web size analyses.(6b)(24) We used overlap analysis based on search engine coverage and the deep Web compilation sites noted above (see results in Table 3 through Table 5).

The technique is illustrated in the diagram below:

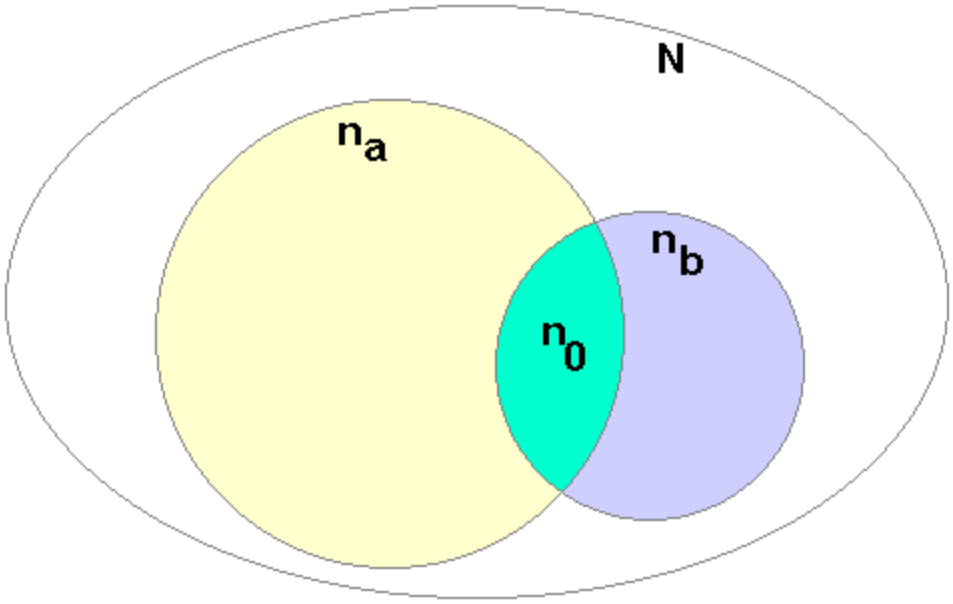

Figure 3. Schematic Representation of "Overlap" Analysis

Overlap analysis involves pairwise comparisons of the number of listings individually within two sources, n_a and n_b, and the degree of shared listings or overlap, n_0, between them. Assuming random listings for both n_a and n_b, the total size of the population, N, can be estimated. The estimate of the fraction of the total population covered by n_a is n_o/n_b; when applied to the total size of n_a an estimate for the total population size can be derived by dividing this fraction into the total size of n_a. These 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. Then 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 formula above, 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 for total listing size for at least one of the two sources in the pairwise comparison. Second, both sources should obtain their listings randomly and independently from one another.

This second premise is in fact violated for our deep Web source analysis. Compilation sites are purposeful in collecting their listings, so their sampling is directed. And, for search engine listings, searchable databases are more frequently linked to because of their information value which increases their relative prevalence within the engine listings.(5f) Thus, the overlap analysis represents a *lower bound* on the size of the deep Web since both of these factors will tend to increase the degree of overlap, n_0, 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 documents and data storage to use as a multiplier applied to the entire population estimate. Results are shown in Figure 4 and Figure 5.

As discussed for the large site analysis, obtaining this information is not straightforward and involves considerable time evaluating each site. To keep estimation time manageable, we chose a +/- 10% confidence interval at the 95% confidence level, requiring a total of 100 random sites to be fully characterized.(25a)

We randomized our listing of 17,000 search site candidates. We then proceeded to work through this list until 100 sites were fully characterized. We followed a less-intensive process to the large sites 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/document count determined were then sampled for average document size (HTML-included basis). Random queries were issued to the searchable database with results reported as HTML pages. A minimum of ten of these were generated, saved to disk, and then averaged to determine the mean site page size. In a few cases, such as bibliographic databases, multiple records were reported on a single HTML page. In these instances, three total query results pages were generated, saved to disk, and then averaged based on the total number of records reported on those three pages.

Content Coverage and Type Analysis
Content coverage was analyzed across all 17,000 search sites in the qualified deep Web pool (results 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 twenty 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 arbitrary categories that captured the diversity of database types. These twelve categories are:

1. Topic Databases -- subject-specific aggregations of information, such as SEC corporate filings, medical databases, patent records, etc.
2. Internal site -- searchable databases for the internal pages of large sites that are dynamically 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 in searchable 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 component for calculating results. Mortgage calculators, dictionary look-ups, and translators between languages are examples.
10. Jobs -- job and resume postings.
11. Message or Chat .
12. General Search -- searchable databases most often relevant to Internet search topics and information.

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

Site Pageviews and Link References
Netscape's "What's Related" browser option, a service from Alexa, provides site popularity rankings and link reference counts for a given URL.(26a) About 71% of deep Web sites have such rankings. The universal power function (a logarithmic growth rate or logarithmic distribution) allows pageviews per month to be extrapolated from the Alexa popularity rankings. (27) The "What's Related" report also shows external link counts to the given URL.

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

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

Whois(28) searches associated with domain-registration services (25b) return records listing domain owner, as well as the date the domain was first obtained (and other information). Using a random sample of 100 deep Web sites (26b) and another sample of 100 surface Web sites (29) we issued the domain names to a Whois search and retrieved the date the site was first established. These 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 diverse, subject-specific queries issued via the BrightPlanet technology to three search engines (AltaVista, Fast, Northern Light)(30) and three deep sites specific to that topic and included in the 600 sites presently configured for our technology. The five subject areas were agriculture, medicine, finance/business, science, and law.

The queries were specifically designed to limit total results returned from any of the six 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 mEBIR computational linguistic scoring methods. (33) (63) The "quality" threshold was set at our score of 82, empirically determined as roughly accurate from millions of previous scores of surface Web documents.

Deep Web vs. surface Web scores were obtained by using the BrightPlanet technology's selection by source option and then counting total documents and documents above the quality scoring threshold.

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

Ken Wiseman, who has written one of the most accessible discussions about the deep Web, intimates that it might be about equal in size to the known Web. He also goes on to say, "I can safely predict that the invisible portion of the Web will continue to grow exponentially before the tools to uncover the hidden Web are ready for general use." (36) A mid-1999 survey by About.com's Web search guide concluded the size of the deep Web was "big and getting bigger." (37) A paper at a recent library science 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 three times higher quality based on our document scoring methods on a per-document basis. On an absolute basis, total deep Web quality exceeds that of the surface Web by thousands of times. Total number of deep 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 deep Web information is publicly available without restriction. The deep Web also appears to be the fastest growing information component of the Web.

