<u>White Paper</u>

The Deep Web:
Surfacing Hidden Value

BrightPlanet.com LLC

July 2000

The author of this study is Michael K. Bergman. Editorial assistance was provided by Mark Smither; analysis and retrieval assistance was provided by Will Bushee.

This White Paper is the property of BrightPlanet.com LLC. Users are free to distribute and use it for personal use..

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.

Summary

BrightPlanet has uncovered the "deep" Web — a vast reservoir of Internet content that is 500 times larger than the known "surface" World Wide Web. What makes the discovery of the deep Web so significant is the quality of content found within. There are literally hundreds of billions of highly valuable documents hidden in searchable databases that cannot be retrieved by conventional search engines.

This discovery is the result of groundbreaking search technology developed by **BrightPlanet** called a LexiBot™ — the first and only search technology capable of identifying, retrieving, qualifying, classifying and organizing "deep" and "surface" content from the World Wide Web. The LexiBot allows searchers to dive deep and explore hidden data from multiple sources simultaneously using directed queries. Businesses, researchers and consumers now have access to the most valuable and hard-to-find information on the Web and can retrieve it with pinpoint accuracy.

Searching on the Internet today can be compared to dragging a net across the surface of the ocean. There is a wealth of information that is deep, and therefore missed. The reason is simple: basic search methodology and technology have not evolved significantly since the inception of the Internet.

Traditional search engines create their card catalogs by spidering or crawling "surface" Web pages. To be discovered, the page must be static and linked to other pages. Traditional search engines cannot "see" or retrieve content in the deep Web. Because traditional search engine crawlers can not probe beneath the surface the deep Web has heretofore been hidden in plain sight.

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. The LexiBot automates the process of making dozens of direct queries simultaneously using multiple thread technology.

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 completed the first documented study to quantify the size and relevancy of the deep Web. Our key findings from this study include the following:

- 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 19 terabytes of information in the surface Web

- The deep Web contains nearly 550 billion individual documents compared to the 1 billion of the surface Web

- More than an estimated 100,000 deep Web sites presently exist

- 60 of the largest deep Web sites collectively contain about 750 terabytes of information — sufficient by themselves to exceed the size of the surface Web by 40 times

- On average, deep Web sites receive about 50% 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 search 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 at least 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 95% of the deep Web is publicly accessible information — not subject to fees or subscriptions.

To put these numbers in perspective, an NEC study published in *Nature* estimated that the largest search engines such as Northern Light individually index at most 16% of the surface Web. Since they are missing the deep Web, Internet searchers are therefore searching only 0.03% — or one in 3,000 — of the content available to them today. Clearly, simultaneous searching of multiple surface and deep Web sources is necessary when comprehensive information retrieval is needed.

The **BrightPlanet** team has automated the identification of deep Web sites and the retrieval process for simultaneous searches. We have also developed a direct-access query engine translatable to about 20,000 sites, already collected, eventually growing to 100,000 sites. A listing of these sites may be found at our comprehensive search engine and searchable database portal, CompletePlanet (see http://www.completeplanet.com).

Table of Contents

List of Figures and Tables

I. Introduction

Internet content is considerably more diverse and certainly much larger than what is commonly understood. Firstly, 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), email, news, Telnet and Gopher (most prominent among pre-Web protocols). This paper does not consider further these non-Web protocols.[1‡]

Secondly, even within the strict context of the Web, most users are only aware of the content presented to them via search engines such as Excite, Google, AltaVista, Snap 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.[2] According to a recent NPD survey of search engine satisfaction, search failure rates have increased steadily since 1997.[3]

The importance of information gathering on the Web and the central and unquestioned role of search engines — plus the frustrations expressed by users in 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 1600's, people had no idea there was a whole world of "animalcules" beyond their vision. Deep-sea exploration has discovered hundreds of strange creatures in the past 30 years 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 this 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 for listing, generally acknowledged to be a minor contributor to total listings. Or, search engines "crawl" or "spider" documents by following one hypertext link to another. Simply stated, when indexing a given document or page, if the crawler encounters a hypertext link on that page to another document, it records that incidence and schedules that new page for later crawling. Like ripples propagating across a pond, in this manner search engine crawlers are able to extend their indexes further and further from their starting points.

The surface Web contains an estimated 1 billion documents and is growing at the rate of 1.5 million documents per day.[18] 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.[4,5] Today, the two largest search engines in terms of internally reported documents indexed are the Fast engine with 300 million documents listed [6] and Northern Light with 218 million documents.[7]

[‡] All document references and notes are shown at the conclusion under Endnotes and References.

Legitimate criticism has been leveled against search engines for these indiscriminate crawls, mostly because of providing way too many results (search on "web," for example, with Northern Light, and you will get about 30 million results!). Also, because new documents are found from links of older documents, documents with a larger number of "references" have up to an eight-fold improvement of being indexed by a search engine than a document that is new or with few cross-references.[5]

To overcome these limitations, the most recent generation of search engines, notably Google and the recently acquired Direct Hit, 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.[5]

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 the posting of new information and its recording on conventional search engines.[8] 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. It is this fundamental aspect of how search engine crawlers work that discloses their basic flaw in today's information discovery on the Web.

