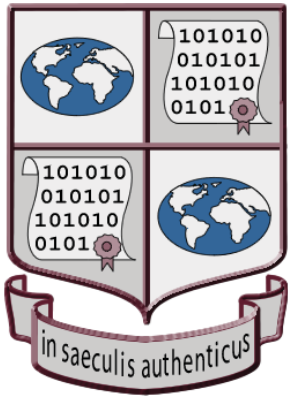
InterPARES Project

International Research on Permanent Authentic Records in Electronic Systems

TEAM ITALY



Studio sulla conservazione delle email

13 gennaio 2009

Natura del documento “email”

Le questioni aperte:

- Quali sono le peculiarità
- Quali sono le buone pratiche di conservazione nel caso di documenti archivistici
- In che misura gli obiettivi di conservazione condizionano la gestione delle email

Obiettivi dello studio

- Analizzare gli aspetti tecnici dei sistemi di posta elettronica e dei metodi di gestione in quanto rilevanti a fini di conservazione
- Rilevanza del problema:
  - Struttura complessa del documento
  - Necessità di tener conto anche delle infrastrutture specifiche di trasmissione (ad esempio il web) (studio preliminare anche relativo ai formati dei messaggi)

Elementi di riflessione

- Scarsa attenzione al nodo tecnico-archivistico della conservazione delle email
- I prodotti di mercato si limitano a trattare il problema della gestione "sicura"
- In pochi enti i sistemi di protocollo trattano l'email in quanto documenti da inserire nella gestione dei documenti
- In Italia il nodo è trattato con molta ambiguità dalla normativa (CAD, articoli 45-47, trasmissione e comunicazione dei documenti; dpcm 31 ottobre 2000)
- La complessità del problema implica che sia affrontato da più punti di vista: tecnologico, organizzativo, giuridico, archivistico-documentario

Il punto di vista dello studio

- Sono state prese in considerazione tutte le funzioni dei prodotti commerciali per la gestione delle email, incluse le cosiddette funzioni di *archiving*
- Si sono analizzati i requisiti e le indicazioni riportate nei principali documenti internazionali di policy e normalizzazione (DoD 5015-02-STD-Electronic Records Management of Electronic Records (2008); MoReq2 (2008); DCC Digital Curation Manual)

La struttura del rapporto - 1

- Capitolo 2: The Internet e-mail infrastructure
  - How does e-mail work
  - End-user access to e-mail
  - Interoperability of e-mail systems: standards and protocols
- Capitolo 3: Format and structure of e-mail messages
  - Message body and attachments
  - Metadata
  - Dynamic contents
  - Multimedia contents

La struttura del rapporto - 2

- Capitolo 4: Security and privacy issues
  - Originality and authenticity of the e-mail message
  - Certified email
  - Security
  - Privacy
- Capitolo 5: Archiving and preserving e-mail
  - Reference model
  - Capturing e-mails
  - Formats for preservation
  - Message classification and metadata extraction
  - Checking and preserving authenticity and integrity
  - Long term preservation

La struttura del rapporto - 3

- 6. Access to the e-mail archive
  - Search and discovery
  - Presentation
  - Access control
  - Audit trail
- 7. Commercial products for e-mail management
  - Clients
  - Integrated systems
  - E-mail archiving
  - State of the art and trends (instant messaging, cell phone, IP telephone calls…)

La prima email

- La prima e-mail inviata e trasmessa da una rete telematica è stata spedita nel 1971 tra due computer situati nella stessa stanza, passando per ARPAnet (la rete da cui è nata Internet)
- Licklider (MIT), il primo teorico della rete globale, ha sottolineato che uno dei vantaggi dei sistemi di messaggistica riguarda la possibilità di poter contattare senza formalità e gerarchie chiunque, anche in ragione della rapidità della comunicazione che avvicina questo sistema alla comunicazione telefonica con il vantaggio di produrre documentazione scritta.

