Intelligent Archiving and Email: Classification, Retention, Filtering & Search

Nick Mehta, Vice President, Symantec

January, 15, 2007
Not All Email Is Created Equal

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

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

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

- **Discovery management**: How can I quickly get to what I need within our mass of email? Finally, as email has increasingly become the “smoking gun” in litigation and regulatory investigations, most large organizations now know that if the email 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 to manage the issues above. These systems typically allow IT to control the growth in email storage costs, while giving the end-users email storage and search in a more user-friendly manner and delivering to legal departments a consistent system for retaining and finding emails across the enterprise.

As IT groups plan or implement these systems, they are beginning to realize, however, that email archiving demands some sizeable considerations:

- **Archive storage size**: While email archives often provide a fast ROI from a storage savings standpoint, they still create a large demand for storage. Given that 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: how long should we be keeping all of this stuff anyways? Many companies and government bodies have had retention policies for traditional paper “records,” yet they struggle to determine what the applicable policies are in email. And they further 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 constantly look for ways to reduce the search time and cost for finding the data that 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 altogether.

What is the root of the remaining challenges? Fundamentally, while every email message may have a sender, a recipient, a subject and a body, not every email message is of the same value to the organization. Consider the following two messages:

Clearly both are important for their own reasons. The left email is a critical company document – an official “record,” if you will. It drives a series of business actions that hopefully will help Acme Corporation compete. And it might be evidence for the future in case this action is investigated for being anti-competitive. In contrast, the right email is also very important to the CEO, but may be less important to Acme Corporation’s future, unless his son is the head of Beta Corporation. Yet most would be treated the same by default in an email archiving environment.

Most email archiving systems work in the following fashion:

- They capture all email from the environment – either immediately (referred to as journaling) or after some period of time (e.g., 30 days).
- They store that email for a period of time defined by the administrator (the “retention period”).
- They index the email, its properties and its attachments so that legal, finance, HR or other groups can later find it.

In email archiving implementations, then, there are three fundamental policy decisions to make:

- What should I archive (versus not archive)?
- How long should I keep it?
- How do I find it later on?
In particular, the decision on how long to keep information is perhaps the most challenging:

<table>
<thead>
<tr>
<th>Email As An Asset</th>
<th>Email As A Liability</th>
</tr>
</thead>
<tbody>
<tr>
<td>▶ User value</td>
<td>▶ Smoking guns</td>
</tr>
<tr>
<td>▶ Business value</td>
<td>▶ Cost of review</td>
</tr>
<tr>
<td>▶ Business history</td>
<td>▶ Storage costs</td>
</tr>
</tbody>
</table>

- **Email as an asset**: On one hand, many business leaders would like to keep email as long as possible. Email is a vital part of doing business and knowledge workers frequently go back to “old emails” to make decisions. Did we commit that to the partner? What did our customer service rep say to the customer? Was any intellectual property leaked to our competitor? The questions are important and the answers are in the email – as long as it is retained. This value exists independent of any regulatory or legal requirements imposed on the organization in question.

- **Email as a liability**: On the other hand, legal and IT professionals often see the downsides to retaining email. First, every additional message retained (even if in an archive), means more storage and IT cost. Second, keeping some messages longer than they have to be kept may create risk for the organization later on (e.g., if the email proves to be incriminating). Finally, the more email that an organization keeps means that there is more to “wade through” when looking for one email in particular years later.

Today most organizations that use email archiving systems fall into one of three camps:

- **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, the truth is that these organizations still “archive” email – just in a very inefficient, ineffective and risky fashion. IT “archives” email by retaining email server backups “for a rainy day.” Users “archive” email out of corporate control in the form of local data copies (e.g., Microsoft .PST files) sprawled across PCs and servers. Management is often shocked to learn that email that was deleted from the email server years ago still remains a skeleton in the closet on a backup tape or laptop. These revelations are ever the more painful.
because they often come up as the company is forced to turn over data it didn’t know it even possessed to an opposing litigant or investigator. This situation is the worst of all worlds, in that some email is kept longer than it has to be retained while other email is 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 purely out of a desire to reduce email storage costs and improve application efficiency. As such, retention policies were often an after-thought. Many of these organizations have archiving systems that keep all email for the same period of time – e.g., 1 year, 3 years or 5 years. Truth be told, most of these customers have not yet gotten to the point where their “clock is up” and they are expiring email actively. Those that reached the end of their retention period often extended it – just “to be safe.”

- **Archive but keep everything forever:** Some early adopters to email archiving drove their implementation based upon regulatory mandates. Given that these regulatory mandates were often vague in scope and length of time, regulated businesses frequently have indefinite retention policies on their archives – waiting for further clarification from the government or other organizations to take the first step before they do.

We believe there is a better way. Intelligent Archiving is the natural evolution of early email archiving software solutions. Rather than treating all email the same, Intelligent Archiving entails:

- **Intelligent Classification:** Deciding which emails are really relevant to which business purposes.
- **Intelligent Filtering:** Throwing out non-relevant emails prior to archival, thereby reducing the size of the archive.
- **Intelligent Retention:** Determining based upon the classification how long to keep archived emails.

