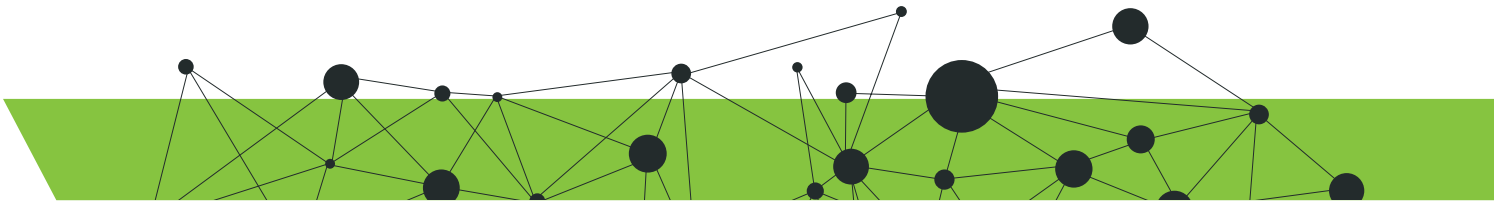
-whitepaper-



UNDERSTANDING THE DEEP WEB IN 10 MINUTES

Learn where it's at, how you can search it, what you'll find there, and why Google can't find everything

by Steve Pederson, CEO
spederson@brightplanet.com
March 2013

BrightPlanet®
deep web intelligence

Understanding the Deep Web in 10 Minutes

Learn where it's at, how you can search it, what you'll find there, and why Google can't find everything

by Steve Pederson

I. Introduction

Don't worry if you don't understand what the term "Deep Web" means. "Deep Web" is a vague description of the internet not necessarily accessible to search engines. The Deep Web is often misinterpreted as the "Dark Web". While browsing the internet, the Deep Web is usually right in front of you, you may just not know it yet. Whether you are searching for unstructured Big Data or trying to answer narrowly targeted questions , it can typically be found somewhere within the millions of Deep Web sources.

Both public and private sector organizations are intrigued by the vast potential of harvesting unstructured content at scale from the internet, tagging entities in the metadata, and curating that semi-structured content into actionable intelligence. There are many questions frequently asked about the process and possibilities for Deep Web harvesting, analytics, and data output. BrightPlanet hopes to answer those common questions in this whitepaper.

In this whitepaper you will discover:

- Where the Deep Web is, and how it compares to the Surface Web and Dark Web
- Why you should care about the Deep Web
- The difference between a search engine and a Deep Web harvest engine
- How data is harvested from the Deep Web
- Deep Web harvest use cases

II. What is the Deep Web?

DEEP WEB VS. SURFACE WEB

The Deep Web is a part of the internet not accessible to link-crawling search engines like Google. The only way a user can access this portion of the internet is by typing a directed query into a web search form, thereby retrieving content within a database that is not linked. In layman's terms, the only way to access the Deep Web is by conducting a search that is within a particular website.

The Surface Web is the internet that can be found via link-crawling techniques; link-crawling means linked data can be found via a hyperlink from the homepage of a domain. Google can find this Surface Web data.

Surface Web search engines (Google/Bing/Yahoo!) can lead you to websites that have unstructured Deep Web content. Think of searching for government grants; most researchers start by searching "government grants" in Google, and find few specific listings for government grant sites that contain databases. Google will direct researchers to the website www.grants.gov, but not to specific grants within the website's database.

Researchers can search thousands of grants at www.grants.gov by searching the database via the website search box. In this example, a Surface Web search engine (Google) led users to a Deep Web website (www.grants.gov) where a directed query to the search box brings back Deep Web content not found via Google search.

DARK WEB AND DEEP WEB - NOT THE SAME THING!

The Dark Web refers to any web page that has been concealed to hide in plain sight or reside within a separate, but public layer of the standard internet.

The internet is built around web pages that reference other web pages; if you have a destination web page which has no inbound links you have concealed that page and it cannot be found by users or search engines. One example of this would be a blog posting that has not been published yet. The blog post may exist on the public internet, but unless you know the exact URL, it will never be found.

