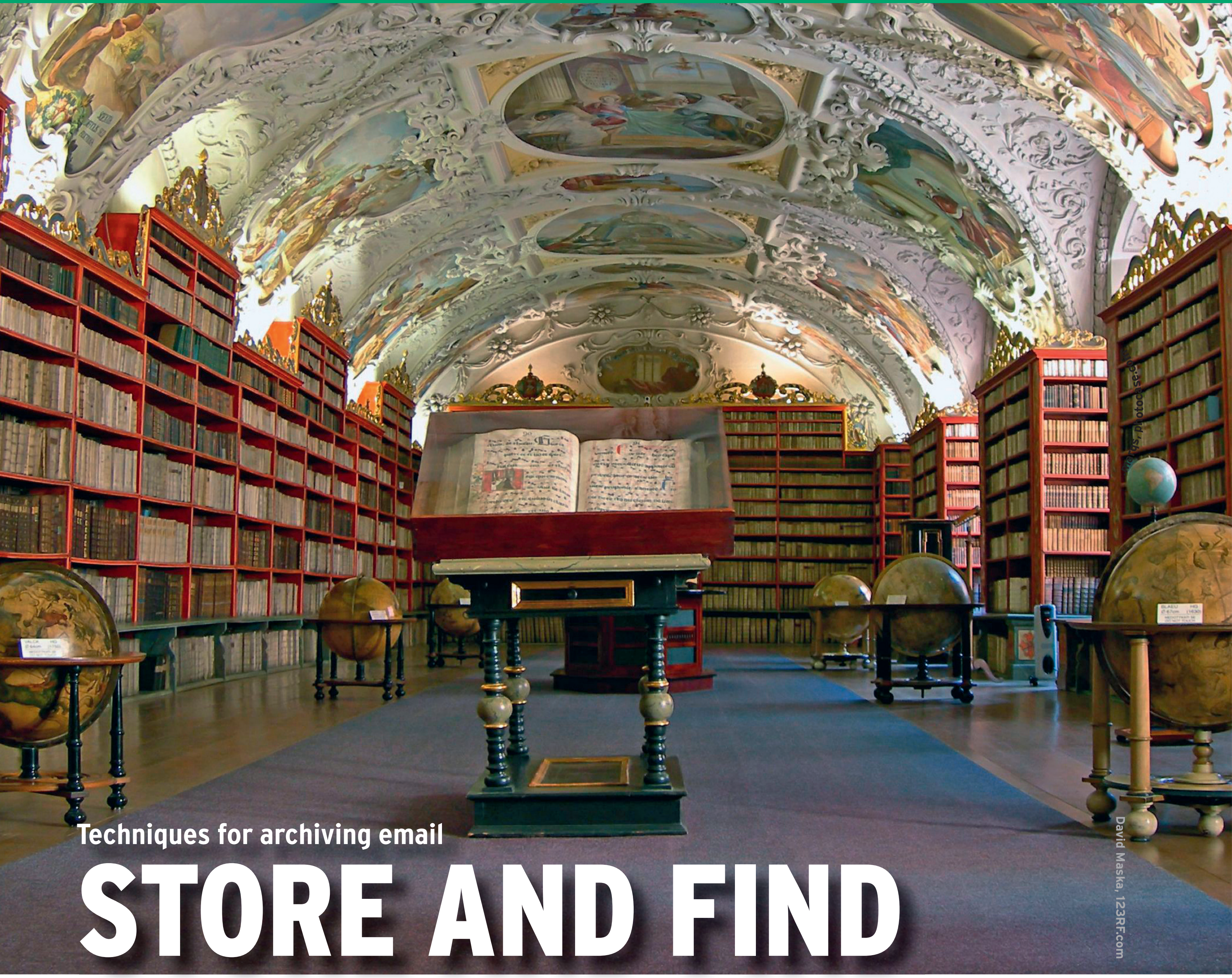


David Maska, 123RF.com

Techniques for archiving email

STORE AND FIND

Email archiving involves more than just backing up your email directories. It is also a question of classifying the email and making it easier for users to find their way around overfilled email folders. **BY JÖRG FRITSCH**

In the past few years, very few technology developments have been as versatile and necessary as email archiving. In fact, it ranks high on the best practices list of many corporations. Many administrators who hear the term "email archiving" for the first time think about data privacy and automatically file the term away in the category of backup. However, email archiving makes sense outside the scope of backup, if you think of it in terms of email management.