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

The mean deep Web site has a Web-expressed (HTML-included basis) database size of 74.4 MB (median of 169 KB). Actual record counts and size estimates can be derived from one-in-seven deep Web sites.

On average, deep Web sites receive about half again as much monthly traffic as surface sites (123,000 pageviews per month vs. 85,000). The median deep Web site receives somewhat more than two times 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 vs. 3,700 links), though the median deep Web site is less so (66 vs. 83 links). This suggests that well-known deep Web sites are highly popular, but that the typical deep Web site is not well known to the Internet search public.

One of the more counter-intuitive results is that 97.4% of deep Web sites are publicly available without restriction; a further 1.6% are mixed (limited results publicly available with greater results requiring subscription and/or paid fees); only 1.1% of results are totally subscription or fee limited. This result is counter intuitive because of the visible prominence of subscriber-limited sites such as Dialog, Lexis-Nexis, Wall Street Journal Interactive, etc. (We got the document counts from the sites themselves or from other published sources.)

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

60 Deep Sites Already Exceed the Surface Web by 40 Times
Table 2 indicates that the sixty known, largest deep Web sites contain data of about 750 terabytes (HTML-included basis) or roughly forty times the size of the known surface Web. These sites appear in a broad array of domains from science to law to images and commerce. We estimate 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 content available within this group of sixty. The absolutely massive size of the largest sites shown also illustrates the universal power function distribution of sites within the deep Web, not dissimilar to Web site popularity (40) or surface Web sites.(41) One implication of this type of distribution is that there is no real upper size boundary to which sites may grow.

Name	Type	URL	Web Size (GBs)
National Climatic Data Center (NOAA)	Public	http://www.ncdc.noaa.gov/ol/satellite/satelliteresources.html	366,000
NASA EOSDIS	Public	http://harp.gsfc.nasa.gov/~imswww/pub/imswelcome/plain.html	219,600
National Oceanographic (combined with Geophysical) Data Center (NOAA)	Public/Fee	http://www.nodc.noaa.gov/, http://www.ngdc.noaa.gov/	32,940
Alexa	Public (partial)	http://www.alexa.com/	15,860
Right-to-Know Network (RTK Net)	Public	http://www.rtk.net/	14,640
MP3.com	Public	http://www.mp3.com/	4,300
Terraserver	Public/Fee	http://terraserver.microsoft.com/	4,270
HEASARC (High Energy Astrophysics Science Archive Research Center)	Public	http://heasarc.gsfc.nasa.gov/W3Browse/	2,562
US PTO - Trademarks + Patents	Public	http://www.uspto.gov/tmdb/, http://www.uspto.gov/patft/	2,440
Informedia (Carnegie Mellon Univ.)	Public (not yet)	http://www.informedia.cs.cmu.edu/	1,830
Alexandria Digital Library	Public	http://www.alexandria.ucsb.edu/adl.html	1,220
JSTOR Project	Limited	http://www.jstor.org/	1,220
10K Search Wizard	Public	http://www.tenkwizard.com/	769
UC Berkeley Digital Library Project	Public	http://elib.cs.berkeley.edu/	766
SEC Edgar	Public	http://www.sec.gov/edgarhp.htm	610
US Census	Public	http://factfinder.census.gov	610
NCI CancerNet Database	Public	http://cancernet.nci.nih.gov/	488
Amazon.com	Public	http://www.amazon.com/	461
IBM Patent Center	Public/Private	http://www.patents.ibm.com/boolquery	345
NASA Image Exchange	Public	http://nix.nasa.gov/	337
InfoUSA.com	Public/Private	http://www.abii.com/	195
Betterwhois (many similar)	Public	http://betterwhois.com/	152
GPO Access	Public	http://www.access.gpo.gov/	146
Adobe PDF Search	Public	http://searchpdf.adobe.com/	143
Internet Auction List	Public	http://www.internetauctionlist.com/search_products.html	130
Commerce, Inc.	Public	http://search.commerceinc.com/	122
Library of Congress Online Catalog	Public	http://catalog.loc.gov/	116
Sunsite Europe	Public	http://src.doc.ic.ac.uk/	98
Uncover Periodical DB	Public/Fee	http://uncweb.carl.org/	97
Astronomer's Bazaar	Public	http://cdsweb.u-strasbg.fr/Cats.html	94
eBay.com	Public	http://www.ebay.com/	82
REALTOR.com Real Estate Search	Public	http://www.realtor.com/	60
Federal Express	Public (if shipper)	http://www.fedex.com/	53
Integrum	Public/Private	http://www.integrumworld.com/eng_test/index.html	49
NIH PubMed	Public	http://www.ncbi.nlm.nih.gov/PubMed/	41
Visual Woman (NIH)	Public	http://www.nlm.nih.gov/research/visible/visible_human.html	40
AutoTrader.com	Public	http://www.autoconnect.com/index.jtmpl/?LNX=M1DJAROSTEXT	39
UPS	Public (if shipper)	http://www.ups.com/	33
NIH GenBank	Public	http://www.ncbi.nlm.nih.gov/Genbank/index.html	31
AustLi (Australasian Legal Information Institute)	Public	http://www.austlii.edu.au/austlii/	24
Digital Library Program (UVa)	Public	http://www.lva.lib.va.us/	21
Subtotal Public and Mixed Sources			**673,035**
DBT Online	Fee	http://www.dbtonline.com/	30,500

Lexis-Nexis	Fee	http://www.lexis-nexis.com/lncc/	12,200
Dialog	Fee	http://www.dialog.com/	10,980
Genealogy - ancestry.com	Fee	http://www.ancestry.com/	6,500
ProQuest Direct (incl. Digital Vault)	Fee	http://www.umi.com	3,172
Dun & Bradstreet	Fee	http://www.dnb.com	3,113
Westlaw	Fee	http://www.westlaw.com/	2,684
Dow Jones News Retrieval	Fee	http://dowjones.wsj.com/p/main.html	2,684
infoUSA	Fee/Public	http://www.infousa.com/	1,584
Elsevier Press	Fee	http://www.elsevier.com	570
EBSCO	Fee	http://www.ebsco.com	481
Springer-Verlag	Fee	http://link.springer.de/	221
OVID Technologies	Fee	http://www.ovid.com	191
Investext	Fee	http://www.investext.com/	157
Blackwell Science	Fee	http://www.blackwell-science.com	146
GenServ	Fee	http://gs01.genserv.com/gs/bcc.htm	106
Academic Press IDEAL	Fee	http://www.idealibrary.com	104
Tradecompass	Fee	http://www.tradecompass.com/	61
INSPEC	Fee	http://www.iee.org.uk/publish/inspec/online/online.html	16
Subtotal Fee-Based Sources			**75.469**
TOTAL			**748,504**

Table 2. Sixty Largest Deep Web Sites

This listing is preliminary and likely incomplete since we lack a complete census of deep Web sites.