Other examples of Dark Web content and techniques include:

- Search boxes that will reveal a web page or answer if a special keyword is searched. Try this by searching "distance from Sioux Falls to New York" on Google.
- Sub-domain names that are never linked to; for example, "internal.brightplanet.com"
- Relying on special HTTP headers to show a different version of a web page
- Images that are published but never actually referenced, for example "/image/logo_back.gif"

Virtual private networks are another aspect of the Dark Web that exists within the public internet, which often requires additional software to access. TOR (The Onion Router) is a great example. Hidden within the public web is an entire network of different content which can only be accessed by using the TOR network.

While personal freedom and privacy are admirable goals of the TOR network, the ability to traverse the internet with complete anonymity nurtures a platform ripe for what is considered illegal activity in some countries, including:

- Controlled substance marketplaces
- Armories selling all kinds of weapons
- Child pornography
- Unauthorized leaks of sensitive information
- Money laundering
- Copyright infringement
- Credit Card fraud and identity theft

Users must use an anonymizer to access TOR Network/Dark Web websites. The Silk Road, an online marketplace/infamous drug bazaar on the Dark Web, is inaccessible using a normal search engine or web browser.

WHY SHOULD YOU CARE ABOUT THE DEEP WEB?

For 2013, it is important to tap into the rich resources existing in the Deep Web. The last time an extensive study was completed estimating the size of the Deep Web was in 2001 — a time when the internet consisted of only approximately three million different domains. The 2001 study revealed that at that time the Deep Web was approximately 400-500 times the size of the Surface Web.

Today's internet is significantly bigger with an estimated 555 million domains, each containing thousands

or millions of unique web pages. As the web continues to grow, so too will the Deep Web and the value attained from Deep Web content.

III. Search Engines vs. Deep Web Harvest Engines

Harvesting is the term BrightPlanet uses when it talks about accessing the Deep Web. It is important to distinguish between traditional searches and Deep Web harvesting. Unlike traditional search technologies, like Google, that index links and allow you to view the results, BrightPlanet takes it a step further and harvests all of the results. The harvest process involves BrightPlanet extracting all of the text based content from each of the results pages and then preparing the content for some type of analysis depending on the needs of end-users.

To understand the major differences between a harvest engine and a search engine, it's important to understand the problem that search engines are meant to solve.

YESTERDAY'S SEARCH ENGINES

The problem search engines tried to tackle dates back to the early 1990s as the internet increased in popularity. Mostly static web pages were being added to the internet, but users needed a way to easily find web pages that contained information.

Search engines like Google, AltaVista, Yahoo!, and Lycos created technologies that crawled through websites and indexed them as a way for users to identify pages of interest. Search engines tried to find the most relevant page containing the answer to what users were looking for.

Questions that were originally asked to search engines in the late 90's were very basic. Students researching class reports replaced encyclopedias with the internet, researchers created basic web pages to share their discoveries, and social sharing consisted of updating your GeoCities page. The 90's internet was non-commercial and viewed with a research purpose.

TODAY'S SEARCH ENGINES

Today's internet is significantly different; millions of web pages are published for all sorts of reasons beyond traditional research.

Search engine companies developed systems able to quickly index millions of web pages in a short time period, therefore allowing users to accurately search the assimilated index. Search engines don't find or store all the content on a web page; they simply lead you to the content's location. This lack of data retention allows search engines to get away with storing minimal information about each individual web page.

Typically, search engines store the most frequently mentioned words, locations of those words, and any metadata (title of the web page, URL of the web page, keywords, etc) when indexing web pages. **The amount of data stored from each page is a crucial difference between search engines and harvesters.**

SEARCH ENGINES AND THE SURFACE WEB

Search engines like Google are really good at finding Surface Web websites; providing answers to basic questions quickly. However, companies and organizations have significantly harder questions than "How late is Burger King open?" Complex questions like those listed below require more than a search engine; they require a Deep Web Harvester®:

- Who is selling my products fraudulently online?
- How many people have won grants on Fetal Alcohol Spectrum Disorders?
- What are clinical trial patients saying about my experimental drug?
- What new information has been published on my competitor's website today?
- Has anything changed in this insurance coverage plan that would affect a pharmaceutical company's stock price?
- What new breast cancer research has been published in the last month? What are people saying about it?