Lo stato dell'arte

- Le e-mail costituiscono oggi il mezzo di comunicazione scritta più diffusa
- Più di 100 miliardi di e-mail sono inviate quotidianamente; saranno 300 nel 2010
- Si scambiano ormai informazioni cruciali (anche se non ancora sempre – almeno nei Paesi di diritto amministrativo – con dichiarato valore giuridico)
- In molti casi le e-mail contengono informazioni che non sono conservate altrove
- Costituiscono il 75% della proprietà intellettuale degli enti

Le criticità nella memorizzazione

- I sistemi non sono predisposti per la gestione e l'archiviazione di grandi quantità di e-mail e allegati per lunghi periodi di tempo
- Le pratiche aziendali tendono a limitare le dimensioni per i dipendenti
- Gli impiegati scaricano e salvano le email ritenute rilevanti a *titolo personale*: le procedure sono sempre informali, non controllate e inaffidabili *(tranne nei pochi casi in cui le procedure documentarie, ad esempio in Italia quelle previste nei manuali di gestione stabiliscano diverse soluzioni)*
- In ogni caso i messaggi così gestiti possono essere utilizzati solo da chi li ha salvati e archiviati (per di più se è in grado di ritrovarli)
- La crescente preoccupazione per l'archiviazione di tale massa di documenti rende il problema cruciale e sempre più al centro delle preoccupazioni dei produttori di software e, in generale, del mercato
- Individuare una soluzione archivistica per l'organizzazione, la classificazione, la identificazione delle email, definire criteri di selezione, stabilire modalità adeguate di ricerca costituiscono ormai delle priorità

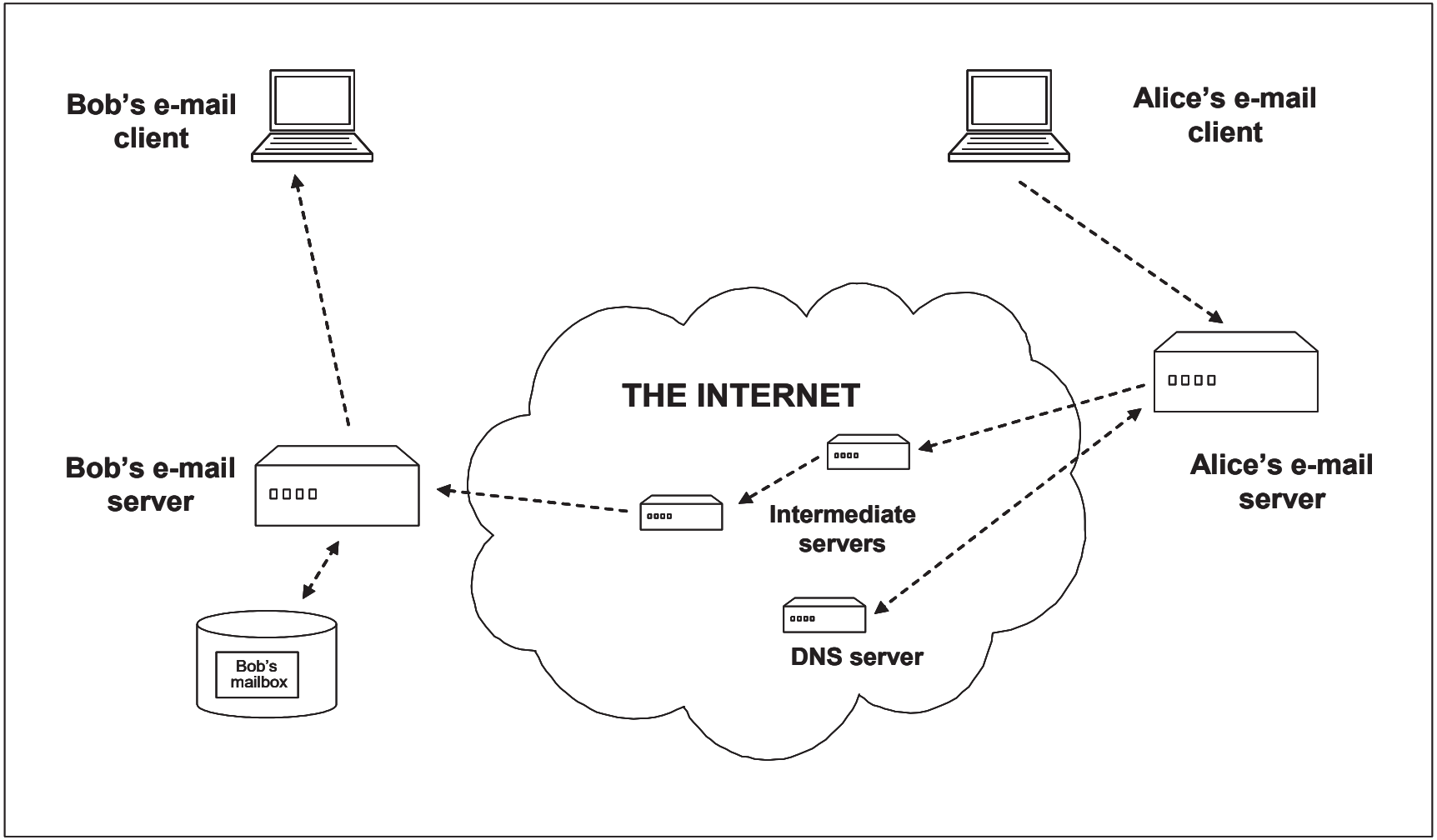
Conformità alla normativa nazionale e internazionale

