# Archiving and preserving e-mail

GIANFRANCO PONTEVOLPE [†] AND SILVIO SALZA [†‡]

pontenvolpe@cnipa.it
salza@dis.uniroma1.it

*(December 30 2008)*

*This report was produced by CNIPA (Centro Nazionale per l'Informatica nella Pubblica Amministrazione), the Agency of the Italian Government for ICT infrastructures in the Italian Public Administration, in cooperation with the InterPARES 3 project, and is part of an activity aimed at proposing guidelines for the management and preservation of e-mail messages as records in the Italian Public Administration.*

[†] CNIPA Centro Nazionale per l'Informatica nella Pubblica Amministrazione, Rome, Italy, www.cnipa.gov.it

[‡] Università degli Studi di Roma "La Sapienza", Dipartimento di Informatica e Sistemistica, Rome, Italy, http://www.dis.uniroma1.it/~salza/

## Note to the reader

The aim of this study is to investigate the technical aspects relevant to the e-mail preservation process. This is important since e-mail messages are a very peculiar kind of electronic document, with a rather complex structure, and because of the need to take into account, to some extent, also the peculiar infrastructure through which they are delivered, i.e. the Internet. To achieve this goal we have considered both the functionalities of the commercial products for e-mail management, included the so-called e-mail archiving systems, and the requirements expressed in several important reference documents. Devising precise and systematic procedures for e-mail preservation is not a goal of this document, and indeed cannot be done in a sufficiently general case, since these procedures may heavily depend on the characteristics of the organization where the process is taking place. However in sect. 5 and 6 we present a reference model and we discuss some practices, based on the current offer of archiving products, and widely adopted by many organizations. This should be taken just as an example aimed at discussing the technical issues, and mostly reflecting the point of view of IT experts. The definition of a more general e-mail archiving and preservation model should be carried out as a separate task, within the InterPARES 3 project, and deserves a more thorough discussion, involving both archivistic and IT competences.

## 1 Introduction

The first e-mail was sent in 1971 between two computers that were sitting side-by-side in the same room, but it went through the ARPAnet (the ancestor of the Internet). It was the first time a message was sent across a computer network in a systematic way.

The impressing remark by J.C.R. Licklider, which we have quoted above, came just a few years later, when e-mail was still restricted to a limited milieu in the scientific community, and the widespread use of it was at least a decade ahead. Licklider, a MIT psychologist who formulated the earliest ideas of a global computer network and greatly contributed to the ARPAnet, had indeed a very neat view of what was to come, and a prophetic feeling about the role that the new medium would have played in human communication.

Presently e-mail is by far the most widely used form of written communication, and it has been estimated that more than 100 billion e-mails are sent daily, and that the number will reach 300 billion by 2010. Moreover, in the last decade it has become more and more evident that in all business, government, and even private activities, a crucial share of the relevant information is exchanged through e-mail messages, and that, in most cases, that information can be found *only* in the e-mail, and nowhere else. For instance It has been estimated that e-mail represents about 75% of corporate intellectual property.

The need of preserving and archiving e-mail has therefore become evident: it would not be wise to preserve the other documents and miss the e-mail, where we know that the largest share of information is concentrated.

As a matter of fact, in the last years, many corporations and government agencies devoted a substantial amount of effort to e-mail archiving, and this has triggered a market which is expected to reach in 2008 half a billion dollars in software licenses and maintenance services.

A more detailed analysis shows several motivations to e-mail archiving.

*Storage concerns*

The volume of e-mail messages that corporations and large organizations must handle is very large, and growing fast. On the other hand e-mail servers have not been designed to store and manage a large amount of messages and attachments for long periods of time.

As a consequence, most organizations enforce size limits to their employees' mailboxes. This often leads users to routinely backup the messages *they* consider *relevant* on their own PCs, before they disappear from their servers. The whole procedure is, of course, informal, uncontrolled and unreliable. Moreover the backed-up messages can only be accessed by the individual users who have stored them (if they are still able to find them).

Up to now, overcoming storage concerns is still the main motivation to e-mail archiving, and hence the strongest market driver.

*Strategic relevance*

E-mail messages have become an increasingly important and strategic resource for the organizations, and hence should be centrally managed and selected for archiving and preservation according to precise and well defined criteria. This contributes to automate and accelerate business processes, and may produce substantial savings by cutting the time spent in locating and retrieving messages.

Moreover, when an archival solution is deployed, e-mail messages can be integrated with other organization data and analyzed to monitor business processes and to extract knowledge which would contribute to devise business strategies.

*Regulatory compliance*

Most companies have been recently fined large amounts of money for failing to preserve corporate e-mail records. In the most evident case, Morgan Stanley was fined in 2005 $ 1.45 billion, in a case dubbed by some as 'legal Chernobyl', for being unable to produce corporate e-mail records, i.e. for failing to reproduce e-mail requested under investigation (back-up tapes lost or unrecoverable). Lower amounts of money have been awarded in other cases, but the overall figure has totaled in the last few years to several billions.

In the US, according to new Federal Rules of Civil Procedure Amendments, the production of electronic information is no longer optional. US companies should therefore be prepared to support electronic discovery, and be able to exhibit in a very short time all records requested by Court, chiefly e-mails, that have played a central role in many recent cases. Though the most evident cases concern private organizations, government agencies have to comply as well.