Document Lifecycle Management and Compliance

For decades, corporations have developed their filing systems into something approaching an art form; thus, they can protect records and retain them for a defined period of time. To cope with these mountains of paper, many policies have been introduced to define how long to retain certain document types before they finally end up in the shredder. Little is left to common sense: In the paper world, everything is strictly organized.

Email management applies these policies and processes to email, thus ensuring harmonized document lifecycle management. Although it is not a question of keeping meaningless messages indefinitely, although some storage vendors might recommend doing so, you can't just adopt quotas and a policy of benign neglect. Just like physical filing, email management requires a policy that, ideally, defines which email the system should keep (and for how long) and which it should delete beyond any possibility of recovery.

Not many corporations have actually achieved this ideal state. Because of a lack of email classification tools, most businesses are unable to distinguish between Internet email (i.e., email that reached the company across the Internet) and internal email and between business and private correspondence. Currently, only two classes typically exist: spam and non-spam. Administrators who want to try out more granular classifications or experience the look and feel of a professional application (e.g., by Titus Labs [1]) can use POPFile [2] on Linux to generate multiple mail classes. For more details on practical email classification, see the "Practical Email Classification" box.

Retention Periods

Compliance rules define how long a corporation must retain email (see the "Compliance Sources" box). Rules in the US are stipulated by HIPAA (Health Insurance Portability and Accountability Act) and SOX (Sarbanes-Oxley). HIPAA mandates archiving of patient-related data for two years after the patient's death. SOX mandates retention of relevant data up to four years after an audit. Major banks or pharmaceutical companies typically are multi-national, so compliance requirements of other countries must be taken into account as well.

Most of these frameworks don't refer to email directly, but to data archiving, and opinions differ among professionals as to which rules and periods actually apply to email.

Mail Server Diet

Even if you find the discussion about document lifecycle management and compliance too abstract, you might enjoy looking into email management from a technical point of view.

Simply doing nothing would be the wrong decision, because it would leave everything to the user's discretion. And administrators will be aware that this is not a good idea. Many users keep everything and never delete their email. Quotas that give the users a limited amount of space for their inboxes are an initial approach, but quotas only really work in the real world for hosting platforms, Internet service providers, and web mail providers. In an enterprise environment, an administrator is more likely to accept an oversized inbox because automatic *Recipient out of quota* messages look unprofessional to customers or business partners.

Where quotas are in effect, users sometimes resort to self-managed offline archives, such as Microsoft Exchange *.pst* files. Archives of this type are problematic because a corporation can never know where its employees are storing their email or whether they are covered by a backup plan. Users tend to manage offline archives locally on laptops and PCs that are not covered by the backup policy. If the laptop is lost or damaged, the company does not have a backup, and it is difficult to assess how much confidential data have been lost. With Exchange, another problem is that users occasionally let their *.pst* files grow to a size that suddenly causes irreparable damage independent of the operating system.

Practical Email Classification

In a production environment, classifying email means adding tags to email subject lines, for example, and adding metadata in the form of additional "X" headers, which mail programs then use to sort and filter messages.

One classification that most users will be confronted with on a daily basis distinguishes between spam and legitimate mails. Another approach commonly considered is adding an additional "internal" tag to internal, confidential email. The enterprise email gateway could then use this tag to prevent confidential data from reaching the Internet, for example, if a message was inadvertently forwarded to an external address. The email client would find the tag in the forwarded message and refuse to dispatch it.

Many classifications are conceivable, but none can be implemented ad hoc in a production environment because they rely on having a policy in place and on user training. The users must know how to apply the policy rules and use the tools to leverage the benefits; user buy-in for the procedure is essential.

Practical Archiving

Basically, there are three methods of archiving email:

- Use an email archiving service (hosted solution).
- Archive incoming and outgoing email by means of network recording or deploying additional smart hosts (appliances) in the email flow.
- Archive all email (including internal messages) by integrating some kind of groupware (e.g., Microsoft Exchange, Axigen [5], etc.). Journaling accounts, or accounts with elevated privileges, are typically required to copy all email to the archiving software or the archiving appliance.

Hosted services (i.e., outsourcing) are only interesting for administrators who simply need to archive incoming and outgoing email and can leave everything else to the users. From a technical point of view, the service provider acts as the DNS MX record for all incoming email and enters its own details as a smart host in all outgoing mail. The biggest provider in this field is Message Labs [6]. This variant is of most interest to corporations who already have accounts with these providers and use their spam and virus filters.

