



IT Knowledge • Business Results

# White Paper

## ***Information Classification: Determining the Best Approach to Preparing Information for Precise Management and Archiving***

By Brian Babineau, Analyst  
Intelligent Information Management

September, 2006

**Table of Contents**

Introduction..... 2  
Intelligent Information Management for Files is a Necessity..... 2  
Information Classification – Preparing Information for Action ..... 4  
Information Classification for Archiving ..... 4  
Conclusion..... 6

## Introduction

---

Most operational and capital budgets for IT departments focus on mission critical applications and the databases that store all of the transactional data that they constantly create. These applications include supply chain management, human capital management, financial reporting and several others. Even e-mail, once an application that was little more than an afterthought from an IT management perspective, now receives a large portion of IT financial and operational resources because it is viewed as a critical productivity and workforce collaboration tool. These applications and the supporting compute and storage infrastructures reside in corporate data centers and are top of mind when it comes to dedicated and discretionary spending. While these aforementioned systems do generate significant storage capacity, often overlooked and the last to receive any significant IT resource allocation are general purpose file servers that house a variety of data types. Microsoft Office documents, Computer Aided Design (CAD) documents, research and development information, corporate logos, images, web content and a myriad of other files that need to be shared are regularly saved on file systems. The capacities of these file systems are constantly growing and will continue to as IT uncovers the ease of management of deploying file-based storage on IP networks. In fact, many IT shops are now deploying databases and other applications on file system storage to take advantage of existing IP networks and to leverage file system expertise within Database and System Administrator groups. Unstructured, file-based data is by far the fastest growing type of data today – far outpacing transactional data growth.

As organizations store more data within file systems, IT must find new ways to gain control and manage all of this data. Further, organizations must take into consideration all of the external impacts, such as regulations and general resource management policies, when managing these files. While e-mails and databases certainly garner headlines when mismanaged, the sheer capacity of file system data can disrupt normal business and IT operations. Files need to be backed up regularly, and with expanding volumes of data, the time it takes to complete data protection processes increases. When a file needs to be recovered, IT must sift through millions of backups to locate the appropriate version. To deal with the growing capacity of files, organizations continue to sink money into file servers and storage just to keep up with data growth and to keep applications online. Unfortunately, the implementation of additional devices requires additional operational funding as incremental backup software licenses must be procured and system administrators must constantly reconfigure applications to align with new system deployments.

When attempting to manage file system information, many organizations fall into a trap that has plagued IT departments for years – managing data blindly. Even the ever popular Information Lifecycle Management (ILM) strategies that talk about managing data based on its value fail to provide insight as to how organizations can locate and group similar data so that management actions such as archival, deletion and migration can be consistently performed. ILM has helped organizations build tiered infrastructures, especially when deploying storage systems. Unfortunately, customers cannot leverage investments in these types of implementation until they locate the data, classify similar files into groups and then take additional action. ESG refers to the process of classifying and taking action against data as Intelligent Information Management, and believes it is the most appropriate approach to gain control over the vast amount of file system data that permeates every organization.

## Intelligent Information Management for Files is a Necessity

---

There are several reasons why organizations must manage information more intelligently, or more importantly, why they cannot afford to ignore the context of the data when taking action with it. First, IT processes traditionally dictate that all file system data is treated the same way. File servers are backed up regularly according to a predefined schedule. Some of these file servers may contain critical application files that are part of a disaster recovery process. In many cases, IT does not know the types of files that are being managed,

protected and replicated. The files may be important to the business or they may be pictures from an employee outing. The files may not have been changed or accessed in over a year, yet the management actions take place consistently and capital resources are used to buy more primary and secondary storage capacity. Recovery operations are often delayed because of the amount of data that needs to be restored from a backup. When IT needs to recover a file, administrators need to sift through all the data that has been backed up, often requiring a user or an application to wait. IT can greatly enhance many data protection operations by identifying older files, segregating business files and creating other data classes that will help to ensure that appropriate backup and business continuance policies are enforced against the right data.

