Symantec Enterprise Vault™
Intelligent Archiving and Email
Classification, Retention,
Filtering, and Search

Nick Mehta
Vice President, Symantec
Symantec Enterprise Vault
Intelligent Archiving and Email
Classification, Retention, Filtering, and Search

Contents
Not all email is created equal .........................................................4
Email archiving considerations .....................................................6
Intelligent Archiving .................................................................8
Sorting the wheat from the chaff: Intelligent classification .................8
Volume challenge .................................................................9
Universality challenge ..........................................................10
Informality challenge ............................................................10
Intelligent classification approaches ........................................10
Putting the intelligence to work: Intelligent filtering, retention, and review ..........14
Intelligent filtering ..............................................................14
Intelligent retention ...............................................................14
Intelligent review .................................................................14
Common organizational practices ................................................14
About Symantec Enterprise Vault .............................................15
Not all email is created equal

With the recognition that email has become as mission critical as any other IT system, most organizations are evaluating their overall policies and systems for email management. Across many industries and public sector organizations, IT professionals are being called on to address the three most common management concerns regarding email:

- **Resource management**—How can the organization keep these systems running and costs under control? With sprawling message stores, longer backup windows, annoying end-user quotas, and out-of-control “rogue” archives (such as Microsoft PST files), IT is struggling to keep email up and running without breaking the budget.

- **Retention management**—How can the organization enforce a consistent retention policy on email? At the same time, IT organizations are being mandated by legal and compliance groups to implement enterprisewide policies on retention for email, rather than leaving it in end users’ hands.

- **Discovery management**—How can the organization quickly retrieve the content it needs within its mass of email? As email has increasingly become the “smoking gun” in litigation and regulatory investigations, most large organizations now know that if the email message is out there, they may be asked to find it.

Given these challenges, tens of thousands of enterprises across the world are evaluating or using email archiving software solutions. These systems typically allow IT to control email storage costs, while giving end users more user-friendly email storage and search and delivering to legal departments a consistent system for retaining and finding email messages across the enterprise.

As IT groups plan or implement these systems, they are realizing, however, that email archiving involves some important considerations:

- **Archive storage size**—While email archives often provide a fast ROI from a storage savings standpoint, they still create a great demand for storage. Since the data in question may be retained for many years, IT departments are seeking ways to optimize their archive storage costs.

- **Archive retention period**—Email archiving projects often force a necessary but challenging discussion within organizations about how long they should be keeping email messages. Many companies and government bodies have retention policies for traditional paper records, yet
they struggle to determine the appropriate policies for email. And they wrestle with how to make those policies practically and consistently work within a complex and growing email environment.

• **Archive search**—Finally, these same groups estimate the amount of data they will have in their archives over time and look for ways to reduce the search time and cost for finding the data they need.

In short, while archiving solutions greatly simplify the issues around storage, retention, and discovery that plague today’s email environments, they do not make those issues go away. What is the root of the remaining challenges? While every email message may have the same fundamental characteristics—a sender, a recipient, a subject, and a body—not every email message is of the same value to the organization. Consider the two messages shown in Figure 1, for example.

![Figure 1. Email message comparison.](image)

Clearly, both are important for their own reasons. The email on the left is a critical company document, an official “record” that drives a series of business actions to help Acme Corporation compete. And it may serve as evidence in the future if these actions are investigated for being anticompetitive. In contrast, the email on the right is important to the CEO but not to Acme Corporation’s future, unless his son is the head of BETA Corporation. Yet most messages are treated the same by default in an email archiving environment.
Most email archiving systems work in the following fashion:

• They capture all email messages from the environment—either immediately (referred to as journaling) or after some period of time (for example, 30 days).

• They store those messages for a period of time defined by the administrator (the retention period).

• They index the messages, their properties, and their attachments so that legal, finance, HR, or other groups can later find them.

**Email archiving considerations**

In implementing email archiving, three fundamental policy decisions need to be made:

• What should the organization archive (versus not archive)?
• How long should the organization keep it?
• How does the organization find it later on?

The decision on how long to keep information is perhaps the most challenging (see Figure 2). On the one hand, many business leaders would like to keep email messages as long as possible. Email is a vital part of doing business, and knowledge workers frequently go back to old email for information, for example, about business commitments, communications, or possible intellectual property leaks. These concerns are important, and the answers are in the email messages—as long as they are retained. This value exists independent of any regulatory or legal requirements imposed on the organization.