If you're not a customer of one of these services but do want to archive email on Linux, you can capture email at any suitable point and push it into the archive. The easiest option is to let a Linux system capture the mail flow on a SPAN port (Switched Port Analyzer, mirror port). The Linux machine would simply run a *tcpdump* command to archive all network traffic on TCP port 25 (SMTP).

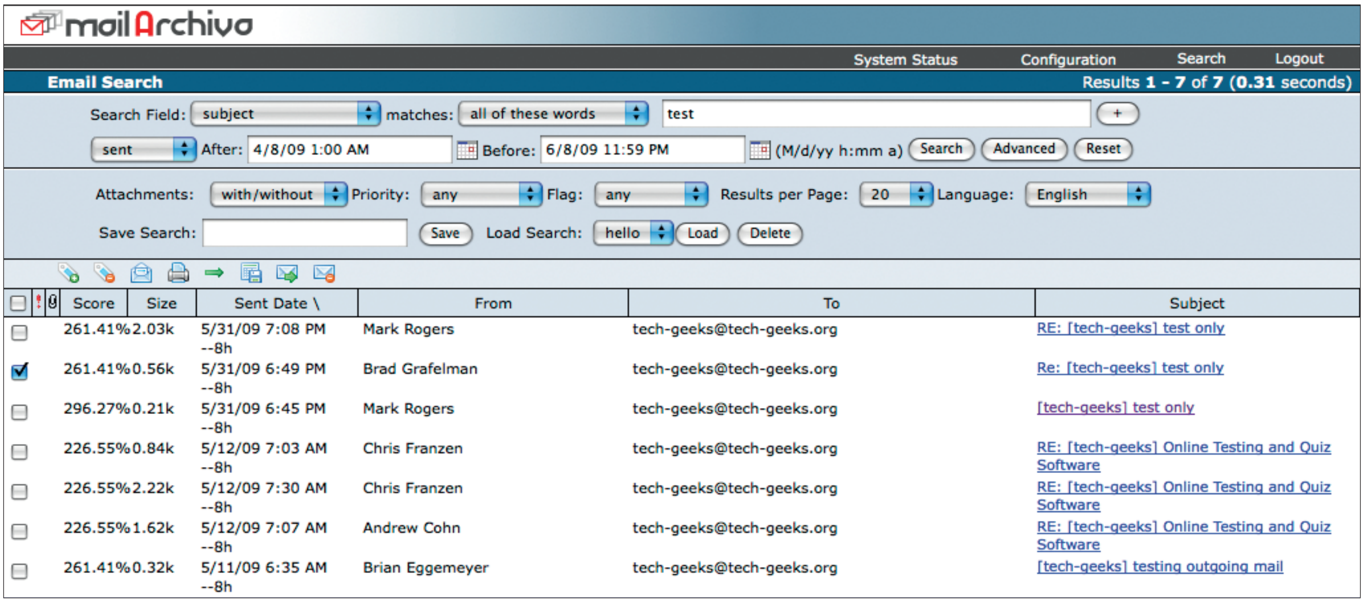
A medium-sized company could use a simple recorder and about 1TB of disk space to archive one or two years' worth of email. The use of tcpdump for ar-

Compliance Sources

Motivation for companies to adhere to compliance rules vary. Companies typically seek compliance to meet legal requirements, to save money, or to ensure that potential customers will take them into consideration.

- Companies listed on the stock exchange must comply with exchange supervisory rules, such as SOX.
- Companies that develop or produce medication or compete for contracts in the healthcare sector must comply with HIPAA rules.
- Companies who process credit card information must comply with a number of requirements imposed by the card processor, or they will be disqualified.
- Companies applying for loans must implement the Basel II rules [3] or expect higher interest rates for failure to comply.

This list is not exhaustive; in particular, one could add rules imposed by government offices, the armed forces, industrial associations, suppliers, and many others.