Regulatory compliance has triggered, in the last few years, many organizations to set-up e-mail archiving systems, and it is in the US a very strong market driver.

*Historical preservation*

Last, but not least, in many circumstances e-mail messages should be archived and preserved as historical records, in the interest of future generations. This is especially true since, as we have already remarked, e-mail has become the most important form of communication between individuals, replacing paper-based correspondence and, in many cases, substituting or integrating telephone conversations.

Historians of future generations may have a better chance to investigate the Internet age than the previous part of the XX century when all quick communication went through the telephone wires, without leaving almost any record in the archives. From our side, we should feel the responsibility of preserving such valuable information.

The purpose of this document is to give a concise but complete account of the main problems connected to e-mail preserving and archiving, point out the main issues and draw up the basic policies and procedures.

This is no trivial task, since e-mail messages are a very peculiar kind of electronic document, with a rather complex structure, and because of the need to take into account, to some extent, also the peculiar infrastructure through which they are delivered, i.e. the Internet.

## 2 The Internet e-mail infrastructure

### 2.1 How does e-mail work

E-mail is a store-and-forward method of exchanging messages on the Internet. This means a message sent by a user goes through an asynchronous process of delivery, typically involving a series of steps. In each step the message is stored by an intermediate server on the network, to be forwarded at a later time, until it finally reaches its destination. Timing depends on the availability of connections on the network.

A schema of the delivery process is shown in Figure 1. The process involves a *sender*, say Alice, and a *destination*, say Bob. Both Alice and Bob use specific applications, called *e-mail clients*, running on their PC to send and receive e-mail. Clients do not communicate directly, but have to connect to *e-mail servers*, i.e. special applications run by Alice's and Bob's organizations or ISPs, that actually take care of carrying on the message delivery.

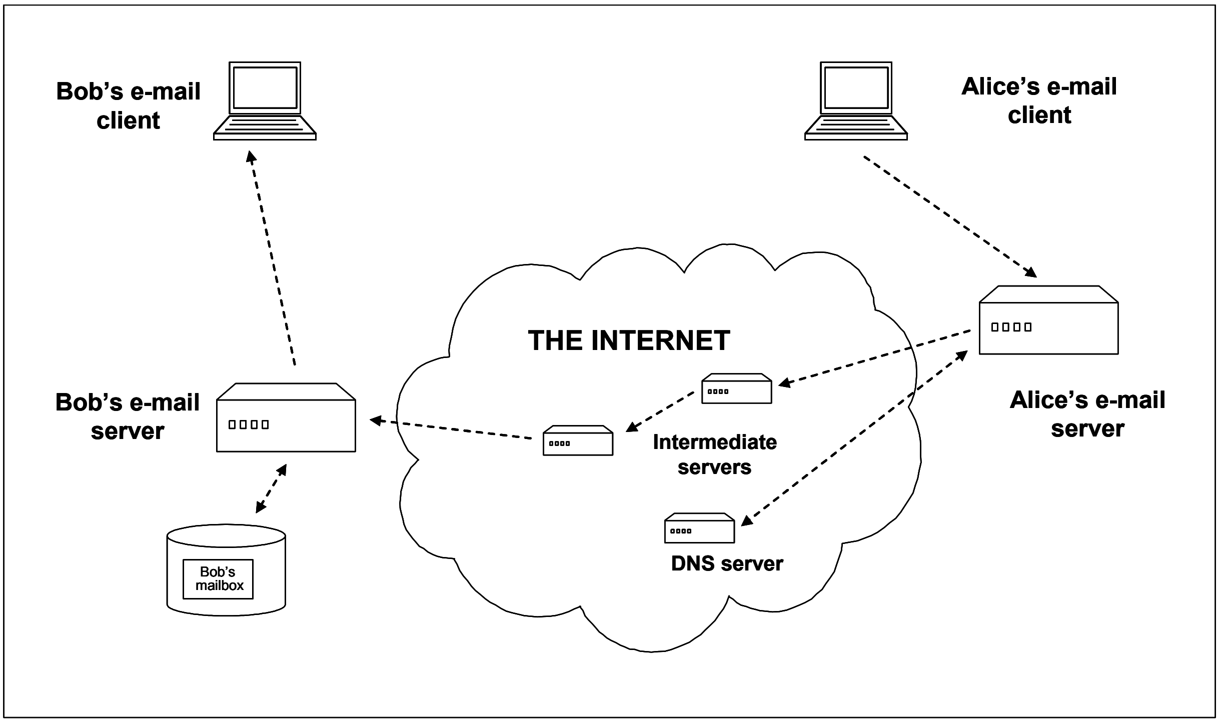

**Figure 1** – Basic e-mail infrastructure

The process goes through the following steps:

- Alice composes the message using her *e-mail client*;
- the message is formatted by Alice's *e-mail client* in a specific i*nternet e-mail format*, and then is sent to her local *e-mail server*;
- Alice's *e-mail serve*r locates the address of Bob's *e-mail server*, exploiting the *Domain Name System (DNS)*, i.e. the distributed directory of the Internet;
- the two *e-mail servers* exchange the message, which may go through a series of intermediate servers on the network, and is finally stored by Bob's *e-mail server* in Bob's personal *mailbox*;
- the message is kept in Bob's mailbox until he reads it and/or downloads it using his *e-mail client*.

