The following contains only Unclassified Information.

Tyson Johnson
tjohnson@brightplanet.com
905-510-0750
“Over 10 years of Deep Web Harvesting expertise, and 9 years working directly with U.S. Intelligence Agencies to fight the War on Terror.”
War on Terror

“OSINT (Open Source Intelligence) can provide up to 90% of the information needed to meet most U.S. intelligence needs.”

Deputy Director of National Intelligence, Thomas Fingar
April 1, 2008, USA TODAY
“Big Data”

Structured Data: Dark Data

Unstructured Deep Web Data

Hybrid Data
“Big Data”

‘Unstructured’ Fast Facts:
- 10s of Millions of photos uploaded on Facebook every hour
- 400 Million tweets on Twitter each day
- 500 Million Internet-connected devices in 2003
- 12.5 Billion Internet-connected devices in 2010

Sum of all stored human knowledge today:
- 1,200 exabytes (1 exabyte = 1 Billion gigabytes)
- If all this information were placed on CDs and they were stacked, the CDs would form five separate piles that would reach the moon (382,500 kilometers for each pile)

from Toronto Star, Saturday, June 22, 2013, WD4
Big Data on the Deep Web

The Web is composed of unstructured data.

Most of the Web's information is buried behind the Surface Web, far down on dynamically generated sites that standard search engines cannot find, the Deep Web.

The Deep Web cannot be searched by traditional search engines because they cannot “see” or retrieve content behind the indexed “surface.” These pages are largely “unknown or hidden” until they are accessed dynamically as the result of a specific query search.

The Deep Web is orders of magnitude larger than the Surface Web.
The Problem

Search Engines return popular links...not content.
Harvesting “Topic Specific” Big Data from the *Deep Web*

Congressman Peter King

Silo Search | Document Rank | Semantic Analytics
---|---|---
Enter keyword query below
"Peter King" between
Search Results | Topic Cluster | Link Analysis
Results for "Peter King" (98 of 68,285)

31 Million Links

98 docs

Congressman Peter King Threat Assessment Silo
Then, Creating New Intelligence from Big Data

Integrate with any database.

Visualize “unknown and hidden” relationships
The Process

Harvesting Big Data for Analytics

DeepHarvester® Workbench

Deep Web Content Silos (Topic Specific)

Analytics (Analyze, Research, Track, Monitor, Alert)

• Beyond the Surface
• Any Content Type
• Any Source
• Any Language
• Big-Data

Harvest

• Tag, Tune, Normalize & Enrich “Unstructured” Content to “Semi-Structured” for Analytics

Curate

• Faceted Search
• Topic Clustering
• Semantic Categorization
• Link Analysis
• Multiple 3rd Party Options

Analyze
A Significant Technology Challenge
(Harvesting Unstructured Content from the Web)

- **Bad HTML**
  - Offer multiple “cleaning” tools

- **Many Document Types**
  - Offer multiple conversion options

- **188 Language/Encoding pairs**
  - Arabic: أنا قادر على أكل الزجاج و هذا لا
  - Urdu: میں کانچ کھا سکتا ہوں او
  - Russian: Я могу есть стекло
  - Spanish: El pingüino Wenceslao hizo
  - Offer multiple languages and encodings

- 4.13% follow w/3c standards (Opera Study)
- Only 50% w/3c validation are actually valid
- Bad HTML, missing tags, wrong syntax, etc.
- Many “cleaning” and output parsing options
- All file format (MS-Word, PDF, PPT, etc.)
- 100s of document encodings
- Full language tokenization support
- Complex problems to resolve!
Deep Web Harvester

Find & Harvest

- 8 Patents
- Auto Scripts Configure new Deep Web Sites
- Returns content not links
- Queries go directly to relevant content
- Skips unrelated content, links, etc.
- "Search" entire site, not just visible content
- Advanced search rules (date range filters i.e.)
- Requires less bandwidth & storage
- Result: Topic Specific harvesting of “unknown and hidden” content

1) Automated Configuration creates Deep Website “Portals”
2) Harvesting relevant content directly to Deep Silos
3) Harvesting from Dynamic D-Bases behind the Surface