Email Search Results **1 - 7** of **7** (**0.31** seconds)

Search Field: subject matches: all of these words test +

sent After: 4/8/09 1:00 AM Before: 6/8/09 11:59 PM (M/d/yy h:mm a) Search Advanced Reset

Attachments: with/without Priority: any Flag: any Results per Page: 20 Language: English

Save Search: Save Load Search: hello Load Delete

□	Score	Size	Sent Date \	From	To	Subject
□	261.41%	2.03k	5/31/09 7:08 PM --8h	Mark Rogers	tech-geeks@tech-geeks.org	RE: [tech-geeks] test only
☑	261.41%	0.56k	5/31/09 6:49 PM --8h	Brad Grafelman	tech-geeks@tech-geeks.org	Re: [tech-geeks] test only
□	296.27%	0.21k	5/31/09 6:45 PM --8h	Mark Rogers	tech-geeks@tech-geeks.org	[tech-geeks] test only
□	226.55%	0.84k	5/12/09 7:03 AM --8h	Chris Franzen	tech-geeks@tech-geeks.org	RE: [tech-geeks] Online Testing and Quiz Software
□	226.55%	2.22k	5/12/09 7:30 AM --8h	Chris Franzen	tech-geeks@tech-geeks.org	RE: [tech-geeks] Online Testing and Quiz Software
□	226.55%	1.62k	5/12/09 7:07 AM --8h	Andrew Cohn	tech-geeks@tech-geeks.org	RE: [tech-geeks] Online Testing and Quiz Software
□	261.41%	0.32k	5/11/09 6:35 AM --8h	Brian Eggemeyer	tech-geeks@tech-geeks.org	[tech-geeks] testing outgoing mail

Figure 1: The Mail Archiva server can use different approaches to integrate with the mail flow, including a Sendmail milter.

chiving purposes sounds complicated and not very user friendly, but this approach works if you don't need to restore email very often (i.e., one to five times a year). The Net VCR appliance by Niksun [7] provides a commercial alternative to tcpdump.

Setting up a Smart Host

A smart host that copies all your email as it passes through is easy to set up: Many appliances that archive email are smart hosts. The only thing that distinguishes them from a do-it-yourself smart host is more convenience in email searching and, in the high-end sector, the kind of media the software uses for archiving.

WORM drives (Write Once, Read Many) are often suggested by service providers: WORM-based appliances will write to DVDs, tapes (this allows you to go on using existing, legacy backup systems) or special arrays. Right now, WORM-based systems are only typical in the financial industry.

The typical issues with do-it-yourself solutions are a lack of user-friendliness and possibly a lack of scalability. Good solutions are easy to use and at least offer some kind of user involvement by giving the user access to their archived email throughout the retention period.

Other solutions tie in directly with the email system. Plugins let users access their own archived email transparently using Outlook or Lotus Notes. Symantec Enterprise Vault [8] is an example of this kind of solution. One decision administrators must make before deploying archiving technology relates to the method the software solution will use to archive the raw data and how it will modify the data, if at all.

Metadata and Pointers

Email not only comprise payload data and usage information, but also meta-

Mail Archiva: An Open Source Solution

Mail Archiva [4] is an open source program that is also offered commercially by the same vendor. The two variants differ with respect to some features. Only the commercial version lets you keep the metadata and reduce space requirements for archiving by using pointers to reference attachments. The open source variant might be sufficient for small to medium-sized companies.

Mail Archiva supports nearly all the functions discussed in this article and can easily be integrated with Microsoft Exchange or Lotus Notes. To allow this to happen, the program uses journaling accounts: These are privileged accounts that the administrator sets up (e.g., on the Exchange Server), which then receive copies of all externally (Internet) and internally (from one local Exchange user to another) routed mail. The system then uses the IMAP protocol to archive the messages.

Journaling accounts are a feature offered by both Exchange and Lotus Notes. To be as generic as possible and support more or less any email solution, Mail Archiva can also use a milter (mail filter, e.g., in Sendmail or Postfix) or a POP- or SMTP-based approach for collecting email (Figure 1).

Easy Configuration

Mail Archiva is fairly easy to configure. Administrators will need to make the most important decisions outside the product, because they relate to the performance and speed of the storage solution (NAS, SAN), the storage medium (WORM?), and the policy. Mail Archiva is one of the solutions that gives the users (i.e., the mail owners) web-based access to their own email. Users can send any email that has been deleted back to themselves at the click of a button. The program simply SMTPs the email to the user's email server in this case, as if the message originated with a mail relay.

Mail Archiva is an extremely lean solution that will generally work with (more or less) any system on the market. The ability to integrate with product-independent resources is useful if you manage a heterogeneous environment or want to roll out a mail system release change. Products that integrate more closely with the email system proper often cause issues when changes need to be made.

In heterogeneous environments, a generic product allows the administrator to offer the same archive system to all end users (whether they use Lotus Notes or an Axigen account).