On the other hand, legal and IT professionals often see the downside to retaining email. First, every additional message retained (even in an archive) means more storage and IT cost. Second, keeping some messages longer than necessary may create risk for the organization later on (for example, if the message proves to be incriminating). Finally, the more email messages that an organization stores, the more they need to “wade through” when looking for one message in particular.
Today most organizations that use email archiving systems fall into one of three groups:

- **No automated archiving system**—The vast majority of IT groups have not yet implemented an email archiving software solution. While they do not have an automated system, these organizations still “archive” email messages—only in a very inefficient, ineffective, and risky fashion. IT archives email messages by retaining email server backups. Users archive email out of corporate control in the form of local data copies (such as Microsoft PST files) sprawled across PCs and servers. Management often learns that email messages that were deleted from the email server years ago still remain on a backup tape or laptop. These revelations often come to light as the company is forced to turn over data it didn’t even know it possessed to an opposing litigant or investigator. This is the worst scenario in that some email messages are retained longer than necessary, while others are deleted too soon, violating corporate or regulatory policies.

- **Archive but keep everything for the same period of time**—IT groups have driven many of the early email archiving deployments to reduce email storage costs and improve application efficiency. As such, retention policies have often been an afterthought. Many organizations have archiving systems that retain all email for the same period of time, such as one year, three years, or five years. Most of them have not yet reached the point where they must expire email actively. And those that have reached the end of their retention period often extend it to be on the safe side.

- **Archive but keep everything forever**—Some early adopters of email archiving governed their implementation based upon regulatory mandates. As these mandates were often vague in scope and length of time, regulated businesses frequently have indefinite retention policies on their archives and await further clarification from the government or depend on other organizations to take the first step.
Intelligent Archiving
Symantec believes there is a better way. Intelligent Archiving is the natural evolution of early email archiving software solutions (see Figure 3). Rather than treating all email the same, Intelligent Archiving entails:

- **Intelligent classification**—Deciding which email messages are relevant to which business purposes
- **Intelligent filtering**—Throwing out irrelevant email messages prior to archival, thereby reducing the size of the archive
- **Intelligent retention**—Determining based upon their classification how long to keep archived email messages
- **Intelligent review**—Tagging email messages during archival with metadata that makes them easier to find in the future

![Intelligent Archiving](image)

**Figure 3. Intelligent Archiving.**

Sorting the wheat from the chaff: Intelligent classification
If not all email is created equal, the question is how to differentiate the enormous number of email messages that organizations send and receive each day. As shown in Figure 1, email can be classified in a number of different categories. These categories can be very basic (business or personal) or very sophisticated, for example:

- 2005 tax records
- Reseller contracts for Germany
- Correspondence with service provider customers
- Employment issues in Asia
Email classification is not a new concept. The area of records management, known by other names, has existed since the dawn of the industrial era. Companies and government organizations have devoted substantial time, money, and personnel to storing official corporate documents in files, filing those files in boxes, storing those boxes in warehouses, and keeping track of the whole process. Indeed, very large businesses (for example, Iron Mountain) exist to outsource the storage and management of these paper records. In many cases, one or more records managers or clerks would be responsible for reviewing, classifying, and managing records and eventually disposing of them.

So why can't that be applied to email? Email introduces three new challenges that make the old approaches inadequate: volume, universality, and informality.

**Volume challenge**

Today, companies and people in those companies receive huge quantities of email messages. In the traditional model of records management, organizations were accustomed to dealing with thousands of official business records; therefore, the threshold for creating a “record” was very high. You 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 you have to do is click Send.

Essentially, email happens at the speed of thought, rather than the speed of print. Accordingly, the volume is daunting. An organization with 10,000 users receiving and sending 100 messages per day with 200 working days per year creates 200 million messages per year. Over a five-year period, that amounts to a billion messages (see Figure 4). To place that number in perspective, consider that Google indexes approximately 4 billion pages on the Web. Yet many large institutions create more content than that in a few years. Can one records clerk—or even 100—keep up with and classify all of that data?

Figure 4. Email volume example.
Universality challenge
So who creates all this data? And can’t they be required to follow a process and workflow? Records used to be created by defined groups—legal, finance, HR, and so on—that could be trained on company policy based on compliance demands. But now everyone from the chief executive officer to the chef in the cafeteria can and does send email messages. Employees across the globe create official “email records.” Contractors and outsourcing partners make this problem even more complex.

