

INTELLIGENT E-MAIL ARCHIVING: CLASSIFICATION, FILTERING, RETENTION, AND DISCOVERY

BY ART GILLILAND

Implementing an e-mail archiving system can bring multiple challenges, including choosing which messages to archive, how long to retain them, and how to find specific messages after they have been archived. The Symantec® intelligent archiving approach enables organizations to efficiently classify, filter, retain, and search for e-mail messages while helping simplify management and control resource costs.



With the recognition that e-mail has become a critical part of IT infrastructure, many organizations are reevaluating their e-mail management policies and systems. Across many industries and public-sector organizations, IT professionals must address three common concerns regarding e-mail: resource management, retention management, and discovery management.

Related Categories:

Backup, recovery, and archiving (BURA)

Microsoft Exchange Server 2007

Regulatory compliance

Storage

Symantec

Visit DELL.COM/PowerSolutions for the complete category index.

Given these challenges, enterprises across the world are evaluating or using software-based e-mail archiving systems. These systems are typically designed to help IT staff control e-mail storage costs while giving end users simplified storage and search functionality and providing legal departments with a consistent system for retaining and finding e-mail messages across the enterprise. As IT departments plan or implement these systems, however, they must account for several important considerations:

- **Storage size:** Although e-mail archives can provide rapid return on investment, they also create a high demand for storage. And because they may have to retain this data for many years, IT departments are seeking ways to optimize their archival storage costs.
- **Retention period:** E-mail archiving can force a necessary but challenging discussion within organizations about how long they should keep e-mail messages. Many enterprises and government bodies have retention policies for

traditional paper records, yet struggle to determine the appropriate policies for e-mail.

- **Search functionality:** Finally, organizations must estimate the amount of data they will accumulate in their archives over time and look for ways to reduce the search time and costs for finding the data they need.

Although archiving systems help greatly simplify the problems with resources, retention, and discovery that plague many e-mail environments, they do not eliminate them. Many of the key challenges that remain stem from the fact that although e-mail messages may share fundamental characteristics—a sender, a recipient, a subject, and a body—they do not all have the same value. For example, Figure 1 shows two e-mail messages from a company CEO. The e-mail on the left is a critical company document, an official record that may drive a series of business actions to help this company compete—and that may serve as evidence in the future if these actions are investigated for being anticompetitive. In contrast, the e-mail on the right is important to the CEO but not to the company's future (unless, perhaps, his son is the head of BETA Corporation). Yet many e-mail archiving environments treat messages as though they all have the same value.

To help overcome this limitation, enterprises should look for ways to intelligently classify, filter, retain, and search for

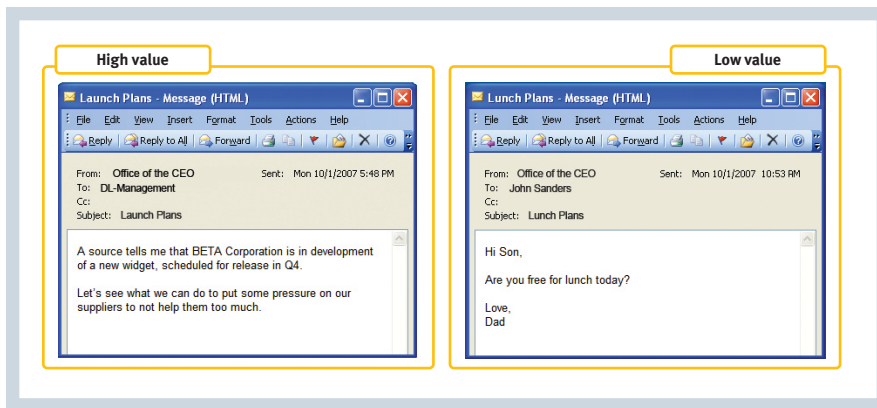


Figure 1. E-mail messages with significantly different enterprise value

messages, enabling them to handle messages differently based on their value, subject, and other considerations. The Symantec intelligent archiving approach can help enterprises implement this type of system and create an efficient automated system for e-mail archiving and retrieval.