Because of data growth, some IT departments have attempted to delete historical or unchanging files just to mitigate storage costs and alleviate backup processes. However, data deletion may not be an option due to the morphing international regulatory landscape and a renewed focus on corporate governance. Record retention regulations, most common in the United States, mandate that certain organizations create and save business records for specified periods of time. The Securities Exchange Act of 1934 Rules 17a-3 and the Health Insurance Privacy and Accountability Act (HIPAA) are examples of record retention rules. Corporate governance regulations, such as Sarbanes-Oxley (United States) and Basel II (European Union), are being put into place to mitigate fraud and other types of malfeasance behavior. These rules require organizations to improve controls over financial reporting and increase the documentation of business processes. Lastly, emerging information privacy and security regulations are making their way through various worldwide governing bodies to improve upon the control of confidential and sensitive information. In the United States alone, there are an estimated 13 information privacy and security related laws currently being debated in Congress. These regulations do not require organizations to retain, secure or monitor all files. Such requirements would be cost-prohibitive, leaving customers without enough resources to be able to run their businesses. Organizations must identify business records, organize sensitive data created by high risk departments such as merger and acquisition teams and then establish information management policies to apply against these data groups. European regulations around privacy and terror defense are seemingly contradictory, but it appears as though the EU will most likely choose safety as the primary requirement, which will demand significantly more record retention.

As a result of the increased regulatory impact on information management, organizations must now produce some of the information in response to legal inquiries. ESG estimates that 46% of organizations have experienced an electronic discovery over the last twelve months where organizations had to produce e-mails, files, medical records, executive memos, spreadsheets and several other files in response to discovery inquiries<sup>1</sup>. As a result, corporate counsel must locate the most relevant information and secure these files for review. This may require files to be retained on immutable storage whereby any evidence cannot be modified or deleted during a legal matter. Organizations must be able to locate certain files quickly and then manage this information according to evidentiary processes that preserve chain-of-custody.

Amongst all of the factors that impact information management, one common solution constantly manifests itself – knowledge about data can facilitate better management. The backup of pictures, deletion of business records, storage of budget files on unsecured file servers all occur because organizations do not know where files are, who owns them, the relative importance of the files to the business and when the files were last accessed. Organizations should begin the process of contextually understanding their files and classifying them. Information classification allows organizations to take advantage of information management software that archives, encrypts, moves, copies and takes other action against the data.

---

<sup>1</sup> ESG Research Report: *Digital Archiving: End-User Survey & Market Forecast 2006 – 2010*, March, 2006.

## Information Classification – Preparing Information for Action

---

Next generation information management solutions, including data archiving, can add value by helping customers classify information – that is, organize data into categories – prior to taking specific action. The classification process enables users to understand more about their files and then set policies to automate the management of these various data groups. Without classification, customers need to manually identify and analyze data to be archived, secured or otherwise managed, a process that is extremely cumbersome when trying to control large volumes of files.

The largest challenge for many organizations when controlling data, especially as information management requirements are constantly evolving, is locating the data. File servers can be found in data centers, remote offices, closets, under desks and many other locations. There are several IT tools, including Storage Resource Management (SRM) products, which can help IT find the data. However, they do not provide for the next critical step for organizations to take which is to classify the data by correlating location and ownership with other attributes of the files, providing additional context so that the proper action can be taken with the information. Attribute Classification helps organizations understand all of the intrinsic information associated with a file, including the location, date of creation, date of last access, owner, owner's organization and format. This intelligence is gathered by analyzing file system structure, directory paths, the directory service and the individual files. Because many organizations store data across multiple file servers and file systems, products that utilize attribute classification must be able to centralize all of the intelligence. This way no file is left stranded or unclassified.

Organizations can begin to group data based on common attributes, such as segregating all spreadsheets owned by the finance department or collecting all human resource files that have not been accessed over the past 12 months. Organizations can also manually group files after attribute classification. However, establishing rules such as finding any spreadsheet with the path '/finance/budgets/final\_versions' and automatically classifying these files is more efficient when dealing with large volumes of data. After classification, organizations can take action against the data groups. In continuing with the example, all spreadsheets created by finance can be moved to a secure file server because these files, per internal policies, are considered confidential with a retention period that satisfies corporate records management policies.

Attribute Classification methods also build an index of all of the intelligence gathered. Customers can use this index to search for files that are described by a specific attribute. For example, a controller may want to locate all of the 'final versions of budget spreadsheets' created over the past 2 years. IT can quickly utilize a product that relies on attribute classification, such as an archiving solution, to quickly scan and retrieve files that meet this criterion. If organizations need to search files for specific content utilizing specific keywords, additional information classification analytics must be performed. Content Classification is the method used for creating an index of the contents of a file and its attributes. The index can then be used to search for and locate a particular file.