![Intelligent Archiving Diagram]
Intelligent Review: Tagging emails during archival with metadata that makes them easier to find down the road.

Sorting the Wheat from the Chafe: Intelligent Classification

If we are convinced that not all email is created equal, the interesting question is how we decide what is what within the huge amount of email organizations send and receive each day. Going back to the previous two example emails, we know that each email can be classified into a number of different categories. These categories can be very basic (e.g., business or personal) or could be very sophisticated – e.g.:

- 2005 tax records
- Reseller contracts for Germany
- Correspondence with service provider customers
- Employment issues in Asia

The truth is that this is not a brand new concept for organizations. The area of “Records Management” has been around since the dawn of the industrial era (though it wasn’t called by that name at the time). Companies and government organizations have devoted substantial time, money and personnel to taking official corporate documents, storing them in files, filing those files in boxes, storing those boxes in warehouses and keeping track of the whole process. Indeed, very large businesses exist (e.g., Iron Mountain) to outsource the storage and management of these paper “records.” In many cases, one of more “records managers” or clerks would be responsible for reviewing records, classifying them and managing and eventually disposing of them.

So why don’t we just do that for email? If only it were that easy. The challenge is that email introduces three new “wrinkles” that make the old approaches inadequate.

Wrinkle #1 – Volume

In short, companies and people in those companies get a LOT of email. In the traditional model of records management, organizations were used to dealing with thousands of official business records. 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), etc. Now all you have to do is hit “SEND.”

Essentially, email happens at the speed of thought, rather than the speed of print. Because of this, 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 5 year period, that is a billion messages.
To put that number in perspective, Google indexes approximately 4 billion pages on the Web. Yet many large institutions with create more content than that in a few years.

Is one records clerk – or even 100 – going to be able to keep up and classify all of that data?

_Wrinkle #2 – Universality_  
So who creates all this data? Can't we just tell them to follow a process and workflow? Can't we just have a meeting and solve this problem?

The truth is that in the “old word,” the answer was “yes.” Records were created by defined groups – legal, finance, HR, etc. – who could be trained and threatened into compliance.

But now everyone from the Chief Executive Officer to the chef in the cafeteria can and do send email messages. Employees exist across the globe and create official “email records.” Contractors and outsourcing partners make this problem even more complex.

Can you really count on this myriad of individuals across countries, languages, time zones and corporate boundaries, to learn documented email retention policies and guidelines? And even if they could learn them, would they follow them. 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.

_Wrinkle #3 – Informality_  
This is perhaps the trickiest problem of all. Email messages, unlike corporate memos or faxes of the past, are notoriously informal. A thread about what you did last weekend can quickly morph into a discussion about this quarter’s sales forecast. A casual joking comment, when taken in context with the rest of the email corpus, can become major evidence in court.
Again, our gut reaction is often that “our employees should be more careful with email.” And that is correct – everyone should be more careful. As they say, never write unless you’re okay with your mom seeing it in the New York Times. Yet, with the universal nature of email described above, formality in email cannot easily be enforced.

Indeed, one of the reasons email is so popular is that work can get done quickly without the need for checks and balances (or even grammar, in some cases).

**Intelligent Classification – What’s Possible?**

In this world of everyone sending too much poorly worded email, how can an organization classify it all? Fundamentally there are three approaches which we will discuss:

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

**Intelligent Classification – Manual Classification**

It may sound hypocritical to talk about manual classification since one of the points of archiving is to take the decision away from the end-users. Yet many organizations have concluded that a blend of automated archiving with some level of user oversight is what is needed.

In this approach, a user lets the archiving system know how to classify an email in the archive from within the email (e.g., Microsoft Outlook) client.

One method involves presenting a folder structure to the end-user in Outlook (in addition to his/her normal personal folders) defined by the IT department. This could map to a subset of the organizations corporate “file plan” or taxonomy of records classifications. As an example, a sales person in Acme Corporation might see the following folders in Outlook:

[Insert graphic of folders with names “Purchase Orders,” “Quotes,” “Sales Contracts,” “Customer Issues”]

The sales person can then drag emails as he sends or receives them from Outlook into these folders. IT and Legal may have defined Sales Contract emails to be stored for 7 years while Purchase Order emails are stored for 3 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 (e.g., 6 months) period of time.

Alternatively, the user could be presented with a pop-up when he sends or read an email messages, presenting a list of similar categories and asking him to choose the category to which this message corresponds. This list can further be pre-filtered to a set of categories that match the user’s group (e.g., 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.

**Intelligent Classification – Automated Classification**

The other side of the coin is to put the decision making in the hands (or circuits) of the system. For many organizations, a perfect automated classification engine would be nirvana. The system would “figure out” what each message is and decide its relevance to the business.

