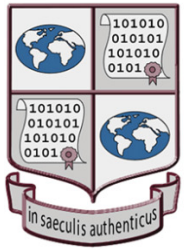


InterPARES 3 Project

International Research on Permanent Authentic Records in Electronic Systems

TEAM Italy

Keeping and Preserving E-mail

General Study 05

Mariella Guercio, Università degli studi di Urbino
Randy Preston, InterPARES Project

InterPARES 3 Symposium
Rio de Janeiro, Brazil
30 September 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, 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 e-mail creation, capture and management (i.e., records management) and permanent preservation (i.e., archival processes)

Scope and Complexity of the Study

Management of e-mail messages requires taking into consideration a number of issues:

- E-mail messages are a very peculiar kind of digital document with a rather complex structure;
- E-mail messages are delivered through a unique infrastructure (i.e., Internet);
- The commercial products used to facilitate the use and so-called "archiving" of e-mail messages have heterogeneous functionalities; and
- Various Internet standards exist to 'guarantee' the interoperability of these heterogeneous systems.

Scope and Complexity (cont.)

- Devising precise and systematic procedures for e-mail records management and/or permanent preservation is outside the scope of this document since such procedures are context-specific and depend on the characteristics of the organization where the process is taking place
- The definition of a more detailed e-mail records management and permanent preservation model requires a more thorough discussion, involving records management, archival and IT competences

Current State of E-mail Usage

- 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 high percentage of relevant corporate information is exchanged through e-mail messages and, in most cases, that information can be found only in e-mail, and nowhere else
- E-mail represents about 75% of corporate intellectual property

Key Motivations

- The need for managing and preserving e-mail has therefore become evident: it would not be wise to manage and preserve the other documents and miss the e-mails, where we know that the largest share of information is concentrated
- Key motivations driving e-mail management activities:
  - Storage concerns
  - Strategic relevance
  - Regulatory compliance
  - Historical preservation

Storage Concerns

- Most e-mail servers are not designed to store and manage large volumes 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 the 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 it is the strongest market driver

Strategic Relevance

- E-mail messages have become an increasingly important and strategic resource for most organizations and, thus, should be centrally managed and selected for maintenance and preservation according to precise and well defined criteria
- By implementing a management solution based on sound records management and archival principles and procedures, e-mail messages can be integrated with other organization data and records and analyzed to monitor business processes and to extract knowledge that can help support business strategies

