On-Line Copyright Protection

FINDbase LLC

1. White Paper

2. Introduction

Freedom of speech on the Internet has always been a topic to raise passionate debate and view. The fact that people can communicate with ease, share information and accomplish more makes the Internet an essential part of life. The sharing of data on the internet has been one of its main benefits and has lead to a huge wealth of available information.

Data is at risk on any network - Internet or Intranet – the magnitude of the risk often being measured by what *happens* to the data. For example, one person stealing a document and reproducing it once has less impact than 2000 people stealing content to start alternative *Disney* sites.

Very few Internet sites require account verification, the vast majority of information is made available under so-called *anonymous copyright* terms. It's analogous to making 10,000 copies of your work, and then throwing them into a hurricane. You have no control over or knowledge of where they go. Enforcing copyright is only possible if you know *where* the copyright material has gone in a manner in which you can *prevent* the access.

Freedom of Speech on the Internet is a privilege, yet many consider it their *right* to take whatever they find and do whatever they wish with it. The dubious justifications for the ***theft*** of music, art, literature and *other peoples property* can be seen on a great many of the bulletin and comment forums on the Internet.

- It is considered *cool* to steal copyright work and republish it. It's still theft.
- Peer pressure encourages many to steal music and distribute it in the form of MP3's. It's still theft.

It is to be wondered how these same people would react to the theft of their expensive music system on the grounds that *music should be free*.

Copyright and Digital Asset Management and Protection became an issue when the first eCommerce site appeared on the Internet. It became of huge concern during the *riches to rags* lifecycles of many eCommerce and dot.com sites. Such sites rely on web site content – if the content is being stolen, they are simply losing more assets.

3. When is Theft not Theft?

Simple; *when you don't get caught.* Since nothing physical has been removed, it is hard for anyone to know what has happened or quantify the loss. For example, if someone breaks into your home and steals your stuff, you would *know* that you had been robbed and call law enforcement. However, if someone periodically breaks into your home and repeatedly steals some breakfast cereal, all you know is that you are purchasing more cereal. It's a *stealth crime* – don't do it often enough to be noticeable and you won't get caught.

Even the most obvious forms of Internet abuse, copyright violations and data theft are overlooked, ignored or simply unseen giving even more encouragement to those breaking the law. Hated by many (usually those breaking the law), the Digital Millennium Act provides some protection to the victims of on-line abuse, yet this modest provision requires evidence of the crime. If you cannot prove a violation has occurred, it is difficult to determine who is to blame.

At a time when so much attention is given to credit card security and personal privacy, virtually none is given to protecting assets at the core of Internet enterprises. The answer is not hard to find. If a clothing store had their window smashed and stock taken, security could catch the culprits and law enforcement called. Possibly even a glazier to repair the damage! But anyone coming to the same stores Web site can obtain all of their pricing, stock levels, inventory analysis and much, much more *without getting caught*. In fact, it is typically possible to download complete inventory histories as a result of many sites leaving old information on their site. Old vegetables smell if they are left on store shelves – old information just occupies space on a web site and goes unnoticed. So, since there is no obvious damage, nothing physical has been removed and nobody could see what was happening; it's business as usual. Especially for the competitor who has downloaded huge amounts of competitive information.

Worse still, those stealing or illegally distributing, for example, music, continue *knowing* that the chances of getting caught are almost zero.

When is theft not theft?
When you don't get caught

eCommerce sites devoting huge amounts of money, time and resources to hyped up concepts such as "the total ordering experience" or needlessly entering the race for more Flash images typically ignore the most basic of business principles;

1. **Asset Protection**
   Does anyone really know *what* took the data *where* today?
2. **Confidentiality and Copyright**
   Is almost a joke if there is no way to stop anyone accessing the site.
3. **Quality**
   Quality Assurance – on an eCommerce site? The publish-and-pray testing

methodology is alive and well. Until a *customer* asks "can the ***I*** afford such poor quality standards?"

FINDbase LLC technologies change this landscape forever.

4.
5. Where did *Your* Data Go Today?