2) Harvesting does NOT ingest unrelated content.
An Automated Force Multiplier

...vs. hundreds of “experts” writing thousands of queries
Intelligence Cycle ROI

**Problem:**
Analysts need to spend more time on analysis & production to add value

Traditional searching is inefficient and time consuming

Confidence on decision-support analysis based on search results is low

**Solution:**
Analysts maximize value by focusing on analysis & production. Collection & filtering is automated

Time efficiencies equal to one FTE per year in hours saved for every two analysts utilizing BrightPlanet technologies

High confidence on decision-support analysis
The Process
Normalizing Unstructured Content to Semi-Structured

DeepHarvester®
Workbench

Deep Web
Content Silos
(Topic Specific)

Analytics
(Analyze, Research,
Track, Monitor, Alert)

- Beyond the Surface
- Any Content Type
- Any Source
- Any Language
- Big-Data

- Tag, Tune, Normalize & Enrich “Unstructured” Content to “Semi-Structured” for Analytics

- Faceted Search
- Topic Clustering
- Semantic Categorization
- Link Analysis
- Multiple 3rd Party Options

Harvest
Curate
Analyze
Creating New Intelligence
Dashboard
Creating New Intelligence

1. Faceted Search
2. Topic Clustering
3. Link Analysis
4. Semantic Ranking
5. People Finder
6. Geo Locate
3rd Party Analytic Technology Options

Creating New Intelligence

Link Analysis

Deep Web

Analytic Silo

Topic Clustering

Faceted Deep Search

Semantic Document Ranking

Geospatial Analysis

Situational Analysis

Semantic Analysis

People Finder
Chinese Infrastructure Silo
Any Language, Any Source, Any Scale

Language Agnostic, Source Agnostic, BIG DATA
New Intelligence
from Chinese Message Board Deep Web Silo

Geo Tagging & Proximity Analysis showing 20 year Hydroelectric Dev. Strategy through 2030.