The procedure is pretty much the same that Alice and Bob follow when they exchange letters. Their local post offices play the same role that local e-mail servers, and the letter delivery may go through additional post offices (intermediate servers). In both cases the delivery time, and delivery itself are not guaranteed.

Internet is a *best-effort network*, and the message, like any other information crossing the network, to reach its destination has to go through several servers run by independent organizations that take no commitment on the availability and the quality of the service. Hence the delivery time cannot be predicted, and the message may even get lost on the way.

Anyway, as we shall discuss later in more detail, all clients and servers involved in the delivery process follow a set of strict rules (protocols). This allows to trace all relevant events, and to record all this information in a rather detailed report that is appended to the message. Moreover, in case of failure, the server may reiterate the delivery, and the sender may ask for delivery reports and receipts to gain evidence that the message has been delivered, and/or that his correspondent has actually read it.

## 2.2 Interoperability of e-mail systems

As we have seen in previous sections, exchanging a message involves interaction among several *agents* (e-mail clients and servers), which are in general *heterogeneous* systems, i.e. based on different hardware and software platforms. Moreover these systems are independently designed and implemented by different parties, potentially without any form of direct coordination.

A main problem in the Internet e-mail system has been therefore to ensure *interoperability*, i.e. correct and reliable communication among these heterogeneous systems. Interoperability is based on two main elements:

- *communication protocols*, i.e. sets of rules governing the communication between agents, which ensure that agents may reliably and correctly interact by means of a common language and of standard procedures;
- *message format*, i.e. a set of formal definitions that specify the structure of the message and how the message and its attachments are encoded, so providing for correct interpretation by different e-mail clients, and guaranteeing that the content of the message is correctly rendered to its recipient.

A further requirement is that interoperability must also be guaranteed across time. That means that when the definition of protocols and message format evolve, they should still guarantee backward compatibility, i.e. new rules should still be compatible with old rules. For example, a message formatted according to an old version of the message format standard should be presented correctly by an e-mail client compliant with the new version of the standard.

Unfortunately this is not always the case, and this is a major problem to be addressed in e-mail archiving, since we must ensure that the messages that we archive remain readable across time, even if standards evolve.

# 3 Format and structure of Internet e-mail messages

Message format and encoding is of crucial importance in e-mail archiving and preservation, for several reasons.

First, in order to archive a message, we first need to determine the message structure and to identify all the elements that compose it:

- *message data*: the sender, the recipients, etc.;
- *delivery information*: e-mail servers that handled the message, date sent, date retrieved, etc.;
- *message text*;
- *attachments*.

## 3.1 Message structure

An Internet e-mail message consists of two major sections:

- *header*, a sequence of lines, at the beginning of the message, generated by the sender e-mail client and by the e-mail servers involved in the delivery process;
- *body*, the rest of the message, that contains the message text in plain ASCII characters, and/or a text containing non-ASCII characters, and binary data in plain ASCII encoding.

In the simplest case, the original message format defined in RFC 822, the message body contains only plain ASCII characters. Such messages are straightforward to handle, and can just be archived in their native format, and then read again with no need for any form of decoding.

Unfortunately, most messages use extended ASCII or Unicode characters, have attachments and/or are in html format. In all these cases the message must be in MIME format. Hence we shall concentrate in the following sections on the structure of MIME messages.

## 4 Security and privacy issues

Security denotes the ability to manage unwanted events, by preventing them or setting up measures for mitigating consequent damage and loss. Hence e-mail security should be addressed considering the whole process in which e-mail occurs, taking into account the environment and the risk conditions.

In this section we will outline the general e-mail security aspects, identifying the main vulnerabilities and the typical risk scenario.

### 4.1 Vulnerabilities

The Internet e-mail infrastructure derives from the one originally designed for the ARPAnet, in which the only strong security requirement was the capability of delivering messages even in the case of partial network failure. Instead, confidentiality, end points authentication and non-repudiation were not considered at the time important issues.

As a consequence, an Internet e-mail message is poorly protected against unauthorized disclosure and can easily be forged. Moreover, no mechanism is provided to detect a loss of integrity. Therefore, to make a comparison, the confidentiality of an e-mail message exchanged through the Internet may be considered comparable to that of a traditional letter mailed without an envelope.

To say the truth, these vulnerabilities are mostly related to the lower-level Internet protocols, mainly the TCP/IP layers, used to ship packets of information through the network. These vulnerabilities could have been handled and fixed at a higher level by e-mail protocols and formats (SMTP and MIME), but, again, this was not actually done, since, at the time these protocols were originally designed, e-mail was mostly used within the scientific community.

More recently, these limits have been overcome by the S/MIME standard, an extension of MIME, which supports an adequate set of cryptographic security services: authentication, message integrity, non-repudiation of origin and confidentiality. At the moment many commercial products support S/MIME, and therefore offer a better security level, but interoperability problems are still frequent and, therefore, full support of S/MIME cannot be considered a standard feature.

### 4.2 Risk scenario

Despite its high degree of vulnerability, the use of e-mail is widespread and users are not concerned about the related security problems. The perceived risk of content disclosure or receiving forged messages is actually very low.

