



A Websense® White Paper

## Deep Content Control™ Keeps Data in the Enterprise

### Introduction:

Information is at the heart of every organization and is often its most critical asset. Through accident or malice, it is sometimes exposed, representing a risk to compliance, business continuity, and an organization's competitive advantage. While organizations focus on the prevention of outside attempts to access sensitive data, few acknowledge – let alone prepare for – the threat from within. Data leakage, i.e., the loss of data more commonly occurs via internal vectors than the theft or destruction of data from intrusion or other external illegal activity. Typically, data leakage results from the negligence or error of employees and/or third party organizations, not an intentional effort to inflict damage.

Most companies have experienced some form of data loss within the last 24 months, according to a recent survey by the Ponemon Institute. 1 In a recent survey 85% of respondents indicated a breach, and 81% of the total number of breaches caused were the result of internal error. 1 The study also revealed that only 6% of leaks were reportedly caused by criminal activity, and another 6% the result of malicious employees. 1 Since the most significant threat to information-centric organizations is not external or malicious, it is probable that intentional data leakage events could be more detrimental than those that result from error.

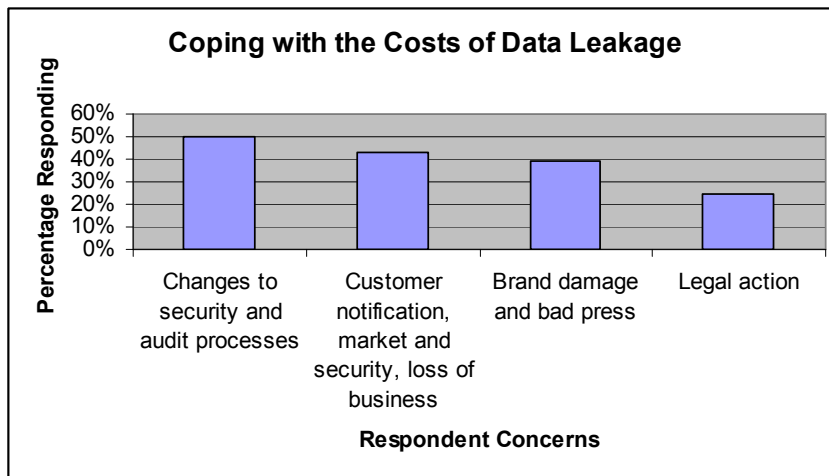
To address this growing problem, many organizations are turning to a new technology, Deep Content Control™. Deep Content Control has increased the effectiveness of information leak prevention (ILP) solutions. It evaluates the meaning of information rather than simply looking for keywords. As the only solution available with this new technology, Websense® Content Protection Suite keeps sensitive data in the enterprise. Rather than signaling false breaches because of misidentification of context-less data, Content Protection Suite analyzes information stored and used, determining the data's meaning and applying powerful security controls and workflows.



<b>Table of Contents:</b>	<b>The Problem of Information Leaks .....</b>	<b>3</b>
	<b>The Importance of Deep Content Control: Content-Context Relationships .....</b>	<b>4</b>
	<b>Websense and the Process of Control.....</b>	<b>6</b>
	<b>Conclusion .....</b>	<b>8</b>

### The Problem of Information Leaks

Sensitive data slips through the cracks, but only occasionally with the help of hackers, malicious employees and other computer users who intend to use company data for personal gain or to cause harm. The economic damage from data leakage incidents is not limited to problem remediation and the consumption of internal resources. While almost half of respondents indicated changes to security and audit processes as a “major cost category,”<sup>1</sup> according to a Forrester Research survey of 28 companies that had experienced breaches, 43% cited “customer notification, market and security response, and loss of business as significant concerns.”<sup>2</sup> Additionally, 39% worried about the extended impact of a breach that would result in bad press and damage to the brand. Only 25% indicated concern about a legal response.



Source: Forrester Research

The impact of a security breach can cascade through the entire organization. But most leaks are preventable, caused by internal lapses rather than impropriety. By tightening security controls and educating the workforce, organizations can reduce the likelihood of negligence or employee error. ILP monitoring tools can serve as a check on procedural controls and education. Further, an internal monitoring program can prevent intentional data leakage while identifying acts of employee impropriety.