Regulatory Compliance

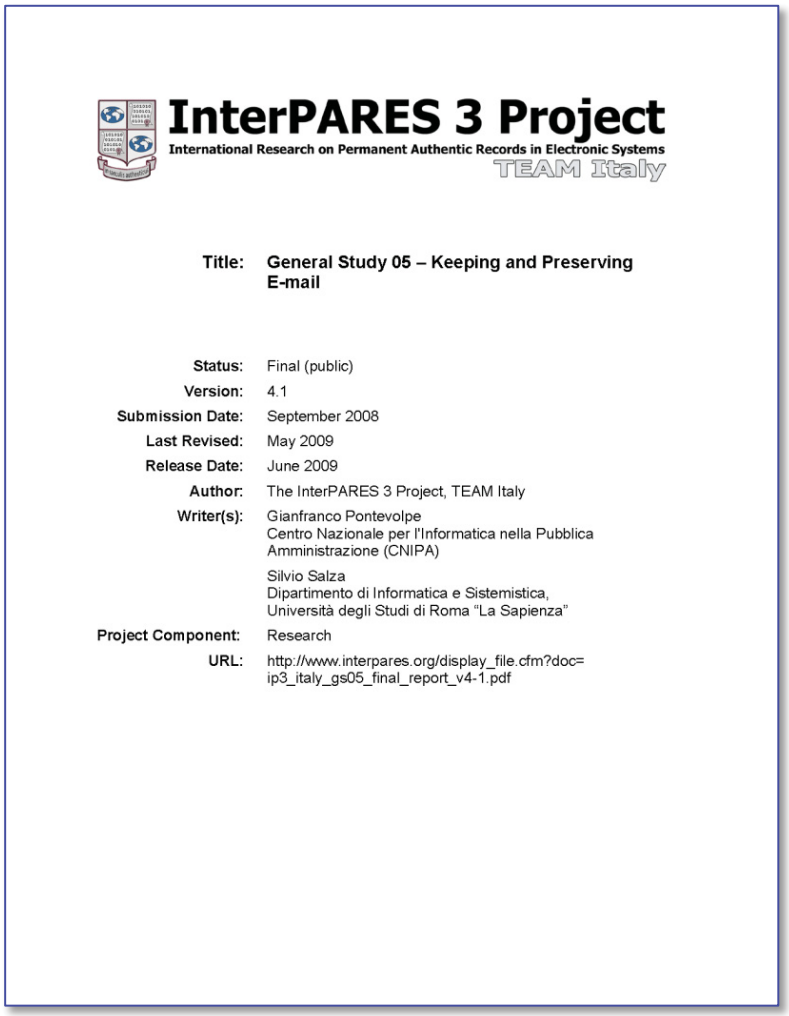
- Companies have been fined large amounts of money for failing to maintain corporate e-mail records
- In North America, the production of electronic information is no longer optional. Companies should therefore be prepared to support electronic discovery, and be able to exhibit in a very short time all records requested by a Court, and only those records (See: Sarbanes-Oxley Act and SEC regulations in the US)
- This requirement of the courts to produce electronic information on demand has implications for security and integrity of the system, description, retrieval, and planned disposition

Historical Preservation

- E-mail messages with archival value should be preserved permanently as historical records, in the interest of future generations
- This is particularly important because e-mail is now the most prevalent form of written communication

Report Contents

1. Introduction
2. Internet e-mail infrastructure (how e-mail 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 messages
4. Security issues: Internet vulnerability, privacy, confidentiality, integrity



InterPARES 3 Project
International Research on Permanent Authentic Records in Electronic Systems
TEAM Italy

Title: General Study 05 – Keeping and Preserving E-mail

Status: Final (public)
Version: 4.1
Submission Date: September 2008
Last Revised: May 2009
Release Date: June 2009
Author: The InterPARES 3 Project, TEAM Italy
Writer(s): Gianfranco Pontevolpe
Centro Nazionale per l'Informatica nella Pubblica Amministrazione (CNIPA)
Silvio Salza
Dipartimento di Informatica e Sistemistica, Università degli Studi di Roma "La Sapienza"
Project Component: Research
URL: http://www.interpares.org/display_file.cfm?doc=ip3_italy_gs05_final_report_v4-1.pdf

Report Contents (cont.)

5. Analysis of the present functions for managing and preserving e-mails: including, 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

Interoperability of E-mail Systems

Interoperability across space is based on two main elements:

1. communication protocols–i.e., sets of rules governing communication between agents; and

2. message format–i.e., set of formal definitions that specify structure of messages and how messages and their attachments are encoded.

- 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

Standardization of Message Format

- 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). SMTP-servers can only handle this type of message.
- To overcome this limitation, the message format has 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; and
  - non-text attachments, including large variety of multimedia files

Structure of E-Mails

An e-mail message consists of two major sections:

1. Header: sequence of lines at the beginning of messages, generated by the sender e-mail client and by the e-mail servers involved in the delivery process; and
2. Body: the rest of the message, which 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 bodies in plain ASCII are straightforward to handle
    - Most messages use extended ASCII or Unicode characters and have attachments and/or are in html format. In all such cases, the message must be in MIME format. For this reason, the report focuses specifically on the structure of MIME messages.