Anyway, a practical remark could be that most business, government and legal processes rely on e-mail, and there is actually no evidence of significant problems arising from content disclosure or message forgery. More serious security concerns are related to other different threats that do not exploit e-mail vulnerability, but take instead advantage of the vulnerability of human behavior: phishing and spam.

Phishing, i.e. the process of acquiring confidential information such as usernames, passwords and credit card data, is a new and very popular form of fraud that uses e-mail as a vehicle. We shall not discuss it, since it is not relevant for the purpose of our study.

Instead spam, i.e. the huge unsolicited stream of e-mail that floods our mailboxes, needs to be carefully analyzed as a delicate issue in e-mail archiving, since it affects the selection of the messages to be archived.

### 4.3 E-mail spam

Every form of communication may occur also if unsolicited. In fact unsolicited messages (mostly advertisements) are frequent in every communication media. It is a sort of 'noise' we have to isolate and discard to get the actual information. The more the level of the noise increases, the more it becomes difficult to cut the noise off, and the more the communication becomes blurred.

In e-mail, Spam is the noise, and it has become in recent years very intense. According to some accounts spam volume exceeded legitimate e-mails in 2007. Even if the goal of spammers is not to block the e-mail service, in reality, among the consequences of the huge volume of spam, there could also be some kind of denial of service.

As every kind of noise, spam can be reduced by using appropriate filters, whose tuning (anti-spam filters) is a very delicate task, since an improper setting may result in mistaking legitimate messages for spam. However, a sophisticated technology has developed, which is able, if properly used, to detect a significant percent of spam with a very low degree of error.

Anti-spam filters drastically reduce the number of messages coming from known spam sources or having typical spam characteristics; however, there are other messages that may be meaningless for the recipient and the organization (e.g. jokes, unsolicited news, service messages, error messages, etc) that still get to the mailbox.

### 4.4 Message authenticity

According to the InterPARES glossary, authenticity is "the quality of being authentic, or entitled to acceptance. As being authoritative or duly authorized, as being what it professes in origin or authorship, as being genuine".

For e-mail messages these characteristics should refer to the original message, i.e. the one sent by the sender's server, and encompass both the message and its metadata (for instance the subject, the sender the date, etc.). To make the point, let us look at some definitions in the RFC 2822 e-mail standard.

The `Date` header "specifies the date and time at which the creator of the message indicated that the message was complete and ready to enter the mail delivery". Namely, it's an information the sender may set up autonomously (usually the mail client set up the `Date` field to the current client system time).

The `From` header specifies "the mailbox(es) of the person(s) or system(s) responsible for the writing of the message". The standard provides also for the case where the mailbox of the author is different from the one of the person who actually sends the message ("if a secretary were to send a message for another person"): in this case the latter mailbox should be specified in the `Sender` header. Therefore, according to the standard, the client should set up the `Sender` header, while the user should set up the `From` header.

For instance, it is easy to forge a message and make it look as if it were coming form another person, just setting up another mailbox name through the client configuration options.

Moreover, in the case of forwarded e-mail, the text of the original mail may be easily modified by the new sender, compromising the forwarded message authenticity.

So an e-mail message can be considered authentic if we can assume the sender shown in the message text and associated to the mailbox indicated in the `From` header, corresponds to the actual sender. In case of forwarded message, we can consider the original message authentic if this condition is satisfied for all messages (original and forwarded) and we trust the forwarder(s). Of course these are necessary but not sufficient conditions.

Anyway these conditions cannot be easily assessed. A misleading setup of the `From` header and `Data` header may be revealed by analyzing the message header and the system data, but most users would not be able to detect this kind of fraud. Manipulation of a forwarded message may be discovered as well, but most users would just trust its authenticity without even taking into account the possibility of text manipulation. To avoid such problem, some e-mail servers track and show user modifications when forwarding a message, but it is still an uncommon proprietary function.

Despite the easiness of forgery, experience shows that e-mails exchanged in common business activities may be nearly always considered authentic. In fact, electronic mails aren't much more vulnerable than traditional letters and, as for paper messages, false e-mails are generally apparent when considered within their contest.

In an archival process, it is more difficult to relate message authenticity to the context, or to perform crosschecks which may reveal inconsistencies. For that reason, message authenticity should be stated by the addressee before starting the recordkeeping process.

Another way to ensure authenticity is to add functionalities, based on trusted authorities, granting message authenticity. Electronic signature is an additional provision granting message authenticity. Other solutions are based on third party services, like the Italian Certified e-mail.

## 5 Archiving and preserving e-mail

In this section we will discuss the organization of the e-mail recordkeeping and preservation process. In doing so, we will propose rather complete and elaborate schemes which are devised, and realistically suitable, only for medium or large organizations. Of course, e-mail maintenance may be an important issue even in the small office and home environment. But we shall not discuss here this case, since the requirements are different, and considerably less complex and less demanding, and other simpler solutions should be envisaged, including outsourcing the whole e-mail maintenance and preservation service.

