

<< ArnoldIT
Home

Articles

Speeches

Services

Features

About

Search AIT

BrightPlanet

An Interview with William Bushee

Search

-
- Only match
-
- whole
-
- words



BrightPlanet was one of the first "deep Web" services. The phrase "deep Web" is catchy but it does not explain what type of information is available to a person with a Web browser. Some Web masters require a user to enter parameters for a query. Those parameters are passed to the query processing system and the matching information is pulled from a database and the results are rendered in a browser. A familiar example is querying Southwest Airlines for its flight schedule. Other types of "deep Web" content may require the user to register. Once logged into the system, users can query the content available to a registered user. A service like Bitpipe requires registration and a user name and password each time I want to pull a white paper from the Bitpipe system.

BrightPlanet can handle both types of indexing task and many more. Bright Planet's technology is used by governmental agencies, businesses, and service firms to gather information pertinent to people, places, events, and other topics.

I spoke with the firm's lead technologist on October 2, 2009. The full text of my interview with William Bushee appears below:

What's your background?

My background has been in search/harvesting/collection/ingest for about 10 years with a primary emphasis on Deep Web harvesting. Going beyond standard search, BrightPlanet is the OSINT leader in harvesting Deep Data ("unknown and hidden") from both the Deep Web and the conventional surface Web. We have developed a patented rule-based expert system for automatically communicating with Deep Web sources – a system that does not require manual scripting. Further, we normalize the harvested content for further analysis and visualization through 3rd party tools.

BrightPlanet is in the process of repackaging and revising our pricing model. Our standard harvest server designed for use by the Intelligence Community allows for customization and additional integration options. Our new package is a service oriented model allowing both wholesale deep harvesting and harvesting as the front-end of our OpenPlanet platform.

BrightPlanet has two secret sauces, 1) the means to harvest Deep Web content and 2) an Integrated OpenPlanet framework of "best of breed" technology partners that provide content enrichment, storage and visualization.

What was the trigger in your career that made search and retrieval a focal point?

My first big trigger into search and retrieval came when we were manually building one-off scripts to federate searches across multiple Deep Web sources, a process that required multiple full-time

interns to simply manage 50-100 sources. I got tired of training new interns so I decided that there must be a better way to automate the process. After spending a short amount of time reviewing various search engine sites, I began to notice patterns that could be used to automate link filtering system – I remember lying out about 20-30 printed pages of search results pages on the folding table in the basement.

At that time, our company was migrating away from data warehousing into unstructured content, primarily from the Web. While we had a great data warehousing solution, we saw an even bigger need working with unstructured content from the Web. Our company had two focuses, harvesting content and indexing unstructured content. The technologies that existed at the time for harvesting were simple spiders, which for the most part worked well because the Web was largely static, nothing like the Web we know today.

A few of us decided crawl technologies would never be able to handle the need to find “unknown and hidden” content from what we called the “Deep Web”. If you remember, BrightPlanet coined the term “Deep Web” about 10 years ago. Some at BrightPlanet took on the harvesting challenges while others began working on an information retrieval solution (IR) to build a text engine. I spent most of my time on harvesting, normalizing and preparing content for indexing, but I did help architect some of the later IR solutions. My experience with IR comes in very handy from time-to-time. Being able to efficiently get data prepared for indexing and knowing what IR solutions need, helps us be smarter during ingest.

Up until that point, all of my background was around 3GL languages, data warehousing and modeling.

Do you have a search related project underway? If so, will you describe it?

We like to differentiate between “surface search” and “Deep Web harvest.” One of my pet projects has always been BrightPlanet’s CompletePlanet Portal. CompletePlanet lists about 70,000 known Deep Web sources, the same sources that we have automatically configured to work with our Deep Harvester. Over the past 4 years our company worked primarily with the U.S. Intelligence Community and we have not had a chance to keep the CompletePlanet Portal updated as new sources are found. As BrightPlanet branches into new commercial markets, it is becoming increasingly important that our library of Deep Web sources is constantly updated and refreshed. One of our projects underway is to replace the underlying infrastructure used to manage sources with a more flexible framework design.

The next update to CompletePlanet will have a few additional features, many more Deep Web Sites (sources) and will be better integrated into our Deep Harvester. The portal will also further strengthen BrightPlanet’s position as the leaders in everything related to Deep Web harvesting.

The number of new companies entering the search and content processing "space" is increasing. What's your view on the "noise" in the search and content processing market sector?

