InterPARES 3 Project

International Research on Permanent Authentic Records in Electronic Systems

TEAM Italy

# Keeping and Preserving E-mail: A Preliminary Report

Mariella Guercio, Università degli studi di Urbino

InterPARES 3 Project, 1st International Symposium
Seoul, South Korea
4 June 2009

# The Aim of the Report

- The report was produced by CNIPA (National Agency for ICT infrastructure in the Italian Public Administrations) as partner of TEAM Italy
  - Authors: Gianfranco Pontevolpe, director of the CNIPA service for digital preservation; Silvio Salza, ICT professor of the University of Rome La Sapienza
- The aim is to investigate the technical aspects relevant to the e-mail creation, capture and management (i.e. records management) and permanent preservation (i.e. archival processes)

# Relevance and Complexity of the Study

- E-mails messages are a very peculiar kind of electronic document with a rather  complex structure
- E-mails require to take into account to some extent also the peculiar infrastructure through which they are delivered (i.e. Internet)
- It has been necessary to consider both the functions of the commercial products for e-mail management and the requirements expressed in many reference documents

# Policies and Models Require Further Analysis

The definition of e-mail records management and permanent preservation models will be carried out as a separate task within InterPARES 3 Project because it deserves a more thorough discussion involving records management, archival and IT competences

# Where It Has Got

- e-mail is by far the most widely used form of written communication
- more than 100 billion e-mails are sent daily, and the number will reach 300 billion by 2010
- a crucial share of the relevant information is exchanged through e-mail messages, and, in most cases, that information can be found *only* in the e-mail, and nowhere else
- e-mail represents about 75% of corporate intellectual property.

# Storage Concerns

- e-mail servers have not been designed to store and manage a large amount of messages and attachments for long periods of time
- most organizations enforce size limits to their employees' mailboxes
- employees backup the messages *they* consider *relevant* on their own PCs, before they disappear from their servers. The whole procedure is informal, uncontrolled and unreliable
- the backed-up messages can only be accessed by the individual users who have stored them (if they are still able to find them)
- overcoming storage concerns is still the main motivation to "e-mail archiving," hence the strongest market driver.

# Regulatory Compliance

- companies have been **fined large amounts of money** for failing to maintain corporate e-mail records, e.g. Morgan Stanley in 2005 $ 1.45 billion, in a case dubbed by some as ‘legal Chernobyl’ (back-up tapes lost or unrecoverable). Lower amounts of money have been awarded in other cases for “spoliation”, but the overall figure has totaled in the last few years to **several billion**
- in NA, the **production** of electronic information is **no longer optional**. Companies should therefore be prepared to support electronic discovery, and be able to **exhibit in a very short time all records** requested by a Court , and only those records (See Sarbanes-Oxley Act and SEC regulations in the US)
- this has implications for **security** and **integrity** of the system, description, **retrieval**, and planned **disposition**
- judicial scrutiny of digital evidence is rooted in proving that results produced are **repeatable**, **objective**, and **verifiable**, whereas industry measures the reliability in terms of **reliability**, **authenticity** and **availability**

# The Report Articulation

1. Introduction
2. Internet e-mail infrastructure (how email works and how end users have access to it); Internet standards for interoperability
3. Format and structure of e-mail messages, with specific attention to the information to be extracted as metadata from the message
4. Security issues: Internet vulnerability, privacy, confidentiality, integrity

# The Report Articulation (cont)

5. Analysis of the present functions for managing and preserving e-mails: strategies to capture messages, preservation formats, classification and extraction of metadata, checking and maintaining authenticity, long-term maintenance
6. Access (search and discovery, protection against unauthorized access and accidental or fraudulent manipulation or destruction)
7. Analysis of commercial products for e-mail management (e-mail servers, integrated systems and e-mail “archiving” systems) and their basic and advanced functions

Appendix: description of the main standards and reference documents at the basis of the report

# Interoperability of E-mail Systems

Interoperability across space 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.

Interoperability must be guaranteed also 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.

# Standardization of message format

- The basic format of e-mail messages is defined by STD 11 (1982), but most applications can now handle the updated version of message format defined in RFC 2822, which is still formally a *Draft Standard*
- E-mail messages should contain only *plain ASCII text* (also called 7-bit ASCII or US-ASCII) characters (1963). SMTP-servers can only handle this type of messages
- To overcome this limitation, the message format has subsequently been extended by the *Multipurpose Internet Mail Extension (MIME)* standard to support:
  - text and headers in character sets other than plain ASCII;
  - messages structured in multiple parts;
  - non text attachments, including a large variety of multimedia files.

# Structure of E-Mails

An 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.

