# Global Information Assurance Certification Paper

# Web Content Security
# The Requirement, Methods Available and Considerations

By Stephen Gray

## Introduction

In most organizations today web access is universal. Almost all employees with access to a PC within an organization have Internet access. With this access comes responsibility to ensure that the Internet access is used in such a way that it:

- Does not adversely affect staff productivity
- Does not contravene acceptable behaviour policies of the organization
- Does not compromise the information or information resources within the organization.

This paper examines the types of web content security, and employee Internet management solutions (EIM) available to organizations to ensure Internet usage is legitimate and safe as possible. It also examines the deployment considerations of different solutions and the relative strengths and weaknesses of these solutions.

## The Requirement

### *Malicious Software*

Nowadays, most organizations have some form of perimeter security in place. These systems include Firewalls and Intrusion Detections Systems (IDS). Many of these systems often make the policy assumption that everything on the outside is dangerous and everything on the inside of the firewall is trusted. This is often reflected in default device policy configurations which allows traffic from inside networks to outside Internet hosts, while blocking all traffic originating from the Internet destined for inside hosts.

Rather than trying to penetrate the plethora of perimeter defences, and risk detection by IDS, hackers can develop malicious software, and attempt to infect a user's host within an organization. The malicious software can be downloaded from the web, or transmitted by email. The user is often unaware of the malicious nature of the software they have downloaded or opened by email. This type of software is often referred to as Trojans. Once installed on an internal PC, the Trojan takes advantage of the aforementioned commonly used default perimeter security policies and initiates connections to external Internet hosts. These Trojans or agents then poll Internet "masters" for instruction, which may include, gathering sensitive information from Internal servers, collecting user activity such as keystrokes, or being used as a member of a distributed denial of service (DDOS) attack.

Some well-known examples of malicious software that operate in this manner include:

- Back Orifice / Back Orifice 2000
- Trinoo

- SubSeven

### *Internet Usage – Employee Internet Management (EIM)*

Another role of Web content security is controlling Internet usage. Many organizations put a price on the cost of virus eradication, or the cost of an intrusion, but often fail to calculate the cost of unproductive and non-business related Internet usage.

For many organizations the cost of lost employee productivity by what is termed "Cyberslacking" is a much greater than the cost of intrusions or virus eradication. Cyberslacking has an impact in the following areas.

- Resources and infrastructure. In a majority of organizations, non-business related Internet usage adds to a significant burden on network resources. Unlike a DDOS attack, which consumes bandwidth and costs a company for the duration of the attack, constant non-business related web usage is always adversely affecting legitimate Internet usage.
- Cost of Internet usage. Many organizations pay for Internet usage by the amount of data downloaded. Non-business Internet usage represents a significant portion of the Internet bill.
- Access to unsuitable Internet sites containing pornography, acts of violence, racial hatred and other controversial content. Such material downloaded from the Internet can lead to sexual harassment charges, adverse publicity and employee dismissals.

## Methods of Controlling Web Access

There are a number of ways available to an organization to control web access. These include.

- Control Lists
- Real-Time Analysis

### *Control Lists (URL Method)*

Devices employing the control list method typically integrate with an organizations firewall, proxy server or Internet caching device. When a web user requests access to a site, the firewall/proxy/cache consults the control list server and determines if the policy for the requesting user will permit or deny access to the requested site. The control list is typically a database of URLs maintained by the vendor of the product. The database of URLs is categorised into a number of different categories. The degree of categorisation and granularity varies from vendor to vendor, but most vendors offer at least the broad categories of Business, Sports Religion, Politics, Entertainment, Computers, Pornography and Racism. Much like virus signatures, the control lists databases are continually updated by the vendors and made available for download to subscribing customers via the Internet.

The strengths of Control List Techniques are:

- Speed. If the URL Matches, the user is either permitted or denied access depending on the policy defined for that particular user or group.

- Scalability. Since URL pattern matching is not overly CPU intensive, control lists can scale to support a large number of concurrent users.
- Precision. Since the URL blockers sites are categorised by the vendors there is little chance of over-blocking or false positives (you may not always agree with the categorisation).

The weaknesses of Control Lists techniques includes:

- Coverage. Not all sites are covered in the database, and some sites not categorised. By in large, most of the popular sites are covered, but there is the chance that some sites will not be detected. Most control list products are of American origin and hence have an American site bias. International sites may not be as comprehensively covered.
- Lack of in-built virus or malicious code protection. Products employing the control list method generally have no control over virus download other than preventing user access to sites that are known to contain viruses, such as hacker sites.

Some of the more popular products employing the Control List method include:

Websense (http://www.websense.com)
SurfControl (http://www.surfcontrol.com)

### *Real Time Analysis Method*

Real Time Analysis is the other method employed by vendors to control web access. When a web user requests a web site, the Real Time Analyser device downloads the web content from the requested site, analyses the content, and determines whether the web content is suitable for presentation to the requesting user.

The Real-Time Analyser performs either (and sometimes both) of the following functions.

- Scans for unsuitable content that contravenes organisational policy
- Detect malicious code.

A Real-Time analyser generally employs Keyword Matching or Pattern Matching to detect unsuitable content. The two methods used are:

- Keyword Matching. Keywords are given weightings depending on their level of offence. Each occurrence of a keyword increments the count by its relative weighting. If the count exceeds a defined threshold, then the web site is not deemed suitable for download and blocked. Keyword matching is particularly prone to giving false positives (over-blocking), eg A Breast Cancer site may be blocked due to the regular occurrence of the word "breast". Regular tuning of the keyword list is required.
- Pattern Matching. Pattern matching the next step in the evolution of Real Time Analysis that aims to overcome the limitations of Keyword matching. Numerous methods such as Bayesian Statistical Classifiers, Neural Net Classifiers, and Support Vector Machines are used to analyse the context of the web content are used to more precisely define the type of web site.