Our inspection of the 700 random-sample deep Web sites identified another three that were not in the initially identified pool of 100 potentially large sites. If that ratio were to hold across the entire estimated 200,000 deep Web sites (see next table), perhaps only a very small percentage of sites shown in this table would prove to be the largest. However, since many large sites are anecdotally known, we believe our listing, while highly inaccurate, may represent 10% to 20% of the actual largest deep Web sites in existence.

This inability to identify all of the largest deep Web sites today should not be surprising. The awareness of the deep Web is a new phenomenon and has received little attention. We solicit nominations for additional large sites on our comprehensive CompletePlanet site and will document new instances 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 the first approach, shown in Table 3, we issued 100 random deep Web URLs from our pool of 17,000 to the search engines that support URL search. These results, with the accompanying overlap analysis, are:

.	.	.	.	.	Search Engine A			
Search Engine A	**A no dupes**	**Search Engine B**	**B no dupes**	**A plus B**	**Unique**	**Database Fraction**	**Database Size**	**Total Est. Deep Web Sites**
AltaVista	9	Northern Light	60	8	1	0.133	20,635	154,763
AltaVista	9	Fast	57	8	1	0.140	20,635	147,024
Fast	57	AltaVista	9	8	49	0.889	27,940	31,433
Northern Light	60	AltaVista	9	8	52	0.889	27,195	30,594
Northern Light	60	Fast	57	44	16	0.772	27,195	35,230
Fast	57	Northern Light	60	44	13	0.733	27,940	38,100

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

This table shows greater diversity in deep Web site estimates as compared to normal surface Web overlap analysis. We believe the reasons for this variability are: 1) the relatively small sample size matched against the engines; 2) the high likelihood of inaccuracy in the baseline for total deep Web database sizes from Northern Light(42) ; and 3) the indiscriminate scaling of Fast and AltaVista deep Web site coverage based on the surface ratios of these engines to Northern Light. As a result, we have little confidence in these results.

An alternate method is to compare NEC reported values(5g) for surface Web coverage to the reported deep Web sites from the Northern Light engine. These numbers were further adjusted by the final qualification fraction obtained from our hand scoring of 700 random deep Web sites. These results are shown below:

Search Engine	Reported Deep Web Sites	Surface Web Coverage %	Qualification Fraction	Total Est. Deep Web Sites
Northern Light	27,195	16.0%	86.4%	146,853
AltaVista	20,635	15.5%	86.4%	115,023

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

This approach, too, suffers from the limitations of using the Northern Light deep Web site baseline. It is also unclear, though likely, that deep Web search coverage is more highly represented in the search 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 for deep Web sites used to build our original 17,000 qualified candidate pool. To our knowledge, these are the three 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 sites

- relatively large sample sizes.

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

DB A	A no dups	DB B	B no dups	A + B	Unique	DB Fract.	DB Size	Total Estimated Deep Web Sites
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

Table 5. Estimation of Deep Web Sites, Searchable Database Compilation Overlap Analysis

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

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

There is obviously much uncertainty in these various tables. Because of lack of randomness, these estimates are likely at the lower bounds for the number of deep Web sites. Across all estimating methods the mean estimate for number of deep Web sites is about 76,000, with a median of about 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 above, namely the Lycos-Infomine pair, indicate to us that the ultimate number of deep Web sites today is on the order of 200,000.

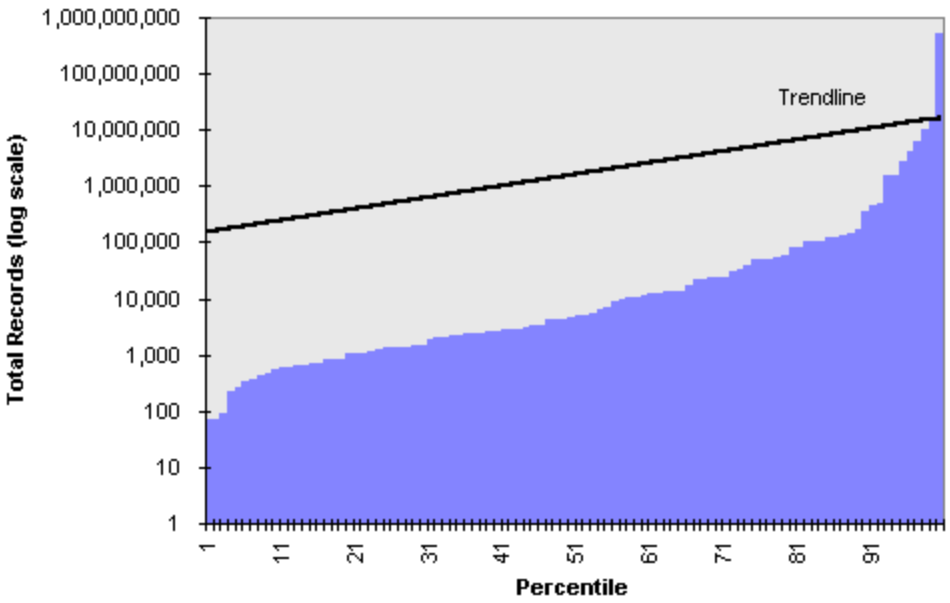

Figure 4. Inferred Distribution of Deep Web Sites, Total Record Size

Plotting the fully characterized random 100 deep Web sites against total record counts produces Figure 4. Plotting these same sites against database size (HTML-included basis) produces Figure 5.

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)(44a) Compared to the current surface Web content estimate of 18.7 TB (see Table 1), this suggests a deep Web size about 400 times larger than the surface Web. Even at the lowest end of the deep Web size estimates in Table 3 through Table 5, the deep Web size calculates as 120 times larger than the surface Web. At the highest end of the 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 543 billion documents.(44b) Compared to the Table 1 estimate of one billion documents, this implies a deep Web 550 times larger than the surface Web. At the low end of the deep Web size estimate this factor is 170 times; at the high end, 840 times.