- Nel settore privato – a livello di multinazionale – l'incapacità di conservare I documenti in forma di e-mail ha comportato il pagamento di multe salatissime: ad esempio la Morgan Stanley nel 2005 ha pagato 1,45 miliardi di dollari, in un caso – soprannominato 'legal Chernobil' in cui i dischi di backup erano andati perduti o divenuti illeggibili. Sommando altri casi (ad esempio incapacità di portare prove documentarie in giudizio in forma di e-mail) si arriva, negli ultimi anni, a diversi miliardi di dollari
- Negli Stati Uniti la nuova normativa sulla procedura civile (ma anche il Sarbanes-Oxley Act e le normative SEC) prevede la produzione di documenti informatici, *per le quali gli enti devono essere in grado di assicurare funzioni di ricerca, esibizione (in tempi rapidi) e quindi gestione e conservazione, anche perché la valutazione della prova documentaria in termini giudiziari è basata sulla possibilità di verificare che i risultati prodotti siano ripetibili, oggettivi e verificabili, così come la qualità delle procedure è valutata in termini di affidabilità, autenticità e disponibilità dei contenuti informativi.*
- Centrale è naturalmente anche l'esigenza della conservazione a fini storici della documentazione prodotta

Come funziona l'e-mail - 1

- Il sistema di e-mail è basato sul metodo *store-and-forward* per lo scambio dei messaggi su internet: il messaggio inviato da un utente è gestito con un processo asincrono di consegna che include una serie di passi, ciascuno dei quali è memorizzato da un server intermedio sulla rete, inoltrato in tempi successivi fino a quando non giunge a destinazione. La tempistica dipende dalla disponibilità delle connessioni sulla rete
- Il processo prevede un mittente, Alice, e un destinatario, Bob. Bob e Alice utilizzano applicazioni specifiche (denominate client di posta elettronica che operano sul PC di ciascun nella spedizione e nella ricezione (ad esempio. Eudora, Outlook). I client non comunicano direttamente, ma devono connettersi a server di posta, ad esempio applicazioni specifiche operative presso gli enti cui Bob e Alice appartengono che gestiscono la consegna dei messaggi

