

# Email Archiving: So, What's Wrong With Saving Everything?

Written by Peg Duncan

Presenters:  
Jason Baron  
Peg Duncan

April 2-4, 2009  
[www.techshow.com](http://www.techshow.com)

## How did we get into such a mess?

### *A short history of e-mail*

In 1971 Ray Tomlinson of ARPANET sent the World's first e-mail, and by 1972 most of the features we would recognize today were there – the memorandum style of address with to: from: cc: subject:, the brevity and informal style, and the tolerance for spelling mistakes and typographical errors. In the 1980's, closed proprietary networks included an electronic mail function, famously figuring in the Iran-Contra affair of the same period. Electronic mail was included in local area networks used for sharing resources, and while individual networks could serve up to a hundred people, there was little connectivity beyond the local network until the arrival in 1995 of enterprise networking systems and enterprise electronic mail based on *de facto* internet protocols. Once the corporate network was permanently connected to the internet, messages could be sent to any recipient anywhere.

Employees were as likely to have encountered e-mail first in the home setting as they were at the office. By 1995, home computers were connecting to the internet and exchanging messages using services like AOL, MCI and Sprint. When the corporate world adopted electronic mail for communications, it inherited the highly personal and informal characteristics that in many ways have lead to the problems most organizations have with it. Worse, the service was installed locally on a “personal computer” over which the employer had almost no control, contributing to the frequent use for private purposes as well as corporate.

### *What e-mail is*

Many e-mail exchanges resemble conversation. There may be short exchanges – a question and response, for example – or there may be long exchanges during which the topics change several times and recipients being added or dropped, with the result that the information exchanged no longer relates to the subject line of the message. Like conversation, most messages are quick and short, taking place over an equally short timeframe. Often the messages are cryptic because they've been dashed off, with simple syntax and vocabulary. As a consequence it is often difficult to determine which “e-mail string” contains the whole story and represents the conclusion of an issue.

E-mail is personal and can reveal a lot about the relationship between the originator and the recipient, particularly in short messages between individuals. Within the corporation it will show who has influence in decision-making that may not be congruent with the organization chart. Employees will use the corporate e-mail system to communicate with friends and family, or with businesses on purely private matters.

The author and recipient(s) of an e-mail message are individuals rather than corporate entities or positions. You do not write to the function or title. There is no sense of corporate organization either in the originating or recipient addresses, and originators do not necessarily include a signature block. Unlike correspondence of the past, “copies” are not routinely created for the files holding all related material – in fact, the messages are kept more like the “chron” files of the past.

But even the chron files in a work unit were kept together. In the case of e-mail, there is no hierarchical, or indeed ANY, relation between accounts – in fact, it is entirely possible for the e-mail accounts belonging to the employees in a work unit to be spread across several different servers, depending entirely on the loading on the servers.

The casual and personal nature of e-mail means that it gets filed as an after-thought, and only if the originator or recipient thinks it might be useful as a reference. An originator might print off “a copy for the files” or transfer a copy to the electronic records management system but not necessarily as part of a standard business process.

### *What e-mail is not*

E-mail bears little or no resemblance to the official correspondence found in the paper files of the past. It is not:

- Grouped together with other documents in a file that is assigned a file number and managed through a lifecycle
- Prepared by a subordinate for review, approval and release by a superior
- Unique. Most e-mail replies contain the original message and possibly the thread of previous exchanges on the same topic, without necessarily containing the entire story.
- Complete in terms of the information about the correspondents – their role and function in the organizational unit at the time of the communication, and therefore some indication of the authority of the message and whether it records a decision
- A standard and predictable work process

## *Why continue to use e-mail?*

However it may have become part of corporate life, it is now pervasive. It has so entirely replaced the use of formal written correspondence for most internal purposes that there is no going back. Therefore, we must either learn how to manage it, or find a better tool.

## **Why not save everything?**

### *Viruses, Spam and the Phishing Industry*

They can be so devastating that they make the international news – think of the ILOVEYOU virus of 2000, “Anna Kournikova” of 2001 and “MyDoom” of 2004. Although viruses were originally spread by other means of file transfer such as downloads, e-mail became a major vector for infection. An indicator of significance is that antiviral software sales exceeded \$2 billion in 2006 and have a compound growth rate of over 10%.<sup>1</sup>

Spam is a nuisance that chews up bandwidth, disk space and productivity – and that’s before you take into consideration how offensive some of it can be. A 2007 report said that spam in e-mail traffic fluctuates between 82-87% of all email traffic.<sup>2</sup> In one instance reported by the spam filter at a Canadian government agency, only 5% of the e-mail traffic received in the 24-hour period was considered “good”, with the remainder caught in the spam filter.

Among the unsolicited mail often found in the spam filter are “phishing” e-mails. “Phishing” targets customers of banks and other financial institutions with the aim of capturing sensitive information such as usernames, passwords and credit card details by masquerading as a trustworthy entity in an electronic communication.<sup>3</sup>

### *Volume*

According to the Radicati Group, corporate users received an average of 18 megabytes (MB) of email per day in 2007; a figure that is expected to grow to over 28 MB per day by 2011. Radicati also found that users sent and received an average of 133 messages per day.<sup>4</sup> That’s 133,000,000 million messages per year in an agency with 4,000 employees, and that excludes weekends.