Log files provide an indication of when data was accessed, but provides no indication of *what* took the data or *how.*

5.1. Measuring Success by Hit Count is DUMB!

That eCommerce sites use hit counts as a measure of success can be considered another factor in many of the *riches-to-rags* stories in the media. Whilst log files show that a site has 50,000 hits per day, and the visit analysis shows the exit pages, nobody really stops to analyze ***what took the data where.*** More to the point – of the 50,000 hits, how many are hackers, competitors, spiders, robots and others wandering around the site without the means to purchase anything? 5,000 per day? 10,000 per day? Who knows. Those relying on Banner Advertising or a revenue model based on visit/hit counts simply *benefit* from exaggerated figures.

5.2. Hit Count Measurement Standards? No Way!!

Log file analysis is well known to produce idiosyncratic results that can be open to ambiguous misinterpretations. It's their very inaccuracy that makes them so popular – especially if you rely on Banner Advertising revenue as many dot.com companies. The recent XXXX attention to Banner Advertisement Fraud just highlights the tip of a problem that *relies* on inaccurate analysis hit counts.

Hit counts have long since been discredited as a means of measuring the number of visitors to a site. Hit counts have been replaced by the concept of *visits.* What's a visit? Its calculated as the entire activity from the moment someone comes to the site to the moment they leave. Obviously the duration of the *visit* determines the number of hits-per-visit and thusly the number of visits-per-day. The simplicity of the concept enables the underlying mathematics to be manipulated to produce the figures you want to present. For example, if the *same* individual makes 60 hits to a site in an hour, taking 1 minute between hits (not unreasonable), the number of visits could range from 1 to 60 depending on what you are intending to represent. Since the number of visits is often a measure of success, creative math's is tempting:-

Hits per hour	Visit Window (minutes)	Visit's per hour
60	1	60
60	10	6
60	60	1

Clearly, the measure of success (ie number of visits) can be adjusted by simply adjusting the *visit window*. Whatever *visit window* is chosen, we are still including all the hits rather than only those from real people. Current tools and techniques make poor filters, which is just as well considering that for many, the entire point of the exercise is to show as many visits as possible.

Findbase technologies filter out non-human accesses and provide a reliable indication of how many *people* are accessing a site.

Accurate *visit counts* are vital to those selling Banner advertising or otherwise basing decisions on these values. Yet the lack of metrics by which these can be measured is incredible given the number of tools available to do just this. Given the importance placed on these statistics, it is incredible that there are no rules or standards defining the calculations. Its rather like the major oil companies having different definitions of for a *gallon*. It is therefore unsurprising that *visit statistics* produced by a given tool are advertised, it is very, very difficult to change to a different tool as there would be an inevitable and potentially embarrassing/costly change in the *visit statistics*. Since the concept of a *visit* is completely open to (miss)interpretation, the situation is unlikely to change.

Findbase technologies filter out unwanted and potentially confusing accesses resulting in a more accurate count of *people visiting a site.*

The conclusion is that it is tempting for many to "enhance" or otherwise encourage more favorable *visit statistics*. This dubious practice is attracting mainstream press attention in addition to the attention of those that actively lose money as a result of bogus statistics – the Banner Advertisement Providers and ISP's. Ultimately, of course, everyone loses.

6.

7. The Ever Present Threats

The threats to on-line assets are almost limitless. From Napster, Gnutella, competitive analysis to plain theft and fraud, the perception is that nothing can be done and nobody will get caught. Fortunately something can be done and people can be caught. It requires more than complaining, it requires pro-active defense of your Digital Assets.

It is hoped that eCommerce sites have security systems fitted to their buildings to prevent unauthorized access and interference. So why not fit access security systems to your Digital Assets? They are far more susceptible to attack than physical assets (the building). Just because we cannot see the problem doesn't mean it does not exist. Detection is the first step towards protection – just look at security systems for buildings. The only difference between Physical Assets and Digital Assets is that Digital Assets are ***easier to steal.***

The threats to Digital Assets are not limited to unwelcome site accesses. They can also come in the form of poor or non-existent *Quality Assurance (QA)*. Lets go back to our clothing store analogy. When they fix the window displays, someone will finally go outside the store, look at what the customers see and make a *quality decision* based on what they see. This is done because the risks are too great *not* to do it. It's a relatively small amount of material to visualize which can be done relatively easily. Is the same technique applied to their Web Site? Who can say – but if it contains many thousands of pages, it is very, very difficult to perform the most rudimentary of checks. How is QA performed on large sites? It's a good question!