### 5.1 Reference model

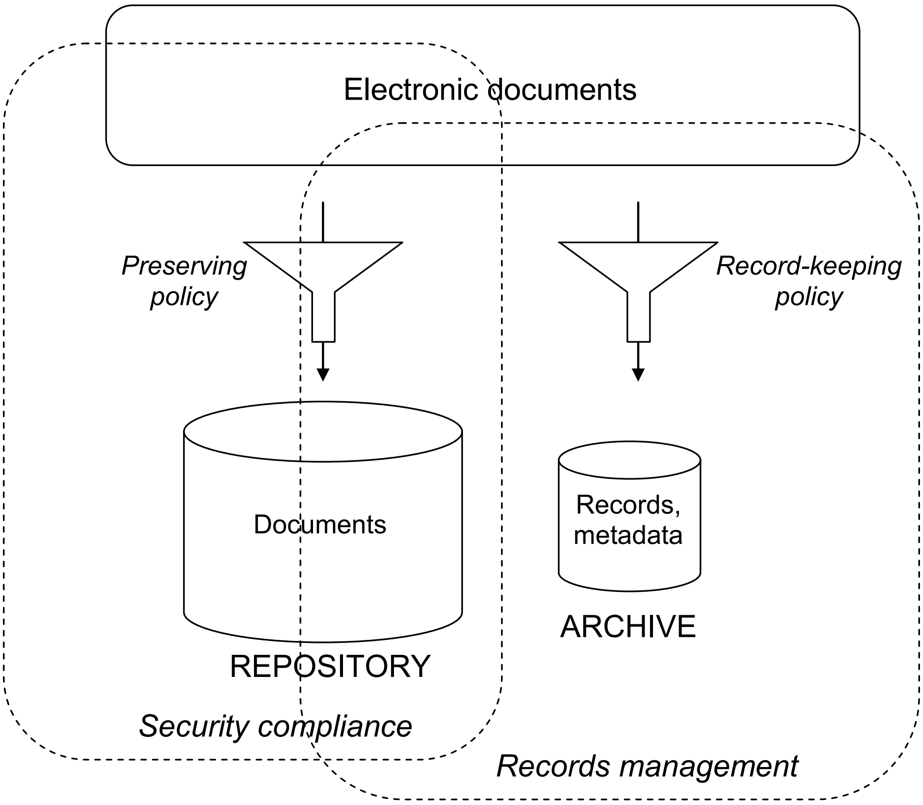

**Figure 6** – The reference model

In order to describe and discuss the policies and procedures used by many organizations for  maintaining and preserving e-mail, we shall refer to the schema shown in figure 6. This model is indeed more general and can be applied to any class of electronic documents.

For any document or class of documents, an organization has to decide whether and how they have to be retained, that is the organization has to define and implement a *retention policy*.

A retention policy is affected by many factors, and may be very simple (e.g. all documents are retained) or quite complex. Factors that usually influence the retention policy are: security policy, compliance with law, legal requirements, storage cost, privacy requirements, support to legal discovery and data mining needs, etc.

Many organizations use a retention policy that does not necessarily require document classification, since, in many cases, their aim is simply to *save* documents, i.e. to avoid losing the information they contain, even if access and retrieval criteria are not yet known, and largely unpredictable, and therefore it is not possible to define appropriate classification criteria. This is mostly driven by security and regulatory compliance requirements.

Beside this basic level of preservation, they often perform another document processing related to their business activities. Documents relevant for the organization are selected, according to a *recordkeeping policy*, and stored together with other information (namely “metadata”) relating them to the organization mission and business. Such metadata allow to group these documents into sets like files, dossiers or other aggregations.

The *recordkeeping policy*, as opposed to the basic *retention policy*, defines rules, procedures and roles for record selection, classification and filing, that is, specifies all the activities having the aim of feeding and managing the recordkeeping system of the organization.

In this schema, the *Repository* and the *Archive* are two logically separated areas, because of the different purposes, and because the procedures for feeding them, as well as the access rules and criteria, are logically different.

Many documents are only be retained in the repository. For instance, it may be important to retain records produced by video surveillance activities for security and compliance reasons, but likely there is no reason to file them in the organization archive.