Only message body in plain ASCII are straightforward to handle, and can just be maintained in their native format, and then read again with no need for any form of decoding.

The majority of 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. So in the report we focus specifically on the structure of MIME messages

# Message Header

- It is a sequence of lines, called header lines or headers, which are produced by the sender e-mail client and by the e-mail servers on the delivery path. The header is terminated by a blank line
- Only a minor part of the information in the message header is displayed by e-mail clients
- E-mail clients generally allow users to inspect the complete header, if they like to investigate the message origin and the delivery process
- There are **four types of header lines**:
  - identity header lines (including thread headers),
  - transmission header lines,
  - security header lines, and
  - format/encoding header lines

# Identity Header Lines

- **date**
- the **author/originator**
- the **addressee/recipients** (cc, bc)
- **organization**
- a **message subject** and/or ID (an identifier that should be unique, at least for each server, and can therefore be used to reference the message, e.g. in other messages)
- a **return-path**, an address to which all *bounce messages*, i.e. notifications and answers generated by a message, should be sent
- the **originator** (or sender): the human or automated agent that is actually sending the message in behalf of the *official* sender, i.e. the one mentioned in the From header line (author/originator)
- **thread** (In-Reply-To/References/Resent-From/Resent-To/Resent – Subject): used in messages that are sent in reply to other messages and in messages used to forward other messages

# Transmission Header Lines

The details about the delivery process:

- **User/agent**
- **Delivered To**
- **Received from/by/with** (server identifiers + ESMPT ID): added to the message each time the message is handled by a server on the delivery path, the first one being the sender's e-mail server, and the last one the recipient's.
- **Timestamp**: associated to each step, specifying the local date/time the message arrived to each receiving server
- **Return-Receipt-To/ Disposition-Notification-To**: specify if the sender requested a receipt, and to which address it should be sent.

# Security Header Lines

- Scanning Agent, e.g. UBC
- Antispam Engine
- Antispam Data
- Spam Report
- Spam Lever
- Spam Flag

# Format/Encoding Header Lines

The structure of the message body and MIME version (always 1.0)

- **Content-Type**: specifies if the message contains one or several parts. In the latter case
- a **Boundary** is also specified: a string that separates the multiple parts of the message in the message body.

If instead the message contains a single part

- **Content-Type** and
- **Content-Transfer-Encoding**

are directly specified in the header.

# Message Body

**Single part**: a single part message is a plain text message with no attachments

**Multipart**: is composed by several parts separated by a *boundary*, i.e. by the string defined in the top-level Content-Type header placed between any two parts.

- Multipart messages can be of several types, specified as *subtypes* in the Content-Type header.
  - multipart/alternative
  - multipart/digest
  - multipart/related
  - multipart/report
  - multipart/signed
  - multipart/encrypted

# Media Type and Maintenance

In the maintenance process, one must guarantee the ability to render any part of a message at any time in the future. One should therefore make sure that:

- all media types that appear in a messages are registered in the archives, together with the information necessary to handle them, even if they are not registered with IANA (*Internet Assigned Numbers Authority*);
- an application is available for each media type registered in the archives; or
- a converted copy of the attachment is preserved as well, in a format that guarantees the possibility of rendering it at a later time.

# Dynamic Content

Problems may arise from dynamic information that may be contained in a message. A common case are external references (e.g. web links), or context-dependent information (e.g. date and time) in attached documents. Such messages are not self-contained and therefore could not be properly rendered at a later time (in some cases even at arrival time!). Therefore, when maintaining these messages, appropriate policies should be followed, either to prevent the insertion of dynamic contents or to ‘freeze’ all dynamic references at arrival (or saving time).

# Vulnerabilities

An 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, 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.

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.

# Vulnerabilities (cont.)

- Despite its high degree of vulnerability, e-mail users are not concerned about the security problems. The perceived risk of content disclosure or receiving forged messages is actually very low.
- A low perception of the risk does not imply that the level of risk is actually low. Furthermore, unauthorized message content disclosure is very difficult to detect, and users are generally unaware of it when it happens
- More serious security concerns are related to threats that take advantage of the vulnerability of human behavior: phishing and spam.

# Authenticity Issues

Commercial products implement mail standards with slight differences, with the aim of simplifying the user interface. A typical approach is the following:

- every header field that could be set up automatically (e.g. Date, From, Reply-to) is usually set up by the client;
- user options are provided for modifying defaults values, and possibly to set up some header values.
- As a consequence, we tend to consider mail header lines as *system data* and, therefore, authentic insofar as the mail system is reliable. **Instead, they should be considered *user data*, like the message text, and therefore authentic only to the extent that we rely on the sender**
  - 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
  - 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.

