



## Deep Web FAQ's

---

### What is the Deep Web?

The Deep Web is that part of the web housing content that is only accessible when “asked for” through a custom query such as BrightPlanet’s DeepHarvester™ (which cannot be accomplished by a simple surface search query such as Goggle). The Deep Web contains unknown or hidden content that resides in databases on the Web, the results from which can only be discovered by a direct query. Deep *Data* itself exists not only on the Open Source Public Web, but on proprietary websites and within private databases.

Without a directed *Deep* query, a Web database does not publish a result and cannot be “searched”. When queried, sites post their results as dynamic Web pages in real-time. Though these dynamic pages have a unique URL address that allows them to be retrieved again later, they are not persistent.

---

### How does the Deep Web differ from the "Surface" Web?

Search engines — the primary means for finding information on the "surface" Web — obtain their listings in two ways. Authors may submit their own Web pages for listing, generally 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.

Thus, to be discovered, "surface" Web pages must be static and linked to other pages. Traditional search engines cannot "see" or retrieve content in the Deep Web, which by definition is dynamic content served up in real time from a database in response to a direct query.

---

### What are Search Engines Missing?

The Surface Web contains only a fraction of the overall content available on-line today. Of the top 5 surface search engines, Google represents only 63% of the total indexed content of the Surface Web alone! (<http://www.comscore.com/press/release.asp?press=2476>)

Limiting search to a single source (like Google); will produce a one-dimensional set of results. Harvesting from many sources, 10 to 20 or even 100, will yield far more documents and far more relevant content. Google, most likely, will not contain the most recent version of a document. Further, there is no way to “refresh” a Google search. Google will often have false



positive hits - content that matches your query but is not relevant to your search. Additionally, Google cannot distinguish a page of links from a page of content.

---

## **Can Goggle, Yahoo!, MSN, Bing and others find Deep Data on the web?**

Surface search results are based on “relevancy by popularity”, ranked by total “hits” by users’ simple search queries. While search engines can “find” deep data, their coverage is often sporadic and intermixed with less relevant (and too much) content. To find exactly the content needed, a user must traverse through “all” content within each surface site (Google, MSN, etc.).

Further, for a researcher to find Deep Data using Surface Search Engines, they must rely on their own content expertise and personal ability to navigate the web “one click at a time”, (link traversal) - a time-consuming process which has become normal behavior when using standard search engines.

---

## **What is the difference between a BrightPlanet “harvest” and a search engine “search”?**

With a standard search engine like Google, you do not have access to the actual content, only the links to content. Conducting a “harvest” will provide fully normalized content that can then be further processed with analytics, reporting or visualization tools.

BrightPlanet can automate custom queries that target Deep Web sites to explicit content needs to provide highly qualified, relevant quality content. Relevant queries will quickly narrow in on a specific answer without a lot of poking and jabbing (clicking) – a time-consuming process which has become normal behavior when using standard search engines.

---

## **Why haven't I heard before about the Deep 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.

What has not been broadly recognized is that information is now being published in a different means on the Web, especially for larger sites or for traditional information providers now moving their content online. The sheer volume of these sites requires the information to be managed from a database, the results of which are "hidden in plain sight" from search engines.



The evolution of the Web to a database-centric design has been gradual and largely unnoticed. Many Internet information professionals have noted the importance of searchable databases to Web content. But **BrightPlanet's** Deep Web white paper is the first to comprehensively define, quantify and characterize this entirely different category of Web content.

---

## **Is the Deep Web the same thing as the "invisible" Web?**

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. We avoid the term "invisible Web" because it is inaccurate. The only thing "invisible" about searchable databases is that they cannot be indexed or queried by conventional search engines. Using our technology, they are totally "visible" to those that need to access them.

The real problem is not the "visibility" or "invisibility" of the Web, but the spider technologies used by conventional search engines to collect their content. For these reasons, we have chosen to call information in searchable databases the Deep Web. Yes, it is somewhat hidden from traditional engines, but clearly available if different technology such as ours is used to access it.

---

## **How large is the Deep Web and how does the content differ from the "surface" Web?**

Public information on the Deep Web is thousands of times larger than the commonly defined Surface Web. 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 5,000 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.

---

## **What is OSINT?**

Open Source Intelligence (OSINT) is an information processing discipline that involves finding, selecting, and acquiring information from publicly available sources and analyzing it to produce actionable intelligence. In the U.S. Intelligence Community (IC), the term "open" refers to overt, publicly available sources (as opposed to covert or classified sources). It is not related to open-source software.