Come funziona l'e-mail – 2
Infrastruttura di base



Come funziona l'e-mail – 3

- Alice compone il messaggio utilizzando il suo client di posta
- Il messaggio è formattato dal client di Alice in un *formato web di posta* e spedito al *server locale di posta*
- Il server locale di posta individua l'indirizzo del server di posta di Bob consultando il *Domain Name System (DNS)*, ovvero la directory distribuita di Internet per i nomi di dominio;
- I due server di posta scambiano il messaggio che può passare attraverso una serie di server intermedi sulla rete prima di essere alla fine memorizzato dal server di posta di Bob nella casella di posta di Bob
- Il messaggio è mantenuto nella casella di Bob fino a quanto non viene letto e/o scaricato utilizzando il suo client di posta
- La procedura è abbastanza simile a quella seguita quando Alice e Bob si scambiano delle lettere tradizionali. Gli uffici postali svolgono la stessa funzione dei server di posta e la consegna può passare attraverso ulteriori uffici di posta (server intermedi). In entrambi i casi il tempo di consegna e la consegna stessa non sono garantiti

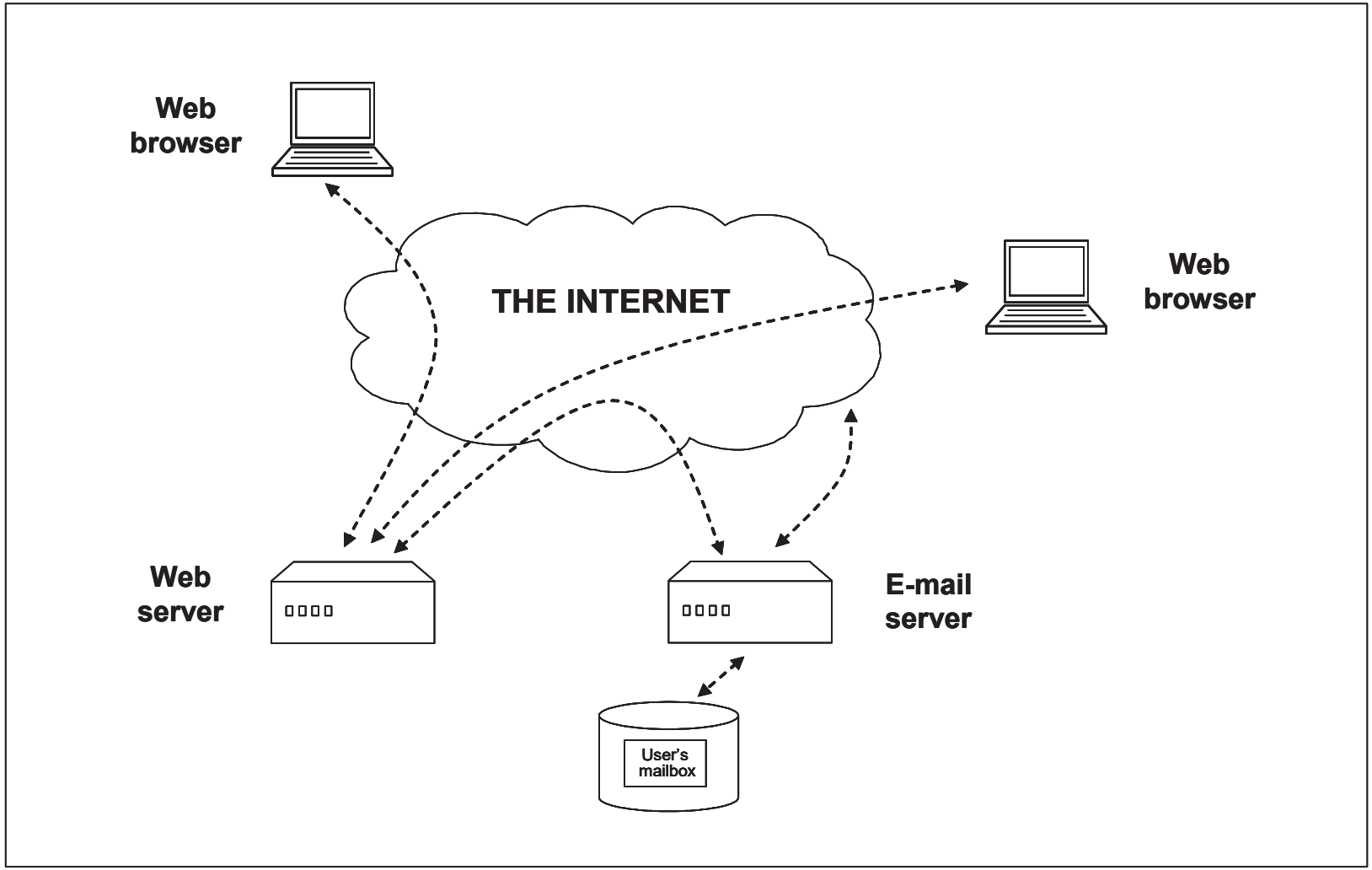
Come funziona l'e-mail – 4

- Internet è una rete basata su un approccio *best-effort (con il miglior impegno possibile*) al fine di raggiungere la sua destinazione; molti server sono gestiti da organizzazioni indipendenti che non assumono impegni specifici sulla disponibilità e la qualità del servizio. Ne derivano tempi incerti di consegna e rischi di perdita