## Information Classification for Archiving

---

Attribute Classification enables organizations to understand and group the large number of files that need to be managed. One management task that can be easily applied against a class is archiving. Customers may archive files to comply with regulations, segregate data for security purposes or expedite backup operations by removing non-transactional data from actively backed-up production storage systems (which also carries the incremental benefit of keeping very expensive storage groomed and performance optimized). The key to archiving is identifying and classifying the files that meet a specific set of criteria. For example, when archiving to remove stagnant data from production systems, it is critical to identify all files that have not changed within a

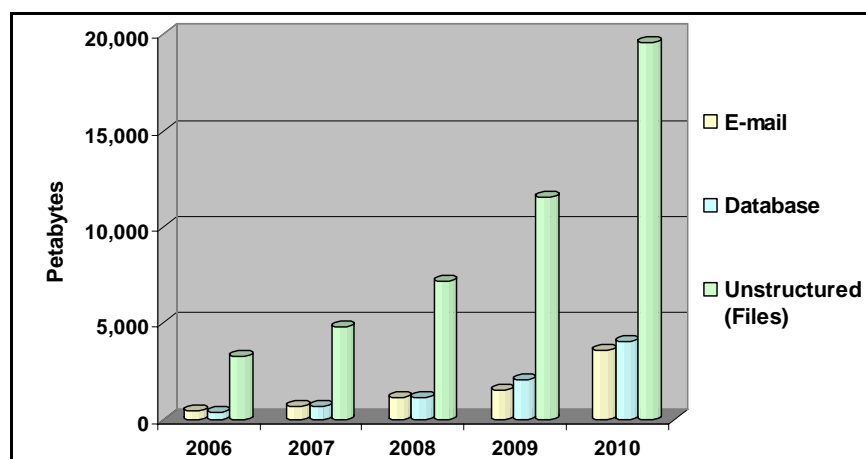
predefined time period. Once these files are classified, IT can transparently and automatically move the files, leaving a link behind so the files can still be accessed.

Archiving applications can utilize Attribute Classification to identify and group data and then take action against it. In some cases, customers may use archiving solutions to move files from one storage system to another. This requirement often arises when files are classified as business records or requested during an electronic discovery. Customers may be forced to store these classes of files on non-erasable, non-rewritable storage media. In addition, archiving solutions can create additional file attributes such as retention periods. By establishing a retention period attribute, customers can prevent the deletion or modification of a file. The retention period attribute also provides another way of locating information as customers can search the archiving application attribute index for 'all files stored with a retention period of three years.' Tracking attributes such as retention periods and locations of files helps organizations audit information management policies and prove to regulators or litigators that the appropriate processes and technologies are in place to store records or evidence in a way that is acceptable for evidence productions.

Customers that archive data using attribute classification can improve data protection operations, comply with regulations, segregate data in response to a legal discovery and leverage investments in tiered storage. These benefits, combined with the growing volumes of file system data, make archiving with attribute classification an easy way for organizations to start managing information more intelligently. ESG Research, as depicted in Figure One, confirms that unstructured content, namely files, will make up the largest amount of data archived.

Figure One: Total Worldwide Digital Archive Capacity by Content Type, 2005-2010

(ESG Research Report – "Digital Archiving: End-User Survey & Market Forecast 2006-2010)



Archiving data via Attribute Classification can also help organizations create classes of files that need further classification. Because Content Classification builds large indices due to its retention of all of the contents of files, it may not be feasible to classify large amounts of files in this manner. Attribute Classification can reduce the amount of information that is fully indexed for search and discovery purposes. The index supports keyword searches and can help identify numerical or text patterns that attorneys often look for when reviewing a subset of potential evidence. For example, a pending legal matter involves a patent infringement and internal counsel must determine if engineers stole intellectual property from their former employer. Initially, the attorneys can use Attribute Classification to locate all design files related to the product involved in the patent infringement case. This group of files can then be archived to a secure file server where only legal personnel can access them. The attorneys can use Content Classification to build a content index and search for specific keywords such as 'steal' or 'competitor'. Archiving with Attribute Classification locates and segregates a group of files where a keyword index can be generated for specific legal discovery purposes. The alternative is a manual

