



## BrightCloud's Classification System Datasheet: Building the BrightCloud Master Database

### Key Internet Statistics (2008)

**Hosts:** 542 million <sup>1</sup>  
**Pages:** over 1,000,000,000,000 (trillion) <sup>2</sup>  
**Corresponding IP Addresses:** 200 million <sup>3</sup>  
**Average number of new websites each day:** over 100,000 <sup>4</sup>  
**Average number of dead websites daily:** 100,000 <sup>4</sup>  
**Annual growth rate:** 15-20% <sup>2</sup>  
**Projected size of the internet in 2012:**  
1 billion websites, 500 million IP addresses

<sup>1</sup> Internet Systems Consortium, <http://www.isc.org/>

<sup>2</sup> [www.google.com](http://www.google.com)

<sup>3</sup> [www.netcraft.com](http://www.netcraft.com)

<sup>4</sup> [www.domaintools.com](http://www.domaintools.com)

### Overview

As shown in the table to the left, the public internet is a large, rapidly changing entity that is growing quickly. Correspondingly, there are a variety of requirements that need to be met in order to enable an enterprise to implement an acceptable internet usage policy- web sites have to be categorized based upon content in order to block access to sites that are not work related, security web sites have to be accurately categorized so that internet threats are not allowed into the end customers' networks. The challenge for a web filtering vendor is how to effectively provide these services while keeping up with the size and growth of the internet.

### The Situation

Traditionally web filtering vendors have attempted to categorize the internet using human classifiers who categorize one site at a time. Humans are a highly accurate means to categorize web sites- for the first hour. But over time, fatigue- or domain expertise, or cultural biases, among other reasons- cause humans to make mistakes with an average error rate of 10-15%, even for obvious categories. One solution is that vendors sometimes have several people categorize the same web site, which reduces that error rate. Unfortunately humans suffer from another problem- it isn't cost effective to employ enough of them to keep up with the size and growth of the internet. A vendor would have to employ a thousand people working 24 hour days just to categorize the new web sites that appear *each day*. Another solution is required.

### The BrightCloud Solution- Multiple Classifiers

As shown in the diagram on the next page, BrightCloud utilizes a variety of input sources to build the BrightCloud Master Database- the largest and most accurate URL database in the world. The main classification sources are:

- **Content Classification-** Content classification typically relies on the words and images on a web page, and compares those to examples that represent a given category. BrightCloud utilizes both automated and human classification. While people are an important classification source at BrightCloud, their role is primarily used to provide examples of URLs in a given category to automated machine learning classifiers, such as BrightCloud's Maximum Entropy Discrimination classifier.
- **Reputation Classification-** BrightCloud also utilizes rich web page information gathered from a variety of sources to provide a clear picture of an individual web page's trustworthiness. BrightCloud's Reputation Classification is utilized both internally to categorize web sites that pose a security risk (for example,



web pages that are part of a phishing attack), as well as accessed externally via the BrightCloud Service.

- **BrightCloud Threat Operations Center**- BrightCloud's TOC is focused on tracking down and identifying URLs that fall in the security categories such as Malware, Spyware\Adware, Phishing and other Frauds, Keyloggers and Monitoring Software. The TOC utilizes a series of tools and technologies to populate the security categories, including AV scanning of URL traffic.

Each of these classifiers feeds in to the BrightCloud Master Database. Utilizing correlation algorithms and other analytical tools, BrightCloud generates a category and reputation score for a given URL that is provided via the hosted BrightCloud Service.

### Leveraging Data in the BrightCloud Master Database

Because the BrightCloud Master Database is the largest categorized database of URLs in the world- six times larger than the URL databases of the market leaders- BrightCloud has a very broad insight into relationships between URLs. For example, some IP addresses host predominantly Adult content, or may be part of a Trojan network.

This relationship can be fed into the BrightCloud reputation classifiers, which also utilize additional information available about web sites such as lifespan or ownership changes, to define a reputation score for a given web site.

Correlation algorithms are used to compute a reputation score for a given domain that can be used to supplement content based classification. End customers can implement policy based on a richer set of information about the URL, or in the event that a content classification is not available, policy can still be implemented. End customers have a richer, more effective policy solution.

### Summary

BrightCloud has adopted a technologies and processes that enable BrightCloud to tackle the tough job of categorizing internet, and keep up with its growth and churn. By leveraging multiple classification sources including content, security, and reputation classifiers, BrightCloud has been able to build the largest and most accurate URL database in the world, bringing end customers the most complete web filtering solution available.

For more information, please visit: [www.brightcloud.com](http://www.brightcloud.com)

### BrightCloud Master Database Classification Sources