Understanding e-mail archiving and the Symantec intelligent archiving approach

E-mail archiving systems typically work by first capturing e-mail messages from the environment, either immediately (referred to as journaling) or after a period of time such as 30 days. They then store those messages for a period of time defined by the administrator, referred to as the retention period. Finally, these systems index the messages, their properties, and their attachments so that legal, finance, human resources, and other groups can find them later.

When implementing an e-mail archiving system, organizations must make three fundamental policy decisions: what they should archive, how long they should retain archived messages and related data, and how they can find messages and data later. Deciding how long to retain information is perhaps the most challenging of these three. On the one hand, many enterprise leaders would like to keep e-mail messages as long as possible: e-mail is vital to enterprise operations, and many employees frequently go back to old messages for information. On the other hand, legal and IT professionals often see the downside of retaining e-mail. First, every additional message retained increases storage and IT

costs. Second, keeping some messages longer than necessary may increase risk for the organization later on. Finally, as the number of archived e-mail messages grows, it becomes difficult to locate individual messages when needed.

Among organizations that use e-mail archiving systems, many of them use systems that fall into one of three groups:

- Non-automated archiving system:** Organizations that lack an automated system still archive e-mail messages—but typically do so in a way that is often inefficient, ineffective, and risky. IT staff may archive e-mail messages by retaining e-mail server backups, while management may discover that messages that were deleted from the e-mail server years ago still remain on a backup tape or notebook computer. These unwanted backups may cause problems if the organization is forced to turn over data it did not even know it possessed to an opposing litigant or investigator.

- Automated archiving system that keeps everything for the same period of time:** IT departments have driven many early e-mail archiving deployments to help reduce e-mail storage costs and increase application efficiency. Many organizations have archiving systems that retain all e-mail for the same period of time, be it one year, three years, five years, or more. Many of these organizations have not yet reached the point where they must actively expire e-mail; however, those that have reached the end of their retention period often extend it, just to be safe.
- Automated archiving system that keeps everything forever:** Some early adopters of e-mail archiving based their implementation on regulatory mandates. Because these mandates were often vague in scope and length of time, some organizations have indefinite retention policies for their archives as they await further clarification from the government or depend on other organizations to take the first step.

The Symantec intelligent archiving approach is a natural evolution of early software-based e-mail archiving systems (see Figure 2). It is designed to provide an automated archiving system that treats e-mail messages differently based on value and enterprise requirements while simultaneously providing simplified search and retrieval functionality through the following components:

- Intelligent classification:** Deciding which e-mail messages are relevant for which purposes

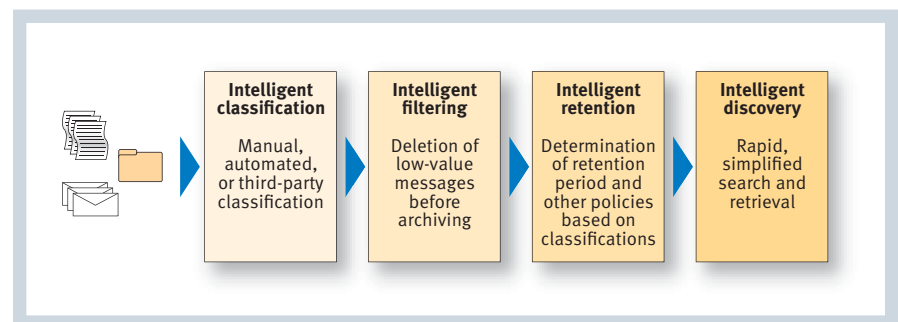


Figure 2. Key components of the Symantec intelligent archiving approach

- **Intelligent filtering:** Discarding irrelevant e-mail messages before archiving, helping reduce the size of the archive
- **Intelligent retention:** Determining how long to keep archived e-mail messages based on their classification
- **Intelligent discovery:** Tagging e-mail messages with metadata during archiving, helping simplify future retrieval

Taking the first step: Intelligent classification

Intelligent classification is critical to the Symantec intelligent archiving approach, enabling organizations to overcome key challenges of classification and differentiate the enormous number of e-mail messages they send and receive each day.