Information leaks can be prevented through the development of a control system that consists of processes, education, and technology. Processes provide a leak prevention framework that governs the data environment and constitutes a platform for employee education (and consequently employee behavior). The staff’s understanding of data leakage and prevention measures is enhanced through the use of technological tools that catch intentional leaks and address leakage due to negligence. ILP solutions are the anti-leak control engine, providing the necessary automation and support to keep the leak control process accurate and employees informed.

Processes and education cannot entirely eliminate human error, however; they only can reduce it. In a high-transaction environment, process owners may find it impossible to monitor comprehensively. Thus, ILP solutions power the process, facilitating monitoring efforts and enabling timely enforcement. To integrate into anti-leak control processes, ILP solutions must discover information automatically and draw immediate relationships through the use of metadata, database schema, and other contextual indicators that convey the meaning of enterprise data.

<sup>1</sup> Trends: Calculating the Cost of a Security Breach. Forrester Research, Inc. April 10, 2007.

<sup>2</sup> Trends: Calculating the Cost of a Security Breach. Forrester Research, Inc. April 10, 2007.

Monitoring follows discovery, providing a framework for catching potential leaks, notifying business unit leaders of risks, and generating reports for use in trend analysis and problem remediation. Ultimately, ILP solutions should be a tool for prevention. However, in an information-centric organization one needs to ensure accuracy and integration into the business processes when evaluating a leak prevention solution; such accuracy is achievable only with a combined content- and context-aware solution.

### The Importance of Deep Content Control: Content-Context Relationships

Preventing information leaks should be a top priority for any information-centric business, particularly since even routine operations could expose sensitive data risk. The prevailing approach has been to scan data transmission and communication (e.g., via e-mail) for key terms through regular expression analysis, but this method tends to be inaccurate and unwieldy. It is too simplistic an approach to a very complex problem. Moreover, remedial (often email-based) ILP solutions flood IT leaders and business managers with event notifications that require a response, even if that response is to resume business as usual. Deep Content Control, conversely, provides an alternative to regular expression analysis. Instead of narrowly looking at the words themselves, Deep Content Control restricts the use or communication of data based on its meaning – content and context rather than content alone.

Regular Expression, full and partial matching	Deep Content Control
<ul style="list-style-type: none"> <li>▪ Evaluates binaries only</li> <li>▪ Searches for targeted expressions</li> <li>▪ Intercepts without context</li> <li>▪ High volume of false positives and negatives</li> </ul>	<ul style="list-style-type: none"> <li>▪ Evaluates data within the context of its source, purpose and use</li> <li>▪ Defines content based on context</li> <li>▪ Limits interception to contextual indications, such as the user, destination, and channel</li> <li>▪ Low volume of false positives – highly accurate</li> </ul>

Deep Content Control represents a combination of content awareness combined with context awareness – specifically, a solution’s ability to interpret what information is and where it is located, who is using it, how they are using it, and where they are sending it. This involves analyzing the data itself as well as the database field in which it is stored (fully qualified to include the table or view and full database name), the data to which it relates, and the systems that use the data. For content to have meaning, a user or system has to be aware of the data as well as how it is being used and stored.

For example, assume that a file of Social Security numbers has been leaked. Ultimately, you want to determine who is responsible. To do so, you follow a chain of reasoning that makes it easier to identify the culprit. You start with what is easiest to determine: From what system was the data taken? With this information, you could use access logs to investigate further. But before you can pinpoint the system that was compromised you need more context. For example, were the stolen Social Security numbers customers’ Social Security numbers or employees’ Social Security numbers? Without the proper context, you know that you have a problem (i.e., stolen sensitive data), but the parameters of the data loss are unclear.

Content without context has only limited meaning, especially in an integrated enterprise in which the same data can reside in multiple systems. Context addresses where data resides, to what it refers and how it is used. Essentially, context makes content meaningful. Consider a series of numbers: 123456789. Is it a Social Security number or an employee ID number? Or, could it be something else? Context provides the answer. If the data came from the payroll system it could be an employee Social Security number, or it could be an internal employee identification number.