DEEP WEB HARVEST ENGINE

Unlike a search engine, BrightPlanet's Deep Web Harvester extracts every single word every time it accesses a web page. Additionally, the Deep Web Harvester stores every single page harvested as a separate version in its database.

For example, BrightPlanet has a list of 100 websites actively harvesting for a customer every four hours. Therefore, the Deep Web Harvester collects a version of every single web page found within the 100 domains every four hours.

To put that into perspective, let's envision that each of those domains is a relatively small website (100 pages). In this scenario, every four hours we harvest content from 10,000 web pages (100 web pages multiplied by 100 domains). In one week, this harvesting process stores 420,000 web pages. BrightPlanet harvested 53 million web pages over a 30-day period for one customer.

A. DEEP WEB HARVEST ADVANTAGES

The concept of a harvest engine has a number of different advantages. The two largest advantages being:

- Analytic capabilities
- Versioning of web pages

Because BrightPlanet harvests the actual raw text from web pages, as opposed to storing metadata and only top keywords, BrightPlanet can integrate its harvested data directly into nearly any analytic technology using our OpenPlanet® Enterprise Platform [see page 7 for more on OpenPlanet].

Combining BrightPlanet's scalable harvesting capabilities with custom analytic technology helps customers visualize, analyze, and ultimately create intelligence from large data sets.

IV. Where do Deep Web websites come from?

SOURCE REPOSITORY: A LIBRARY OF 85,000 (AND GROWING) DEEP WEB SOURCES

The Deep Web is at least 400-500 times the size of the Surface Web. It is continuously growing, and that means new Deep Web sources are also growing. BrightPlanet harnesses Deep Web sources by sorting and indexing them in its Source Repository.

The Source Repository is a library of Deep Web sources/websites that BrightPlanet has collected over 10 years of web harvesting on behalf of clients.

New sources are added and updated daily. There are currently over 85,000 Deep Web sources, grouped

by source type, in BrightPlanet's Source Repository. Examples of source type groups include Law, Healthcare, Pharmaceuticals, Social Media, Major Media, Newspapers, Finance & Economics, and Politics to name just a handful of the over 60 groups.

HOW YOU CAN LEVERAGE THE SOURCE REPOSITORY

End-users do not need to worry about communication with sources; those processes are all done automatically by BrightPlanet. You just need to identify the information you are trying to find and from what sources, and BrightPlanet can harvest it on your behalf.

BrightPlanet commonly works with its end-users to harvest content from custom Deep Web sources. End-users define hundreds or thousands of Deep Web sources for BrightPlanet to query with many keywords at once. Once new sources are entered into the Source Repository, they will be indexed and saved for future harvests.

Here are just a few examples of how the Source Repository can be leveraged:

Newspapers

The Newspapers group in the Source Repository includes every newspaper in the U.S. In a matter of seconds, BrightPlanet could harvest topic specific content from every newspaper in the U.S. Instead of searching newspaper website after newspaper website, the information could be harvested instantly. Additionally, the papers are sorted by state so you could limit the search to certain states if it better fits your needs.

Law

There are several categories within this group. One of those categories is Courts. This group includes sources that would allow you to search Court rulings at all levels of the judicial branch; state, local, and federal, instantly.

Finance & Markets

Buy the rumor; sell the news. Users can find both rumors and news faster than the competition by harvesting from the News, Finance Blog/Website, Finance Message Board, and industry-specific blogs and message board source groups.

Health & Pharmaceutical

There are dozens of possibilities for leveraging the Source Repository for the health and pharmaceutical sectors: fraud, diversion, health websites, disease-specific websites, and message boards to name a few. For example, if you wanted to search for any mentions of a new multiple sclerosis drug, selecting the M.S. Message Boards, M.S. Blogs/Websites, Health Blogs/Websites, and Health Message Boards source groups yields access to 75 reliable Deep Web sources for you to instantly search.

VIEWING THE CONTENT IN DEEP WEB INTEL SILOS

Another solution BrightPlanet offers, to help sort and view harvested Deep Web data, is Deep Web Intel Silos. For the purposes of this paper, we'll talk about how healthcare research has leveraged Deep Web Intel Silos.