Message Header

- A sequence of lines, called header lines or headers, which are produced by the sender's e-mail client and by the e-mail servers on the delivery path
- Four types of header lines:
  1. Identity (including thread headers)
  2. Delivery
  3. Security
  4. Format/Encoding
- Typically, only a small part of the information in the message header is displayed by e-mail clients
- E-mail clients generally allow users to view the complete header, if necessary, to investigate the message origin and the delivery process

Complete vs. Typical Header

```
Return-path: <luciana@interchange.ubc.ca>
Received: from mta1.interchange.ubc.ca (mta1.interchange.ubc.ca [142.103.145.69])
by store2.interchange.ubc.ca (iPlanet Messaging Server 5.2 HotFix 1.21 (built Sep  8 2003))
                with ESMTP id 0KPW00JRY226JG@store2.interchange.ubc.ca
                for rpreston@interchange.ubc.ca; Sat, 12 Sep 2009 19:39:42 -0700 (PDT)
Received: from mr7.mail-relay.ubc.ca (mr7.mail-relay.ubc.ca [137.82.45.13])
by smtp.interchange.ubc.ca (iPlanet Messaging Server 5.2 HotFix 1.21 (built Sep  8 2003))
                with ESMTP id 0KPW00AGG226KD@smtp.interchange.ubc.ca
                for rpreston@interchange.ubc.ca (ORCPT rpreston@interchange.ubc.ca); Sat, 12
Received: from webeleven.randomlink.com (unknown [216.18.17.1])
                by mr7.mail-relay.ubc.ca (Postfix)
with ESMTP id 09A1A1C22A
                for <rpreston@interchange.ubc.ca>; Sat, 12 Sep 2009 19:39:42 -0700
Received: (qmail 23974 invoked by uid 110); Sat, 12 Sep 2009 19:32:05 -0700
Received: (qmail 23956 invoked from network); Sat, 12 Sep 2009 19:31:57 -0700
Received: from mr6.mail-relay.ubc.ca (137.82.45.11)
by www.mintrec.com
with SMTP; Sat, 12 Sep 2009 19:31:56 -0700
Received: from mta1.interchange.ubc.ca (mta1.interchange.ubc.ca [142.103.145.69])
                by mr6.mail-relay.ubc.ca (Postfix)
with ESMTP id 748A415839
for <randy@interpares.org>; Sat, 12 Sep 2009 19:39:28 -0700 (PDT)
Date: Sat, 12 Sep 2009 19:39:26 -0700
From: Luciana Duranti <luciana@interchange.ubc.ca>
Subject: Fwd: IP3 in November
To: Randy Preston <randy@interpares.org>
Message-id: <24384_1252809582_1252809582_0KPW00AYG21RU8@smtp.interchange.ubc.ca>
MIME-version: 1.0
X-Mailer: QUALCOMM Windows Eudora Version 7.1.0.9
Content-type: text/plain; charset=us-ascii; format=flowed
Content-transfer-encoding: 7BIT
Delivered-to: 97-randy@interpares.org
X-Spam-Checker-Version: SpamAssassin 3.1.3 (2006-06-01)
                on webeleven.randomlink.com
X-Spam-Level: X
X-Spam-Status: No, score=0.1 required=7.0 tests=AWL autolearn=ham
                version=3.1.3
X-Ubc-Received: from LD_panasonic.interchange.ubc.ca (216-19-179-18.dyn.novuscom.net [216.19.179.18])
by smtp.interchange.ubc.ca (iPlanet Messaging Server 5.2 HotFix 1.21 (built Sep  8 2003))
with ESMTPA id  <0KPW00AYF21QU8@smtp.interchange.ubc.ca>
for randy@interpares.org; Sat, 12 Sep 2009 19:39:28 -0700 (PDT)
X-UBC-Scanned: Sophos PureMessage 5.5.5.374460, Antispam-Engine: 2.7.1.369594, Antispam-Data: 2009.9.13.23037
X-UBC-Relayed: Relayed through mail-relay.ubc.ca
X-PerlMx-Spam: Probability=11%, Report= LINES_OF_YELLING_3 0.671,
 X_MAILER_EUDORA_7109 0.05, BODY_SIZE_3000_3999 0, BODY_SIZE_5000_LESS 0,
 BODY_SIZE_7000_LESS 0, FORGED_MUA_EUDORA 0, RDNS_GENERIC_POOLED 0,
 RDNS_SUSP 0, RDNS_SUSP_GENERIC 0, __ANY_QUALCOMM_MUA 0, __C230066_P2 0,
 __C230066_P5 0, __CT 0, __CT_TEXT_PLAIN 0, __EUDORA_MUA 0, __HAS_MSGID 0,
 __HAS_X_MAILER 0, __LINES_OF_YELLING 0, __MIME_TEXT_ONLY 0, __MIME_VERSION 0,
 __OEM_PRICE 0, __SANE_MSGID 0, __STOCK_PHRASE_7 0, __TO_MALFORMED_2 0
X-Spam-Flag: No
Original-recipient: rfc822;rpreston@interchange.ubc.ca
```