Figure 1 indicates that searching the Web today using search engines is like dragging a net across the surface of the ocean. 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 in plain sight beneath the surface of the Web.
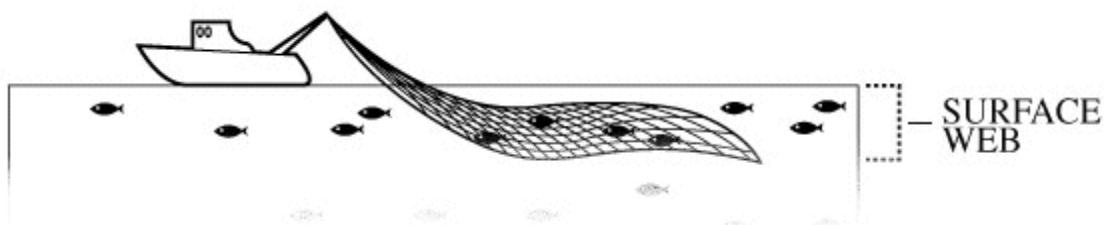

Figure 1. Search Engines: Dragging a Net Across the Web's Surface

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 results were persistent and constantly available, they could easily be crawled by conventional search engines. For example, in July 1994, Lycos went public with a catalog of only 54,000 documents;[9] yet, today, with estimates at 1 billion documents,[18] the compound growth rate in Web documents has been on the order of more than 200% annually![10]

Sites which were required to manage tens to hundreds of documents could easily do so by posting all pages within a static directory structure as fixed HTML pages. However, beginning about 1996, three phenomena took place. First, database technology was introduced to the Internet through such vendors as Bluestone's Sapphire/Web and later Oracle and others. 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 Census Bureau, Securities and Exchange Commission and Patents and Trademarks 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 can not be seen can not be defined, and what is not defined can not 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 using conventional search engines — no longer applies to the information content of the Internet.

As early as 1994, Dr. Jill Ellsworth first coined the phrase "invisible Web" to refer to information content that was "invisible" to conventional search engines.[11] 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.[12] 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 or queryable by conventional search engines. Using our technology, they are totally "visible" to those that need to access them.

Thus, the real problem is not the "visibility" or "invisibility" of the Web, but the spidering technologies used by conventional search engines to collect their content. What is required is not Superman with x-ray vision, but different technology to make these sources apparent. For these reasons, we have chosen to call information in searchable databases the "deep" Web. Yes, it is somewhat hidden, but clearly available if different technology is employed to access it.

The deep Web is qualitatively different from the "surface" Web. Deep Web content resides in searchable databases, the results from which can only be discovered by a direct query. Without the directed query, the database does not publish the result. Thus, while the content is there, it is skipped over when the traditional search engine crawlers can't probe beneath the surface.

This concept can be shown as a different harvesting technique from search engines, as shown in Figure 2. By first using "fish finders" to identify 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.
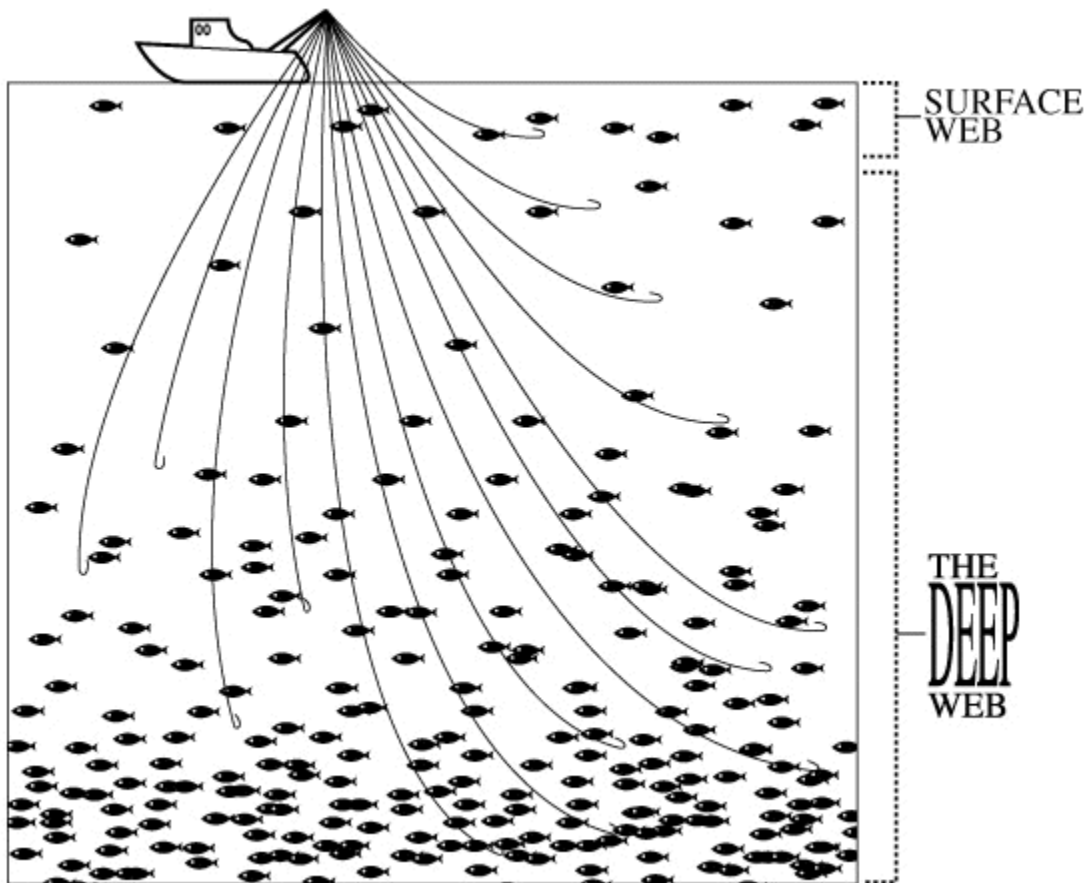
Figure 2. Harvesting the Deep and Surface Web with a Directed Query Engine

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 LexiBot technology is uniquely suited to tap the deep Web and bring its results to the surface. The simplest way to describe the LexiBot is a "directed query engine." The LexiBot 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.

Of course, search engines are themselves searchable databases. Therefore, surface Web results are easily integrated with LexiBot deep Web searches. By definition, however, search engines are limited to surface Web documents that can be discovered by crawling. We maintain the distinction in this paper between deep Web searchable databases and surface Web search engines.