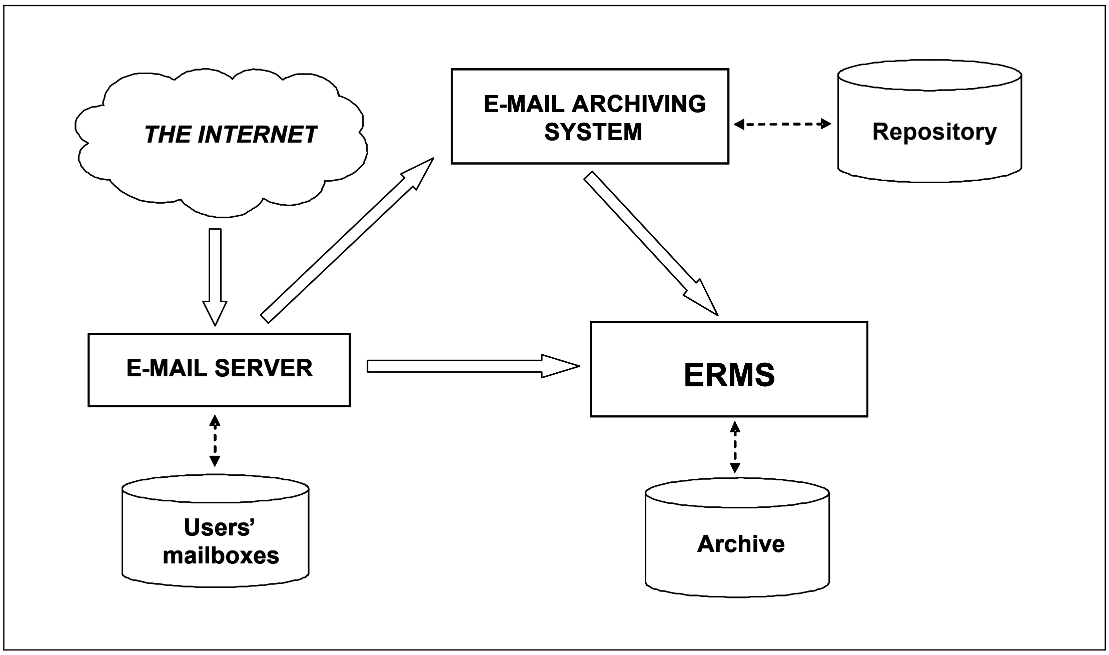

**Figure 7** – Architecture of the e-mail preservation and archival system

This two-level retention and recordkeeping model is particularly interesting in the e-mail case. On one hand, potentially all messages are worth to be saved, at least for regulatory compliancy, legal discovery and data mining. On the other hand, it is not reasonable to file all messages in the organization recordkeeping system, both because most of them may just not deserve to be kept, and because the cost connected to filing (classification etc.), may not be affordable, given the huge volume of messages.

Therefore, a quite reasonable solution is to save most (all) messages in the repository ~~,~~ a task that can be conveniently performed with automatic procedures, and to file in the archive only those messages that fit specific criteria, based on the organization mission and workflow.

This kind of schema has also the advantage of overcoming the delicate problems connected with message privacy and confidentiality discussed in section 4.6. As the repository and the recordkeeping system level may have different access policies, and generally do, a strict policy can be used for the preservation level (e.g. access only to administrators, who could have seen e-mail anyway), thus overcoming the privacy and confidentiality barriers.

As for the architecture of the e-mail retention and recordkeeping system, most organizations adopt a three level architecture, like the one shown in figure 7, that is based on the current availability of commercial products, and takes also into account the fact that the organization may already manage an ERMS (Electronic Records Management System).

The e-mail system (e-mail server or corporate server) stands as a first level of storage both for inbound and outbound messages, but it typically has storage limitations, restricted capabilities of associating additional information (e.g. metadata) to the messages, and does not allow the definition of elaborate access control schema and the deployment of audit procedures (see sect. 6.3 and 6.4). In addition, this system is designed for short-term storage and may not be suitable also to meet short-term retention requirements. Finally, it may be not within the control of the organization, a more frequent situation as outsourcing the e-mail service is becoming an attractive and cost-effective option.

For all these reasons, a separate system is needed to implement the repository level, and these functions may be conveniently carried out by commercial "e-mail archiving" products (see sect. 7.3), which, despite their name, are more oriented towards the initial retention of e-mail than its maintenance, at least according to the terminology in our reference model. Therefore such systems may be unfit to appropriately cover also recordkeeping functions.

From this situation comes the need of having a distinct subsystem, an ERMS, to manage the maintenance function, a solution that has the further advantage of a natural integration of e-mail in the organization recordkeeping system. However, if e-mail maintenance requirements are less demanding, and the organization does not yet use an ERMS, the retention and maintenance functions may collapse into one function at the system level, and can both be carried out directly by the e-mail maintenance system.

## 5.2 Capturing e-mails

Capturing e-mails is the first, and perhaps the most delicate, phase in the e-mail maintenance process. It can basically be performed in two ways:

- *server-based capture*: incoming and outgoing messages are systematically captured when they get to the e-mail server, potentially after being filtered according to predefined rules;
- *client-based capture*: messages are captured with the cooperation and consensus of the user, which interacts through the e-mail client.

Server-based capture is in principle the most simple and desirable option, since it allows the screening of all inbound and outbound traffic, and to perform the selection of the messages to be captured according to uniform rules specifically devised to comply with the organization policy. In this way, if the rules are correctly defined, no information relevant to the organization is lost.

A simple solution to this problem, adopted by many organizations, is to inform users that all messages going through their mailboxes that comply with certain rules are going to be captured, and to ask them to use 'personal' mailboxes, out of the capture mechanism, for their private e-mails.

In other cases, asking the user's consensus for every specific message capture may be necessary, and a client based capture scheme has to be implemented. But pure client-based capture has several drawbacks, since message selection relies on the decision of individual users, who may fail to apply correctly and uniformly the organization's selection criteria.

Referring to the two-level retention and maintenance reference model discussed in the previous section, server-based capture is certainly suitable for the retention level, since it is performed automatically, without putting any burden on the user, and overcomes the privacy and confidentiality issues (if appropriate access schemes are implemented). Moreover it has the further advantage of preserving the messages as soon as they arrive on the server.

Instead, capture at the maintenance level is likely to require the intervention of the user, both because of privacy and confidentiality and to determine if the message needs or deserves to be filed into the recordkeeping system. A 'mixed approach' that takes advantages from both capture schemes is the following:

- a first level message selection is performed at server level, filtering out all ephemeral and non relevant messages;
- candidate messages are proposed to the user who is their sender or recipient, and the user is asked for consensus;

- individual users retain the capability of independently capturing any message they are sending or receiving.