7.1. Detection is the Key to Defense

Allowing those you want to your site is the most effective means of defense. It's easier and more secure than specifying all the permutations of what you don't want.

Most eCommerce sites want real people visiting their sites. Real people are more likely to generate revenue than hackers, robots and those stealing your data. Findbase technologies can filter out these categories and more, putting you in control over who accesses your data and how.

7.2. Be Proacative !

Findbase technologies do more than *just* prevent unwanted guests. It provides invaluable information as to who accessed your site, using what, and from where. For example, its not uncommon to find "unofficial" fan-clubs on the Web. From Starwars to hampsters, there *will* be a fan-club on the Web. If these sites re-publish material from other sites we have the potential for copyright violations. Given the easy and speed of data dissemination on the Web, it is difficult and costly to determine who is using copyrighted material. Findbase technologies changes this. Part of the detection and prevention of unwanted access to a site is the ability to record information about *who and what* is making the access. This information can then be used as part of a *Web Mining* exercise to determine if the *who* is using other copyrighted materials. This proactive searching for other violations is a vital step in the prevention of future attacks.

The clear message ***You Can Be Stopped and Found*** makes life more difficult for those deliberately violating your copyright. Simply and easily, Findbase technologies enable you to not just prevent your digital assets from being misused, they enable you to proactively research other potential violations.

In real terms, *Internet Anonymity* is a myth as it is possible to get a great deal of information about those accessing an Internet site. Whilst it is possible to masquerade identities, the majority of attacks are from those without the skill or equipment to disguise their activities. The arguments surrounding the continuation of *Internet Anonymity* is as dubious as Bank Robbers arguing that their masks and guns should be

allowed in Banks. Banks have detection devices installed to monitor activities and highlight potential trouble. Findbase provides the detection systems you need to protect your assets.

Until now, the Internet has had little in the way of detection and tracking making life easy for those indulging in activities of dubious nature. Arguments that detection devices infringe on *Freedom of Speech* is are often from those finding constraints being placed on activities that should not have been allowed to continue for so long. Those evangelizing the *Open-Everything* philosophy are unlikely to agree to an *open-money* scenario – where everyone could help themselves to *their* money.

8. webQA

In every major industry sector, Quality Assurance (QA) programs are in place to ensure that the product or service actually meets the expectations of its consumers. QA procedures ensure that the correct amounts are dispensed from ATM's. QA procedures are in large part responsible for ensuring that Oil Refineries don't explode and airlines keep don't fall from the sky. To question QA procedures is an invitation to disaster; if the product/service is of highest quality, it will pass the most rigorous of QA procedures. If it's not – it won't. High Quality means that when you purchase an electrical item, you have a high confidence that it will work as expected and have a long life expectancy.

High Quality means that an eCommerce site will transact your purchase and deliver the goods on time and for the purchased price.

In relation to the Internet,

8.1. How are QA standards applied to the Web?

It depends on how we are to define QA

8.2. How Does It Work?

A *Client Side Map* of links is produced defining a *master map* of the web site. Each link is then cross-referenced against the *meaning* of the object it points to. This *master map* defines all the links in the site against the meaning of all the objects the links refer to. During this process, broken links, duplicates are highlighted, the resulting map being represented in the form of a hierarchy and level tree.

Now, when the site is changed, a new map can be produced and compared with the *master* – and the differences highlighted.

A major source of problems can be described as *confused links.* These are links which remain relatively static, but point to items that can change. Typical examples are pricing information on eCommerce sites; the product description is unlikely to change, but the link to the price could "point" to different prices. A verification for broken links will not

determine that such links are pointing to the *correct* price information. Typically such checks are performed manually and can be impossible if there are thousands of items. However, because webQA has a correlation between the link and the *meaning* of the object it points to, it is a straightforward process to compare any changes to the *master map* of the site. Errors and inconsistencies are quickly highlighted and fixed.