There are millions of documents available on the Deep Web for healthcare research that current methods of online research have no way of finding or collecting. Deep Web Intel Silos can create collections of nearly any open-source content. For healthcare research, BrightPlanet creates disease and healthcare

topic-specific research silos to which researchers subscribe.

Unlike a traditional static database like PubMed or LexisNexis, where the dataset is predefined by the organization offering access, topic-specific research silos start with a base set of data and add additional sources requested by subscribers. This allows for collaboration between research institutions.

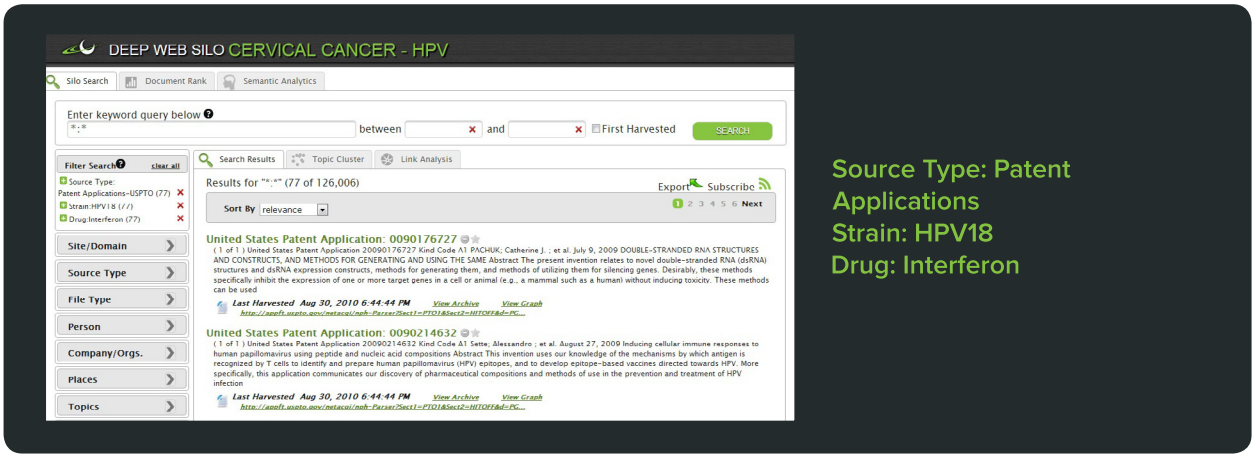
As more and more researchers request sources to be added to research silos, and as BrightPlanet continuously monitors these sources in key topic areas, research silos develop into some of the most comprehensive topic-specific research databases worldwide.

Since the subject matter experts, healthcare researchers in this case, identify the sources and source types they want to draw from, and dictate how they want harvested documents tagged and sorted. Tagging documents becomes crucial when creating intelligence from large datasets; the big challenge everyone has with Big Data. The final Deep Web Intel Silo dataset contains only relevant, searchable data with customizable drill-down search facets.

REFINING A LARGE SET OF RELEVANT DATA INTO ACTIONABLE INTELLIGENCE

Let's say a research silo contains 126,000 harvested documents related to the broad topic of cervical cancer. If the researcher is only interested in patent applications mentioning the drug Interferon with the HPV18 strain, the user can create an advanced search focused only on patent applications.

By narrowing the search to only patent applications, the huge dataset is reduced to 77 relevant patent applications mentioning HPV18 and the drug Interferon. Any additional search queries the user performs will comb through only those 77 super-relevant documents.



THE VALUE OF SILO SERVICES

Deep Web Intel Silos are individual repositories for topic-specific content, and are updated with new and relevant information from the harvester in real-time. Each silo is filled with high-quality Deep Web resources – databases, RSS feeds, and more – that lie beyond the reach of traditional search engines. They also include standard analytical tools like raw data views, topic clustering, and link analysis. Additional custom analytical modules can easily be added to meet your reporting needs.