- Peraltro tutti i client e i server coinvolti nel processo di consegna seguono un set di regole rigide (*protocolli*). Questo consente di tracciare tutti gli eventi rilevanti e documentare queste informazione in un rapporto piuttosto dettagliato che è allegato al messaggio. Inoltre, in caso di fallimento, il server può reiterare il processo e il mittente può richiedere rapporti di consegna e ricevuto per avere la prova che il messaggio sia stato consegnato e/o sia stato effettivamente letto *(o meglio aperto)*.

Le modalità di accesso dell'utente finale alle e-mail - 1

- **Client di posta**: forniscono interfacce utenti amichevoli e funzioni per organizzare e memorizzare messaggi, gestire directory, ecc. I messaggi sono generalmente scaricati e memorizzati sul computer dell'utente, con problemi per chi viaggia che ha necessità di accedere alla propria posta da computer differenti.
- **Webmail**: gli utenti accedono alle e-mail mediante un servizio offerto dal loro ISP (Internet Service Provider o fornitore di accesso) or da un'organizzazione terza quale Hotmail o G-mail. Un programma di navigazione (Internet browser, tra cui Explorer, Mozilla, ecc.) connette l'utente al web server, dove è in funzione una speciale applicazione. Il web server opera come elemento di intermediazione e gestisce la connessione con il server di posta. I messaggi non sono scaricati sul computer dell'utente, ma gestiti direttamente e archiviati sul web server.