1. >>>>**WHAT WE ARE SAYING:**
2. **ITS MORE THAN JUST LOG FILE ANAL, WE CAN**
   - do data changes
   - delay data (throttling)
   - anal *in real time with no parsing*
   - perform URL mapping ---DONE
   - we have a client side map – we can understand *what* has accessed the data DONE
   - filtering before the data is requested. DONE
   - WE ARE A WEB-DATA-BROKER
3. **We Can CONFUSE SHOPPING BOTS**
   - By sending back info that is cheaper…. We are cheaper ☺
4. The usual analogies
   - track the offenders, look for what they are doing elsewhere.
5. Keeping in the people
6. Threats by roaming 'Bots
7. Threats from Shopping 'Bots
8. Threats to b2b from Bots' and other b2b'ers
9. Employment Sites
10. Auction Sites
11. Copyright Protection and Findbase
12. DOSbS attacks (to banner ads)
13. Other stuff

---

9. Solutions

Knowing your Data (QA)

10. Findbase Applications

>>>>> put in a short overview for the BP. THEN do one for the RIAA with screen shots. THEN do the eCommerce one for BEA etc etc. Then webQA, NLP etc. Then Banner Ad's

This section provides brief outlines of FINDbase technology applications.

10.1. Auction, Career, Automobile Sites

INFOtector can be used to protect data in on-line Auction sites, Career Resources, Automobile sales, Realty – ***anywhere where the bulk of the value of the site is the data.*** Such sites rely on banner advertising and subscriptions for revenue that is in turn entirely based on the number of visitors viewing the data. If this data is appropriated by other parties, the value of the site – and thus its revenue potential – is decreased.

This has been amply demonstrated in a well publicized actions between the various members of the on-line Auction community. Installing INFOtector would protect this data from (miss)appropriation and thus protect the revenue assets of the sites. It would also have the beneficial effect of reducing a competitor's ability to compete.

Equally well known are the problems surrounding the on-line Career sites which advertise resumes and job vacancies. Some estimates indicate that there are over 15,000 sites on the Internet and the amount of "unofficial data sharing" between these sites is immense. The copyright on the bulk of the material on such sites is often spread amongst many thousands of owners of the resumes (etc) making the prevention of bulk downloads paramount. INFOtector can in most cases completely eliminate the bulk downloading of these assets ***and*** in combination with the *Extraction and Analysis* technology, can mine the web finding unauthorized copies. Such automatic checks can be invaluable in determining who and what to keep out of a site being protected and can in many cases provide enough information for enforcement action.

There are numerous other instances where sites rely almost exclusively on material with an uncertain copyright. Examples are:-

1. Realty – where the copyright of the property descriptions rests with the owner(s) and almost anyone can make use of the material on the web. INFOtector can prevent such accesses.
2. Automobile sales. It would be possible for a well-connected competitor to use the information on other car sale sites to adjust their stock inventories and prices. The various car-finder organizations use these sites to find vehicles for potential purchasers. Again, INFOtector can protect these assets.

10.2. ECommerce Analysis and Comparison

10.3. Banner Advertizement

10.4. RIAA – the Gnutella threat

10.5.

10.6. WebQA

10.7.

10.8. NLP searching

10.9.

10.10. Realtors etc

11.

12. Findbase Solutions

INFOtector

webQA

Client Extraction

Link Mapping

13. The Products

14. Conclusion

15. INFOtector Operation

INFOtector is *not* log-file analysis which simply identifies what requests have happened. Each incoming data request is validated against user configurable parameters to determine if the request is wanted or not. All log-analysis can do is say what happened.

The separation of *wanted* from *unwanted* data accesses are performed by a comparison engine intercepting incoming data requests. This analysis can be performed on any combination of individual data items, entire data hierarchies or the entire data repository. For example, web sites typically have a linked structure that can be represented by a tree. Each file, combination of files and/or tree hierarchy can have different levels of protection applied to it. Since every item in a repository can have individual protection parameters applied to it, INFOtector maintains a *client side map* of all elements that need protection and their interrelationships.