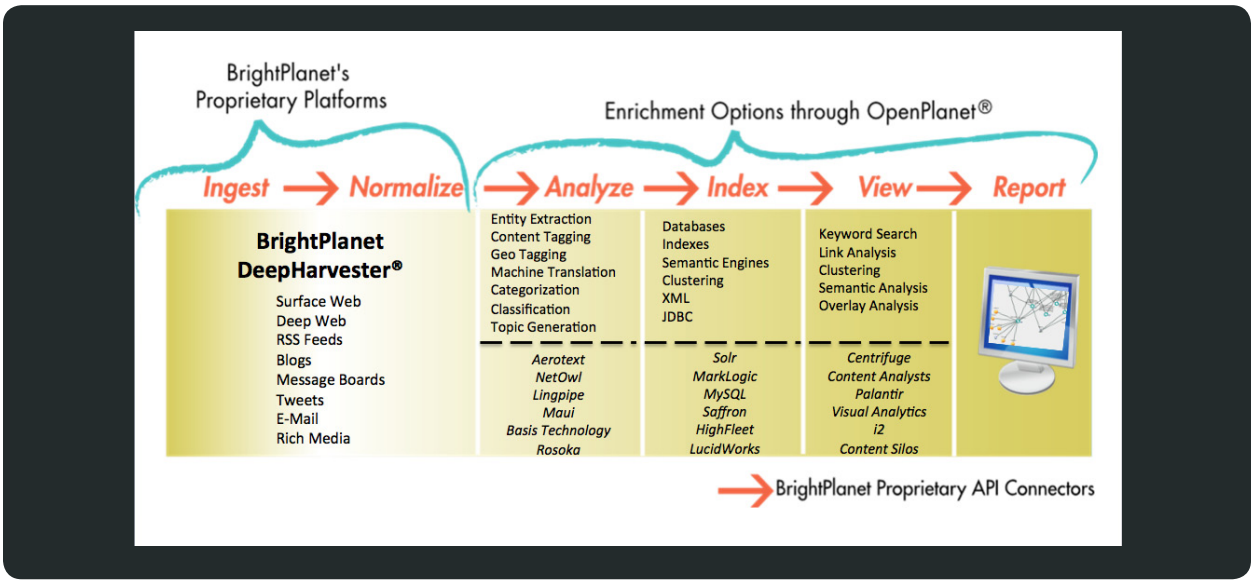
The true value of silo services lies with BrightPlanet's Deep Web Researchers. These are highly skilled content managers who take the complexity out of Deep Web research. Think of them as your personal guides to the Deep Web, discovering and harvesting the resources that fill your silo with relevant, timely content. Our researchers work hard to deliver the best results available, leaving you time to do what you do best: analyze, interpret, and create actionable intelligence.

LARGE SCALE – OPENPLANET

Many customers only require access to harvested content to make searching capabilities simpler for them. For these specific customers, access to a Deep Web Intel Silos fulfills their needs. Customers wanting to make additional conclusions from harvested data can easily integrate the data into any number of analytic capabilities.

Through its tenure with the U.S. intelligence community, BrightPlanet has learned that a single end-to-end harvest platform takes anywhere from six months to three years to set up, depending on the scale and number of components. While this is a good business opportunity for system integrators who can bill hourly, it is not a desirable solution for commercial deployments demanding a higher level of integration without custom development. BrightPlanet saw this need for open integration early on and spent two years developing an open platform called OpenPlanet to overcome these limitations.

The OpenPlanet platform is based on a simple workflow that completely separates the harvesting and analytic components of data collection and analysis. This concept allows BrightPlanet to easily swap in and out different analytic technologies with no knowledge of where the data comes from. Allowing customers to integrate multiple datasets, not just harvested web data, with multiple analytic technologies in one workflow without significant development.