### *Duplicates*

---

<sup>1</sup> Magic Quadrant for Enterprise Antivirus, 2006, Gartner RAS Core Research Note G00141873, Arabella Hallawell, Peter Firstbrook, 31 August 2006, RA1 12152006

<sup>2</sup> MAAWG Email Metrics Report #6 Q2 2007, available at [http://www.maawg.org/about/MAAWG20072Q\\_Metrics\\_Report.pdf](http://www.maawg.org/about/MAAWG20072Q_Metrics_Report.pdf)

<sup>3</sup> Phishing – Wikipedia. Available at <http://en.wikipedia.org/wiki/Phishing>

<sup>4</sup> Email statistics reported by the Radicati Group are cited in the article Email as Information Asset, in the July/August issue of AIIM’s Infonomics publication. Available at <http://www.aiim.org/Infonomics/ArticleView.aspx?id=34895>

Every email has at least one sender and one recipient, although there are frequently more. Emails get forwarded to other recipients, and replies inevitably include the text of the original message. Moreover, the same attached document may be sent more than once to different recipients, especially during the drafting stage when the document is shared as part of the consultation process.

### *Jokes and email chain letters*

‘nuff said.

### **What else is broken in e-mail?**

#### *Inbox size limits*

Although “storage is cheap”, an IT organization can still run out of space on individual servers and many corporations establish inbox size limits with the aim preventing a system crash. The other reason often cited is that size limits will encourage employees to clean out the non-record type e-mails and delete all but the final version of an e-mail exchange that contains all the previous messages. The e-mails that are truly records are therefore saved in the Electronic Records Management System or printed off. Unfortunately, no one has found a way to guarantee 100% compliance, and even where employees are willing, different interpretations of the content may have some employees deleting messages that others feel have substance. Or – they delete all messages earlier than a certain date.

#### *Personal Archives*

The mostly likely outcome of inbox size limits is the creation of personal archives, termed PSTs in the Microsoft world. These archives are saved on personal shares on the servers (if one is lucky!), but may also be stored on the local workstation or saved on CDs. CDs go missing, and files on local workstations will not be backed up, leading to their possible loss in the event of a disk failure.

### *Encrypted and password protected files*

Password protection and encryption work to secure a file during transmission and to avoid unauthorized access. However, several years later when the password is forgotten and the key lost, even authorized users will have trouble getting access.

### *Attachments in obsolete software*

Office documents that are not in current use are not converted when software is upgraded to a newer release or replaced entirely with a new package. (Something to watch for during mergers.) Will the document attached to an e-mail be readable in 10 years, or will portions of the contents be unreadable?

## **Unhealthy IT practices**

### *Treatment of files and e-mail accounts upon severance*

Until recently, the electronic documents on the workstation and the e-mail stores were destroyed when an employee left the organization, based on the assumption that anything that had been important would have been printed and put on the file. The files might not be directly destroyed, but the workstation would be reassigned to another user and the files on the hard drive deleted. In the case of e-mail, the account would be de-activated and would probably not be transferred to the new server during an upgrade.

Realizing that there may be useful information in the e-mail stores of a departed employee, some organizations assign the old e-mail account and files to the employee's successor or supervisor, who then struggles to understand the idiosyncrasies of the file folder hierarchy set up to help the former employee manage the e-mail.

### *Using backup as an archive*

Backups are designed for disaster recovery. Although the technology is changing, servers in many cases are still backed up to tape, a medium that deteriorates over time. When the disk becomes unreadable or the entire server is lost, the most recent set of backup tapes is used to recreate the disk. If those tapes are unreadable, an earlier backup version may be restored. Some IT units keep the last backup of the month "just in case," but all other tapes are returned to a pool for re-use.

The point to keep in mind is that backup happens every night, but recovery happens extremely rarely. For this reason, the design is focused on improving backup rather than on ease of recovery/retrieval. As storage capacity increases, the technology changes to increase the speed and efficiency of backup; tapes created at one point may be unreadable by the newer technology five years down the road.

By contrast, archives are designed for long term storage of information that may be required for business, legislative or historical reasons in the future. Where active data is stored and frequently

accessed using desktop applications available in the ordinary course of business, archive data is stored in a stable long-term format that will still be understandable 10 years hence. The archive may be kept in “near off-line” storage and migrated over time to other storage devices as long-term storage technology advances.

### **“Functional Requirements” for E-mail archiving tools**

1. Capture and archive all e-mail messages entering and leaving the corporation and messages between users within the enterprise as unique, indexed records. Only one copy of unique messages will be kept. (Single-instance storage (SIS) technology.)
2. Allow for policies to determine how long certain kinds of e-mails need to be kept, and for automatic deletion of certain types of e-mails such as notifications from work management systems, listservs, and other low-risk non-record messages.
3. Allow for policies to set journaling for certain key custodians, particularly in the event of a legal hold.
4. Move older, less frequently accessed e-mail to near on-line storage disk and tape.
5. Leave stubs or links to the older e-mail messages in the active e-mail data store so that messages can be found. Prune the active e-mail store based on policies established by Records and IT in consultation with business managers and legal counsel.
6. Offer personal data store migration tools and temporary offline local store options to eliminate the store of e-mail messages outside the control of the archive system.
7. Provide or integrate with a robust records management solution to manage the life cycle of the records to ensure proper retention and deletion.
8. Offer a powerful search tool to help employees find e-mail messages.

### **Better alternatives to e-mail**

Introduce and enforce usage of collaboration tools that replace “messaging” with common work spaces for sharing information and tracking business flows so that record keeping becomes a by-product of the work process; note that RSS feeds on such sites can give notification of changes and updates without requiring an email message

References:

1. Magic Quadrant for E-Mail Active Archiving, 20 May 2008, Carolyn DiCenzo, Kenneth Chin, Gartner RAS Core Research Note G00157611
2. A Survey Of Federal Agency Records Management Applications, 2007, available at: <http://www.archives.gov/records-mgmt/resources/rma-study-07.pdf>
3. Guidance concerning the use of e-mail archiving applications to store e-mail, NARA Bulletin 2008-05, July 31, 2008. Available at: <http://www.archives.gov/records-mgmt/bulletins/2008/2008-05.html>

© Peg Duncan 2009