BrightPlanet is more interested in providing content to those who already do search rather than compete directly with search providers. We focus on harvesting content and providing a framework for enrichment which allows others to build better search and content processing solutions, so these new players are actually helping us.

BrightPlanet has been doing Deep Web harvesting for almost 10 years. While there are a few new players entering the Deep Web harvest space, we are largely solving different problems. The new players and added press coverage about the Deep Web has been helpful to add focus to the need of better ways to zero in on true intelligence within content, not just better ways to search.

Obviously, I follow a lot of the companies with similar technology or similar solutions. While there are some great products out there, you’ll find their missions are different than ours. We harvest, normalize and enrich deep data from the Web at scale, providing content for search and visualization systems. I try to keep in touch with our counterparts and push projects their way if it is a better fit. I really like the work that [Kapow Technologies](#) has done with their robot maker.

What are the functions that you want to deliver to your customers?

There are two distinct problems that BrightPlanet focuses on for our customers. First we have the ability to harvest content from the Deep Web. And second, we can use our OpenPlanet framework to add enrichment, storage and visualization to harvested content. As more information is being published directly to the Web, or published only on the Web, it is becoming critical that researchers and analysts have better ways of harvesting this content.

However, harvesting alone won't solve the information overload problems researchers are faced with today. The answer to a research project cannot be simply finding 5,000 raw documents, no matter how good they are. Researchers are already overwhelmed with too many links from Google and too *much* information in general. The answer needs to be better harvested content (not search), better analytics, better enrichment and better visualization of intelligence within the content – this is where BrightPlanet's OpenPlanet framework comes into play.

While BrightPlanet has a solid reputation within the Intelligence Community helping to fight the "War on Terror" our next mission is to be known as the commercial and academic leaders in harvesting relevant, high quality content from the Deep Web for those who need content for research, business intelligence or analysis.

What are two or three of the key features you are / will be implementing?

That's a great question. BrightPlanet is scheduled to release a completely new version of our Deep Harvester in October 2009. This release adds a new OpenPlanet framework used to add content enrichment tools into the document workflow. We have realized that most customers are interested in a better and faster way to add different enrichment tools, technologies like entity extraction, entity resolution, machine translation, geo-tagging, language linguistics, indexing and visualization engines, into the framework without the need to build one-off solutions. By providing enriched content using specific analytic tools for each project, we are able to provide a faster and more flexible solution than these one-off custom integrations.

Through OpenPlanet, the Deep Harvester will integrate with technologies such as [Basis Tech](#), [LingPipe](#), [Apache Tika](#), [Mark Logic](#), [Jackrabbit](#), [AeroText](#), [Visual Analytics](#), [Palantir](#) and others depending on the need and scope of the customer. BrightPlanet's proprietary connectors within our OpenPlanet framework allow integration to partner APIs, allowing Deep Web content to be processed, stored or viewed by various combinations of tools. As important, this integration allows the framework to provide results much faster, providing well defined integrations that might have otherwise taken months to build. And, as new tools become available, they only need to be added to our framework once, allowing a lot more plug-and-play integration.

In addition to our Deep Harvester release, we will be releasing the Deep Web Source Repository later this year. The Deep Web Source Repository, a library of pre-configured Deep Web Sites, will be the backbone to our CompletePlanet Portal and will also provide customers better access to pre-configured, ready to use, Deep Web sources... providing even more Deep Web tools for our customers.

There's a push to create mash ups--that is, search results that deliver answers or reports. What's your view of this trend?

That's a tough question. I'm not sure mash ups are what people need to find answers. The overall concept is very innovative and it does have its purpose for pushing content around the Web. BrightPlanet has the ability to publish harvested content as XML which could then be used within a mash up.

Of all the mash up solutions I have seen and used, I think that the openkapow solution is by far the best overall solution. I still haven't come up with a reason to persist a mash up beyond the initial test and experiment phase.

What differentiates your approach from the systems available? Do you support other vendors' solutions, or are you a stand-alone solution?

BrightPlanet actually is both. As we just talked about, we provide our own standalone Deep Harvester solution and we provide the OpenPlanet framework to add additional vendor solutions into the document workflow. We strive to support accepted open standards. Whenever possible, BrightPlanet leans toward integration over re-invention.

Our Deep Harvester is a proprietary homegrown solution, however, we do use a number of 3rd party solutions as part of our document normalization. We decided early on that we did not want to be experts on decompressing PDF files to extract text or determining word boundaries in Chinese, those solutions are best left to the experts.

Our OpenPlanet framework is built entirely on the principle of interoperability between various vendor solutions. Companies should not try to be experts in too many things.