Date: Sat, 12 Sep 2009 19:39:26 -0700
From: Luciana Duranti <luciana@interchange.ubc.ca>
Subject: Fwd: IP3 in November
To: Randy Preston <randy@interpares.org>

Message Body

- Single part: plain text message with no attachments
- Multipart: message composed of 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 / mixed
    - Multipart / alternative
    - Multipart / digest
    - Multipart / related
    - Multipart / report
    - Multipart / signed
    - Multipart / encrypted

Single Part Message Structure

```
Message-ID: <006401c91467$186fb1d0$6602a8c0>
From: "Silvio Salza" <salza@dis.uniroma1.it>
To: "Silvio Salza" <salza@dis.uniroma1.it>
Subject: Sample single part message
Date: Fri, 12 Sep 2008 01:35:37 +0200
Organization: =?iso-8859-1?Q?Universit=E0_di_Roma?=
MIME-Version: 1.0
Content-Type: text/plain;
charset="iso-8859-1"
Content-Transfer-Encoding: quoted-printable

Message from the University of Rome
Messaggio dall'Universit=E0 di Roma
```

Multipart Message Structure

```
MIME-Version: 1.0
Content-Type: multipart/alternative;
boundary="---separator---"

This is a multi-part message in MIME format.

---separator---
Content-Type: text/plain; charset="iso-8859-1"
Content-Transfer-Encoding: quoted-printable

Message from the University of Rome

---separator---
Content-Type: text/html; charset="iso-8859-1"
Content-Transfer-Encoding: quoted-printable

< message text in html >

---separator---
```

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;
- 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

- Common example involves 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 may not be properly rendered at a later time (in some cases even at arrival time!).
- Policies are needed to:
  - prevent insertion of dynamic content; or
  - 'freeze' all dynamic references at arrival (or when saved).

Vulnerabilities

- An e-mail message is poorly protected against unauthorized disclosure and is easy to forge.
- Moreover, no mechanism is provided to detect a loss of integrity.
- Confidentiality of an e-mail message exchanged through the Internet is comparable to that of a traditional letter mailed without an envelope.
- Extended S/MIME standard attempts to overcome these limitations.
- However, interoperability issues are still a problem.

Vulnerabilities (cont.)

- The perceived risk of content disclosure or receiving forged messages is actually very low.
- However, this does not imply that the actual level of risk is 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 can be set up automatically (e.g., Date, From, Reply-to) is usually set up by the client; and
  - user options are provided for modifying defaults values, and possibly to set up some header values.

Authenticity Issues (cont.)

- Tend to consider mail header lines as system data and, therefore, authentic insofar as the mail system is reliable.
- Should instead be considered user data, like the message text, and therefore authentic only to the extent that we can rely on the sender. However...
  - it is easy to forge a message and make it look as if it were coming form another person; and
  - 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's authenticity.