With myriad individuals across countries, languages, time zones, and corporate boundaries, organizations are challenged to disseminate and enforce documented email retention policies and guidelines. And with many employees now sending and receiving several hundred messages per day, any additional step beyond “send” in the email process is one step too many.

Many organizations have determined that they are not willing to stake their reputation and financial security on trusting every user to follow the process.

Informality challenge
Informality is perhaps the trickiest problem of all. Email messages, unlike corporate memos or faxes, are notoriously informal. A thread about last weekend’s activities can quickly morph into a discussion about this quarter’s sales forecast. A casual comment, when taken in context with the rest of the email corpus, can become major evidence in court.

Again, a gut reaction is often that “our employees should be more careful with email.” Yet, given the universal nature of email, formality cannot easily be enforced. And one of the reasons email is so popular is that work can get done quickly without the need for checks and balances.

Intelligent classification approaches
This section details three approaches to classifying email:

- Manual classification: Force users to do it as part of archiving.
- Automated classification: Have the archiving system do it.
- Third-party classification: Have another system do it.

Manual classification
Though one of the points of archiving is to take the decision away from end users, many organizations have concluded that a blend of automated archiving with some level of user oversight is needed. In this approach, a user lets the archiving system know how to classify an email message in the archive from within the email client (for example, Microsoft® Outlook®).
One method involves presenting a folder structure defined by the IT department to the end user in Outlook (in addition to his or her normal personal folders). This could map to a subset of the organization's corporate “file plan” or taxonomy of records classifications. For example, a salesperson in Acme Corporation might see the folders shown in Figure 5 in Outlook.

![Example folders displayed in Outlook.](image)

The salesperson can then drag email messages as he or she sends or receives them into these folders from Outlook. IT and Legal might have defined “sales contract” email messages to be stored for seven years, while purchase order email messages are stored for three years. Different groups of users could see different sets of folders, depending on their job structure. And messages left in the inbox or other folders could be kept for a default period of time, such as six months.

Alternatively, the user could be presented with a pop-up window when he or she sends or reads an email message. Such a window could display a list of categories and ask the user to choose the category to which this message corresponds. This list could also be further prefiltered to a set of categories that match the user’s group (for example, job function).

The advantage of manual classification is that sometimes only the user knows the true value of an email message. At the same time, this approach creates more work for users and can lead to inaccuracies due to user error or malicious intent.
Automated classification
The opposite approach is to put the decision making into the hands (or circuits) of the system. For many organizations, a perfect automated classification engine would be able to “figure out” what each message is and decide its relevance to the business.

Most classification engines today use a combination of approaches to analyze a message and determine type of content. Such approaches include:

• Evaluating senders and recipients (and the groups in which they reside) to determine probable content type. For example, messages from the legal department usually contain legal content.

• Evaluating message direction. For example, messages sent externally have a higher degree of scrutiny and retention.

• Evaluating messages for keywords or phrases. For example, messages and attachments are searched for the “confidential” disclaimer to identify data that could be stored as intellectual property.

• Evaluating messages for patterns. For example, messages are searched for ###-##-#### to identify Social Security Numbers and flag those as “patient information” for a hospital (with different retention rules).

• Evaluating messages for a combination of criteria. For example, messages sent from the finance department with a spreadsheet attachment are likely to be “financial documents.”

• Evaluating messages based upon machine learning. For example, “train” the system with 100 examples of intellectual property and have it learn how to detect IP in the future.

These are just a few examples of how automated classification can work. In contrast to manual classification, the automated approach places a limited burden on end users and decreases the risk of data being misclassified. However, classification systems, like other automated systems, are subject to the same false positives that occur in other domains. For example, many companies place their confidential disclaimer at the bottom of nearly every document (confidential or not), so the disclaimer is no longer useful for flagging intellectual property.
**Third-party classification**

Another approach to archiving is to depend on email classification performed by another system. Many organizations are looking at deploying records management systems to categorize and manage records across multiple content types (such as paper, email messages, documents, and so on). These systems can be integrated with leading email archiving solutions to allow the archive to store and optimize email messages while enabling the records management system to drive retention decisions that are consistent across different types of data (email and other).

For example, users could define their records as “file plan” in their records management system and use that system to present folders in Outlook. When a user drags an email message from the inbox to one of those records folders, the records management system would get a reference to the copy of the email message stored in the archive (but not the message itself). It would then notify the email archive to store the message for the period associated with that records folder. In this way, the user’s email is stored efficiently in one place (in the archive), with federated management being driven by the records management system.