More context is necessary. If entire records were leaked, more context is available and the investigation continues. True content awareness will reveal that data is a real Social Security number stored in your database, having fingerprinted it during the discovery process. A simple pattern matching solution could not facilitate such content awareness because it fails to accurately differentiate one piece of content from another.

In order to determine the meaning of data being communicated, the relationships among data, systems, and users must be examined. This includes other data elements stored in a particular record or database table. Metadata and database schema provide the next layer of context, defining data by where it is stored and other data elements to which it is related. Outside the databases in which data is stored, the systems and applications that house or access databases also contribute to the meaning of the data itself. In particular, the relationship between one piece of data and another may be of particular importance. Going back to the example of the customer Social Security number, it may prove inconsequential in a large organization if one, or even a group are transmitted outside the organization unsecured. However, if those same Social Security numbers are paired with corresponding customer names, birthdates, and account numbers, the criticality of the incident and response may be much more significant. Thus, metadata helps us understand another important relationship – data to data – that only Deep Content Control technology can provide.



Deep Content Control includes destination awareness and control – the ability to monitor and protect data by its destination. For example, with Deep Content Control administrators can build a policy to say that employee John Smith is prohibited from emailing confidential information via web-based mail sites (i.e., webmail), from posting it to blogs, or from sending sensitive data to sites in a specific foreign country. With Deep Content Control, destination controls can be applied to specific users, groups, and data sets, providing governance over who can send what information where.

To prevent data leaks, ILP solutions should protect company information wherever it resides, including data “at rest” (i.e., data stored on servers, PCs, laptops and external storage devices), data in use by applications, and data being communicated (e.g., by e-mail, FTP, HTTP, and print). Anti-leak solutions must identify all sources of information in the enterprise, monitor its use and protect that data from impropriety. Given the size of most businesses – and the increasing rate at which data is produced and stored – automation is necessary. Leak prevention solutions must be

able to discover, monitor, and protect information with minimal manual intervention – a key driver for a reasonable cost of ownership.

Discovery is the backbone of most ILP solutions, which focus on finding data in the enterprise in order to apply regular expression analysis to prevent leaks. The primary challenge of leak prevention, though, is categorization, particularly in real-time. Many solutions claim to be able to locate data; however, effective control can be achieved only when the anti-leak solution understands the meaning of the data discovered, through a combination of internal relationships and how it is used by systems and end-users. Deep Content Control results from ascertaining the meaning of data and integrating web intelligence to cross-reference context with system, transmission, destination, and end-user information. In short, Deep Content Control reveals who is sending what information where and how.

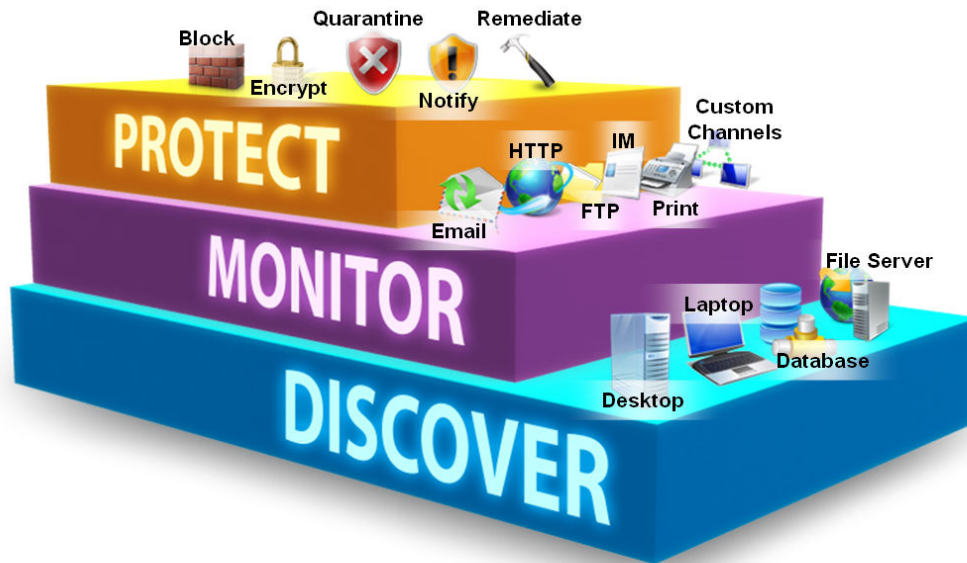
Outside of systems and applications, user access to data provides an additional element of context. The use of role-based access control (RBAC) to determine an individual's purpose for using a particular application contributes to the total context of data stored in the enterprise; who has access to data helps to explain what it means. A system administrator would have universal access to all data in the application, perhaps including the ability to modify metadata and some aspects of the database schema from within the application itself. A management-level user would have the ability to view and override data entered or modified by staff members, and staff members have spheres of influence within an application that are limited strictly to specific information that supports the job function.