(Harvested from a 4 year old message board.)
## Intellectual Property

**Patents AND Non Patent Literature Prior Art**

**Ranked by Concept & Keywords**

<table>
<thead>
<tr>
<th>Rank</th>
<th>Claim 21</th>
<th>Document Title</th>
<th>Document Type</th>
<th>Final URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>no title</td>
<td>Known Art - Examined</td>
<td>PDF</td>
<td><a href="http://bps2k8-0031b:8330/LocalHarvest/Examined/TPCHEM_INTERNET_INTID.PDF.doc">URL</a></td>
</tr>
<tr>
<td>2</td>
<td>Epoch Communications Newsroom</td>
<td>Known Art - Examined</td>
<td>PDF</td>
<td><a href="http://bps2k8-0031b:8330/LocalHarvest/Examined/TPCHEM_INTERNET_INTID.PDF.doc">URL</a></td>
</tr>
<tr>
<td>3</td>
<td>no title</td>
<td>Known Art - Examined</td>
<td>PDF</td>
<td><a href="http://bps2k8-0031b:8330/LocalHarvest/Examined/TPCHEM_INTERNET_INTID.PDF.doc">URL</a></td>
</tr>
<tr>
<td>4-10</td>
<td>FLEX Design and Management Strategy for Sc Known Art - Unexamined</td>
<td>Known Art - Examined</td>
<td>PDF</td>
<td><a href="http://bps2k8-0031b:8330/LocalHarvest/Examined/TPCHEM_INTERNET_INTID.PDF.doc">URL</a></td>
</tr>
<tr>
<td>11</td>
<td>no title</td>
<td>Known Art - Examined</td>
<td>PDF</td>
<td><a href="http://bps2k8-0031b:8330/LocalHarvest/Examined/TPCHEM_INTERNET_INTID.PDF.doc">URL</a></td>
</tr>
<tr>
<td>12</td>
<td>no title</td>
<td>Known Art - Examined</td>
<td>PDF</td>
<td><a href="http://bps2k8-0031b:8330/LocalHarvest/Examined/TPCHEM_INTERNET_INTID.PDF.doc">URL</a></td>
</tr>
<tr>
<td>13</td>
<td>no title</td>
<td>Known Art - Examined</td>
<td>PDF</td>
<td><a href="http://bps2k8-0031b:8330/LocalHarvest/Examined/TPCHEM_INTERNET_INTID.PDF.doc">URL</a></td>
</tr>
<tr>
<td>14</td>
<td>no title</td>
<td>Known Art - Examined</td>
<td>PDF</td>
<td><a href="http://bps2k8-0031b:8330/LocalHarvest/Examined/TPCHEM_INTERNET_INTID.PDF.doc">URL</a></td>
</tr>
<tr>
<td>15</td>
<td>no title</td>
<td>Known Art - Examined</td>
<td>PDF</td>
<td><a href="http://bps2k8-0031b:8330/LocalHarvest/Examined/TPCHEM_INTERNET_INTID.PDF.doc">URL</a></td>
</tr>
<tr>
<td>16</td>
<td>no title</td>
<td>Known Art - Examined</td>
<td>PDF</td>
<td><a href="http://bps2k8-0031b:8330/LocalHarvest/Examined/TPCHEM_INTERNET_INTID.PDF.doc">URL</a></td>
</tr>
<tr>
<td>17</td>
<td>no title</td>
<td>Known Art - Examined</td>
<td>PDF</td>
<td><a href="http://bps2k8-0031b:8330/LocalHarvest/Examined/TPCHEM_INTERNET_INTID.PDF.doc">URL</a></td>
</tr>
<tr>
<td>18</td>
<td>no title</td>
<td>Known Art - Examined</td>
<td>PDF</td>
<td><a href="http://bps2k8-0031b:8330/LocalHarvest/Examined/TPCHEM_INTERNET_INTID.PDF.doc">URL</a></td>
</tr>
<tr>
<td>19</td>
<td>no title</td>
<td>Known Art - Examined</td>
<td>PDF</td>
<td><a href="http://bps2k8-0031b:8330/LocalHarvest/Examined/TPCHEM_INTERNET_INTID.PDF.doc">URL</a></td>
</tr>
<tr>
<td>20</td>
<td>no title</td>
<td>Known Art - Examined</td>
<td>PDF</td>
<td><a href="http://bps2k8-0031b:8330/LocalHarvest/Examined/TPCHEM_INTERNET_INTID.PDF.doc">URL</a></td>
</tr>
<tr>
<td>21</td>
<td>no title</td>
<td>Known Art - Examined</td>
<td>PDF</td>
<td><a href="http://bps2k8-0031b:8330/LocalHarvest/Examined/TPCHEM_INTERNET_INTID.PDF.doc">URL</a></td>
</tr>
<tr>
<td>22</td>
<td>no title</td>
<td>Known Art - Examined</td>
<td>PDF</td>
<td><a href="http://bps2k8-0031b:8330/LocalHarvest/Examined/TPCHEM_INTERNET_INTID.PDF.doc">URL</a></td>
</tr>
<tr>
<td>23</td>
<td>no title</td>
<td>Known Art - Examined</td>
<td>PDF</td>
<td><a href="http://bps2k8-0031b:8330/LocalHarvest/Examined/TPCHEM_INTERNET_INTID.PDF.doc">URL</a></td>
</tr>
<tr>
<td>24</td>
<td>no title</td>
<td>Known Art - Examined</td>
<td>PDF</td>
<td><a href="http://bps2k8-0031b:8330/LocalHarvest/Examined/TPCHEM_INTERNET_INTID.PDF.doc">URL</a></td>
</tr>
<tr>
<td>25</td>
<td>no title</td>
<td>Known Art - Examined</td>
<td>PDF</td>
<td><a href="http://bps2k8-0031b:8330/LocalHarvest/Examined/TPCHEM_INTERNET_INTID.PDF.doc">URL</a></td>
</tr>
<tr>
<td>26</td>
<td>no title</td>
<td>Known Art - Examined</td>
<td>PDF</td>
<td><a href="http://bps2k8-0031b:8330/LocalHarvest/Examined/TPCHEM_INTERNET_INTID.PDF.doc">URL</a></td>
</tr>
<tr>
<td>27</td>
<td>no title</td>
<td>Known Art - Examined</td>
<td>PDF</td>
<td><a href="http://bps2k8-0031b:8330/LocalHarvest/Examined/TPCHEM_INTERNET_INTID.PDF.doc">URL</a></td>
</tr>
<tr>
<td>28</td>
<td>no title</td>
<td>Known Art - Examined</td>
<td>PDF</td>
<td><a href="http://bps2k8-0031b:8330/LocalHarvest/Examined/TPCHEM_INTERNET_INTID.PDF.doc">URL</a></td>
</tr>
<tr>
<td>29</td>
<td>no title</td>
<td>Known Art - Examined</td>
<td>PDF</td>
<td><a href="http://bps2k8-0031b:8330/LocalHarvest/Examined/TPCHEM_INTERNET_INTID.PDF.doc">URL</a></td>
</tr>
<tr>
<td>30</td>
<td>no title</td>
<td>Known Art - Examined</td>
<td>PDF</td>
<td><a href="http://bps2k8-0031b:8330/LocalHarvest/Examined/TPCHEM_INTERNET_INTID.PDF.doc">URL</a></td>
</tr>
<tr>
<td>31</td>
<td>no title</td>
<td>Known Art - Examined</td>
<td>PDF</td>
<td><a href="http://bps2k8-0031b:8330/LocalHarvest/Examined/TPCHEM_INTERNET_INTID.PDF.doc">URL</a></td>
</tr>
<tr>
<td>32</td>
<td>no title</td>
<td>Known Art - Examined</td>
<td>PDF</td>
<td><a href="http://bps2k8-0031b:8330/LocalHarvest/Examined/TPCHEM_INTERNET_INTID.PDF.doc">URL</a></td>
</tr>
</tbody>
</table>