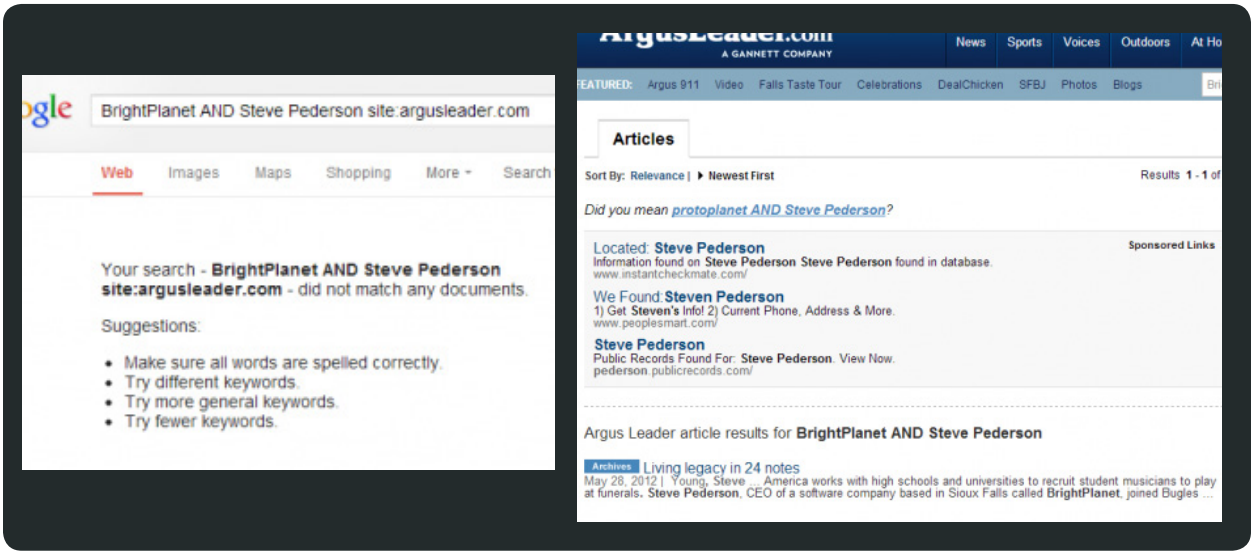
V. Deep Web Harvest End-User Examples

A DEEP WEB HARVEST VS. SEARCH ENGINE USE CASE

The following example shows the kind of Deep Web content search engines may be missing.

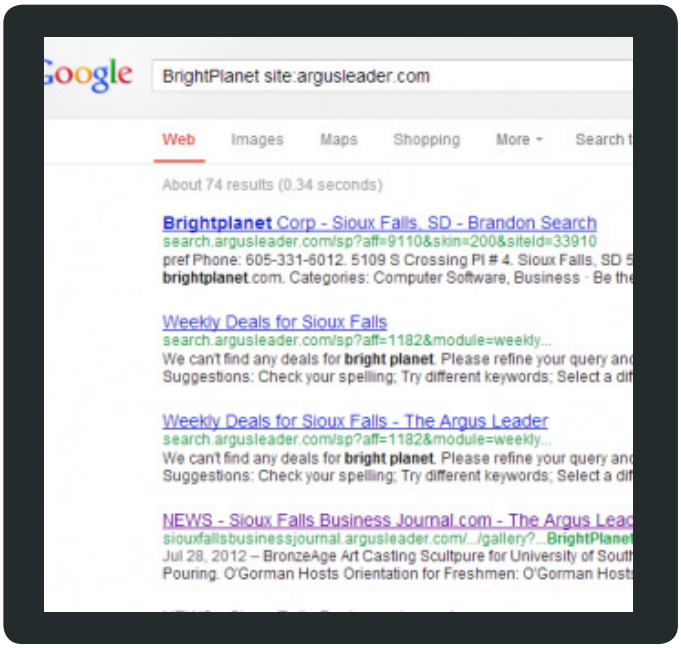
The *Argus Leader*, the local newspaper of Sioux Falls, South Dakota, did an article about BrightPlanet's CEO, Steve Pederson (an avid bugler) titled "Living Legacy in 24 Notes." The article at one point in time had been on the homepage of the Argus Leader, a location that is reachable by a Surface Web search engine like Google. A few days after the article was featured on the homepage, the article was pushed into archive format, and thus only reachable via a query through the search box located on Argus Leader's site; it left the Surface Web and entered the Deep Web.