Study Objectives

The objectives of this study are thus to:

1. Quantify the size and importance of the deep Web
2. Characterize the deep Web's content, quality and relevance to information seekers
3. Discover automated means for identifying deep Web search sites and directing queries to them, and
4. Begin the process of educating the Internet search public for this heretofore hidden and valuable information storehouse.

As with any newly-discovered phenomena, we are at the beginning steps in defining and understanding the deep Web. Daily, as we have continued our investigations, we have constantly 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

We already noted 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 by definition 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 Javascripts),[13] this study does not evaluate specific information within this code. We do, however, include these codes in our quantification of total content (see next section).

Finally, _none_ of the estimates for the size of the deep Web herein include either specialized search engine sources — which may be partially "hidden" to the major traditional search engines (see p. 24) — 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, [14] or somewhat larger than the known size of the surface Web.

II. Methods

This section describes the survey and evaluation methods used to quantify the size of the deep Web and to characterize its contents. Data for the study were collected between March 13 and 30, 2000.

A Common Denominator for Size Comparisons

All deep and surface Web size figures herein 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 (see further [13]). 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 LexiBot reports document sizes on this same basis.

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

In actuality, data storage from deep Web documents will therefore be considerably less than the figures reported herein.[15] 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 the LexiBot

All retrievals, aggregations and document characterizations in this study used **BrightPlanet**'s LexiBot technology. The LexiBot uses multiple threads for simultaneous source queries and then document downloads. The LexiBot completely indexes all documents retrieved (including HTML content). After download and indexing, the documents are scored as to relevance using four different scoring algorithms, prominently vector space modeling (VSM) and standard and modified extended Boolean information retrieval (EBIR).[16]

Automated deep Web search site identification and qualification also used a modified version of the LexiBot 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.[4] An update to their study employing a different methodology was published in *Nature* magazine in 1999, using analysis from February 1999.[5] This study documented 800 million documents within the publicly indexable Web, with a mean page size of 18.7 kilobytes (KBs) exclusive of images and HTTP headers.[17]

In partnership with Inktomi, NEC updated its Web page estimates to 1 billion documents in early 2000.[18] We've 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 resulting baseline figures used for the size of the surface Web in this paper.

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 8-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 60 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, and, then,
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, in descending order of importance and confidence in deriving the total document count:

1. Emails 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 standalone 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 [19]
5. A site which failed all of these tests could not be characterized as to size, and was dropped from the results listing.

The resulting top 60 sites in Table 2 resulted from the detailed investigation of more than 100 total sites.

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 pageviews 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.[20] After harvesting, this pool resulted in 45,732 actual unique listings after proprietary tests for duplicates. Cursory inspection indicated that in some cases the subject page was one link removed from the actual search form. Proprietary criteria were developed to determine when this might be the case. The LexiBot was used to retrieve the complete pages and fully index them for both the initial unique sources and the one-link removed sources. Some 43,348 resulting URLs were actually retrieved.

We applied an initial 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 filter resulted in 17,579 pre-qualified URLs.

Subsequent hand inspection of 700 randomized 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.

The testing of hand-qualified sites has resulted in an automated test within the LexiBot for qualifying search sites with 98% accuracy. Additionally, automated means for discovering further search sites has been incorporated into our internal version of the LexiBot based upon this learning.

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.[4,21] We used overlap analysis based on search engine coverage and between the deep Web compilation sites noted above (see results in Table 3 through Table 5).

The technique is illustrated in the diagram below:
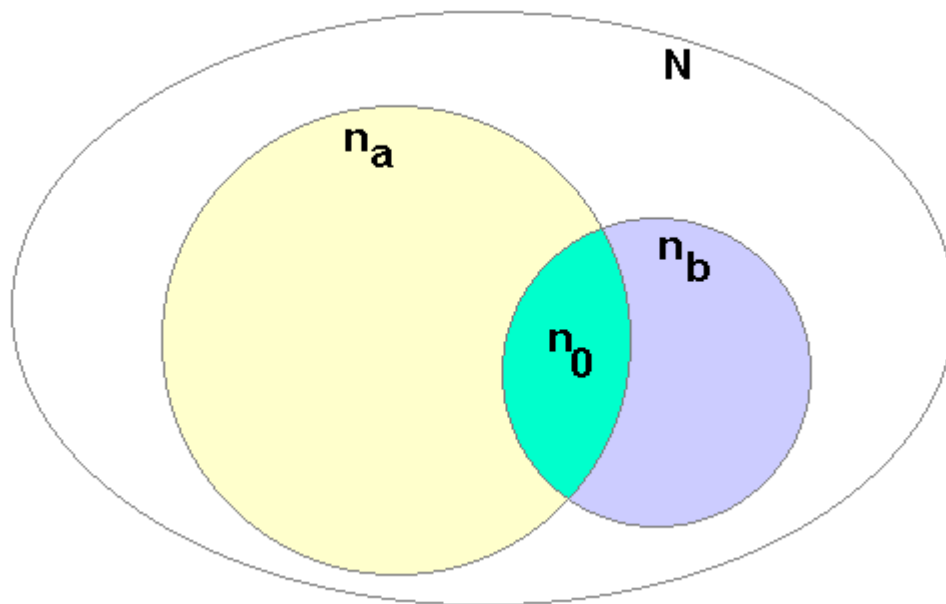

Figure 3. Schematic Representation of "Overlap" Analysis

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. For compilation sites, which have been purposeful in collecting their listings, their sampling has been 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.[5] Thus, the overlap analysis herein 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.[22]

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:

1. Total record counts as reported by the site itself. This involved inspecting relating pages on the site including help sections, site FAQs, etc.
2. 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 standalone 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

3. A site which failed all of these tests could not be characterized as to size, and was dropped from full site characterization.