### Known Prior Art

- **Conventional Sources** (yellow)

### New Unexamined & Unknown Content

- **Open Source Art and USPTO Content Harvested by BP** (green)

---

*Image of a spreadsheet showing ranked claims for intellectual property.*
Portfolio Management
Product Analysis and Ranking

Challenge:
To automate the process of identifying and ranking products for potential infringement and/or licensing opportunities against any sized patent portfolio.

1. Potentially hundreds or thousands of patents and claims

2. Potentially hundreds or thousands of competitor products

3. Automating the process for identifying and ranking those documents that may be the most interesting
I.P. Claim Mapping

Ranked Products vs. Competition/Portfolio Claims

Documents ranked specific to threaded claims, integrated directly into the SLW Claim Map using a near seamless format.

Panoramic Chart for Competitors Products vs. Portfolio Claims

<table>
<thead>
<tr>
<th>Title</th>
<th>Settable spotting</th>
<th>Biodegradable downhole tools</th>
<th>Cement</th>
<th>Misc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patent</td>
<td>7,119,084</td>
<td>7,220,879</td>
<td>7,382,263</td>
<td>7,382,263</td>
</tr>
<tr>
<td>Assignee Name</td>
<td>Halliburton Energy</td>
<td>Halliburton Energy Services, Inc.</td>
<td>Halliburton Energy</td>
<td>Halliburton Energy</td>
</tr>
<tr>
<td>Total Claims</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>25</td>
</tr>
<tr>
<td>Claims (Independent)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Technology Category (Alphabetical, Desc.)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Settable or Cement Compositions</td>
<td>29%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sensing Fluids</td>
<td>10%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Biodegradable downhole tools and methods</td>
<td>8%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Patent Score</th>
<th>Links to Documents</th>
<th>Technology Categories, Scope Concepts &amp; Keywords</th>
<th>Threaded Independent Claim</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patent Document Score</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

23
Social Media Harvesting
Message Boards, Blogs, Twitter, Facebook, LinkedIn

Itchin for some meth (@ The Cave)
4sq.com/v2ldsM

from Sioux City
Iowa, US
View Tweets about this place

The Pioneers of the Deep Web
The Case for Online Pharma Fraud

“We have more fakes than real drugs in the market”

Christopher Zimmermann
World Customs Organization
Anti-Counterfeiting and Piracy Coordinator
Pharma Case Study #1

Online Pharmacies
Counterfeiting, Diversion, Theft
Customer Problem

U.S. pharmaceutical companies are losing billions of dollars each year to online counterfeiting, fraud, and diversion.