Similarly, if an organization is using a gateway email monitoring solution (such as the Symantec™ Mail Security appliance), the policies and categories from this system could be passed to the archive through an x-header in the email message. In this way, the gateway could flag an email message as containing “intellectual property” and direct the archive to store it for an extended period of time.

**Summary**

Table 1 summarizes the strengths and weaknesses of each approach.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Description</th>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manual</td>
<td>User classifies by dragging into folder or by selecting options from a pop-up</td>
<td>User may be the only one who truly knows the value of email</td>
<td>More work for user; user may deliberately or inadvertently misfile email</td>
</tr>
<tr>
<td>Automated</td>
<td>System classifies by analyzing message properties and content</td>
<td>No burden on end user; repeatable process</td>
<td>No rule is perfect</td>
</tr>
<tr>
<td>Third-Party</td>
<td>Third-party records management or gateway system classifies email</td>
<td>Leverages existing classification system</td>
<td>Underlying system is likely automated or manual as well</td>
</tr>
</tbody>
</table>
Putting the intelligence to work: Intelligent filtering, retention, and review

After sorting through email messages and classifying them, archiving solutions typically can take three actions based upon classification: filtering, retention, and review.

Intelligent filtering

For many organizations, not everything needs to be archived, such as personal email. Company broadcast email may not need to be archived for every mailbox. Filtering out these noncritical messages can result in a lower total cost of ownership for your archive implementation.

For example, many organizations filter out common Internet email newsletters (for example, WSJ.com or ebay.com). Others choose not to archive server “bounce” messages and other internal content.

Intelligent retention

As discussed previously, intelligent classification can be used to retain items based upon their category. Company records managers can define a set of categories that map to distinct retention periods. The classification system can then help decide the length of time appropriate for each message based upon the categories listed. This reduces the risk of keeping some messages too long while retaining others for too short a period of time.

Intelligent review

Finally, classification systems can tag messages with metadata for more effective searching. Some organizations review email on a daily basis (for example, to comply with NASD and SEC regulations) and may want to filter from review messages that are clearly personal. Others may want to review any message sent externally to a competitor’s email domain that likely contains intellectual property. Still others may want to tag email messages from Legal as “possibly privileged” to reduce the search time involved in future discovery requests.

Common organizational practices

So far, this paper has presented a number of tradeoffs. Yet, this is the reality of email management today. Each IT department must assess its own business objectives and employee behavior and decide which approach is appropriate. Some common organizational practices include:

- Archiving all email for at least the same period of time that backup tapes were retained. If you kept your email tapes for 90 days, archive your email and keep it for at least 90 days. In this way, you preserve a similar level of retention but in a more easily searchable form.
• Placing “holds” on all email that is subject to outstanding investigations so that it is not deleted within the normal (for example, 90-day) window (a requirement under new amendments to the United States Federal Rules of Civil Procedure).

• Asking users to drag email into “records folders” in Outlook—managed by the archiving solution or by a third-party records management system—to classify email that needs to be stored for longer than the default period.

• Pushing out these folders only to users who tend to be process oriented (for example, legal, finance, HR) and accepting a manual filing requirement.

• Applying a default policy using automated classification for the other groups of users (for example, sales email are kept by default for three years) that are less likely to be open to manual filing.

• Enforcing an overriding policy to retain email containing sensitive information (such as intellectual property) that has been flagged at the archiving system or the gateway.

Regardless of the direction an organization takes, it would be well-served to add intelligence to its archiving policies. Intelligent Archiving enables companies to optimize storage, records retention, and records discovery and to capture business value from their archiving implementation.

About Symantec Enterprise Vault
Symantec Enterprise Vault provides a software-based Intelligent Archiving platform to store, manage, and discover corporate data from email systems, file server environments, instant messaging platforms, and content management and collaborations systems. Because not all data is created equal, Enterprise Vault utilizes intelligent classification and retention technologies to capture, categorize, index, and store target data to enforce policies and protect corporate assets, while reducing storage costs and simplifying management. It also provides specialized applications such as Discovery Accelerator and Compliance Accelerator to mine archived data in support of legal discovery, content compliance, knowledge management, and information security initiatives.

Visit our Web site: www.symantec.com/enterprisevault
About Symantec
Symantec is a global leader in infrastructure software, enabling businesses and consumers to have confidence in a connected world. The company helps customers protect their infrastructure, information, and interactions by delivering software and services that address risks to security, availability, compliance, and performance. Headquartered in Cupertino, Calif., Symantec has operations in 40 countries. More information is available at www.symantec.com.