Le modalità di accesso dell'utente finale alle e-mail – 2. Webmail



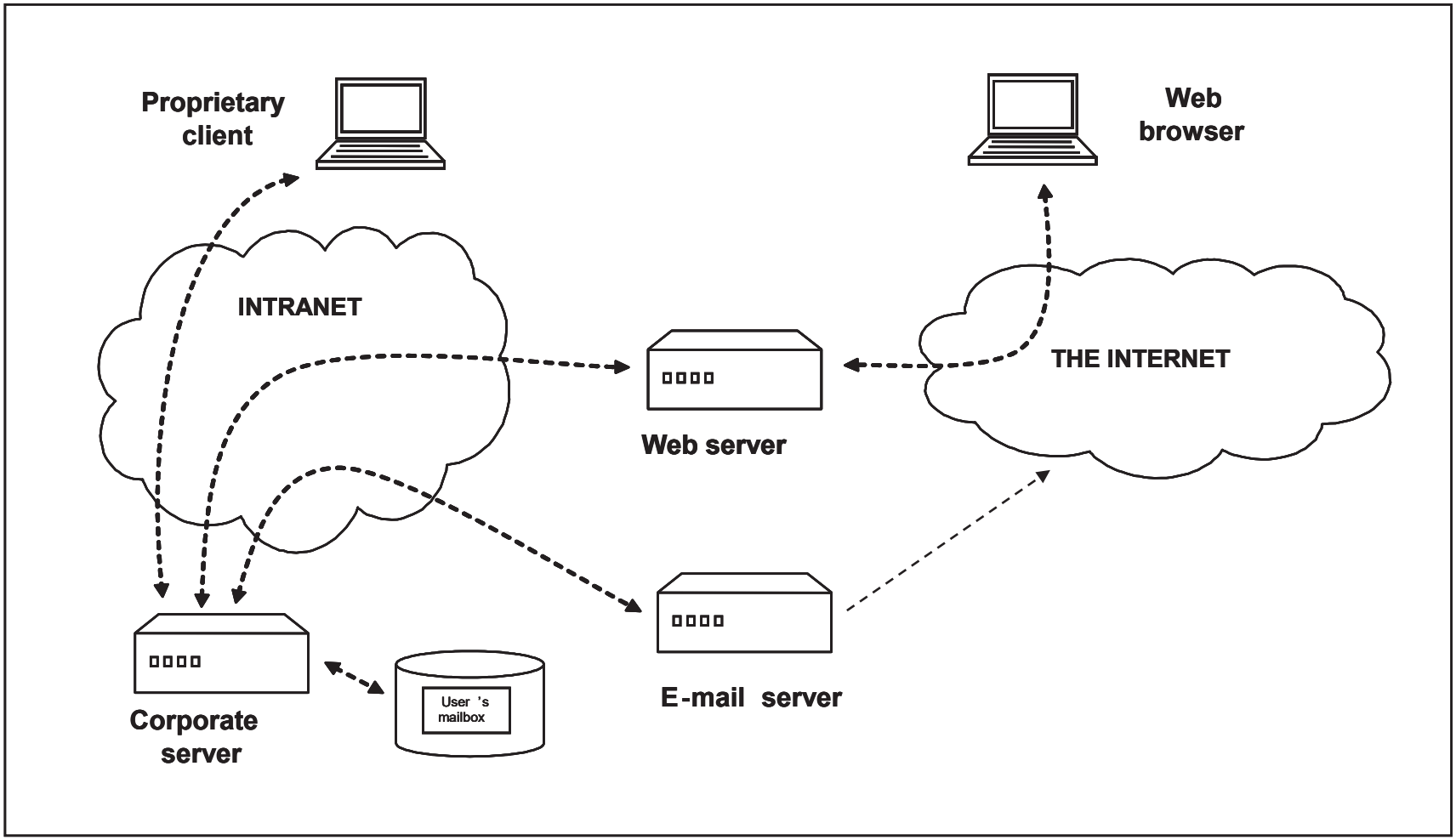
Le modalità di accesso dell'utente finale alle e-mail - 3

- **Sistemi integrati**: sono basati sull'idea di integrare l'accesso alle e-mail in un ambiente collaborativo più ampio che include altre funzioni tra cui funzioni di direct messaging, rubrica, contatti, agenda, supporto per il cellulare e accesso web, ma anche memorizzazione dei messaggi su un server centrale

  Gli utenti utilizzano sul loro computer applicazioni client proprietarie (ad esempio Microsoft Exchange o Lotus Notes) che li connettono al server dell'ente che a sua volta si connette con il server di posta.

  Questi sistemi hanno anche una interfaccia web opzionale, funzionalmente equivalente al webmail, che consente l'accesso via internet con un browser web. Tuttavia l'interfaccia primaria è proprietaria ed è quella utilizzata nell'intranet dell'ente.

Le modalità di accesso dell'utente finale alle e-mail – 4. Sistema integrato



Interoperabilità dei sistemi di posta elettronica

L'interoperabilità nella trasmissione è basata su due elementi principali:

- *Protocolli di comunicazione*, ovvero set di regole che governano la comunicazione fra operatori,assicurando che possano interagire con affidabilità e correttezza grazie a un linguaggio comune e a procedure standard;
- *Formato del messaggio*, ovvero un set di definizioni formali che specificano la struttura del messaggio e le modalità con cui sono codificati il messaggio e gli allegati, in modo da rendere possibile una loro corretta interpretazione da parte di client di posta diversi e garantire che il contenuto del messaggio sia correttamente presentato a chi lo riceve
- L'interoperabilità deve essere garantita anche nel tempo. Questo significa che quando la definizione di protocolli e il formato del messaggio si modificano, deve essere comunque garantita una compatibilità retroattiva, ovvero le nuove regole devono essere ancora compatibili con le vecchie