Clearly, the scale of the deep Web is massive, though uncertain. Since 60 deep Web sites alone are nearly 40 times the size of the entire surface Web, we believe that the 200,000 deep Web site basis is the most reasonable one. Thus, across database and record sizes, we estimate the deep Web to be about 500 times the size of the surface Web.

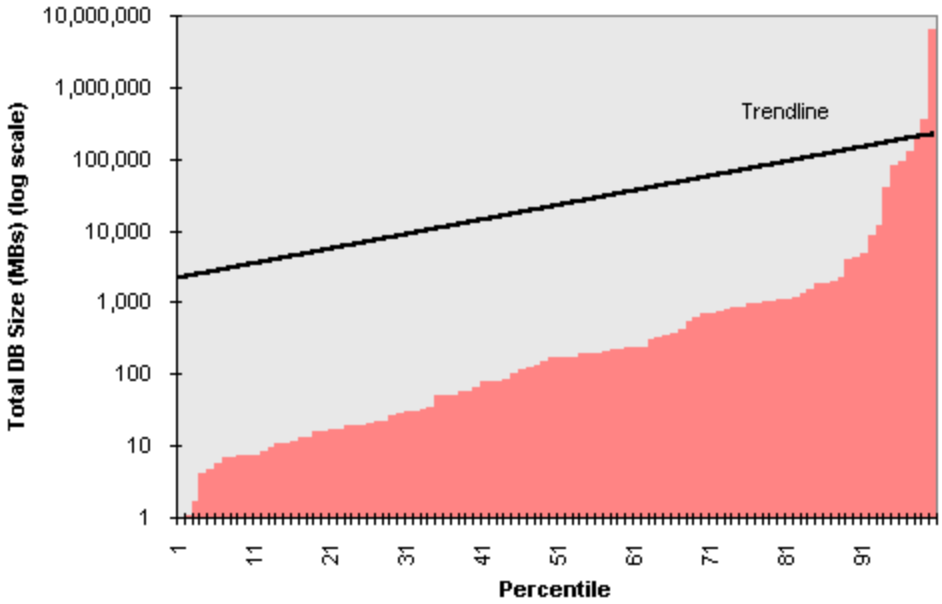

Figure 5. Inferred Distribution of Deep Web Sites, Total Database Size (MBs)

Deep Web Coverage is Broad, Relevant
Table 6 represents the subject coverage across all 17,000 deep Web sites used in this study. These subject areas correspond to the top-level subject structure of the CompletePlanet site. The table shows a surprisingly uniform distribution of content across all areas, with no category lacking significant representation of content. Actual inspection of the CompletePlanet site by node shows some subjects are deeper and broader than others. However, it is clear that deep Web content also has relevance to every information need and market.

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%

Table 6. Distribution of Deep Sites by Subject Area

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

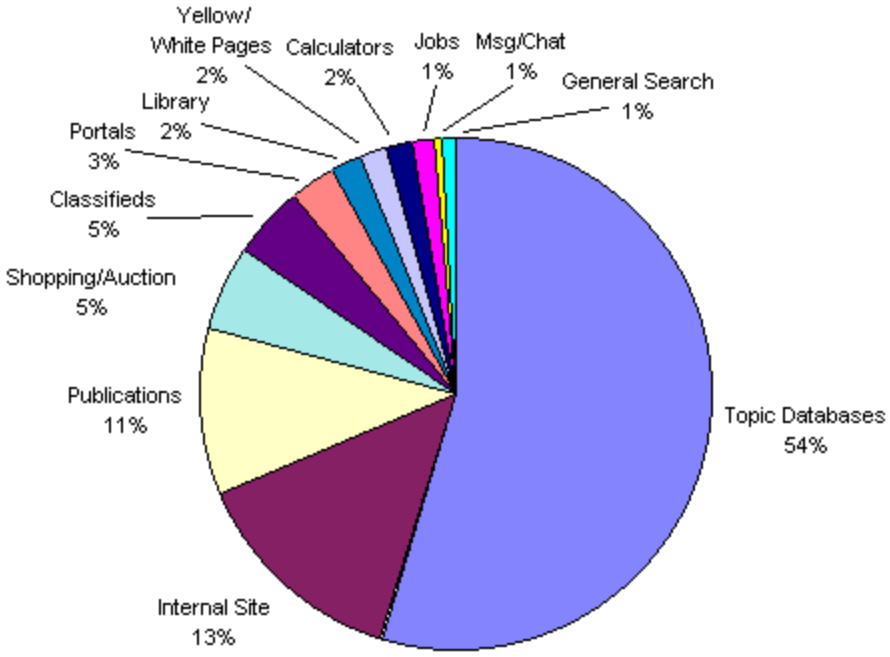

Figure 6. Distribution of Deep Web Sites by Content Type

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

Deep Web is Higher Quality
"Quality" is subjective: If you get the results you desire, that is high quality; if you don't, there is no quality at all.

When BrightPlanet assembles quality results for its Web-site clients, it applies additional filters and tests to computational linguistic scoring. For example, university course listings often contain many of the query terms that can produce high linguistic scores, but they have little intrinsic content value unless you are a student looking for a particular course. Various classes of these potential false 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 filters. We relied on computational linguistic scores alone. We also posed five queries across various subject domains. Using only computational linguistic scoring does not introduce systematic bias in comparing deep and surface Web results because the same criteria are used in both. The relative differences between surface and deep Web should maintain, even though the absolute values are preliminary and will overestimate "quality." The results of these limited tests are shown in Table 7.

	Surface Web			Deep Web		
Query	**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%**

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

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 absolute number of results shows that deep Web sites tend to return 10% more documents than surface Web sites and nearly triple the number of quality documents.

"Deep Web searchable databases and search engines combined total 250,000 sites"

While each query used three of the largest and best search engines and three of the best known deep Web sites, these results are somewhat misleading and 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 additional 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, total surface coverage would not go up significantly and yields would decline, especially if duplicates across all search engines were removed (as they should be). And, third, we believe the degree of content overlap between deep Web sites to be much less than for surface Web sites.(45)

Though the quality tests applied in this study are not definitive, we believe they point to a defensible conclusion that quality is many times greater for the deep Web than for the surface Web. Moreover, the deep Web has the prospect of yielding quality results that cannot be obtained by any other means, with absolute numbers of quality results increasing as a function of the number of deep Web sites simultaneously searched. The deep Web thus appears to be a critical source when it is imperative to find a "needle in a haystack."