### 5.3 Archival formats

As for any digital record, maintenance and preservation of an e-mail message must ensure two conditions:

- the original structure and all the information contained in the message must be retained;
- future users must be able to access the information in the message in its original form, i.e. perceiving it in the same way the original users (sender and recipients) have seen it.

This means that not only the content, but also the structure and the appearance of the message must be preserved.

Therefore the RFC 2822/MIME format should always be the primary maintenance format for e-mail messages. Moreover, this solution is easy to implement, since this is the format used by many e-mail servers to store messages internally.

The RFC 2822/MIME format guarantees that all the information is retained, and the structural integrity is maintained, but the rendering of the information in its original form is directly granted only for messages in plain ASCII, which are today a very strict minority. Instead, messages exploiting the full MIME format, i.e. with attachments in a variety of media types, rely on external applications to be decoded and rendered.

As a principle, to grant access to the records one should preserve the original hardware-software environment, or, at least, for every media type registered in the recordkeeping system, provide for maintaining applications having certified compatibility with the original ones. Because of technical obsolescence, this is, of course, no easy task, especially on the long term.

Actually, to carefully assess the relevance of this problem we must make clear distinction between two different kinds of scenario:

- *short-term preservation*, when messages must be maintained and accessed for a short period of time, typically up to 10 years;
- *long-term preservation*, when messages must be maintained and accessed for a long period of time, typically more that 10 years.

At the moment, the large majority of organizations is mostly interested in the first scenario, mostly because of regulatory compliance, and commercial “e-mail archiving” products, which we shall discuss in more detail in sect. 7.3, are designed to meet these needs. We shall therefore discuss in this section only the short-term scenario. The long-term scenario, which deserves a different and more complex approach, will be discussed in sect. 5.6.

In the short-term scenario, access to maintained messages with attachments in a variety of media types does not pose special problems and can be granted rather easily, by means of a few very simple provisions.

In fact, we may conveniently assume that, when a message is registered in a recordkeeping system by an organization, i.e. just after it has been sent or received, the current hardware-software environment in the organization allows the user who has sent or received the message to read it, with all its attachments. In the short term, the same kind of access can therefore be granted also to all ‘recently’ registered messages, directly through the e-mail client interface.

What must be done is simply to make sure that software applications for media types in all currently kept messages are preserved, as well as the hardware-software platform needed to run them.

Moreover, to conveniently support presentation and search and discovery actions (see sect. 6.1), which may include searching by content, it can also be useful to store copies of the

attachments, converted in standard searchable print-image format (e.g. PDF), as separate records linked to the message.
Summarizing, in the short-term recordkeeping scenario:

- messages are maintained in RFC 2822/MIME format to preserve the authenticity;
- attachments are extracted as binary files, and stored in the recordkeeping system as separate records, linked to the main record;
- attachments are also optionally converted to a print-image format and kept as separate records, linked to the main record, to support search and discovery actions;
- a database of media types in all currently maintained messages and the corresponding software application are maintained;
- actions are taken to guarantee the availability within the organization of all the necessary applications and of the hardware-software platforms needed to run them.

Both the retention and recordkeeping policy should define how long messages have to be kept in the repository/recordkeeping system and how they should be managed after the short-tem preservation period. In general there are three possibilities:

a) messages are discarded after the period of interest;
b) all or part of the messages whose preservation term has expired are transferred to other organizations for long term preservation;
c) messages continue to be stored in the repository/recordkeeping system, but preservation and access are no more fully guaranteed: the only commitment is some kind of 'best effort preservation'.

These alternatives are not mutually exclusive and the choice depends on many factors, such as regulatory compliance, mission of the organization, storage constraints, etc.

### 5.4 Message classification and metadata extraction

There are basically two options in implementing message classification, which may be considered independently or in a combination:

- messages are classified by means of automatic classification tools;
- users (i.e., senders or recipients of messages) are requested to provide a classification code or naming convention.

The first option is typically used in a server-based capture scenario (see section 5.2), since the selection process naturally provides some degree of classification, based on the rules that have been used to choose to retain the message. Simple categorization schemes are based on main message metadata, but some sophisticated tools may exploit also message content, i.e. information in the text and the attachments.
However, at the state of the art, automatic classification procedures have an insufficient level of reliability, and therefore, when considering the recordkeeping level, automatic classification could be inappropriate, at least in more demanding environments as government agencies, or institutions that have explicit legal recordkeeping responsibilities. On the other hand, giving the users full responsibility may put too much of a burden on them.
As for the message capture, a mixed approach could be the appropriate option. The server-side system may propose a set of possible classifications, derived from the message content and metadata, and the user is only requested to make the final choice, or, if necessary, to override system proposals by introducing his own.

## 5.5 Checking and preserving authenticity and integrity

As we have seen in sect. 4.4, assessing the authenticity of an e-mail message is a nontrivial task, since the e-mail infrastructure, and chiefly the Internet through which messages are transmitted is unprotected, and manipulation of digital records may occur, as for traditional paper documents.

In general, moving from the traditional paper environment to the electronic environment, does not improve the situation and does not decrease the need to care about document authenticity, according to appropriate policies, which depend on the document's purpose, the organization's mission, the risk level, and the environment.