These two images demonstrate the differences between the Deep Web and the Surface Web. The image on the left is a search of what Google has indexed. The query (BrightPlanet AND “Steve Pederson” site:argusleader.com) tells Google that the only results we want are from the Argus Leader domain. The search returns zero web pages that have been indexed by Google containing both BrightPlanet AND “Steve Pederson”.



The image on the right proves that results containing both terms do exist. This search is performed using the search box provided by the *Argus Leader* website. The reason why this search returns results is because the search box points to the newspaper’s database, a Deep Web source. Archived content can only be accessed via the Argus Leader’s search, making that content exclusive to the Deep Web. Google does not direct queries into any site searches, as it only finds documents via link following. The “Living Legacy in 24 Notes” news article has fallen into the Deep Web.

When BrightPlanet collects Deep Web content, it is exactly this type of search placed directly into the search forms that BrightPlanet can execute at a very large scale; issuing thousands of search queries into thousands of Deep Web sites and pulling all the content back for analysis. Imagine being able to query every single online newspaper web search form within the United States simultaneously.



The other major advantage of using a Deep Web harvest over a search engine is efficiency. Doing a search for the query BrightPlanet on the Argus Leader web page will return the same one article. Doing a search for BrightPlanet within the Argus Leader domain on Google will return 74 results (see image on left).

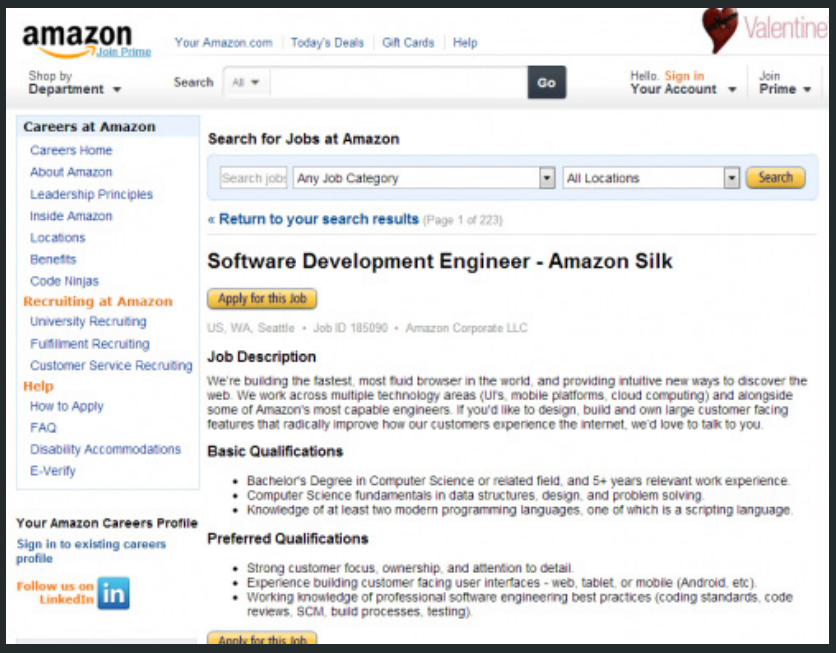
The extra 73 results return links that no longer contain BrightPlanet on the actual page, as Google is still searching an old version of the page. When Google crawls through a site, it filters through millions of links, often picking up irrelevant content. When BrightPlanet performs a Deep Web search on a site, it only harvests the relevant content related to your queries.