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 100 sites which 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 out 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 12 arbitrary categories that captured the diversity of database types. These twelve categories are:

- Topic Databases — subject-specific aggregations of information, such as SEC corporate filings, medical databases, patent records, etc.
- Internal site — searchable databases for the internal pages of large sites that are dynamically created, such as the knowledge base on the Microsoft site
- Publications — searchable databases for current and archived articles
- Shopping/Auction
- Classifieds
- Portals — these were broader sites that included more than one of these other categories in searchable databases
- Library — searchable internal holdings, mostly for university libraries
- Yellow and White Pages — people and business finders
- 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
- Jobs — job and resume postings
- Message or Chat
- 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, provided as a service from Alexa, provides site popularity rankings and link reference counts for a given URL.[23] About 71% of deep Web sites have such rankings. The universal power function allows pageviews per month to be extrapolated from the Alexa popularity rankings.[24] 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 proxy was necessary.

Whois [25] searches associated with domain registration services return records listing domain owner, plus the date the domain was first obtained (among other information). Using a random sample of 100 deep Web sites and another sample of 100 surface Web sites [26] 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 LexiBot to three search engines (AltaVista, Fast, Northern Light)[27] and three deep sites specific to that topic and included in the 600 presently configured for the LexiBot. 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.[28] The specific LexiBot configuration settings are documented in the endnotes.[29]

The "quality" determination was based on an average of the LexiBot's VSM and mEBIR computational linguistic scoring methods. The "quality" threshold was set at the LexiBot score of 82, empirically determined as roughly accurate from millions of previous LexiBot scores of surface Web documents.

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

III. 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 (see 'For Further Reading'). Estimates of size and importance have heretofore been anecdotal at best, and certainly underestimates as to 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.[30]

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."[31] A mid-1999 survey by About.com's Web search guide concluded the size of the deep Web was "big and getting bigger."[32] A paper at a recent library science meeting suggested that only "a relatively small fraction of the Web is accessible through search engines."[33]

When we began this study, we anticipated that the deep Web may have been on the order of the size of the known surface Web. The more we have probed, the more we have discovered as to the size, importance and accessibility of the deep Web.

The deep Web is about 500 times larger than the surface Web, with, on average, about three times higher quality 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 100,000 today, and is growing rapidly. 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 mean deep Web site has a Web-expressed (HTML included basis) database size of 74.4 megabytes (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 50% greater monthly traffic than 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; greater results require 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. Indeed, about one-third of the large sites listed in the next section are fee based.

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 60 known, largest deep Web sites contain data of about 750 terabytes (HTML included basis), or roughly 40 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 85 billion.

Roughly two-thirds of these sites are public ones, representing about 90% of the content available within this group of 60. 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 [34] or surface Web sites.[35] One implication of this type of distribution is that there is no real upper size bound 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

Name	Type	URL	Web Size (GBs)
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
Balckwell 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. Largest Known Top 60 Deep Web Sites

By nature, 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 a further three that were not in the initially identified pool of 100 potentially large sites. If that ratio were to hold across the entire estimated 100,000 deep Web sites (see next), 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 today to identify all of the largest deep Web sites 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 (see further 'Comments and Data Revisions Requested').

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:

Engine A	A no dups	Engine B	B no dups	A + B	Engine A			Tot Est Deep Web Sites
					Unique	DB Fract.	DB Size	
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, 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 [36]; 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 was to compare NEC reported values [5] 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 in Table 4.

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 same 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 in Section II.

Our third approach is more relevant. It 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 (see site identifications in [20]). To our knowledge, these are the three largest listings extant, excepting our own CompletePlanet site.

This approach has the advantages of: 1) providing an absolute count of sites; 2) ensuring final LexiBot qualification as to whether the sites are actually deep Web search sites; and 3) 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	DB A Unique	DB Fract.	DB Size	Tot Est 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 Section II discussed, there is certainly sampling bias in these compilations, since they were purposeful and not randomly obtained. Also, despite this, there is a surprising amount of uniqueness between the compilations.

The Lycos and Internets listings are more similar in focus in that they are commercial sites. The Infomine site has been 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 undercount 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 100,000.

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.
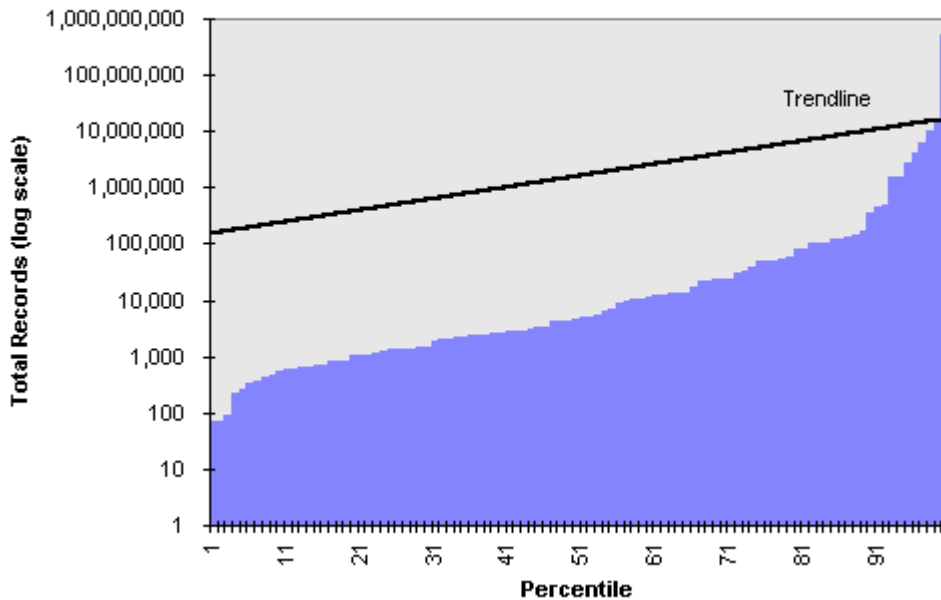