Recent studies estimate the counterfeit drug industry to cost over $200 Billion a year to pharmaceutical providers.
Technical Challenge

• Over 50,000 websites (*found in the past 12 months*) exist that are actively advertising drugs online.

• Over 100,000 new domains are registered on a daily basis.

• Pharmaceutical companies need to target hubs to stop thousands of websites at a time.
Methodology

Step 1: Identifying Sites Selling Drugs Online

- Monitoring Newly Registered Domains
  - Pharma needs a way to monitor new domains registered on a daily basis, looking for new OLP’s
  - Harvest from domains with certain keywords in domain name

- Spam Email Lists
  - Monitor known email spam lists to identify new domains

- Source Discovery
  - Source Discovery Completed by Investigators
Methodology

Step 2: Tagging Sites for Indicators

Over 35 Indicators

- Whols Data
- MetaData
- E-Commerce Information
- Document Properties
Solution
Silo and Reports to Contribute to Evidence Packages

Over 1 Million Documents Harvested from Online Pharmacies
Over 40K Starting Domains
Duplicate e-mail address found on different domains
Enrich “Unstructured” Content to “Semi-Structured” for Analytics
“Art of the Possible”

Link Analysis to Identify Network Relationship of Domains

Graph Changes Over Time

Online Pharma
Week of 10-15-2012 to 10-22-2012
Previously Unharvested Domains
Pharma Case Study #2

Tracking Theft of Trade Secrets Online
Another Pharma Problem

Pharmaceutical companies are losing billions of dollars because of diversion and theft of experimental compounds being sold on international message and trade boards.

How can they be identified, disrupted, and dismantled?
Technical Challenge

At any given time, pharmaceutical companies are creating and testing experimental compounds that have billions of dollars invested in them.

With thousands of online message and trade boards domestically and internationally, it becomes nearly impossible to monitor the web for these chemicals at scale.
Methodology

Step 1: Harvesting from Message and Trade Boards

• Harvest content from over 1,000 message and trade boards (Ex: alibaba.com)
  – Harvest all content from each site at scale and tag content
• Identify new message and trade boards
  – Newly registered domains
  – Spam email lists
  – Source discovery
Methodology

Step 2: Tagging Sites for Indicators

Over 35 Indicators
- Product Name
- Experimental Chemical Names
- WhoIs Data
- Chemical Formula Breakup
- CAS#
“Art of the Possible”

Graph Changes of New Domains & Compound Elements Over Time

Top 10 Domains
MKP4-New MK Document Report
Week of 9-2-2012 to 9-8-2012

Visual Analytics to Determine Relationships Among Harvested Domains
“Art of the Possible”

- Filter websites based on CAS numbers
- Integration with current analytic technologies
- Custom output to use within your own database
Pharma Case Study #3

Competitive Intel
Monitoring the Competition Online
Customer Problem

Having a competitive advantage during the development of a new drug is the difference between a successful product and failed development effort. You need to know what your competition is saying and doing.
Technical Challenge

Many websites, especially those of the top pharmaceutical companies, are millions of Web pages in size. It’s crucial to know what is new and changed to stay ahead of the competition.

- Pfizer Domain Size: 29,000 Pages
- Bristol-Myers Squibb: 306,000 Pages
- Merck Domain Size: 939,000 Pages
- Johnson & Johnson: 22,900 Pages
Methodology

1.) Harvest all content from the Top 25 pharmaceutical websites.