More can be done about integrity, i.e. checking and guaranteeing that the information contained in the record is complete and unaltered in all its parts. Strictly speaking, for an electronic record, this means that the original binary file is preserved, and not a single bit is changed. However, in some contexts, the definition of integrity may be slightly more flexible, only requiring that the essential parts of the message are unaltered.

For incoming messages, beside all the information in the message header (see section 5.4), retaining and maintaining the message in the RFC 2822 format ensures that all the data about the sender, the transmission path and the dates are preserved. Moreover, as the message is saved in its original format, exactly as it was delivered to the receiving server, this is an essential element for any future control.

For outgoing messages, the e-mail server log files, that should be retained as well, along with the bounce messages and the thread metadata, help to assess when the message was actually sent, and if and when it was delivered to its recipients.

A stronger assessment of authenticity can be made through the use of the electronic signature, which is becoming widely adopted, especially when e-mail messages are used to transfer documents with legal value or records of business transactions.

Encryption is used in e-mail to protect the confidentiality of the content during message transmission, and may be performed either by the e-mail system or by the user.

System encryption often occurs at front-end level, via VPN (Virtual Private Network). In this case both the user and the mail server are unaware of the cryptographic process and the retention/maintenance activity is not influenced by the encryption process.

As said before, encryption may also be performed at e-mail-client or e-mail-server level according to the S/MIME standard, but interoperability problems still hamper the diffusion of this kind of protection. Therefore, quite often, encryption and decryption are performed by end users by means of cryptographic functions of commercial products (e.g. cryptographic options of Microsoft Word, Adobe Acrobat, PGP). Decryption of such messages requires in general the cooperation of the user who is the message recipient, and presumably the owner of the key.

Obviously, messages should always be retained and/or maintained in the form in which they were intended to be manifested, therefore, if transmitted in encrypted form, they should be decrypted before retention/maintenance takes place.

Maintaining the records in the decrypted form is important for long-term preservation, since encryption is likely to reduce the ability to access the records in the long term, because of unavailability of decryption software and/or the loss of the decryption key.

Another practice used to protect the integrity of the records, both the whole message and its attachments, if stored separately, is to generate *digests* (i.e. digital fingerprints) for all these objects. These should be kept separately, linked to the corresponding records, and possibly electronically signed by archive administrators.

## 5.6 Long-term preservation

As we pointed out in sect. 5.3, most organizations are only interested in short-term preservation of e-mail messages, i.e. with a time horizon up to 10 years. Long-term preservation, at least as far as e-mail is concerned, is a problem that concerns only a limited number of large organizations, generally at national level, like National Archives in many countries and a few others. In these cases, e-mail messages are managed together with many other types of digital records, and their preservation may benefit from large scale factors and the support of an efficient and complete structure.

Long-term preservation poses two kinds of problems:

- preserving the message integrity in the long time;
- preserving the ability to access all the information contained in the messages and in their attachments.

The first problem is a general one in digital information preservation, and really there is nothing specific to the e-mail case. It is just a matter of recording binary files on physical supports, controlling technical obsolescence and monitoring the quality of the recordings to decide when new copies of the records should be produced, eventually with new technologies. Therefore we will not discuss this matter here, and one should refer to the general literature.

The second problem, as we have seen, deals with media types and the preservation of hardware-software environments necessary to handle them, and has some specific aspects in the e-mail case:

- the variety of media types is extremely large, compared with the limited number of formats a typical ERMS has to deal with;
- there is a total lack of control on the document formation process: in some cases, e-mail users pick-up attachment formats at their wish, while in some other environments organizations may be able to strongly recommend, or even enforce, the use of formats suitable for long-term preservation.

The approach of preserving the applications and the hardware-software environment needed to run them that we have discussed in sect. 5.3 for the short-term scenario, is realistically out of question for the long term scenario, unless we wish to transform National Archives into ICT museums.

Pragmatically, the only solution considered reasonable is to convert the messages and all their attachments, *as soon as they enter the recordkeeping system*, in standard formats that is realistically possible to support on the long term.

More precisely:

- messages are anyway maintained in RFC 2822/MIME format to preserve their authenticity; in a future time, if applications are still available, the attachments could still be accessed in the ‘native’ way;
- attachments that are ‘printable’ are converted in a supported standard print-image format, maintained as separate records and linked to the main record;
- attachments that are ‘not printable’ (e.g. sound, movie etc.) are converted in the best suitable supported standard format, maintained as separate records and linked to the main record;
- information about the original format and the details of the conversion process are registered as message metadata for all converted objects; this provides some kind of assessment of the conversion procedure, and allows to eventually understand to what extent the integrity of the record may have been compromised;
- a database of all converted records and their formats is maintained;
- when a supported format approaches obsolescence, all records in that format are converted in a new ‘equivalent’ supported format.

As a final remark, we shall point out that, in the long-term scenario, since messages are mostly preserved for historical purposes, the main goal is usually to preserve the *integrity of the information in the message* at a semantic and semiotic level, even if the i*ntegrity of the message* can be compromised by a format conversion, since this process may introduce slight changes in message rendering.

A future user, reading in 2050 the converted copy in PDF/A v. 47.1 of an attachment originally in MS Word 2003 format, may get all the information he needs, and be comforted about the authenticity of this information by the assessment of the archivist who in 2031 performed the last conversion. Anyway, if he is not satisfied with this, the alternative for him could just be the contemplation of a binary file.