Management Issues

Important to distinguish between the e-mail application (transitory / short-term storage) and the recordkeeping system (medium / long-term storage)

- Most e-mails are transitory and will only be kept in the e-mail application
- Usually, e-mails kept in the e-mail application are not classified or registered
- E-mails transferred to the recordkeeping system are (or should be) classified and registered

Management Issues (cont.)

Capturing e-mails (three options):

1. server-based capture: incoming/outgoing messages systematically captured when they get to the e-mail server, potentially after being filtered according to predefined rules.
    - Better suited to management of transitory e-mails
2. client-based capture: messages are captured with the cooperation and consensus of the user, who interacts through the e-mail client.
    - Better suited to management of corporate e-mails to be transferred to the recordkeeping system
3. mixed capture: capitalizes on unique advantages of both server-based and client-based capture schemes.

Management Issues (cont.)

Declaring e-mails as records:

- Regardless of the scheme adopted, in most cases, users will need to be involved in the classification of the records and in manually attaching additional identity and integrity metadata.

  InterPARES Recommendation: These functions should be entrusted jointly to the user (records creator) and to the recordkeeping system under the control of the trusted records officer (or system administrator).

Maintenance and Preservation Issues

Fundamental concerns:

- Maintenance and/or preservation of an e-mail message must ensure two conditions:
  1. the original structure (intellectual form) and all the information contained in (and attached to) the message must be retained; and
  2. future users must be able to access the information in the message in its original documentary form.
- This means that not only content, but also structure / form and composition data of the message must be maintained and preserved.

Maintenance and Preservation Issues

Three different e-mail records scenarios:

1. 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;
2. Long-term maintenance: when e-mail records must be maintained and accessed for a long period of time by the creator, typically more than ten years; and
3. Permanent preservation: when e-mail records are determined by the creator to be inactive, and are determined by the designated records preserver to have archival value.

Short Term Maintenance

This involves...

- maintaining messages in RFC 2822/MIME format to help ensure their authenticity;
- either extracting attachments as binary files, and storing in the recordkeeping system as separate records, linked to the main record, or converting attachments to a print-image format (.pdf) and keeping as separate records, linked to the main record;
- keeping a database of media types used in all stored messages and of the corresponding software application(s) needed to access them; and
- taking actions to guarantee the availability of all the necessary applications and of the hardware-software platforms needed to run them.

Long-term Maintenance/Preservation

In part, this involves preserving the..

- integrity of e-mails (same as for any digital record), which is a matter of saving the digital components of the records in non-volatile storage on reliable digital media; controlling for technical obsolescence, etc.; and the
- accessibility of e-mails, which involves consideration of:
  - the variety of media types and subtypes used in the creation of digital documents in general (which may be included as attachments); and
  - the general lack of control over the creation process in most e-mail environments.

Long-term Maintenance/Preservation

Pragmatically speaking, the only solutions currently considered reasonable are to...

1. convert messages and all their attachments, preferably as soon as they enter the recordkeeping / preservation system into standardized data or file formats that are realistically possible to support over the long term;

2. maintain messages in RFC 2822/MIME format;

3. convert 'printable' attachments into a supported standardized print-image format, maintained as separate records and linked to the main record;

Long-term Maintenance/Preservation

4. convert 'non printable' attachments (e.g., sound, movie, etc.) into the most suitable supported standardized format, maintained as separate records and linked to the main record;
5. convert messages into a new supported format whenever an existing data or file format approaches obsolescence; and
6. register information about the original data or file format and the details of all conversion processes used as message metadata for all converted records or individual digital components.

Long-term Maintenance/Preservation

Some other considerations...