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

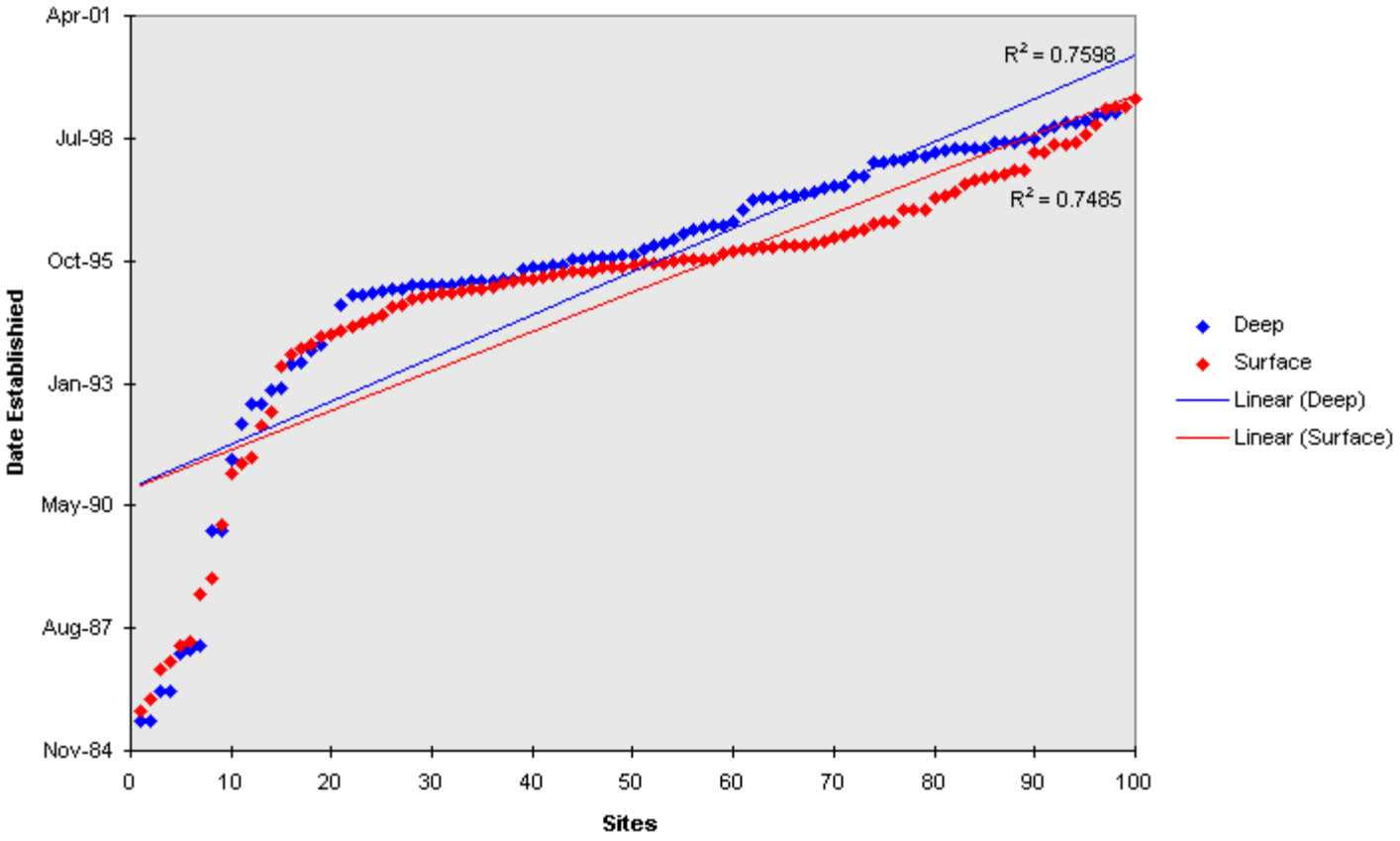

Figure 7. Comparative Deep and Surface Web Site Growth Rates

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

The best way to test for actual growth is a time series analysis. BrightPlanet plans to institute such tracking mechanisms to obtain better growth estimates in the future.

However, this limited test does suggest faster growth for the deep Web. Both median and average deep Web sites are four or five months "younger" than surface Web sites (Mar. 95 v. Aug. 95). This is not surprising. The Internet has become the preferred medium for public dissemination of records and information, and more and more information disseminators (such as government agencies and major research projects) that have enough content to qualify as deep Web are moving their information online. Moreover, the technology for delivering deep Web sites has been around for a shorter period of time.

Thousands of Conventional Search Engines Remain Undiscovered
While we have specifically defined the deep Web to exclude search engines (see next section), many specialized search engines, such as those shown in Table 8 below or @griculture.com, AgriSurf, or joefarmer in the agriculture domain, provide unique content not readily indexed by major engines such as AltaVista, Fast or Northern Light. The key reasons that specialty search engines may contain information not on the major ones are indexing frequency and limitations the major search engines may impose on documents indexed per site.(11b)

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

.	.	.	.	.	Search Engine A		.
Search Engine A	**A no dupes**	**Search Engine B**	**B no dupes**	**A plus B**	**Unique**	**Search Engine Fraction**	**Search Engine Size**	**Est. # of Search Engines**
FinderSeeker	2,012	SEG	1,268	233	1,779	0.184	2,012	10,949
FinderSeeker	2,012	Netherlands	1,170	167	1,845	0.143	2,012	14,096
FinderSeeker	2,012	LincOne	783	129	1,883	0.165	2,012	12,212
SearchEngineGuide	1,268	FinderSeeker	2,012	233	1,035	0.116	1,268	10,949
SearchEngineGuide	1,268	Netherlands	1,170	160	1,108	0.137	1,268	9,272
SearchEngineGuide	1,268	LincOne	783	28	1,240	0.036	1,268	35,459
Netherlands	1,170	FinderSeeker	2,012	167	1,003	0.083	1,170	14,096
Netherlands	1,170	SEG	1,268	160	1,010	0.126	1,170	9,272
Netherlands	1,170	LincOne	783	44	1,126	0.056	1,170	20,821
LincOne	783	FinderSeeker	2,012	129	654	0.064	783	12,212
LincOne	783	SEG	1,268	28	755	0.022	783	35,459
LincOne	783	Netherlands	1,170	44	739	0.038	783	20,821

Table 8. Estimated Number of Surface Site Search Engines

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

M. Hofstede, of the Leiden University Library in the Netherlands, reports that one compilation alone contains nearly 45,000 search site listings.(46) Thus, our best current estimate is that deep Web searchable databases and search engines have a combined total of 250,000 sites. Whatever the actual 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 includes specialty search engines in its listings.