Figure 4. Inferred Distribution of Deep Web Sites, Total Record Size

Multiplying the mean size of 74.4 MB per deep Web site times a total of 100,000 deep Web sites results in a total deep Web size projection of 7.44 petabytes, or 7,440 terabytes.[37] 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 100,000 total deep Web sites results in a total record count across the deep Web of 543 billion documents. Compared to the Table 1 estimate of 1 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 100,000 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.
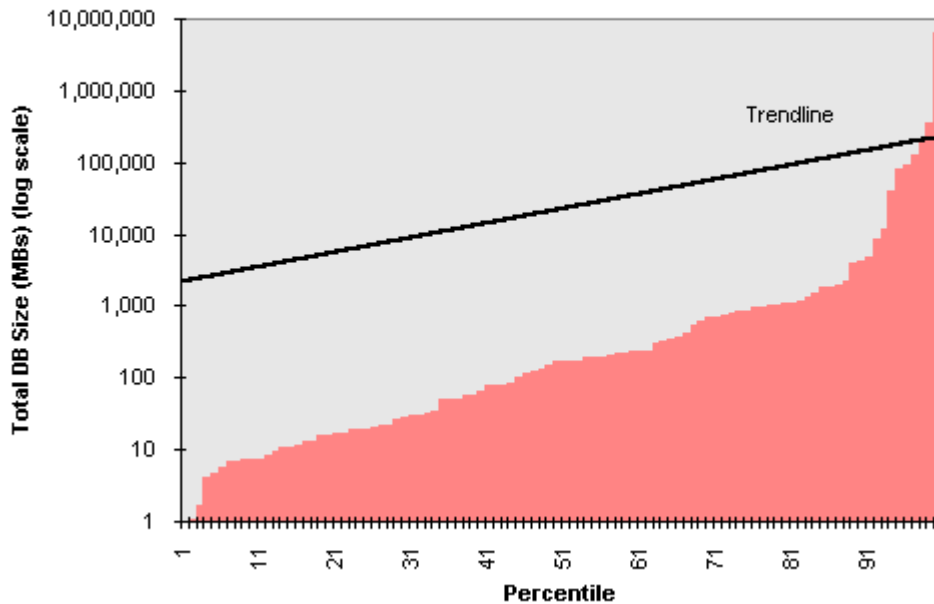
Figure 5. Inferred Distribution of Deep Web Sites, Total Database Size (MBs)

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 obviously 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.

Agriculture	2.7%	Employment	4.1%	Law/Politics	3.9%	References	4.5%
Arts	6.6%	Engineering	3.1%	Lifestyles	4.0%	Science, Math	4.0%
Business	5.9%	Government	3.9%	News, Media	12.2%	Shopping	3.2%
Computing/Web	6.9%	Health	5.5%	People, Companies	4.9%	Travel	3.4%
Education	4.3%	Humanities	13.5%	Recreation, Sports	3.5%		

Table 6. Distribution of Deep Sites by Subject Area

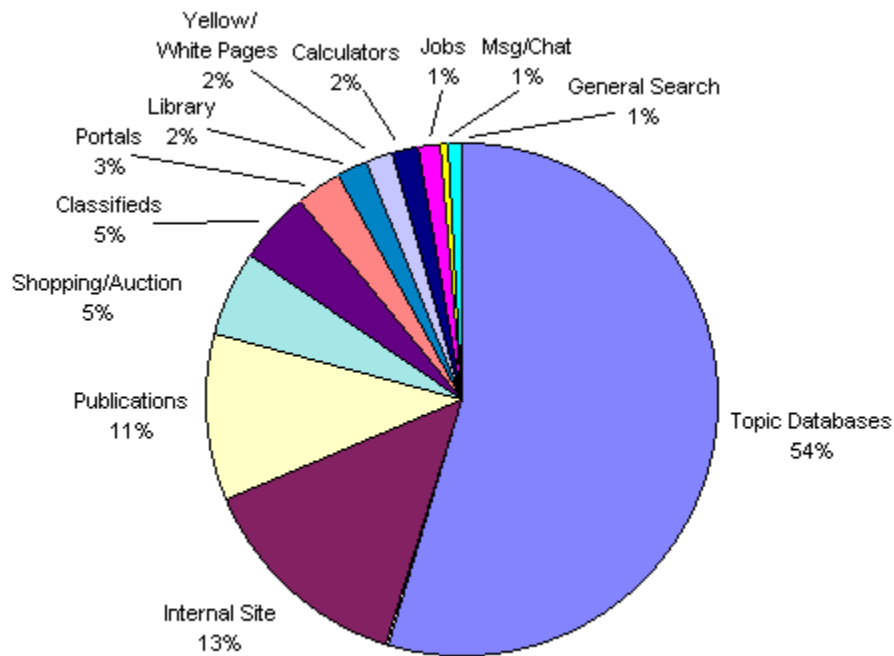Figure 6 displays the distribution of deep Web sites by type of content.

Figure 6. Distribution of Deep Web Sites by Content Type

More than half of all deep Web sites feature topical databases (see p. 11 for definitions). Nearly 80% of all deep Web sites include this category and large internal site documents and archived publications. Purchase transaction sites — including true shopping sites with auctions and classifieds — account for another 10% or so of sites. The other 8 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** pre-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 the actual "content" resides in course titles (*e.g.*, Agronomy Engineering 101) of little value. Various classes of these potential false positives exist and can be applied through learned business rules.

