

The Company

BrightPlanet is the leader in harvesting high quality content from inaccessible **Deep Web** and Surface Web sources.

With over 10 years of Deep Web extraction expertise, the company has developed a heuristic, rule-based expert system for communicating with Deep Web sources that does not require one-off scripts to be built by hand.

BrightPlanet's patented software employs technologies harvesting data/documents from:

- (1) The conventional *Surface Web*,
- (2) The larger and more authoritative **Deep Web**,
- (3) *Proprietary* Data sources
(such as LexisNexis and Dow Jones/Factiva),
- (4) Customers' own *Internal /Private* Data sources.

Content is harvested, federated and normalized, (*regardless of its source language, document encoding, format, or storage mechanism*) to provide qualified, relevant data for analysts and analytic technologies.

What is Deep Data?

Deep Data is unknown or hidden data that cannot be easily found using conventional search engines. Deep Data is content that exists deep within sites which require Deep Web harvesting techniques to uncover. Deep Data exists on the Open Source Public Web, on proprietary websites and within private databases.

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 (which cannot be accomplished by a simple surface search query such as Google).

Typically, this content cannot easily be found using link traversal techniques as employed through traditional search engine crawlers. Based on some studies, the Deep Web is at least 1000 times greater than the Surface Web, leaving the bulk of all "searchable" information out of a common search engines' reach. (*Exploring a 'Deep Web' that Google can't Grasp*: NYT 2-23-09 <http://www.nytimes.com/2009/02/23/technology/internet/23search.html>)

What is OSINT?

After four years of direct experience working with U.S. Intelligence Agencies, BrightPlanet has achieved a strong reputation and is acknowledged as *the* resource for Deep Web Harvesting in the Intelligence Community (IC).

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.

Can Google, Yahoo!, Bing and others find Deep Data on the web?

Surface search results are based on "relevancy by popularity" and returned by the total "hits" by a user's 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, Yahoo!, Bing, 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 Google 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.

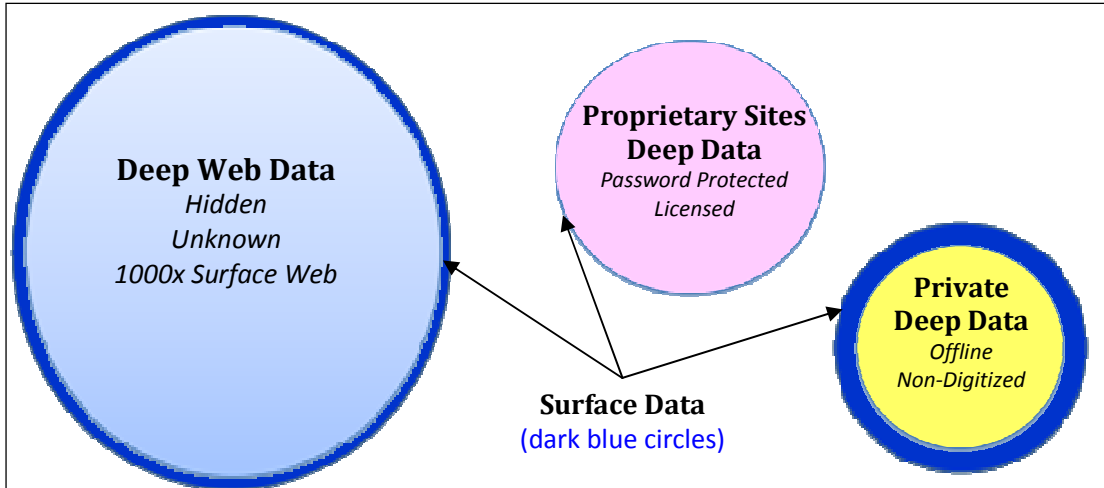
What is the difference between a BrightPlanet "harvest" and a search engine "search"?

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.



Harvesting Deep Data At the Scale of the Internet



Providing Normalized, Relevant Content for Research & Analytics