So what can classification engines do today? Most use a combination of approaches to analyze a message and determine what type of content it is. These include:

- Evaluating senders and recipients (and the groups in which they reside) to determine probable content type – e.g., messages from the Legal Department tend to be legally oriented.
- Evaluating message direction – e.g., messages sent externally have a higher degree of scrutiny and retention.
- Evaluating messages for keywords or phrases – e.g., searching messages and attachments for the “confidential” disclaimer to identify data that could be stored as “intellectual property.”
- Evaluating messages for patterns – e.g., searching messages 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 – e.g., messages sent from the Finance Department with a spreadsheet attachment are likely to be “financial documents.”
- Evaluating messages based upon machine learning – e.g., “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 of the examples of how automated classification can work. In contrast to Manual Classification, the automated approach puts limited burden on end-users or risk that they mis-classify data. However, classification systems like other automated systems, are subject to the same “false positives” that we see in other domains. As an example, many companies put their confidential disclaimer at the bottom of nearly every document (confidential or not) so the disclaimer is no longer always useful in flagging intellectual property.

**Intelligent Classification – Third-Party Classification**

A third approach is for the archiving solution 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 (e.g., paper, email, documents, etc.).

These systems can be integrated with leading email archiving solutions to allow the archive to store and optimize the email while enabling the Records Management system...
to drive retention decisions that are consistent across different types of data (email and otherwise).

As an example, a user could define their records “file plan” in their Records Management system and use that system to present folders to the user in Outlook. When a user drags an email from the inbox to one of those records folders, the Records Management system would get a reference to the copy of the email stored in the archive (but not the email itself). It would then notify the email archive to store the email 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 as containing “intellectual property” and direct the archive to store it for an extended period of time.

**Intelligent Classification – Summary**

The table below 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 only one who truly knows value of email</td>
<td>More work for user User may inadvertently or inadvertently misfile email</td>
</tr>
<tr>
<td>Automated</td>
<td>System classifies by analyzing message properties and content</td>
<td>No burden for 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>Leverage existing classification system</td>
<td>Underlying system is likely automated or manual as well</td>
</tr>
</tbody>
</table>

**Putting the Intelligence to Work: Filtering, Retention and Review**

So now that you’ve sorted through all of this email and figured out what’s important, what do you do with that knowledge? Archiving solutions can typically take three actions based upon classification.
Intelligent Filtering: Should I Archive This At All?

For many organizations, not everything needs to be archived. Personal emails may not need to be archived. Company broadcast emails may not need to be archived for every mailbox. And filtering out these non-critical messages can result in a lower Total Cost of Ownership for your archive implementation.

As an example, many customers take lists of common Internet email newsletters (e.g., WSJ.com or ebay.com) and filter those out altogether. Others take server “bounce” messages and other “internal spam” and choose to not archive this content.

Intelligent Retention: How Long Should I Keep This?

As discussed previously, Intelligent Classification can then 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: How Do I Find This Later On?

Finally, classification systems can then tag the messages with “metadata” for more effective searching in the future. Some organizations review email on a daily basis (e.g., for compliance with NASD and SEC regulations) and may want to filter “out” messages that are clearly personal (e.g., the ebay newsletters above) from review. Other organizations may want to proactively review any message sent externally to a competitor’s email domain that contains likely “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.

Wrapping It Up: Common Customer Practices

So far, this paper has presented a lot of tradeoffs – a lot of gray and very little black and white, if you will. Yet this is the reality of email management today. There is no clear silver bullet – from the law or from vendors – that makes the problem the same for every organization. Although it is cliché to say, each IT department must assess its own business objectives and employee behavior and decide what’s appropriate.

That being said, some common customer practices that we have seen include:

- Archiving all email for at least the same period of time that you previously retained your backup tapes. If you kept your email tapes for 90 days, archive your email and keep it for at least 90 days. You are preserving a similar level of retention but you are now keeping it in a more easily searchable form.
- Placing “holds” on all email that is subject to outstanding investigations so that this email is not deleted within the normal (e.g., 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 these folders only out to users that tend to be process-oriented (e.g., legal, finance, HR) and accepting of a manual filing requirement.

Applying a default policy using automated classification for the other groups of users (e.g., sales email are kept by default for 3 years) that are less likely to be open to manual filing.

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

While these approaches are not universally adopted by all companies by any means, these represent some of the common trends we are seeing in our customer base.

In summary, regardless of what direction a given organization takes, customer are well-served to add intelligence to their archiving policies to make the optimal balance of storage optimization, records retention and fast discovery and to capture the business value that they desire 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, 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 in order to enforce policies and protect corporate assets while reducing storage costs and simplifying management. Enterprise Vault then provides specialized applications such as Discovery Accelerator and Compliance Accelerator to mine archived data to support legal discovery, content compliance, knowledge management and information security initiatives.

Visit our web site: [www.symantec.com/enterprisevault](http://www.symantec.com/enterprisevault)

**About Symantec**
Symantec is the global leader in information security providing a broad range of software, appliances and services designed to help individuals, small and mid-sized businesses, and large enterprises secure and manage their IT infrastructure. Symantec’s Norton brand of products is the worldwide leader in consumer security and problem-solving solutions. Headquartered in Cupertino, Calif., Symantec has operations in more than 40 countries.

**Symantec Corporation World Headquarters**
20330 Stevens Creek Boulevard
Cupertino, CA 95014 USA
408 517 8000
800 721 3934
www.symantec.com