Our measurement of deep vs. surface Web quality did not apply these more sophisticated filters, instead relying on computational linguistic scores alone. We also posed a limited number of five queries across various subject domains.[28] Nonetheless, using only computational linguistic scoring does not introduce systematic bias in comparing deep and surface Web results. The relative differences between surface and deep Web should maintain, though the absolute values are preliminary and should 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, on average for the limited sample set, that there is about a three-fold improved likelihood for obtaining quality results from the deep Web as for the surface Web. 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.

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 absolutely 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). And, third, we believe the degree of content overlap between deep Web sites to be much less than for surface Web sites.[38]

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 prospects of yielding quality results that can not 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 is Growing Faster than the 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 in Figure 7 with superimposed growth trend lines.
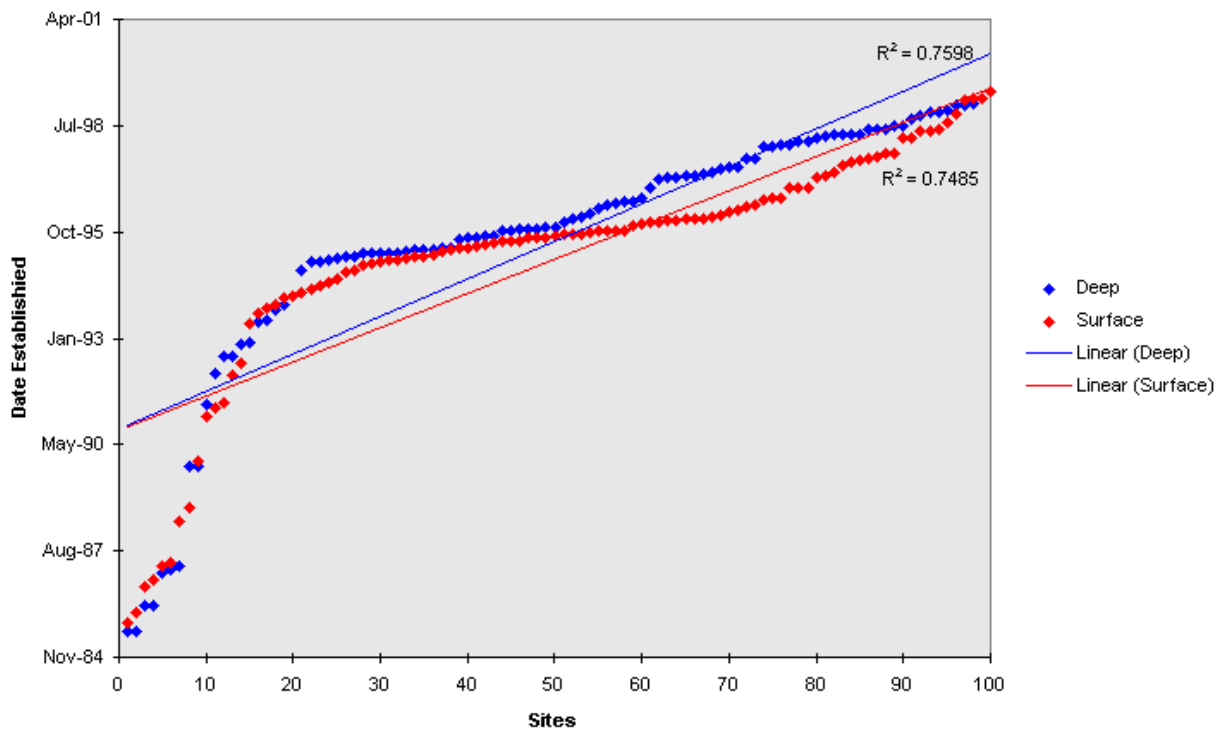
Figure 7. Comparative Deep and Surface Web Site Growth Rates

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 means to test for actual growth is site sampling over time and then trending results. **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 observation should not be surprising in light of the Internet becoming the preferred medium for public dissemination of records and information, and that many existing collectors and authors of such information such as government agencies and major research initiatives are increasingly moving their information online.

Thousands of Conventional Search Engines Remain Undiscovered

Finally, while we have specifically defined the deep Web to exclude search engines (see next section), many specialized search engines provide unique content not readily indexed by the major engines such as AltaVista, Fast or Northern Light. The key reasons that specialty search engines

may contain information not on the major ones is indexing frequency and rules the major engines may impose on documents indexed per site.[8]

Thus, using similar retrieval and qualification methods as for the deep Web, we conducted pairwise overlap analysis for four of the larger search engine compilation sites on the Web.[39] The results of this analysis are shown in the table below.

| SE A | A no dups | SE B | B no dups | A + B | Search Engine A | | | Est. No |
					Unique	SE Fract.	SE Size	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.) 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.

IV. Commentary

There is no "bright line" that separates various content sources on the Web. There are circumstances where "deep" content can appear on the surface, and, as for specialty search engines, when "surface" content can appear to be deep.

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

The Gray Zone Between the Deep and Surface Web

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, that provides full-text searching of SEC corporate filings.[40] We issued a query on 'NCAA basketball' with a restriction to only review annual filings filed between March 1999 and March 2000. Six results were produced; the first listing is for Broadcast.com, Inc. Clicking on that listing produces full text portions for where this query appeared in that Broadcast.com filing. (With another click, the full filing text can also be viewed.) The URL resulting from this direct request is:

http://www.10kwizard.com/fil_blurb.asp?iacc=899188&exp=ncaa%20basketball&g=

Note two things about this URL. First, our query terms appear in it. Second, the 'iacc=' shows a unique record number, in this case 899188. 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 through this manner 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 extends the other direction. In this example, let's take the case of the Open Directory Project, an effort to organize the best of surface Web content using voluntary editors or "guides."[41] The Open Directory is laid out similar to Yahoo!, that is, through a tree structure with directory URL results at each branch node. The results pages are static, basically being 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, each of which is a static page.[42] In fact, there are more than this number because of help files, etc. The key point, however, is that, technically, 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

Clearly, the engines themselves are imposing decision rules with respect to either depth or breadth of surface pages indexed for a given site. There was also broad variability in the timeliness of results from these engines. Specialized surface sources or engines should therefore be considered when truly deep searching is desired. Again, the "bright line" between deep and surface Web shows 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's take for example our own listing of 17,000 deep Web sites. Within this compilation, we have some 430,000 unique terms and a total of 21,000,000 "tokens." 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. Yet this is only a small indicator of the scope that might be contained within an individual, large source database. For example, one of the largest corpus of text terms we know is the British National Corpus containing more than 100 million unique terms,[43] and that is likely not the largest extant.