The Client Side Map

This map is a complete hierarchy of all data elements (eg: web site pages) that require protection and how they interrelate. This relationship is represented in different ways by the configuration program. The *client side map* is also used directly for specific applications which are described in other sections.

>>>PROTECTION PARAMETERS<<<

Yeah, right, put in all the math. SCREEN SHOTS!

Protection Options

Unwanted data items can have the following actions applied to them:

1. **Reject** – the requester is given an *item not found* (ie: 404) error.
2. **Delay** – the data item is delayed for a period of time before being returned to the requester.
3. **Same Item** – the current data item or page is returned to the requester.
4. **Different item** – a different item is returned to the requester.

Each option has specific applications which are described in other sections of this document. For example, the *delay* option can be used to manage data bandwidth by *throttling* the amount of data being dispensed. The *Different Item* option can be used to provide the requested with an item describing why their request has been denied.

Each option can be applied individually or in combination with other data items being protected providing the ultimate in flexibility.

16.
17. INFOtector Architecture

The INFOtector family is a 100% software solution that seamlessly interacts with an existing web server. There is no need to change existing web content and there is a bare minimum of installation to the web server being protected.

INFOtector can reside within a web server platform, completely standalone, or in *Enterprise* mode, protecting clusters of Web Servers. The figure (right) shows how INFOtector is installed into an existing Web Server. INFOtector receives incoming data requests for Validation Processing. Validated requests are parsed to the Web Server software. All other data requests are handled by INFOtector. This configuration is suitable for standalone information servers and/or those that have low traffic levels.

INFOtector – Enterprise Edition

INFOtector's scalability protects the future growth that is vital for any information server deployment. Installing an INFOtector into each server provides full protection, but it not suitable if the servers share the same data or if the incoming data requests can be split between servers. For example, it is not uncommon for eCommerce sites to use more than one computer system to service data requests. Such systems often share the same data sources and input requests are shared between the servers. Protection of these system is performed by installing an INFOtector into each server and utilizing a centralized co-ordinating INFOtector system. The central INFOtector system analyses the data requests to the individual servers, maximizing the protection against attack and distributed intrusion. The system requirements of the controlling INFOtector depend on the total number of hits per second from all the servers. Additional INFOtector systems can be added in parallel for massively distributed systems.

The connection between the Servers and the controlling INFOtector can be a local network or an internet connection. INFOtector's ability to coordinate information requests from many distributed and remote servers using network connections provides the highest degree of flexibility and scalability.

The INFOtector family of products comprises the *Standard* edition and the *Enterprise* Edition. The level of protection offered is the same between family members, the difference is in the installation and operation.

INFOtector *Standard*

INFOtector *Standard* is a software product installed directly onto the platform supporting the web server. The *client side map* is produced for the server being protected. It is common for a server to service a number of different Web sites and/or data repositories. INFOtector can protect each of these sites/repositories individually or in combination by use of different *client side maps*. For example, individual *map*s would be used to

represent the pages to be protected in a number of different web sites. The number of sites being protected is limited by the system resources of the web server, a typical limitation being disk space.

INFOtector *Enterprise*

The *Enterprise* version makes use of centralized XXXX's accepting requests from individual Servers. The *Client side maps* for each server is stored on the central XXX which is accessed by the corresponding server via a network connection. Like the *Standard* version, the *Enterprise* version can support multiple sites/repositories on the same server – it is the *Client Maps* that are stored on the XXX. This architecture provides many advantages over the *Standard* edition:-

1. **Scalability**. A single XXX can support many distributed web sites/repositories allowing INFOtector to be easily used by ISP's and distributed networks of servers such as eCommerce sites
2. **Speed and Throughput.** By concentrating much of the data storage and co-ordination on the XXX, the individual servers have more resources to service data requests. This is especially effective for load balancing and data throttling which can be performed by the XXX.

The inter XXX-server network connection can introduce a propagation delay, but this is typically very small, although it is dependant on the speed of the connecting network.

INFOtector Administration

Applet