Gli standard Internet - 1

- Gli standard Internet sono sviluppati e promossi dall'*Internet Engineering Task Force* (*IETF)*, che coopera strettamente con i principali organi di normalizzazione internazionale, quali ISO/IEC e il *World Wide Web Consortium (W3C)*, la principale organizzazione internazionale per la normalizzazione del WWW
- Il processo di standardizzazione è fortemente cooperativo ed è basato su documenti specifici chiamati *Request For Comments (RFC)*. Gli RFC sono documenti in bozza, per lo più proposte di standard, pubblicate da IETF e messe a disposizione sulla rete come 'request for comments'. A ciascun RFC è assegnato un numero univoco e non è mai annullato o modificato. Se si richiedono modifiche, si predispone un nuovo RFC con un nuovo numero che sostituisce il precedente

Gli standard Internet - 2

- "Non tutti gli RFC sono standard" (come stabilito da RFC 1796, che affronta il processo di standardizzazione). Alcuni sono solo promemoria, osservazioni che le persone vogliono condividere, documenti di ricerca, o proposte preliminari su materie che riguardano Internet e i sistemi basati su Internet. L' IETF assegna perciò a ciascun RFC una valutazione, chiamata *status*.
- Gli RFC maturi sono valutati *Standard Track*, e sono ulteriormente divisi in *Proposed Standard*, *Draft Standard* e *Internet Standard*. Gli Internet Standard (STD) si riferiscono ciascuno a un RFC (o a un insieme di RFC) e ottengono un numero univoco come nel caso degli RFC. Tuttavia, diversamente dal numero di RFC, quando lo standard si modifica il numero STD non cambia, ma semplicemente si riferisce a un nuovo RFC che sostituisce quello originario.

Standardizzazione della trasmissione di e-mail

- L'interoperabilità tra serve e tra client è assicurata mediante *SMTP, Simple Mail Transfer Protocol*, che è un Internet Standard STD 10 (1982), basato su RFC 821, ma il protocollo utilizzato dalla gran parte delle applicazioni di posta elettronica è quello noto come *ESMTP* (*Extended SMTP*) e definito in RFC 2821, pubblicato ad aprile 2001
- SMTP specifica il modo in cui il client di posta interagisce con il server di posta e consegna il messaggio e il mondo in cui i server di posta (spesso chiamati *SMTP-server*) interagiscono reciprocamente in modo che il messaggio passi attraverso una serie di operatori e raggiunga la destinazione finale.
- Questo standard stabilisce il formato di base dei messaggi che possono essere gestiti dai server SMTP. E' un formato semplice costituito dal testo in *plain ASCII* (detto anche ASCII a 7-bit o US-ASCII), adatto solo per l'inglese e per poche altre lingue. Questo limite è superabile definendo modalità speciali per 1) codificare qualunque contenuto più ricco in caratteri plain ASCII, consentendo in tal modo l'utilizzo di un più ampio set di caratteri nel testo del messaggio e 2) includere nei messaggi testo formattato e contenuti multimediali

Standardizzazione della comunicazione client-server

Si utilizzano due protocolli principali per leggere e recuperare le e-mail.

- **Post Office Protocol version 3 (POP3).** Originariamente destinato a sostenere connessioni temporanee (ad esempio connessioni dial-up). Quando l'utente finale si connette scarica dal suo PC i nuovi messaggi, che sono cancellati dal server. Una copia del messaggio può essere lasciato sul server, ma è un'opzione utilizzata raramente dato che molte applicazioni non sono in grado di distinguere tra i nuovi messaggi e quelli che sono stati già scaricati.
- **Internet Message Access Protocol (IMAP)**. (RFC 3501 – *Proposed Standard*). E' pensato in modo specifico per le esigenze degli utenti che viaggiano in quanto consente l'accesso alle e-mail da computer diversi e da un server remoto. L'utente può creare contenitori sul server per organizzare i messaggi. Mantenere tutti i messaggi sul server è una condizione positiva per la loro conservazione

POP3 e IMAP sono utilizzati anche per altri schemi di accesso alle e-mail. Nel caso di webmail, il server li utilizza per il recupero dei messaggi, mentre SMTP spedisce i messaggi. Allo stesso modo nei sistemi integrati, il server utilizza protocolli standard per connettersi al server di posta (SMTP server), mentre protocolli proprietari sono utilizzati nella comunicazione con l'utente finale. In entrambi i casi i messaggi sono mantenuti sul server.