Key classification challenges

E-mail can be classified into a number of different categories, ranging from the very basic (“Business” or “Personal”) to the very sophisticated (“2005 Tax Records,” “Reseller Contracts for Germany,” and so on). This type of classification is not new—records management has existed for over a century, with enterprises and government organizations devoting substantial time, money, and personnel to storing official documents in files, placing those files in boxes, storing those boxes in warehouses, and keeping track of the entire process. However, e-mail has introduced three challenges that make previous records management approaches inadequate: volume, universality, and informality.

Volume. Many organizations receive huge quantities of e-mail messages. In traditional records management models, because organizations might deal with thousands of official records, the threshold for creating a record was very high. Someone had to print or write a document, submit it to a records clerk (or have it be part of a defined process), and so on. Now, all it takes to create a record is for someone to click the Send button.

Essentially, e-mail can happen at nearly the speed of thought, rather than at the speed of print—and the volume increases accordingly. An organization with 10,000 end users sending and receiving 100 messages per day with 200 working days per year would create 200 million messages per year. Over a five-year period, that amounts to a billion messages.

Universality. Under traditional records management processes, records were typically created by defined groups—legal, finance, human resources, and other departments—that could be trained on organizational policies based on compliance demands. Now, anyone in an organization across the globe might be sending e-mail messages, and contractors and outsourcing partners can increase the complexity of this problem even further.

With myriad individuals across countries, languages, time zones, and enterprise boundaries, organizations are challenged to disseminate and enforce documented e-mail retention policies and guidelines. Many organizations are not willing to stake their reputation and

financial security on trusting every user to follow the process.

Informality. Informality is perhaps the trickiest problem of the three. E-mail messages, unlike memos or faxes, are notoriously informal. A thread about last weekend’s activities can quickly transform into a discussion about this quarter’s sales forecast, and a casual comment, when taken in proper context, can take on major importance.

Approaches to intelligent classification

There are three primary methods of intelligently classifying e-mail: manual (having end users classify their messages), automated (having an archiving system classify messages), and third-party (having another system classify messages, either manually or automatically). Figure 3 summarizes the advantages and disadvantages of each.

Manual classification. Although a key advantage of archiving is avoiding the need for end users to make classification decisions, many organizations have concluded that blending automated archiving with some human oversight is necessary. In this approach, a user lets the archiving system know how to classify an e-mail message in the archive from within the organization’s e-mail software, such as the Microsoft® Office Outlook® client. One method involves presenting a folder structure defined by the IT department to the user in Outlook (in addition to his or her normal personal folders). For example, a salesperson using Outlook might see the folders shown in Figure 4.

The advantage of manual classification is that end users can often judge a message’s value more accurately than an automated algorithm. At the same time, however, this approach increases work for end users and can lead to inaccuracies from user error or malicious intent.

Automated classification. The opposite approach is to rely on the archiving system to classify messages. For many organizations, an ideal automated classification engine could identify what each message is and its relevance to the enterprise.

	Description	Advantages	Disadvantages
Manual classification	End users classify messages by dragging them into a folder or by selecting options from a pop-up window	May provide better judgment of message values than an automated algorithm	Requires additional work for end users, and end users may inadvertently or deliberately misclassify messages
Automated classification	The archiving system classifies messages by analyzing message properties and content	Uses a consistent, repeatable process with a low burden on end users	Has potential for false categorizations
Third-party classification	A third-party records management or gateway system classifies messages	Can take advantage of existing classification systems and handle multiple types of content	Same as for manual and automated systems, depending on the underlying approach