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?

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 indexes — 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,' if you will, 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 question of "quality" has been raised throughout this study. The reason for this prominence is that searchers do not want more, but want answers. This has been a long-standing problem for the surface Web and without appropriate technology will be a problem for the deep Web as well.

Effective search should both identify the relevant information desired and present it in order of potential relevance or quality. Sometimes, information needs are to find the "needle in a haystack." Other times, information needs are to find the "best" answer.

These are daunting requirements in a medium such as the Internet and not able to be solved simply from awareness that the deep Web exists. However, if useful information that is obtainable is excluded from the search, clearly both requirements are hindered.

Nonetheless, Table 10 attempts to bring together these disparate metrics into a single framework using the key conclusions from this paper.[44] The availability of "quality" information is a function both of actual "quality" and coverage of desired subject matter.

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

Analysis herein suggests that, at initial unveiling of **BrightPlanet**'s CompletePlanet site, that including deep Web sources can improve quality yield by a factor of more than 2,000 when searching against a single surface Web search engine; or 400-to-one when compared with a combined surface Web metasearch. By the time that CompletePlanet lists all deep Web sites, including deep Web sites could raise quality yield against the metasearched surface Web for a factor of nearly 2,100 to one.

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 1999 Web sites were no longer available in 1999 and that 45% of existing sites were half-finished, meaningless or trivial.[45] Lawrence and Giles' NEC studies suggest that individual major search engine coverage dropped from a maximum of 32% in 1998 to 16% in 1999.[5]

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, the subject of commentary by many pundits.

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. Further, because of the nature of deep Web sites, publication or authorship responsibility is clear. Deep Web sites that do not promote quality will be recognized as such. The anonymity of surface Web search listings is removed.

Directed queries to deep Web sources allow users to make this authoritative judgment on their own. Search engines, because of their indiscriminate harvesting, do not. By careful selection of searchable sites, users can make their own determinations as to quality, even though a solid metric for that value is difficult or impossible to assign independently.

Likelihood of Deep Web Growth

All signs point to the deep Web as the fastest growing component of the Web and eventually dominating it.[46]   For anecdotal examples, at least 240 other major libraries have their catalogs on line;[47] UMI, a former subsidiary of Bell & Howell, has plans to put more than 5.5 billion records of complete document images online;[48] and major astronomy data initiatives are moving toward putting petabytes of data online.[49]

It is, of course, impossible to know of the plans of major data producers with respect to publishing information on the Internet. A different measure for these prospects comes from International Data Corporation, an independent industry research company. IDC estimates that the amount of disk storage capacity sold annually grew from 10,000 terabytes in 1994 to 116,000 terabytes in 1998, and is expected to increase to 1,400,000 terabytes in 2002, representing an expected compound annual growth rate of 86% from 1998 to 2002.[50] The filing from which this was drawn states its belief that the growth rate is reflective of the overall growth in electronic data in the increasingly information-based economy.

The Bottom Line

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 has got to evolve to encompass the complete Web.

Directed query technology is the only means to integrate deep and surface Web information. But because deep information is only discoverable via directed query and can not be comprehensively indexed, it is unclear today whether a "global" solution can be found. Server-side implementations, such as search engines, which offer a central index lookup and quick response, do not appear easily implementable (or at all) for deep Web sources. The sources are numerous and diverse; the information needs are specific; and can only be obtained for the request at hand.

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. Today's technology suggests the only way to accomplish these twin imperatives within limited bandwidth is via a client-side tool under the direct control of each user. Pre-assembled storehouses for selected content according to the 80:20 rule 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.[51]

Client-side tools are not universally acceptable because of the need to download the tool and issue effective queries to it.[52] These observations suggest a splitting within the Internet information search market: search directories for the "quickest," hand-qualified surface content; search engines for more robust surface-level searches; and client-side tools for totally integrated search where comprehensiveness and quality are imperative.

A final comment: It is **BrightPlanet**'s desire that this study stimulates new observations and tools. We need to create greater awareness within the sophisticated Internet search public about the importance of the deep Web. Authors of deep Web sites need to be aware of the unique function they provide. Site authors need to publish the number of documents they contain and provide other quantified measures so as to allow the general search public to better understand their importance.

Comments and Data Revisions Requested

Early attempts to analyze the size of the known surface Web were fraught with difficulties (see especially discussion at the beginning of [4]) and today there still remains much uncertainty as to the known Web's total size and importance. By its nature, the Web is a chaotic, dynamic, rapidly-growing and distributed medium that does not easily lend itself to comprehension nor documentation. As a first-ever study, we certainly expect our analysis of the deep Web to suffer from the same limitations.

We hope that other researchers investigate the deep Web with alternative methodologies and data. **BrightPlanet** would be pleased to provide detailed backup for all of its analysis to qualified researchers desirous of improving our understanding of the deep Web.

We also welcome critical comments and suggestions from the interested search public. **BrightPlanet** is committed to updating and improving this study of the deep Web with subsequent revisions and enhancements.