The combination of application composition (e.g., schema and metadata), the purpose of the application or system, and RBAC-based user access provides a wider definition of each data item stored in the enterprise. The use of the entire enterprise to support the meaning of every data item in the enterprise results in the infusion of meaning. As a result, ILP solutions with Deep Content Control can interpret communicated data and data at rest more accurately, allowing normal business communication without disruption and protecting sensitive data from outside use. The remediation of a data leakage event requires the identification of leaked content and an understanding of what that content means. With a firm grasp of both the content and context involved, it is possible to investigate effectively, find the source of the leak, and take action.

### **Websense and the Process of Control**

The Websense Content Protection Suite keeps sensitive information secure. By monitoring data and communications protocols, Websense Content Protection Suite ascertains the meaning of data in real-time and enforces regulatory compliance and corporate governance policies to protect it. Data used through the course of normal business proceeds without interruption so long as it is used and secured properly. Otherwise, Deep Content Control is combined with a variety of enforcement protocols, including encryption, quarantine, notification, block, and remediation, to automatically secure critical information assets.

Deep Content Control is a core, proprietary component of the Websense Content Protection Suite. By monitoring all communication protocols to track enterprise data, the Websense Content Protection Suite is able to monitor and protect both structured and unstructured data, whether stored on laptops, desktops, or servers, to ensure that it is not communicated to an inappropriate destination.



Unlike other leak prevention solutions, Websense Content Protection Suite includes patented PreciseID™ technology, for accurate fingerprinting of confidential information, also referred to as “content awareness”. PreciseID creates mathematical representations, analyzing hundreds of megabytes every minute, and stores these fingerprints for correlation with data being used or transmitted. It uses advanced techniques, developed originally for sonar technology, to evaluate minute differences between like pieces of information. The accuracy of this data analysis separates Content Protection Suite from other content-aware solutions.

PreciseID not only accurately identifies the uniqueness of the information, it includes a reference to the data’s broader context. Fingerprint evaluation, consequently, replaces the regular expression analysis. The result is faster content analysis and a lower likelihood of false positives and false negatives. The use of the data fingerprint enables Content Protection Suite to catch only the data most likely to be sensitive. The fingerprint actually includes contextual information with content, for comprehensive monitoring and control.

Websense Content Protection Suite discovers data on all networked systems, whether in use or at rest, and automatically creates unique fingerprints. The solution then monitors all communication protocols to track the use of data by employees and authorized outsiders in order to prevent leaks. Routine data that does not violate a policy proceeds without issue, while problematic data is intercepted and action is taken.

Reporting is the intersection at which automated execution and human intervention meet. Reporting capabilities range from real-time alerts that signal an imminent breach to long-term trends that identify internal weaknesses and broken business processes that need to be resolved. Websense Content Protection Suite addresses the tactical and strategic components to facilitate a comprehensive approach to leak prevention.

Alerts constitute real-time, event-based reporting at a granular level. Content Protection Suite also offers activity reporting capabilities, snapshots of individual users, data sets, compliance requirements and other information to help IT security professionals plug leaks before they occur, to investigate specific events, and to monitor the security of the enterprise. Historical data supports trend analysis, allowing for ongoing refinement of an organization’s data security plan. Trend identification is a powerful tool in data leak prevention.