2.) Release data to the client in the form of a customizable, actionable alert.
Solution

Daily Email Alert of New and Modified Content

Deep Web Monitor

Top Alerts For "Device Reimbursement Terms"

New Alerts: 11

Term Matches For This Alert: [acth, acyclovir, botox, filgrastim, glatiramer acetate, memantine, namenda, natalizumab, risperidone, rituximab, testim]

NAMENDA (MEMANTINE HYDROCHLORIDE) TABLET [BRYANT RANCH PREPACK]
Drug Class SetId Label Type: Human Drugs Animal Drugs RxNorm Names Not yet provided
NAMENDA (memantine hydrochloride) tablet [Bryant Ranch Prepack] Permanent Link:
How Supplied Patient Counseling Information Supplemental Patient Material Boxed Warning Patient Package Insert Highlights Full Table of Contents Medication Guide Rx Only DESCRIPTION
Namenda® (memantine hydrochloride) is an orally active NMDA white to off-white powder and is soluble in water. Namenda is available as tablets or as an oral solution. Namenda is available for oral administration as capsule-shaped, film-coated tablets containing 5 mg and 10 mg of memantine hydrochloride of CYP450 enzymes is unlikely. Co-administration of Namenda with the AChE inhibitor donepezil HCl does not affect the pharmacokinetics of either compound. Drugs Eliminated via Renal Mechanisms: Memantine is eliminated in part by tubular secretion. In vivo new on 02/06/2013 2:30 PM
Inclusion Term Matcher: [Memantine, Namenda]

LOSARTAN POTASSIUM AND HYDROCHLOROTHIAZIDE TABLET, FILM COATED [LAK...

of anionic exchange resins. Single doses of either cholestyramine or colestipol resins bind the hydrochlorothiazide and reduce its absorption from the gastrointestinal tract by up to 85 and 43 percent, respectively, Corticosteroids, ACTH, or glucocorticin.
False rumors and misconceptions can doom a new drug by sparking widespread public concern before it is launched.

Companies need the ability to monitor the online reputation. Biogen Idec wanted to monitor its new multiple sclerosis drug Tecfidera (dimethyl fumarate), that was slated for launch in 2013.
Technical Challenge

Thousands of Web sources where patients, clinical trial nurses, caregivers, investors, and financial analysts discuss treatments daily.

Collecting, analyzing, and visualizing sentiment from thousands of diverse sources (websites, social media, message boards from both finance and health sectors) is difficult at scale.
Methodology

1. Finds websites where products are being talked about (websites, blogs, message boards, other social media)

2. Harvests individual Web pages mentioning specific products (over 5,200 Web pages in 2012) into a Deep Web Intel Silo.
Methodology

Search entire data set with custom queries

Drill-down by Source Type and individual domain

Research why people are switching off a competitor’s drug
Methodology

A Deep Web Researcher analyzes the web page, extracting opinions and verifying date, sentiment, and uniqueness.

Step 3

Unique sentiments are summarized, categorized, and pushed into a chronologically ordered report. Visualizations help show how sentiment changes over time.

Step 4

A Deep Web Researcher analyzes the web page, extracting opinions and verifying date, sentiment, and uniqueness.
Solution – Product Launch Report

Rank mentions of competitor treatments

Graphs to help visualize changes in volume and sentiment

Extracted opinions organized by Source, Sentiment, and Date.

Top 10 M.S. Treatment Mentions (found on the same web page as keyword)
1. Tyzabni - 1,085
2. Copaxone - 1,003
3. Gilenya - 908
4. Avonex - 900
5. Rebif - 779
6. Betaseron - 647
7. Aubagio - 473
8. Rituxan - 116
9. Novantrone - 57
10. Tysabri - 30

Sentiment: M.S. Message Boards

New M.S. Message Board posts (6 months trailing)
OEM Trade Board Reviews

OEM Trade Board Report Summary

- Ongoing monitor for stolen or counterfeit products on B2B boards (Kijiji, eBay, Alibaba, etc)
- Harvesting at scale
  - Products within proximity to manufacturing
  - Specific models of products (pre-release)
  - Accessories (batteries, etc)

Proactively identify suspect product online, create “targeting packages” to begin executing buys online, leading to loss reduction / loss recovery