Table 1: Alternative Technologies

Requirements	Technology/Product	Benefits
Informal conversations	Instant Messaging (IM)	Brief, informal conversations do not stress the mail server and inboxes
Collaboration	IM, Web EX, Zoho, Fuze, Groove, Joomla, Sharepoint, Alfresco	More dimensions to collaboration than with email (shared desktops, interactive)
Document exchange	Sharepoint portal, Alfresco	Easily indexed, managed project, spaces, full document lifecycle management

data and headers that say where the message originated, which route it took to reach your company, or how it left the company.

The available technologies and products differ greatly in this respect. In many cases, additional metadata are generated by the archiving process; some solutions completely discard the metadata because they only store the route to the smart host or service provider. Other solutions let you view, but not extract, the metadata.

Because choosing an archiving system will have long-term effects, it is important to check your company's email policy to see which features you need. One major criterion is the way the system archives email. Some solutions write email messages as files to the filesystem (with a SAN or NAS), whereas others store email messages in a database.

Most systems that use databases replace attachments with pointers: If the same attachment is sent with multiple mail messages, the attachment is only stored once rather than duplicated. The same thing applies to messages with multiple recipients in the To, CC, or BCC fields. A combination of pointers and compression can save a huge amount of disk space.

With regard to disk space, you have another area of concern: Some solutions copy user email to the archive and still let users manage their "live email." In this case, the archive is simply used for auditing purposes, and the users are just as exposed to the flood of email as they would be without an archiving service. Other solutions move email into the archive and thus save each message once only on the enterprise network.

E-Discovery: Email Archiving and Auditing

An archive is not there just to keep the mail server lean and fast or to help users with email management tasks. E-discovery is typically heard in the context of email archiving, and e-discovery tools keep evidence accessible and confidential. In other words, the software indexes the stored data and then supports full-text searching.

Besides email and instant messages, e-discovery includes messages transmitted via other technologies. Some major banks go so far as to include the voice messages from VoIP phone mailboxes in their document lifecycle and use this material as documented evidence or for audit purposes.

An electronic signature for each email in the archive is not typical – this would add overhead for PKI, PKI policies, and key escrow to email archiving (see the "Key Escrow" box). Some products sign the archive on a daily basis, but this doesn't improve confidentiality unless you have a robust enterprise policy in place with respect to the keys used for this purpose.

Email – A Legacy Technology?

Although email is widespread as a communications technology, the technology itself has not developed greatly in the past eight to 10 years. Users today use email for chatting, collaboration, and exchanging documents.

Alternatives that support collaboration and communication, and add the ability to archive files or messages, are listed in Table 1. Besides the benefits stated in column three of the table, these newer tools (at least theoretically) reduce the amount of email that users need to send.

Some experts predict that this development will make email a legacy technology within the next five to 10 years. Although it is hard to say whether this will happen, after years without any major changes in the use of email, email archiving is at least an innovation that improves and facilitates the use of the medium. ■

Key Escrow

Once an enterprise has set up a Public Key Infrastructure (PKI), the keys required to decipher information encrypted by end users can be deposited securely – this is key escrow. The decision for or against key escrow is typically taken at the start of the enterprise PKI planning phase.

E-discovery is a popular field of application for key escrow. Other applications support users who have lost their keys or decrypt business-critical information that was encrypted by users who are unreachable (e.g., because they have changed jobs or in case of extended sickness). A number of different implementations for these critical areas help mitigate the risks and worries.

Key escrow is also referred to as "key recovery" or "data recovery."

INFO

[1] Titus Labs Message Classification: *http://www.titus-labs.com/software/Classification_default.html*

[2] Tutorial on mail classification with POPFile: *http://www.howtoforge.com/popfile_ubuntu_feisty*

[3] Basel II: *http://www.bis.org/publ/bcbs107.htm*

[4] Mail Archiva: *http://www.mailarchiva.com*

[5] Axigen Mailserver: *http://www.axigen.com*

[6] Message Labs: *http://www.messagelabs.com*

[7] Niksun Net VCR: *http://www.niksun.com/product.php?id=3*

[8] Symantec Enterprise Vault: *http://www.symantec.com/business/enterprise-vault*

THE AUTHOR

Jörg Fritsch majored in Chemistry and has worked in software development and IT security ever since. He joined the NATO C3 Agency as an Engineer of Communication & Information Security in 2003. Jörg has published various articles on load balancing, TCP/IP, and security.