- 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.
- RFC 2822/MIME should always be the primary long-term maintenance or permanent preservation data format for e-mail messages

RFC 2822/MIME Issues

Advantages:

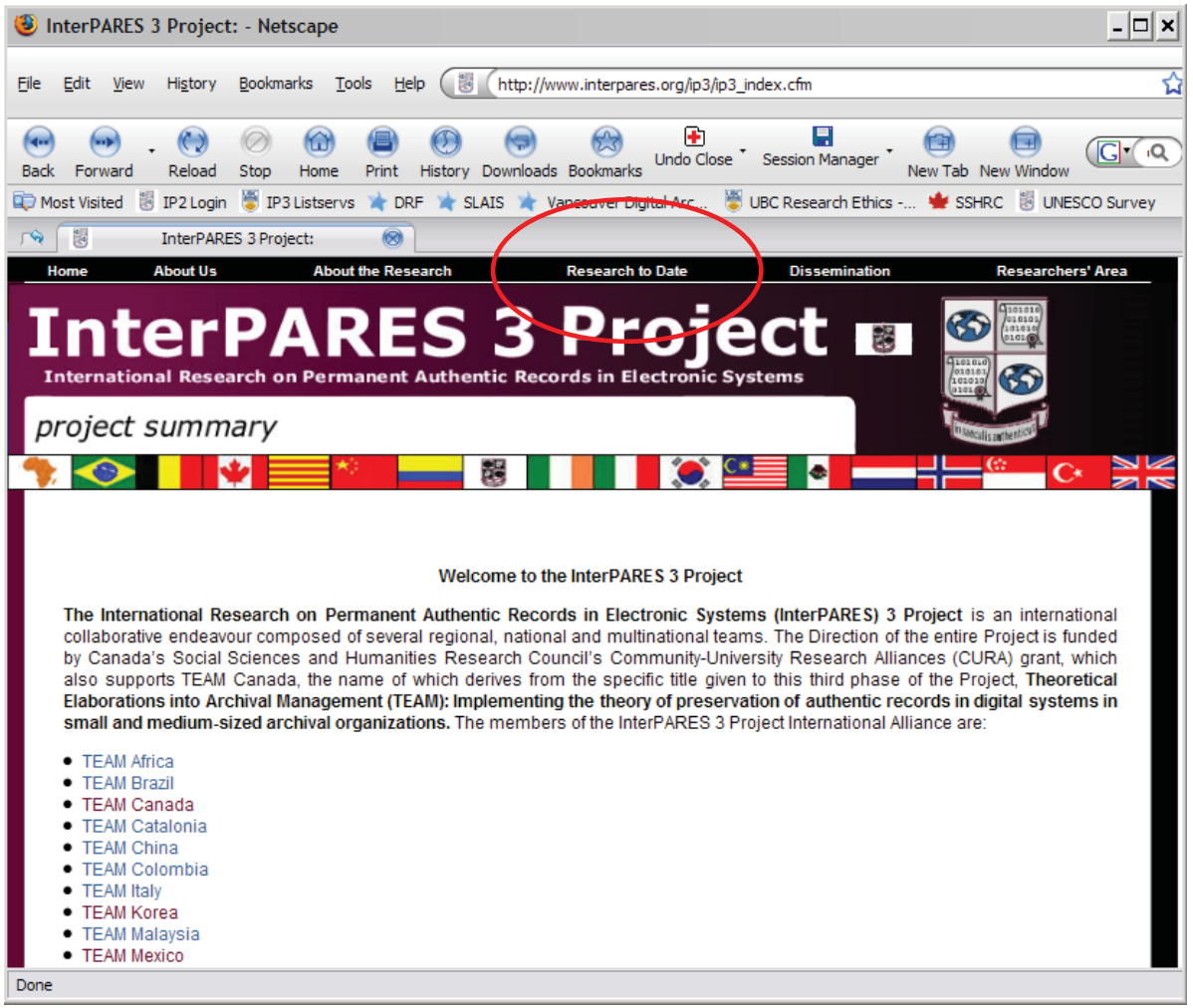
- Easy to implement
- Guarantees that all information (content data) is retained and that structural integrity (form data) is maintained

Cautions:

- Rendering of messages in their original documentary form (using composition data) is guaranteed only for messages created in plain ASCII, which are today a small minority of all messages
- Messages exploiting the full MIME format (i.e., with attachments in a variety of media types), rely on external applications to be decoded, reconstituted and manifested to the user

For More Information

'Research to Date'
on the InterPARES 3 Web site...