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

Original Deep Content Now Exceeds All Printed Global Content
International Data Corporation predicts that the number of surface Web documents will grow from the current two billion or so to 13 billion within three years, a factor increase of 6.5 times;(47) deep Web growth should exceed this rate, perhaps increasing about nine-fold over the same period. Figure 8 compares this growth with trends in the cumulative global content of print information drawn from a recent UC Berkeley study.(48a)

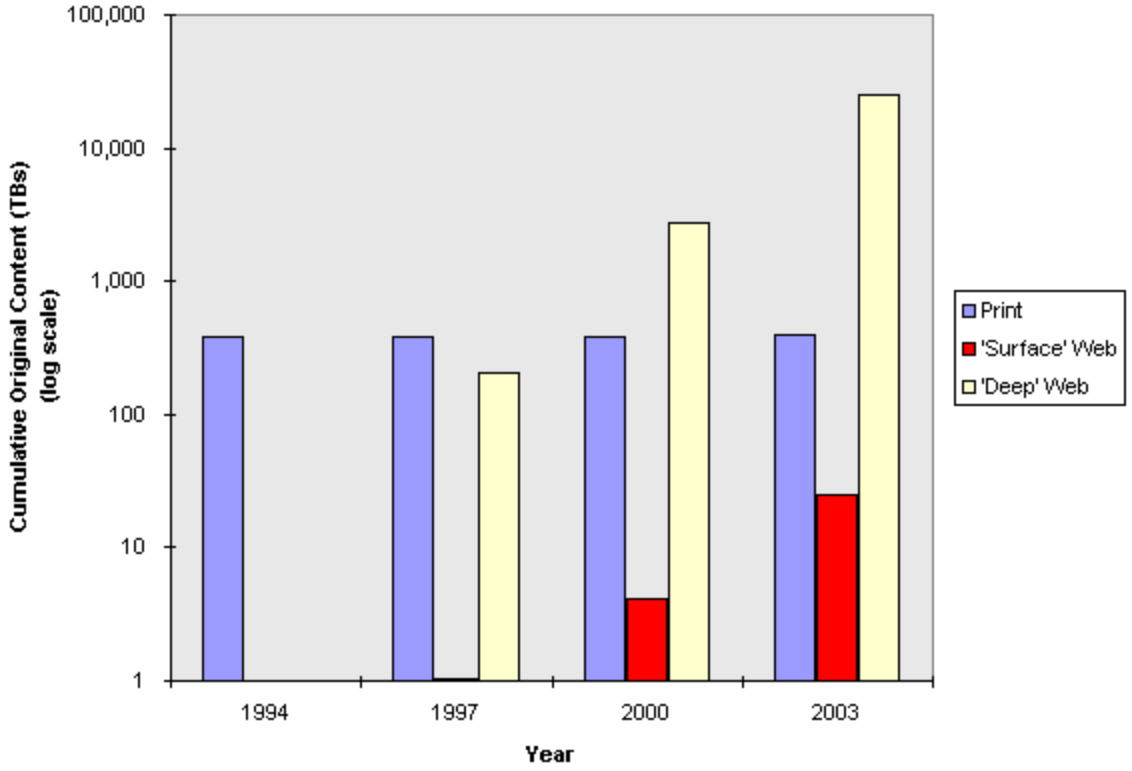

Figure 8. 10-yr Growth Trends in Cumulative Original Information Content (log scale)

The total volume of printed works (books, journals, newspapers, newsletters, office documents) has 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 deep Web 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 and will continue to dominate it. (49) Even today, at least 240 major libraries have their catalogs on line;(50) UMI, a former subsidiary of Bell & Howell, has plans to put more than 5.5 billion document images online;(51) and major astronomy data initiatives are moving toward putting petabytes of data online. (52)

These trends are being fueled by the phenomenal growth and cost reductions in digital, magnetic storage.(48c) (53) International Data Corporation estimates that the amount of disk storage capacity sold annually grew from 10,000 terabytes in 1994 to 116,000 terabytes in 1998, and it is expected to increase to 1,400,000 terabytes in 2002.(54) Deep Web content accounted for about 1/338th of magnetic storage devoted to original content in 2000; it is projected to increase to 1/200th by 2003. As the 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, when "surface" content can appear to be deep.

Surface Web content is persistent on static pages discoverable by search engines through crawling, while deep Web content is only presented dynamically in response to a direct request. However, once directly requested, deep Web content comes associated with a URL, most often containing the database 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. 10Kwizard provides full-text searching of SEC corporate filings (55). We issued a query on "NCAA 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 full-text portions for the query string in that annual filing. With another click, the full filing text can also be viewed. The URL resulting from this direct request is:

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

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

Now, if we were doing comprehensive research on this company and posting these results on our own Web page, other users could click on this URL and get the same information. Importantly, if we had posted this URL on a static Web page, search engine crawlers could also discover it, use the same URL as shown above, and then index the contents.

It is by doing searches and making the resulting URLs available that deep content can be brought to the surface. Any deep content listed on a static Web page is discoverable by crawlers and therefore indexable by search engines. As the next section describes, it is impossible to completely "scrub" large 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 sites. For instance, the Open Directory Project, is an effort to organize the best of surface Web content using voluntary editors or "guides." (56) The Open Directory looks something like Yahoo!; that is, it is a tree structure with directory URL results at each branch. The results pages are static, laid out like disk 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 is a static 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 specified based on URL. The query "URL:dmoz.org" (the address for the Open Directory site) was posed to these engines with these results:

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%

Table 9. Incomplete Indexing of Surface Web Sites

Although there are almost 250,000 subject pages at the Open Directory site, only a tiny percentage are recognized by the major search engines. Clearly the engines' search algorithms have rules about either depth or breadth of surface pages indexed for a given site. We also found a broad variation in the timeliness of results from these engines. Specialized surface sources or engines should therefore be considered when truly deep searching is desired. That bright line between deep and surface Web shows is really shades of gray.

The Impossibility of Complete Indexing of Deep Web Content
Consider how a directed query works: specific requests need to be posed against the searchable database by stringing together individual query terms (and perhaps other filters such as date restrictions). If you do not ask the database specifically what you want, you will not get it.

Let us take, for example, our own listing of 38,000 deep Web sites. Within this compilation, we have some 430,000 unique terms and a total of 21,000,000 terms. If these numbers represented the contents of a searchable database, then we would have to issue 430,000 individual queries to ensure we had comprehensively "scrubbed" or obtained all records within the source database. Our database is small compared to some large deep Web databases. For example, one of the largest collections of text terms is the British National Corpus containing more than 100 million unique terms.(59)

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

Possible Double Counting
Web content is distributed and, once posted, "public" to any source that chooses to replicate it. How much of deep Web content is unique, and how much is duplicated? And, are there differences in duplicated content between the deep and surface Web?