What is your view of NLP [natural language processing]?

NLP is getting quite a lot of media attention. Semantics and NLP will play critical roles. However, the adoption has been slower than I hoped. Our OpenPlanet framework allows our Deep Harvester to easily adopt as the technologies mature and as new solutions make their way to market.

I think we are going to see a lot more collaboration of semantic solutions. Instead of one NLP solution, we are going to see two, three or four daisy chained together. BrightPlanet fits best into a semantic search solution with our ability to harvest and normalize high quality content in the first place and to provide the framework for the solutions – feeding the beast, so to speak.

A number of vendors have shown me very fancy interfaces. The interfaces take center stage and the information within the interface gets pushed to the background. Are we entering an era of eye candy instead of results that are relevant to the user?

There is definitely a lot of eye candy today and I love it! I really like visualization systems like Visual Analytics and Palantir. As the quantity of data increases (and the overall quality decreases), users need to rely more on visualization techniques than standalone search techniques. I am a big believer in creating better visualization techniques.

Visualization of large datasets is tricky. Thomson Reuters has some interesting clustering and visualization systems that I have never seen elsewhere as part of their patent analysis systems. I also liked the work Centrifuge is doing. However, the industry needs more robust open-source solutions.

What text processing functions do you offer?

BrightPlanet is continually improving its technology around content normalization, marking or stripping non-relevant content from Web pages. For the most part, we rely on 3rd party analytic and storage solutions to do text processing. Years ago we developed our own text engine solution and it worked very well. We kept adding more and more features but ultimately we were unable to compete with the big players like Mark Logic and Microsoft Fast. About a year ago we decided to abandon those efforts and focus our efforts on harvesting. We have since begun partnering with our former competitors to provide solutions for text processing and indexing. The Deep Harvester has connectors to push data into existing text engine and database solutions like [Mark Logic](#), [Solr](#) or MySQL through our OpenPlanet framework.

What is it that you think people are looking for from semantic technology?

I think people are simply looking for better ways to sift through an overwhelming amount of “potentially” relevant content. Semantic technology has always been about getting to the meaning or the context of the information, which is not a simple problem to solve.

People want the ability to vet their harvest/search results from 10,000, 100,000 or even one million documents down to the those on-target of the user’s “intent”. This “intent” is critical and hopefully semantic processing and NLP will help solve the problem of finding “content in context” while resolving ambiguity at the same time – a tall order.

What are the hot trends in search for the next 12 to 24 months? How will you take advantage of

them; for example, go public, partner, sell to a larger firm, etc.

Two hot trends over the next 12 months are going to be cross-document intelligence and cloud computing.

One thing that I don't think a lot of people appreciate is the value of intelligence that is found by collecting a lot of small pieces of information from a large set of documents and find cross-document intelligence. For example, there is not going to be a document online that says "this group will hijack this location on this date using this plan." However, some or all of those pieces of information may be found in known Deep Web sites and internal databases. Using technologies like Visual Analytics, Saffron Technology, Palantir and others, will greatly help in sifting through large amounts of data and visualize the relationships while combining information from many data sources into a single view. The key here is quality data, not just any data but good quality data from good quality sources, combined with better analytic tools and robust visualization systems.

I know cloud computing is getting to be one of those phrases that no one wants to hear but it will continue to change how we work. The real power behind cloud computing, especially services like EC2 and Google Apps, is not virtualization but process distribution. Processes used by search engines to respond to search requests translate well to cloud computing because they fit nicely into MapReduce techniques. On the other hand, document workflow with thirdparty solutions cannot as easily take advantage of MapReduce solutions. BrightPlanet has begun evaluation how to best distribute high-CPU intense processes like linguistic processing, entity extraction and machine translation through a cloud distributed workflow to scale and speed up the document processing systems. One of the biggest hurdles is going to be working out the licensing of 3rd party software, surprisingly, most have not thought about how to charge for a license that might only exist for a fraction of a day. We will be blazing some new ground in this area later this year.

ArnoldIT Comment

BrightPlanet surfaced in an ArnoldIT.com review of specialized systems to extract data and information not easily acquired by traditional crawlers. The system performed well and earned high marks from the ArnoldIT.com engineers because BrightPlanet's technology "played well with others". Too many specialized systems are difficult to integrate into the heterogeneous systems that many clients use for intelligence activities. If you are in the business or government intelligence market, you will want to give BrightPlanet a test drive. We recommend the system for Deep Web content access.

Stephen E. Arnold, October 6, 2009