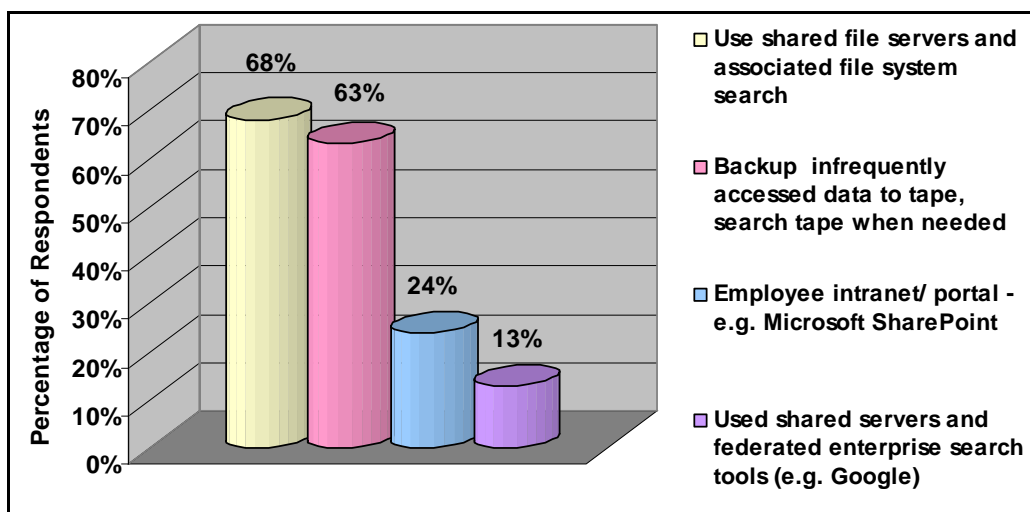
review by attorneys of printed copies of all files or the creation of a large Content Classification index and the conducting of multiple searches to locate relevant files.

## Conclusion

As the amount of information created increases, especially file system data, organizations struggle just to store it. Managing the data is often an afterthought and data protection processes are disrupted while compliance with external regulations overhangs as a 'to-do' item. IT has attempted to retrofit several traditional IT processes to improve management and control of millions of files. Figure Two indicates that IT has modified backup operations and currently tries to use native file system tools to locate and take action on their data. These methods are ineffective and rarely scale across multiple file servers and large amounts of data. In order for organizations to improve the way they manage information, they need to understand the context of the data. This contextual understanding is made possible by classification based on attributes of the data, allowing management policies to be enforced against a group of similar files.

Figure Two – Current Methods of Managing, Storing, Searching and Retrieval of Unstructured Content (Files)

(ESG Research Report – *Digital Archiving: End-User Survey & Market Forecast 2006-2010*- Respondents could select multiple answers)



Attribute Classification enables customers to group information based on file ownership, location, format and other key descriptors. The benefits of information classification are simple: more precise, consistent information management across volumes of data. Organizations can respond to electronic discovery events faster, meet governance requirements and archive older data from production systems to speed up backup and recovery procedures.

Intelligent Information Management, including archiving files, requires organizations to contextually understand the information they create and how it needs to be managed. Classification is an essential part of the process to locate, group and prepare information for action. Without classification, the risk of data mismanagement increases. This risk could equate to data loss, increased legal expenses, higher storage-related capital costs and fines for failing to comply with regulations. There are far too many benefits for organizations not to evaluate information management solutions that incorporate information classification technology, and far too much risk and cost associated with doing nothing. Investing in an Intelligent Information Management solution with attribute classification can easily be justified when measured against the cost of keeping data forever in the face of increased regulation and electronic discoveries while trying to keep up with data protection operations.

All trademark names are property of their respective companies. Information contained in this publication has been obtained by sources The Enterprise Strategy Group (ESG) considers to be reliable but is not warranted by ESG. This publication may contain opinions of ESG, which are subject to change from time to time. This publication is copyrighted by The Enterprise Strategy Group, Inc. and is intended only for use by Subscribers or by persons who have purchased it directly from ESG. Any reproduction or redistribution of this publication, in whole or in part, whether in hard-copy format, electronically, or otherwise to persons not authorized to receive it, without the express consent of the Enterprise Strategy Group, Inc., is in violation of U.S. copyright law and will be subject to an action for civil damages and, if applicable, criminal prosecution. Should you have any questions, please contact ESG Client Relations at (508) 482-0188.