"Surface Web sites are fraught with quality problems"

This study was not able to resolve these questions. Indeed, it is not known today how much duplication occurs within the surface Web.

Observations from working with the deep Web sources and data suggest there are important information categories where duplication does exist. Prominent among these are yellow/white pages, genealogical records, and public records with commercial potential such as SEC filings. 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 uniquely valuable. These mostly fall within the categories of topical databases, publications, and internal site indices -- accounting in total for about 80% of deep Web sites -- and include such sources as scientific databases, library holdings, unique bibliographies such as PubMed, and unique government data repositories such as satellite imaging data and the like.

But duplication is also rampant on the surface Web. Many sites are "mirrored." Popular documents are frequently appropriated by others and posted on their own sites. Common

information such as book and product listings, software, press releases, and so forth may turn up multiple times on 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 discovery. The deep Web is not as easily discovered, and while mostly public, not as easily copied by other surface Web sites. These factors suggest that duplication may be lower within the deep Web. But, 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 not a long list of hits, but the right list. Searchers want answers. Providing those answers has always been a problem for the surface Web, and without appropriate technology will be a problem for the deep Web as well.

Effective searches should both identify the relevant information desired and present it in order of potential relevance -- quality. Sometimes what is most important is comprehensive discovery -- everything referring to a commercial product, for instance. Other times the most authoritative result is needed -- the complete description of a chemical compound, as an example. The searches may be the same for the two sets of requirements, but the answers will have to be different. Meeting those requirements is daunting, and knowing that the deep Web exists only complicates the solution because it often contains useful information for either kind of search. If useful information is obtainable but 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) defining quality as both actual quality of the search results and the ability to cover the subject.

Search Type	Total Docs (million)	Quality Docs (million)
Surface Web		
Single Site Search	160	7
Metasite Search	840	38
TOTAL SURFACE POSSIBLE	1,000	45
Deep Web		
Mega Deep Search	110,000	14,850
TOTAL DEEP POSSIBLE	550,000	74,250
Deep v. Surface Web Improvement Ratio		
Single Site Search	688:1	2,063:1
Metasite Search	131:1	393:1
TOTAL POSSIBLE	655:1	2,094:1

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

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

Surface Web sites are fraught with quality problems. For example, a study in 1999 indicated that 44% of 1998 Web sites were no longer available in 1999 and that 45% of existing sites were half-finished, meaningless, or trivial.(61) Lawrence and Giles' NEC studies suggest that individual major search engine 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 provide the authority necessary for users to judge the quality of information. The Internet lacks such authority.

An intriguing possibility with the deep Web is that individual sites can themselves establish that authority. For example, an archived publication listing from a peer-reviewed journal such as *Nature* or *Science* or user-accepted sources such as the *Wall Street Journal* or *The Economist* carry with them authority based on their editorial and content efforts. The owner of the site vets what content is made available. Professional content suppliers typically have the kinds of database-based sites that make up 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 searchable sites, users can make their own determinations about quality, even though a solid metric for that value 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. Searching must evolve to encompass the complete Web.

Directed query technology is the only means to integrate deep and surface Web information. The information retrieval answer has to involve both "mega" searching of appropriate deep Web sites and "meta" searching of surface Web search engines to overcome their coverage problem. Client-side tools are not universally acceptable because of the need to download the tool and issue effective queries to it.(62) Pre-assembled storehouses for selected content are also possible, but will not be satisfactory for all information requests and needs. Specific vertical market services are already evolving to partially address these challenges.(63) These will likely need to be supplemented with a persistent query system customizable by the user that would set the queries, search sites, filters, and schedules for repeated queries.

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

* * *

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

Endnotes

1. Data for the study were collected between March 13 and 30, 2000. The study was originally published on BrightPlanet's Web site on July 26, 2000. (See http://www.completeplanet.com/Tutorials/DeepWeb/index.asp.)Some of the references and Web status statistics were updated on October 23, 2000, with further minor additions on February 22, 2001.

2.A couple of good starting references on various Internet protocols can be found at 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, May 14, 1999. See http://www.gvu.gatech.edu/user_surveys/survey-1998-10/tenthreport.html.

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/.

5a, 5b, 5c, 5d, 5e, 5f, 5g. "Sizing the Internet, Cyveillance, http://www.cyveillance.com/web/us/downloads/Sizing_the_Internet.pdf.

6a, 6b. S. Lawrence and C.L. Giles, "Searching the World Wide Web," *Science* 80:98-100, April 3, 1998.

7a, 7b. S. Lawrence and C.L. Giles, "Accessibility of Information on the Web," Nature 400:107-109, 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 be issued to get an actual document count from its data stores. See http://www.northernlight.com NL searches used in this article exclude its "Special Collections" listing.

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

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

13a, 13b. This analysis assumes there were 1 million documents on the Web as of mid-1994.

14. See http://www.tcp.ca/Jan96/BusandMark.html.

15. See, for example, G Notess, "Searching the Hidden Internet," in Database, June 1997 (http://www.onlineinc.com/database/JunDB97/nets6.html).

16. Empirical BrightPlanet results from processing millions of documents provide an actual mean value of 43.5% for HTML and related content. Using a different metric, NEC researchers found HTML and related content with white space removed to account for 61% of total page content (see 7). Both measures ignore images and so-called HTML header content.

17. Rough estimate based on 700 million total documents indexed by AltaVista, Fast, and Northern Light, at an average document size of 18.7 KB (see reference 7) and a 50% combined representation by 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 disk storage space 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 reported sizes when an HTML document is saved to disk from a browser. The 1999 NEC study also reported average Web document size after removal of all HTML tag information and white space to be 7.3 KB. While a more accurate view of "true" document content, we have used the HTML basis because of the equivalency in reported results from search engines themselves, browser document saving and our technology.

21. Inktomi Corp., "Web Surpasses One Billion Documents," press release issued January 18, 2000; see http://www.inktomi.com/new/press/2000/billion.html and http://www.inktomi.com/webmap/

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

23. The compilation sites used for initial harvest were:

- AlphaSearch -- http://www.calvin.edu/library/searreso/internet/as/
- Direct Search -- http://gwis2.circ.gwu.edu/~gprice/direct.htm
- Infomine Multiple Database Search -- http://infomine.ucr.edu/
- The BigHub (formerly Internet Sleuth) -- http://www.thebighub.com/
- Lycos Searchable Databases -- http://dir.lycos.com/Reference/Searchable_Databases/
- Internets (Search Engines and News) -- 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 Public Web Search Engines," paper presented at the Seventh International World Wide Web Conference, Brisbane, Australia, April 14-18, 1998. The full paper is available at http://www7.scu.edu.au/programme/fullpapers/1937/com1937.htm.