Reports can be assembled on an ad hoc basis or scheduled to run periodically, reaching recipients by e-mail. By planning report distribution, it is possible to raise the profile of data security across the business, accelerating return on investment and generally reducing the risk of future leaks.

Content Protection Suite monitors the data management environment and signals imminent breaches. In the event of a breach, Content Protection Suite alerts the appropriate manager or process owner. Immediate investigation and remediation can result. Its proactive, automated enforcement capabilities keep data in the enterprise when an incident occurs.

The near-term use of ILP solutions is for policy enforcement, which delivers immediate value to the business. Upon identifying a breach, the impacted manager can investigate the situation through the role-based, web interface – without the need for the IT department's immediate involvement. Within minutes, the manager can obtain insight into the incident and begin remediation, enabling the business units to address what is fundamentally a business problem, and lowering the cost of ownership.

Following a verified breach, incident management involves both technical and organizational remediation measures. The technical source of the leak must be improved (if necessary) to prevent similar occurrences in the future. Further, the employees responsible for the leak may require action – from an informal correction of behavior to criminal prosecution (in the extreme).

Content Protection Suite reporting provides the historical data necessary to develop policy controls for technical gaps and enhance security configurations. As all activity requires an audit trail, Websense Content Protection Suite equips the business to justify controversial decisions that could have public, employee, or legal implications.

## Conclusion

The risk of leaked data is not limited to malicious employees or outside hackers. Instead, routine error is most likely to cause internal, private data to leave the enterprise. Yet, most data-driven businesses focus on the smaller threat. While external attempts to access proprietary information may seem more nefarious, flawed operations are more frequent and addressable, providing a higher return on an investment in a leak prevention solution. A modest effort can yield substantial results to a problem that most companies experience every day.

Deep Content Control provides a critical capability not offered by traditional leak prevention solutions. At a minimum, primitive solutions focus on the evaluation of regular expressions, which take the words and phrases used only at face value. At a maximum, they employ rudimentary content aware technology to gain a somewhat accurate detail of the data. Deep Content Control considers both the content and context of the information very granularly and accurately, through an evaluation of related information and other indicators, including metadata, database schema, end-user, and destination.

Websense Content Protection Suite includes Deep Content Control to gain the added intelligence of who is sending what information where and how. Through a robust reporting engine, Websense Content Protection Suite monitors and protects data security policies, constituting a comprehensive leak prevention solution. The short-term effect is that fewer leaks occur and breaches can be remedied quickly. Long-term, the organization adopts a climate of information control that reduces costs and bolsters the organization's competitive position.

Leak prevention solutions enable business management of the problem (rather than the need to commit extensive IT resources), streamline evaluation and control processes, and reduced breach incidents to keep data inside the enterprise. The Websense Content Protection Suite lowers the cost of protecting information and recovering from a breach, while enhancing competitive advantage. For a business-centric approach to information security, Deep Content Control is a must-have component to protect your data, your customers, and your business.

### About Websense, Inc.

Websense, Inc. (NASDAQ: WBSN), protects more than 25 million employees from external and internal computer security threats. Using a combination of preemptive ThreatSeeker™ malicious content identification and categorization technology and information leak prevention technology, Websense helps make computing safe and productive. Distributed through its global network of channel partners, Websense software helps organizations block malicious code, prevent the loss of confidential information and manage Internet and wireless access. For more information, visit [www.websense.com](http://www.websense.com).

### About The Author

David Meizlik is the Product Marketing Manager for Security Solutions at Websense, Inc., the leading provider of web and content security solutions based in San Diego, California. His responsibilities include product positioning, go-to-market strategy and development, market and competitive analysis, program development, and management of all outbound marketing activities for Websense security products. Meizlik earned a bachelors degree from the Marshall School of Business at the University of Southern California, and went on to receive a graduate degree in communications management and technology from the USC Annenberg School for Communication.

© 2007 Websense, Inc. All rights reserved. Websense and Websense Enterprise are registered trademarks of Websense, Inc. in the United States and certain international markets. Websense has numerous other unregistered trademarks in the United States and internationally. All other trademarks are the property of their respective owners. 08.26.07