Figure 3. Comparison of intelligent classification methods

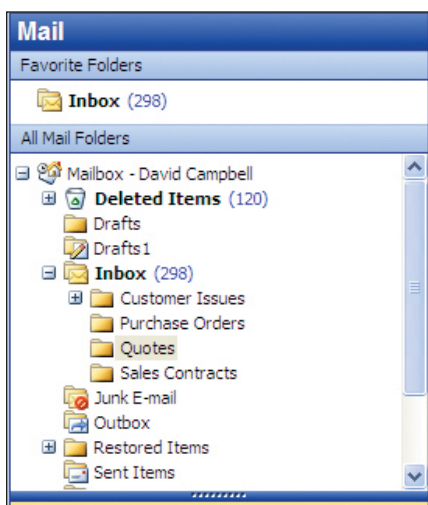


Figure 4. Example folders displayed in Microsoft Office Outlook

Classification engines typically use a combination of approaches to analyze a message and determine the type of content, including using information such as the following:

- **Senders and recipients:** Messages from the legal department, for example, typically contain legal content.
- **Keywords or phrases:** Messages and attachments with a “confidential” disclaimer, for example, typically identify data that could be stored as intellectual property.
- **Patterns:** Messages containing numbers in the form ###-##-####, for example, typically contain Social Security numbers, and might be classified as patient information for a hospital and require special retention rules.

In contrast to manual classification, the automated approach places a limited burden on end users and decreases the risk of data being misclassified from user error or malicious intent. However, like other automated systems, classification systems are subject to false categorizations.

Third-party classification. Many organizations are deploying systems to categorize and manage records across multiple content types. They can then integrate these systems with leading e-mail archiving systems to allow the archive to store and optimize e-mail messages

while enabling the records management system to drive retention decisions that are consistent across different types of data.

Putting classification to work: Filtering, retention, and discovery

After sorting through e-mail messages and classifying them, Symantec intelligent archiving can take three actions: filtering, retention, and discovery.

Intelligent filtering

For many organizations, not everything needs to be archived—messages sent to all employees may not require archiving for every mailbox, and personal e-mail may not require archiving at all. Filtering out these noncritical messages helps reduce the total cost of ownership of archiving systems.

Intelligent retention

Intelligent retention bases retention policies on classifications the organization has established. For example, records managers can define a set of categories that map to distinct retention periods. The system can then determine the appropriate retention period for each message based on these categories. This approach helps reduce the risk of keeping some messages too long while not retaining others long enough.

Intelligent discovery

Classification systems can tag messages with metadata to enable effective search and retrieval. Some organizations review e-mail on a daily basis and may want to filter out messages that are clearly personal. Others may want to tag e-mail messages from the legal department as “possibly privileged” to help reduce the time required for future searches.

Implementing intelligent archiving with Symantec Enterprise Vault


When implementing an archiving system, each IT department must assess its own objectives and requirements and decide on an appropriate approach. But regardless of the direction they ultimately choose, many organizations would

be well served to add intelligence to their archiving policies.

The Symantec Enterprise Vault™ application provides a software-based intelligent archiving platform to store, manage, and search for enterprise data from e-mail systems, file server environments, instant messaging platforms, and content management and collaboration systems. Because not all data has equal value, Enterprise Vault utilizes intelligent classification and retention technologies to capture, categorize, index, and store target data to help enforce policies, protect enterprise assets, reduce storage costs, and simplify management. It also integrates specialized applications such as Discovery Accelerator and Compliance Accelerator to help enterprises mine archived data in support of legal discovery, content compliance, knowledge management, and information security initiatives.

In addition, Dell and Symantec have partnered to deliver a comprehensive Secure Exchange solution designed to protect critical Microsoft Exchange environments. Featuring state-of-the-art components from Dell and Symantec, this end-to-end solution is based on a modular reference architecture independently validated by Symantec for high performance, flexibility, and scalability. Enterprises can combine this solution with Symantec Enterprise Vault to facilitate intelligent e-mail archiving and regulatory compliance in Exchange environments.

Deploying a robust, efficient e-mail archiving system

Symantec intelligent archiving enables organizations to efficiently classify, filter, retain, and search for e-mail messages. By implementing intelligent archiving with Symantec Enterprise Vault, enterprises can create a robust e-mail archiving system to help simplify management and control resource costs while meeting regulatory and enterprise requirements. 

Art Gilliland is vice president of product management at Symantec.