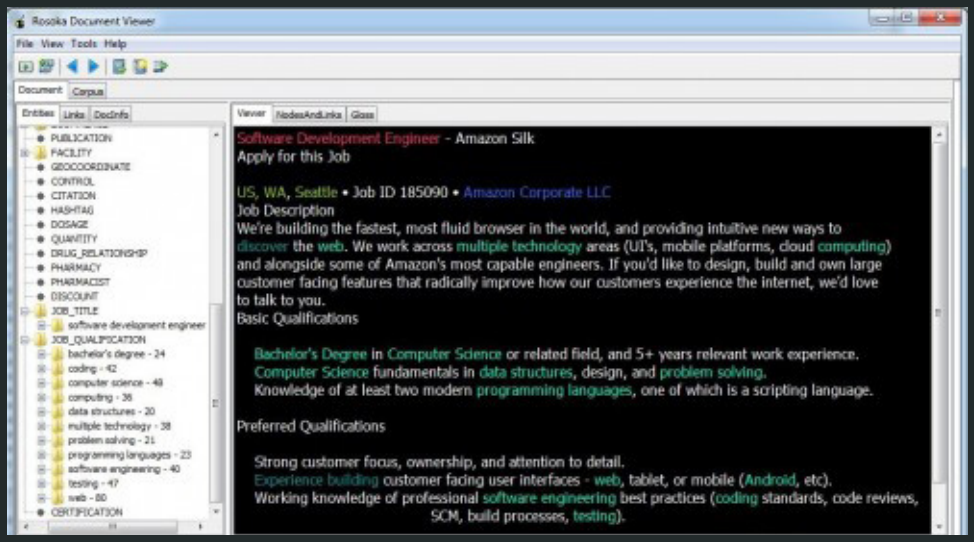
AN OPENPLANET USE CASE

An interesting example of following data from the harvest stage through the OpenPlanet Platform is a recent project BrightPlanet completed for a management consulting firm; BrightPlanet delivered bi-weekly exports of all job postings from every Fortune 200 company.

First, BrightPlanet harvested all job postings from Fortune 200 companies. The raw text of each job posting wasn't enough to give insight into the hiring actions of the companies, so BrightPlanet worked with the end-user to enrich the content with custom tags. BrightPlanet wrote custom rules and dynamically tagged and extracted the locations of the job postings, job titles, job qualifications, and required certifications.

The deliverable for the end-user was a CSV file consisting of the company, job title, location, important qualifications, URL, and the raw text of each posting. The end-user uploaded the data into their own database for analysis, and the management consulting firm was able to add value to its product offerings.





WHO WE ARE

Since our inception 13 years ago, BrightPlanet has worked closely with the U.S. Department of Defense harvesting open-source information for the U.S. government's "War on Terror". The Intelligence Advanced Research Projects Activity (IARPA) has made significant investment in 'Sensemaking' initiatives; and BrightPlanet Corporation and their partner companies have successfully applied IARPA methodology and enabling technologies to create Big Data solutions.

Now, the company's patented Deep Web Harvester and Deep Web Intel Silo Services are serving the needs of companies and organizations that need help in harvesting and analyzing Big Data from the Deep Web. The company partners with third party, 'best of breed' technologies agnostically, to provide custom solutions for nearly any analytic need.

MORE INFORMATION

BrightPlanet provides free resources such as white papers, eBooks, blog posts and videos online at the Deep Web University. Subscribe and keep up-to-date on the latest Big Data news.

To learn more about how BrightPlanet solutions can help you harvest Big Data from the Deep Web to create actionable intelligence, please visit our website or contact BrightPlanet to schedule a demonstration of the Deep Web Harvester and Deep Web Intel Silo Services.

Email: contact.website@brightplanet.com
Web: www.brightplanet.com

The images on page 9 show a web page (top) and what it looks like once it is normalized and the entities are extracted (bottom). The image on the bottom is displaying the web page in Rosoka's Document viewer. The highlighted text in the second image displays entities that have been extracted from the text of the job posting.

Even though the final deliverable for the end-user was not an analytic interface or report, it's easy to see the insight you could quickly draw from the job posting output. For example, users could quickly identify which companies had postings for computer programmers that require Java programming skills. Many valuable insights can be drawn from the data set because of the extracted and enriched data BrightPlanet provided.

VI. Conquer Big Data by Pairing Internal Data with Unstructured Deep Web Data

Big Data doesn't just come from within company walls. Structured enterprise data is only one part of the hybrid data spectrum. Unstructured data found on the internet is the other part, and there are trillions of unstructured web documents ready to be harvested.

Search engines are a good starting point, but they only skim the surface of available content. The more-efficient source qualification and filtering capabilities of Deep Web harvests lead to less time searching, leaving more time for creating actionable intelligence.

Now that you know what the Deep Web is and the many ways to get data from it, what actionable intelligence can BrightPlanet help you harvest from the Deep Web? The Deep Web grows exponentially every year; start tapping into it for your business or institution.