# Maintenance Issues

- Distinction between the e-mail application and the recordkeeping system
- Most e-mails will only be kept in the application or repository
- E-mails may be first captured 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 the most simple and desirable option, since it allows the screening of all traffic, and to perform the filtering 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.

# Maintenance Issues (cont.)

- Likely to require the intervention of the user to determine if the message needs 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.
- Regardless of the scheme adopted, the user should be involved in the classification of the records and in manually entering additional metadata.

  *According to the InterPARES recommendation this function should be entrusted jointly to the user and to the recordkeeping system under the control of the system administrator*

# Maintenance and Preservation Issues

- Maintenance and preservation of an e-mail message must ensure two conditions:
  - the original structure (*intellectual form*) and all the information contained in (and attached to) the message must be retained;
  - future users must be able to access the information in the message in its original (*documentary*) form, i.e. *manifested* to future users in the same way it was *manifested* to the original users (sender and recipients).
- This means that not only the content, but also the *structure/form* and the *composition data* of the message must be maintained and preserved.
- The RFC 2822/MIME format should always be the primary maintenance or *permanent preservation data* format for e-mail messages. Moreover, this solution is easy to implement, since this is the format used by many e-mail servers and clients to store messages internally.

# Maintenance and Preservation of MIME

- The RFC 2822/MIME format guarantees that all the information (*content data*) is retained, and the structural integrity (*form data*) is maintained, but the rendering of the information in its original (*documentary*) form (*using the composition data*) is guaranteed only for messages created in plain ASCII, which are today a small minority of all messages. 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 reconstituted and manifested to the user.
- A future user can therefore access an attachment in the MIME encoded form, but may be unable to actually access its content, unless the corresponding application is available. This is indeed a well known problem in digital record preservation, since all digital records rely on an appropriate hardware-software environment to be correctly rendered.

# Short Term Maintenance

- 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 (.pdf) 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 is 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.

# Different kinds of e-mail records scenarios

- *short-term maintenance*, when e-mail records must be maintained and accessed for a short period of time by the creator, typically up to ten years;
- *long-term maintenance*, when e-mail records must be maintained and accessed for a long period of time by the creator, typically more that ten years
- *permanent preservation*, *when e-mail records are determined by the creator to be inactive (i.e., no longer needed or used in the creator's day-to-day course of business), and are determined by the designated records preserver to have archival value and, accordingly, are transferred to the custody and control of the designated records preserver to ensure 'permanent' (or indefinite) preservation of the records and to ensure ongoing access to the records by all (authorized) users into the foreseeable future*.

# Long-term Preservation

- Preserving *integrity*: not a different issue for e-mail: it is a matter of saving the digital components of the records in non-voltile storage on reliable digital media; controlling the technical obsolescence…
- Preserving *accessibility* has some specific aspects in the e-mail case:
  - the variety of media types and subtypes used in the creation of digital documents in general;
  - there is a general lack of control over the document creation process in most e-mail environments: in some cases, e-mail users may include attachments in any registered and supported MIME media type, while in some other environments organizations are able to strongly recommend, or even enforce, the use of data formats more suitable for long-term maintenance/permanent preservation .

# Long-term Preservation (cont)

- Pragmatically, the only solution considered reasonable is to convert the messages and all their attachments, *preferably as soon as they enter the recordkeeping system (*or the permanent preservation system in the case of the permanent preservation scenario), into standardized data or file formats that are realistically possible to support over the long term;
- messages should be maintained in RFC 2822/MIME format;
- attachments that are ‘printable’ should be converted into a supported standardized print-image format, maintained as separate records and linked to the main record;

# Long-term Preservation (cont.)

- attachments that are ‘not printable’ (e.g. sound, movie etc.) should be converted in the most suitable supported standardized format, maintained as separate records and linked to the main record;
- when a supported data or file format approaches obsolescence, all records in that format should be converted into a new supported format;
- Information about the original data or file format and the details of all conversion processes to which the records have been subjected should be registered as message metadata for all converted records (or for their individual digital components, if relevant); this provides some kind of assessment of the conversion procedure, and allows future users to assess to what extent the integrity of the record may have been compromised.

# Long-term Preservation (cont.)

- 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 *integrity of the message* is “compromised” by a format conversion that introduces slight changes in the rendering of the record’s documentary form.
- A future user, reading in 2050 the converted copy in PDF/A v. 47.1 of an attachment originally created in MS Word 2003 .doc file format, may get all the information s/he needs, and be comforted about the trustworthiness of this information by the assessment of the archivist who, in 2031, performed the last conversion.