25a, 25b. See, for example, http://www.surveysystem.com/sscalc.htm, for a sample size calculator.

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

27. See reference 38. Known pageviews for the logarithmic popularity rankings of selected sites tracked by Alexa are used to fit a growth function for estimating monthly pageviews based on the Alexa 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 randomized to remove (partially) ranking prejudice in the order Northern Light lists results.

30. These three engines were selected because of their large size and support for full Boolean queries.

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

32. The BrightPlanet technology configuration settings were: max. Web page size, 1 MB; min. page size, 1 KB; no date range filters; no site filters; 10 threads; 3 retries allowed; 60 sec. Web page timeout; 180 minute max. download time; 200 pages per engine.

33. The vector space model, or VSM, is a statistical model that represents documents and queries 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 Organization and 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 vectors; see further Salton, Gerard, Fox, Edward A. and Wu, Harry, (Cornell Technical Report TR82-511) 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 http://www.invisibleweb.com.

36. K. Wiseman, "The Invisible Web for Educators," see http://www3.dist214.k12.il.us/invisible/article/invisiblearticle.html

37. C. Sherman, "The Invisible Web," http://websearch.about.com/library/weekly/aa061199.htm

38.I. Zachery, "Beyond Search Engines," presented at the Computers in Libraries 2000 Conference, March 15-17, 2000, Washington, DC; see 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 deep Web search sites. Subsequent customer projects have allowed us to update this analysis, again using overlap analysis, to 200,000 sites. This site number is updated in this paper, but overall deep Web size estimates have not. In fact, still more recent work with foreign language deep Web sites strongly 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," 1999; see http://www.parc.xerox.com/istl/groups/iea/www/growth.html

42. The Northern Light total deep Web sites count is based on issuing the query "search OR database" to the engine restricted to Web documents only, and then picking its Custom Folder on Web search engines and directories, producing the 27,195 count listing shown. Hand inspection of the first 100 results yielded only three true searchable databases; this increased in the second 100 to 7. Many of these initial sites were for standard search engines or Web site promotion services. We believe the yield of actual search sites would continue to increase with depth through the results. We also believe 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 their effect on accuracy.

43. 1024 bytes = I 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 = 1,024,000,000,000,000 bytes or 10^{15}.

44a, 44b. Our original paper published on July 26, 2000, use d estimates of one billion surface Web documents and about 100,000 deep Web sea rchable databases. Since publication, new information suggests a total of about 200,000 deep Web searchable databases. Since surface Web document growth is no w on the order of 2 billion documents, the ratios of surface to Web documents (400 to 550 times greater in the deep Web) still approximately holds. These tren ds would also suggest roughly double the amount of deep Web data storage to fifteen petabytes than is indicated in the main body of the report.

45. We have not empirically tested this assertion in this study. However, from a logical standpoint, surface search engines are all indexing ultimately the same content, namely the public indexable Web. Deep Web sites reflect information from different domains and producers.

46. M. Hofstede, pers. comm., Aug. 3. 2000, referencing http://www.alba36.com/.

47. As reported in Sequoia Software's IPO filing to the SEC, March 23, 2000; see http://www.10kwizard.com/filing.php?repo=tenk&ipage=1117423&doc=1&total=266&back=2&g=.

48a, 48b, 48c. P. Lyman and H.R. Varian, "How Much Information," published by the UC Berkeley School of Information Management and Systems, October 18. 2000. See http://www.sims.berkeley.edu/research/projects/how-much-info/index.html. The comparisons here are limited to archivable and retrievable public information, exclusive of entertainment and communications content such as chat or e-mail.

49. As this analysis has shown, in numerical terms the deep Web already dominates. However, from a general user perspective, it is unknown.

50. See http://lcweb.loc.gov/z3950/.

51. See http://www.infotoday.com/newsbreaks/nb0713-3.htm.

52. A. Hall, "Drowning in Data," Scientific American, Oct. 1999; see http://www.sciam.com/explorations/1999/100499data/.

51. As reported in Sequoia Software's IPO filing to the SEC, March 23, 2000 ; see http://www.10kwizard.com/filing.php?repo=tenk&ipage=1117423&doc=1&total=266&back=2&g=.

54. From Advanced Digital Information Corp., Sept. 1, 1999, SEC filing; see http://www.tenkwizard.com/fil_blurb.asp?iacc=991114&exp=terabytes%20and%20online&g=.

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

56. Though the Open Directory is licensed to many sites, including prominently Lycos and Netscape, it maintains its own site at http://dmoz.org. An example of a node reference for a static page that could be indexed by a search engine is: 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 through a static 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 categories.

58. See previous reference. This number of categories may seem large, but is actually easily achievable, 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 node choices, on average about 30 second-level node choices, etc. Not all parts of these subject trees are as complete or "bushy" as other ones, and some branches of the tree extend deeper because there is a richer amount of content to organize. Nonetheless, through this simple progression of subject choices at each node, one can see how total subject categories - and the static pages associated with them for presenting result - can grow quite large. Thus, for a five-level structure with an average number or node choices at each level, Open Directory could have ((15 * 30 * 15 * 12 * 3) + 15 + 30 + 15 + 12) choices, or a total of 243,072 nodes. This is close to the 248,000 nodes actually reported by the site.

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

60. Assumptions: SURFACE WEB: for single surface site searches - 16% coverage; for metasearch surface searchers - 84% coverage [higher than NEC estimates in reference 4; based on empirical 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; 200,000 potential deep Web sources; 13.5% quality retrieval from all deep Web searches.

61. Online Computer Library Center, Inc., "June 1999 Web Statistics," Web Characterization Project, 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 Boolean constructs or queries. Also, most studies suggest users issue on average 1.5 keywords per query; even professional information scientists issue 2 or 3 keywords per search. See further BrightPlanet's search tutorial at http://www.completeplanet.com/searchresources/tutorial.htm.

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

Some of the information in this document is preliminary. BrightPlanet plans future revisions as better information and documentation is obtained. We welcome submission of improved information and statistics from others involved with the Deep Web.

Links from this article:

10Kwizard 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 http://www.bluestone.com/

Excite http://www.excite.com

Google http://www.google.com/

joefarmer 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/