The primary strength of the Real Time Analysis method is:

- Maximum Coverage. A Real Time Analyser will catch the majority of sites it is meant to block if the thresholds are set appropriately.
- Some Real Time Analysers also include virus/malicious code detection as a complementary function

The weaknesses of Real Time Analysis are:

- Scalability. The process of downloading each web page for analysis is processor intensive, and therefore can impact adversely on response times. Adding more hardware (CPUs, Servers), possible in a cluster arrangements can alleviate this, however this adds to the cost of implementing such a solution.
- Lack of Precision. Real Time Analysis is adequate for blocking pornographic and offensive material, as the textual content is quite predictable, but difficult to configure to optimize for productive Internet usage. Restriction of other categories of sites (News, Sports, Online broking) are much more difficult to block generate text lists for.

An example of a web content security products employing Real Time Analysis is *WebSweeper* from Baltimore Technologies (www.mimesweeper.com )

## Web Based Virus Management

Apart from virus scanning at the desktop, there are two common methods used to scan Internet web content for viruses in within an organization,

- HTTP Redirection
- Proxy Scanning

### *HTTP Redirection*

HTTP redirection is a technique whereby any web pages requested by an Internal user are redirected by the firewall to a virus-scanning machine to verify the content is virus free before presenting the data to the client browser. CheckPoint, the manufacturers of Firewall-1, names their implementation of this technology Content Vectoring Protocol (CVP). It is also available for virus scanning of SMTP based email. Numerous anti-virus vendors have products that integrate with the leading commercial Firewalls.

While email redirection virus scanning works well for email messages where slight delays to scan messages are acceptable, virus scanning of web content can impose noticeable delay to web page retrieval. The problem is that web pages typically contain 20 or more individual components (mostly graphics) , which will be scanned by the virus scanner. Multiply this by the number of users Web browsing at any one time, and the virus scanner has a significant workload. Once the number of concurrent users reaches a critical limit , the performance impact can becomes large enough to cause web browsing to become unusable. Users may then look to alternative means of Internet access , such as a dialling into an ISP using a modem connected directly to their PC.

## *Proxy Scanning*

The other common technique is proxy scanning. This technique typically involves placing a virus scanning web proxy between the corporate proxy server and the Internet. The corporate proxy server is configured to forward all web requests to the virus scanning proxy. The configuration arrangement can be reversed, where user's browsers are configured to target the scanning proxy, which is configured to chain proxy its request to the corporate proxy. In environments without a corporate proxy, user's browsers must be configured to forward web requests to the virus scanning proxy.

The nature of web content is such that virus scanning web proxies suffer the same scalability issues as virus scanning by firewall content redirection. Some virus scanning proxies provide update messages to the user browser to let them know that the page is being scanned rather than allowing the browser (or user) to give up as nothing is happening in the browser window.

The advantage of a web proxy scanning solution is that once the web content has been scanned, it is stored in cache, making web browsing faster for subsequent access to scanned pages. Typically the cache performance of virus scanning proxies does not approach that of dedicated WWW proxy and Internet caching appliances, as the virus scanning proxy is usually preoccupied with its primary task of virus scanning web content.
In order to scale to support large user populations, vendors of scanning proxies recommend clustering and or load-balancing techniques be used to ensure adequate web browsing performance.

## *Web Virus Scanning Deployment Considerations*

Web virus scanning is risk versus reward proposition.
By scanning at the gateway, latency is introduced and scalability issues arise as Internet bandwidth and user population increases. This has to be weighed up against the comprehensive coverage a gateway antivirus solution can provide and the reduced risk of internal virus infection.

Virus infection risks from web downloads is mitigated a number of ways:

- **Desktop Virus protection**. If the virus scanner deployed on the workstation is up to date and also capable of blocking or cleaning web content, the need for a centralised web scanning solution is reduced.
- **Prevention of access to virus prone sites**. Most Employee Internet Management (EIM) products using the Control List method have a category, which includes Hacking and Warez sites. These are the dominant sources of web-based viruses. By preventing access to these sites with an EIM product, the risk of web virus download is reduced.

## Conclusion

There are numerous products, employing the above techniques, available to organizations wishing to control the information brought into their organization by staff. These solutions permit enforcement of company Internet access policy and help ensure that Internet access is legitimate business related. There is no single solution that will be suitable for every organization, and the choice of product will depend on a number of factors including:

- Company Internet Usage Policy and Management buy-in
- Number of concurrent web users
- Internet bandwidth
- Budget
- IT security staff resources and skills to manage the environment.
- Existing content security and anti-virus environment in place

Fortunately, most vendors of content security products provide downloadable evaluation versions (and implementation guidelines) for their products. This enables IT departments to determine the suitability of such products before committing to a particular vendors product.

## Resources.


1. Christiansen, Christian A "Employee Internet Management" (IDC White Paper) http://www.websense.com/products/resources/wp/idc.pdf
2. Hockey, Alyn "Creating High Availability and Highly Scalable Solutions for WEBsweeper 4", 15 October 2000. http://www.mimesweeper.com/products/collateral/pdfs/wsw4scalability.pdf
3. Checkpoint Software Technologies. http://www.checkppoint.com
4. Surf Control. "The Cost of Non-Business Browsing: An Illustration" 2000 http://www.surfcontrol.com/news/white_papers/pdfs/SC_Cost_of_Browsing.pdf
5. Websense WhitePaper "Internet Usage and the Workplace: It's a Now a HR Issue". http://www.websense.com/products/resources/wp/hr_wp.pdf