We have established special sections of **BrightPlanet**'s CompletePlanet Web site for submission of new deep Web sites, corrections and updates to our largest sites listing, and moderated reviews of individual sites. We also are committed to rapidly grow our compilation of search sites — both deep and surface — to aid the search public in mastering information retrieval on the Web. Our commitment is for CompletePlanet to remain the "Source of all search."

Comments and submissions in these regards should be sent to: deepweb@brightplanet.com.

For Further Reading

For additional information about the deep Web, here is a sampling of some of the better third-party documents available. As starting points, we recommend:

- http://www3.dist214.k12.il.us/invisible/article/invisiblearticle.html
- http://websearch.about.com/internet/websearch/library/weekly/aa061199.htm

Further background information may be found at:

- http://www.internetworld.com/print/1997/02/03/industry/spiders.html
- http://www.searchenginewatch.com/sereport/99/07-invisible.html

And, some interesting introductions to growing scientific and publishing content coming on line can be found at:

- http://www.sciam.com/explorations/1999/100499data/ (large science databases)
- http://www.elsevier.com/homepage/about/resproj/trchp2.htm.

About BrightPlanet

*BrightPlanet.com LLC is an Internet content company with unique technologies for content discovery, retrieval, aggregation, qualification and classification. Its Enterprise Services division provides pre-qualified content to B2B/C portals and content mining tools to individual companies. Its Professional Services division provides search content tools to individual information professionals and authoritative Web sites on discovering and exploiting Internet content. **BrightPlanet** is a privately held company founded in 1999 and is based in Sioux Falls, SD.*

*The two principal owners of **BrightPlanet** are VisualMetrics Corp., Iowa City, IA, and Paulsen Marketing Communications, Inc., Sioux Falls, SD. VisualMetrics is a systems software technology company specializing in computational linguistics, data warehousing and mining, and developing user-intelligent tools for the Internet. The LexiBot technology is an extension of the commercially proven and award-winning search tool Mata Hari®, first developed by VisualMetrics in 1998. Paulsen Marketing Communications brings nearly 50 years of marketing, advertising and public relations experience to the venture. **BrightPlanet***'s corporate Web site is found at* http://www.brightplanet.com.

***BrightPlanet** is the developer of the CompletePlanet Web site, a comprehensive listing of 22,000 deep and surface Web search sites. This compilation is presented under a subject taxonomy of more than 2,000 specific subject headings germane to every information need and market. CompletePlanet also offers tutorials and other assistance on how to more effectively obtain quality results from the Internet. Its Web site is found at* http://www.completeplanet.com.

BrightPlanet's LexiBot may be downloaded from* http://www.lexibot.com. *This desktop search agent presently can search up to 600 deep and surface Web sites simultaneously. It is available for a free, 30-day demo. LexiBot licenses are $89.95, with volume discounts available.*

References and Endnotes

[1] 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/.

[2] 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.

[3] "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/.

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

[5] S. Lawrence and C.L. Giles, "Accessiblity of Information on the Web," Nature **400**:107-109, July 8, 1999.

[6] See http://www.alltheweb.com and quoted numbers on entry page.

[7] 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 herein exclude its 'Special Collections' listing.

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

[9] See http://www.wiley.com/compbooks/sonnenreich/history.html.

[10] This analysis assumes there were 1 million documents on the Web as of mid-1994.

[11] See http://www.tcp.ca/Jan96/BusandMark.html.

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

[13] 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 5). Both measures ignore images and so-called HTML header content.

[14] 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 5), and a 50% combined representation by these three sources for all major search engines. Estimates are on an 'HTML included' basis.

[15] 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.

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

[17] 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 LexiBot.

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

[19] 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 agricuture'] an empirical coverage factor is calculated.

[20] 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/search.phtml
- 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.

[21] 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

[22] See, for example, http://www.surveysystem.com/sscalc.htm, for a sample size calculator.

[23] See http://cgi.netscape.com/cgi-bin/rlcgi.cgi?URL=www.mainsite.com/dev-scripts/dpd.

[24] See reference 34. 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.

[25] See, for example among many, BetterWhois at http://betterwhois.com.

[26] 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.

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

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

[29] The LexiBot 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.

[30] See the Help and then FAQ pages at http:www.invisibleweb.com.

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

[32] C. Sherman, "The Invisible Web," http://websearch.about.com/internet/websearch/library/weekly/aa061199.htm

[33] 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

[34] Alexa Corp., "Internet Trends Report 4Q 99," see http://www.alexaresearch.com/top/report_4q99.cfm

[35] 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

[36] 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 3 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 beileve 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.

[37] 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}.

[38] We have not empiricallly 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.

[39] The four sources used for this analysis were:

- FinderSeeker – http://www.finderseeker.com/
- Search Engine Guide – http://www.searchengineguide.com/
- A Collection of Special Search Engines (Netherlands) – http://www.leidenuniv.nl/ub/biv/specials.htm
- LincOn.com – http://www.lincon.com/srclist.htm

[40] See http://www.10kwizard.com/.

[41] 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://www.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.

[42] 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.

[43] See http://info.ox.ac.uk/bnc/.

[44] Assumptions: SURFACE WEB: for single surface site searches - 16% coverage; for metasearch surface searchers - 84% coverage [higher than NEC estimates in refernece 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; 100,000 potential deep Web sources; 13.5% quality retrieval from all deep Web searches.

[45] Online Computer Library Center, Inc., "June 1999 Web Statistics," Web Characterization Project, OCLC, July 1999. See http://www.oclc.org/oclc/research/projects/webstats/.

[46] As this analysis has shown, in numerical terms the deep Web already dominants. However, from a general user perspective, it is unknown.

[47] See http://lcweb.loc.gov/z3950/.

[48] See http://www.infotoday.com/newsbreaks/nb0713-3.htm.

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

[50] 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=.

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

[52] 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.