

Content filtering technologies are only as good as the database of both good and bad content found on the Internet. Collecting data about the ever changing Internet enables us to continuously and proactively modify our content database. CyberPatrol's proprietary SiteCAT content engine is a proven technology that has been continuously improved over the past ten years.

The SiteCAT content engine is used by the full CyberPatrol product line-CyberPatrol Parental Controls, SiteSURV and SiteSURV*plus*.

Innovative Modular Engines

SiteCAT analyzes web sites on a number of levels using individually activated modular engines that include; Regular Expression Engine, Script Engine (dynamically loads C# script files), Keyword Engine, and Link Analysis Engine. Each is capable of reaching decisions using embedded scoring algorithms. SiteCAT uses its own Core Result Engine to categorize sites using all the gathered information. Using modular engines makes it easy for CyberPatrol to develop and add new modules as Internet threats change.

Site Review Tool (SRT)

SiteCAT's Site Review Tool (SRT) is used by company and partner researchers to review sites and classify them when automated means cannot determine the classification. The Site Review Tool is a self contained web browser and voting tool. Users load a site list from the SiteCAT Master Database and review each site one at a time voting on the categories to place that site into. They can automatically review the source, reference links, and any included information found by SiteCAT. This human reviewed information is immediately incorporated into the machine analysis to improve the classification of web sites.

What are web crawlers?

A web crawler is an application that is used to visit and download pages of web sites. CyberPatrol is constantly looking for sites with our Web Crawlers which are hosted in data centers in across the US.

SiteCAT continually scans the Internet and categorizes web sites at the rate of over one million sites per day. The SiteCAT database currently contains over 22 million domains representing approximately 220 million individual URLs. As new domains are found through links in existing known sites they are automatically added to the database and scheduled to be crawled. We continually scan the Internet categorizing web sites assuring the most complete and up to date content database available.

The crawling of websites is the most prevalent way for CyberPatrol to learn about new websites. When a website links to a domain we have not seen before, the domain is automatically added to our database so it can be crawled and further analyzed. Next, the sites linked to by that site are also crawled and the list quickly grows. This is sometimes called Web Spidering.

Why are web site pages downloaded?

During the web crawling process actual web pages are downloaded. Each page from a site is then analyzed for the links it contains as well as the general intent of the site. The number of pages on each site can vary depending upon how many pages are present. CyberPatrol always attempts to download a minimum of five pages from a domain before any determination of the intent or type of content found on the site is made. If the site only has a single page we will work with the available content. Obviously, the most ideal situation would be to have numerous pages to evaluate to ensure correct categorization.

Does this impact the site's servers?

The CyberPatrol crawlers are very careful when visiting websites to ensure they do not add to existing load or impact the users' experience. In fact a normal user or the GoogleBot will pull far more content than the CyberPatrol crawlers. Images are never downloaded from the pages since they are usually the largest part of the site bandwidth usage which allows CyberPatrol to easily gather the information they need.

Why do the stats continuously change?

Crawler Stats

The Crawler Stats are broken down into the past 24 hours, 7 days and 30 days. This represents the total number of Domains that were crawled during that period.

The number of sites crawled in a given day can vary quite a bit. The speed of our internet connection, the number of crawlers concurrently running, and the complexity of the pages themselves all contribute to the speed of the crawlers. If the crawlers hit a large block of domains with single pages, for example, they go much faster than a site with many pages and links.

Web Filter Stats

The Web Filter Stats are split into Good Domains, Bad Domains, Unknown Domains, and Total Lookups. These numbers can be a bit confusing due to the way the CyberPatrol client code works. The time period interval for this is per the past 24 hours. The number continuously changes throughout the day. The numbers represents the number of unique domains our users have asked about. So if 100 people asked about CNN.com, it would only appear once even though there were 100 requests made.

Domains

Good Domains are domains we include on our allowed list. But the CyberPatrol Parental Control client will only have to look this domain up once, and then it is cached client side for a period of time. This makes local look ups fast.

Bad Domains are domains found on one of our default block lists. Again, CyberPatrol Parental Control client will only have to look this domain up once, and then it is cached client side for a period of time making local look ups fast.

Unknown Domains are domains looked up that did not have an exact match, or we have not categorized within the current database. This doesn't mean the site has not been visited, but perhaps we found conflicting evidence on the site or were unable to reach them. These domains

are stored in absolute format and may appear high at times. For example, if a user visited ads.images.cnn.com and it was not already in the database, it would be stored as a unique value. Then if another user visited money.cnn.com it would also be stored. We do not reduce the domain to its base value (cnn.com in this example) until we have crawled the location and verified that it is indeed part of the same website.

Total Lookups is the number of domains looked up by clients during this period. This number is not an addition of the above. This is the total number of questions the CyberPatrol Lookup servers have been asked during the past 24 hours.

As you can see, web crawlers are one small piece of the proprietary technology that makes up SiteCAT. The information and statistics it gathers are a real indicator of the environment of the Internet on a day to day basis.

The modular categorization engines along with actual human review make this technology innovative and it delivers accurate content. CyberPatrol products have been designed to incorporate this comprehensive, under the hood capability, to provide strong, accurate filtering for our customers. SiteCAT is the core technology for all CyberPatrol products.