

THE DEEP WEB

You don't know what you're missing.



BrightPlanet®

Pioneers in Harvesting the Deep Web

What is Deep Data?

Deep Data, sometimes referred to as invisible or hidden data, is content that is locked away behind Web search forms in databases that standard crawl techniques cannot easily crawl through. A specialized Deep Web harvester, like that developed by BrightPlanet, is necessary for extracting this content through directed query techniques. Without a query, Deep Web content remains locked up and inaccessible to traditional Web search engines.

Deep Data is there for the taking from the Open Source Public Web as well as from proprietary websites and private databases. But all Deep Data requires an exacting expertise and specialized harvesting tool to find and extract it.

Where is the Deep Web?

The Deep Web is a far vaster mother lode of information than the Surface Web. Some say the Deep Web is as much as a thousand times greater in size. And it's not just a matter of quantity, but quality as well. The Deep Web consists of both structured and unstructured content compiled by experts, researchers, analysts and automated processing systems working behind the scenes at an array of institutions throughout the world.

Ordinarily, this content cannot easily be found by using typical link traversal techniques as employed by traditional search engine crawlers. So this richest of content lies untapped without a dedicated harvesting technology that can "ask" a custom—even complex—query that is beyond the capability of link traversal systems.

(Exploring a 'Deep Web' that Google can't Grasp: NYT 2-23-09 http://www.nytimes.com/2009/02/23/technology/internet/23search.html?_r=1&ref=business)

Who is BrightPlanet?

BrightPlanet is the pioneer in harvesting high quality content from the **Deep Web** and making it accessible for those who need the valuable, untapped resources that lie beneath the Surface Web.

The company has more than a decade of experience working with the Intelligence Community in its War on Terror to target and access data hidden beyond the reach of typical Web search engines.

BrightPlanet is now bringing its patented Deep Web harvesting technology to the commercial and research community through multiple service solutions.

In bypassing a requirement for one-off scripts that need to be built by hand, BrightPlanet's software employs technologies specially designed to harvest data/documents from:

1. The rich, but largely unexplored Deep Web
2. *Proprietary* Data sources
3. Customers' *Internal /Private Data* sources
4. And the conventional *Surface Web*

This powerful solution not only harvests content, but also federates and normalizes it regardless of its source language, document encoding, format, or storage mechanism. The result is high-quality, relevant data for wide-ranging uses by analysts and analytic technologies.



THE DEEP WEB

You don't know what you're missing.

What OSINT Is.....and Isn't.

Open Source Intelligence, often referred to as OSINT, is an information processing discipline that involves finding, selecting and acquiring information from publicly available sources and analyzing it to produce actionable intelligence for business purposes.

Open Source Intelligence (not to be confused with open source software licensing like Linux) is neither covert nor classified information. Instead, it is overt or public information that is often hidden nonetheless because of the vast and impenetrable nature of the Deep Web. BrightPlanet has years of experience helping the Intelligence Community to penetrate the Deep Web and harvest mission-critical content buried within it.

How Come My Little Search Engine Cannot Find Big, Deep Data?

While mainstream search engines can locate some Deep Data, their coverage is often sporadic and intermixed with content that is less relevant or more popular than useful. To find *exactly* what you want, users of these little-engines-that-really-can't must traverse through all the content within each surface site *as well as* conduct their own additional queries and link traversals on each targeted finding.

Furthermore, for researchers to find Deep Data through traditional search engines, they must rely on their own content expertise and personal skills to navigate the Web one-click-at-a-time (link traversal), a grossly inefficient process of conducting research.

Where Is This Deep Data Gold?

Limiting Web searches to a single, one-dimensional source produces results that are both superficial and skewed. BrightPlanet overcomes this issue by harvesting from many sources — 10...20...even 100s — which yield a far more streamlined set of documents with far more relevant content and better coverage.

Traditional search engines that are tethered to the Surface Web often contain content that is out-of-date, leaving their users with no way to refresh or update their indexes. Traditional search engines also produce a lot of false positive results (documents that may match your query but do not match what you are searching for).

When Can BrightPlanet Go to Work for Me?

Immediately.

Standard search engines do not deliver access to actual content, only links *to* content. A BrightPlanet harvest on the other hand will provide a fully normalized version of the content that can then be further processed with analytics, reports or visualization tools. You get the actual content, not just the link.

BrightPlanet automates custom queries that target Deep Web sites to meet your explicit content needs, resulting in relevant, high-quality content from the Deep Web. Without the usual hunting and pecking, topic-specific queries can quickly narrow in on topic-specific answers--throwing lots of light on big, dark problems and unseen opportunities.