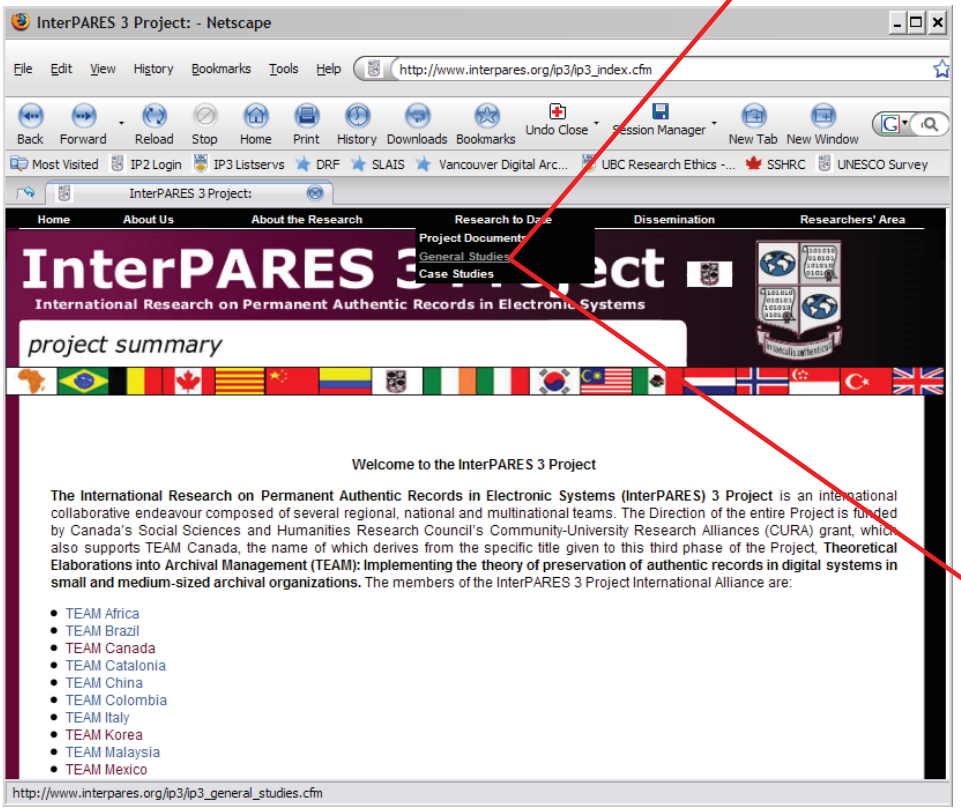


For More Information

'General Studies'

For More Information

General Study 05



InterPARES 3 Project
International Research on Permanent Authentic Records in Electronic Systems
TEAM Italy

Title:	**General Study 05 – Keeping and Preserving E-mail**
Status:	Final (public)
Version:	4.1
Submission Date:	September 2008
Last Revised:	May 2009
Release Date:	June 2009
Author:	The InterPARES 3 Project, TEAM Italy
Writer(s):	Gianfranco Pontevolpe Centro Nazionale per l'Informatica nella Pubblica Amministrazione (CNIPA)
	Silvio Salza Dipartimento di Informatica e Sistemistica, Università degli Studi di Roma "La Sapienza"
Project Component:	Research
URL:	http://www.interpares.org/display_file.cfm?doc=ip3_gs05_e-mail_final_report_v4-1p.pdf