I. Introduction

Being in the business of Deep Web harvesting, we find ourselves answering the same questions regularly. Online search is such a part of our daily lives; Kleenex is to facial tissue as Google is to search. Google has become synonymous with search in many people’s minds. However, what Google does and what Deep search engine technology indexes billions of webpages over time, therefore allowing users to search an already assembled index. Search engines don’t find or store all the content on a web page; they simply link users to the content’s location. The amount of data stored from each page is the crucial difference between search engines and content harvesters.

Our Deep Web Harvester extracts every single word every time it accesses a webpage. Additionally the Deep Web Harvester stores every single page harvested as a separate version.

Solving specific problems vs. being everything to everybody

You may be wondering why someone would need every single word to be extracted every time? Search engines like Google are really good at finding Surface Web websites; providing answers to basic questions like “How late is Burger King open?” quickly. However, companies and organizations have significantly harder questions that can only be answered by web data beyond the scope of a search engine. 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?

Put these questions in the Google search bar and the answer will not be found. Our technology can help answer these questions by harvesting Big Data from the Deep Web, connecting the results, and delivering them in the way that is useful for the end-user.

Specific sites vs. the entire web
Instead of searching the entire web, the harvester performs directed harvests on specific websites.
End-users can define hundreds or thousands of Deep Web sources for BrightPlanet to query with many keywords at once.
If an end-user does not know the exact websites they want to search, they can just identify the information they are trying to find or problem they are trying to solve and BrightPlanet will help them find and define the web sources. Sources can continue to be added as the project evolves and the end-user decides which data is most useful.

III. Deep Web Alerts vs. Google Alerts

Page changes vs. mentions
Google indexes billions of webpages and sends alerts based on any new pages that may contain a mention of the selected keywords or phrases. Google Alerts does a very good job at trying to monitor the Surface Web for new webpages. However, Google’s alerting service will never send an alert on pages that already exist within their index, only new pages. This means pages that didn't previously contain the specific keywords will never be found via Google’s alerting service. BrightPlanet’s Deep Web alerts are meant to tackle a completely different problem.
Oftentimes end-users know the pages or website they want to monitor and they want to be alerted when it changes and what is changing. BrightPlanet’s technology harvests the content from specific webpages the users choose and sends alerts when any textual changes have been made on those specific pages. The harvests and alerts can be scheduled and sent every minute to every week depending on what the end-user would like.

Examples of why this is valuable:
1. Monitor and track textual changes to a corporate website
2. Monitor government filings and receive real-time alerts when changes are made
3. Track competitor’s product literature and be alerted when new features are added
4. Monitor when news stories on sites get changed or updated

If the user prefers, BrightPlanet's technology can also monitor specific parts of pages; the entire page does not need to be harvested every time. For example, if a user just wanted to track the “Events” section of their favorite organization's website, they could receive email alerts every time just that section was changed.

Customizable emails vs. generic

Google Alert emails have a generic layout that is used for every user. The alerts sent by BrightPlanet are customizable. Everything from the title to the content can be customized for the specific user.

At its simplest, BrightPlanet's alert emails include past versions of the monitored page's content and the new versions of the pages content with the specific content that was modified highlighted. This ability to pick specific sections of pages to harvest allows users to be alerted if specific relevant portions of pages are modified.

Near real-time vs. recurring

Google Alerts are sent on a daily basis and BrightPlanet's Deep Web alerts can be sent in near real-time. An email alert can be sent within 60 seconds of a page change. If the user prefers, the alerts can be sent on a recurring basis they specify, but the technology has the ability to send them in near real-time unlike Google Alerts.

BrightPlanet Deep Web alerts at work

To illustrate the use of BrightPlanet's Deep Web alerts below is an example where the technology was set up to monitor changes made to an inventory of a car dealership.

Car dealership inventory

Deep Web alerts were set up to harvest the inventory page of a car dealership. The campaign was set up to send daily alerts showing the changes that have been made to the page. A competitor could use this to know what has been added to the lot and purchased.

Below are screenshots of email alert portions. The top screenshot shows the text for a car that was removed highlighted in red; presumed sold. The bottom two screenshots show cars that were added to the inventory page highlighted in green.
This is a very entry-level Deep Web alert. The alerts just contain the text from the specific page with the modifications highlighted. As was mentioned above, these e-mail alerts can be customized based on the needs of the user.

IV. The Value of Defining Your Own Dataset

There are many web monitoring and listening platforms, both free and commercial options, available online. The majority of these companies advertise their ability to monitor and scour hundreds of thousands of sources simultaneously. These turnkey solutions allow the user to submit lists of keywords and the platform returns thousands of hits from blogs, message boards, social media websites, news websites, etc. This type of solution works great for projects that simply need to track mentions of specific keywords, particularly keywords and sources that are used globally. However, this type of turnkey model using a standard database doesn't work for everybody. Many users need to define their own dataset rather than blindly trusting one provided to them.

Who should define their own dataset and what value does it bring?

Marketing professionals monitoring a brand in a specific region

Marketing professionals wanting to track a brand in a specific region must often sift through thousands of irrelevant false hits from all across the globe. Instead, select the specific state and local newspapers and blogs to monitor and select custom geographic areas on Twitter to collect tweets from.

Investment managers monitoring specific companies

Investment managers know certain sources of information are especially relevant when trying to monitor and track specific companies. Take advantage of defining sources and datasets by picking sources such as press releases, government regulatory sites (USPTO, FDA, SEC, etc.), news, and investor relations portions of sites.

Health researchers

Health researchers want to stay on top of what is happening in their field without having to read through every article in their favorite journals and health research news websites. They can harvest from journals and news sources they select and create keywords to alert them to only the content relevant to them.

Any data analyst

BrightPlanet offers near real-time alerting of new and modified content from any web source for those that can identify where and what they are looking for.

Benefits of defining your own source

Get more data

BrightPlanet’s Deep Web harvesting technology collects content from both the Surface Web and the Deep Web; delivering more data than a traditional collection technology.
Harvest content, not links

When BrightPlanet harvests data all the text-based content on the page is harvested and stored offline rather than just giving the end-user a link to the URL. This gives the end-user a few advantages that include:

1. Pages that have been harvested and then removed will still have a record
2. Ability to integrate harvested data with third party analytics
3. Ability to monitor pages for changes made to content

Harvest content at big data scale

BrightPlanet's Deep Web Harvester is highly scalable with the capability to issue hundreds of queries (aka searches) to hundreds of websites simultaneously.

V. Conclusion

Frequently asked questions (FAQs) help to make answering common questions more efficient. Since people often think online search equals Google, we spend a lot of time educating contacts on the difference between Deep Web harvesting and search tools like Google. Google and other search tools like it are sufficient for many projects and tasks but when people need to answer harder questions than “How late is Burger King open?”, Deep Web harvesting is available to help.

After reading this whitepaper, is there a project you have that would benefit from harvest instead of search, page changes versus mentions, defined sources versus the entire web, and/or real-time alerts? Deep Web harvesting and BrightPlanet’s Deep Web technologies can help answer more complicated questions and help you take data from the Deep Web and make it actionable.

Check out our other free resources on the Deep Web and Big Data to learn more about Deep Web search and what is possible.