Standardizzazione del formato del messaggio

- Il formato di base dei messaggi è definito da STD 11 (1982), ma la maggior parte delle applicazioni può trattare una versione aggiorgnata del formato definita in RFC 2822, che è ancora formalmente un *Draft Standard*
- I messaggi di e-mail dovrebbero contenere soltanto caratteri *plain ASCII text* (ASCII a 7 bit o US-ASCII) (1963). I server SMTP possono gestire solo questo tipo di messaggi
- Per superare questa limitazionel il formato del messaggio è stato di conseguenza esteso dallo standard *Multipurpose Internet Mail Extension (MIME)* al fine di supportare:
  - Testi e intestazioni in set di caratteri diversi dal ASCII;
  - Messaggi strutturati in più parti;
  - Allegati non testuali, inclusa una ampia varietà di file multimediali

Codifica MIME - 1

- Non è ancora formalmente un Internet standard. E' definito da una serie di RFC collegati il cui status è ancora quello di *Draft Standard*
- E' basato sull'idea di codificare caratteri non ASCII e potenzialmente qualunque tipo di informazione allegata al messaggio; lo schema di codifica è aggiunto al messaggio al fine di permettere la decodifica del messaggio al momento del recupero.
- Ogni attività di codifica e decodifica è condotta dai client di posta al momento della spedizione e del recupero dei messaggi. Il messaggio, quando trasmesso, è costituito solo di caratteri plain ASCII per cui non è richiesta nessuna estensione a SMTP e ai server SMTP per trattare i messaggi MIME
- E' per sua natura estensibile e la sua definizione include un meccanismo per *registrare,* se necessario, nuovi tipi di data, denominati tipi *Internet media* o tipi MIME. La registrazione di nuovi tipi di dati è gestita da un autorità indipendente *Internet Assigned Numbers Authority (IANA)*, un'entità che supervisiona, tra l'altro, l'allocazione degli indirizzi IP e la gestione di DNS

Codifica MIME - 2

- Un notevole numero di tipi Internet media sono stati registrati. Questo consente di allegare a una e-mail qualunque tipo di file, in particolare testi formattati, contenuti multimediali, ecc. I dati binari sono codificati utilizzando uno schema ben conosciuto denominato BASE64, in caratteri plain ASCII.
- Una ulteriore recente estensione di MIME è *Secure/Multipurpose Internet Mail Extension (S/MIME)*, che definisce uno standard per la cifratura e la firma a chiave pubblica di e-mail incapsulate in MIME

La struttura di E-Mail

Un messaggio e-mail consiste di due sezioni principali::

- *Header* (intestazione) una sequenza di righe, all'inizio del messaggio, generate dal client di posta che spedisce l'e-mail e dai server di posta coinvolti nel processo di consegna;
- *body*, il resto del messaggio che contiene il testo in caratteri plain ASCII e/o testo contenente caratteri non-ASCII e dati binari codificati in plain ASCII

Nel caso più semplice, il corpo del messaggio contiene soltanto caratteri palin ASCII.. Tali messaggi sono semplici da trattare, possono essere mantenuti nela forma nativa e successivamente letti senza bisogno di altre forme di decodifica.

Sfortunatamente, la gran parte dei messaggi utilizzano caratteri ASCII o Unicode estesi, hanno allegati e/o sono in formato html. In tutti questi casi il messaggio deve essere in formato MIME.

Si analizzerà quindi la struttura delle e-mail

Intestazione del messaggio (header)

- E' una sequenza di righe prodotte dal client di posta e dai server coinvolti nel processo di consegna. L'intestazione si conclude con una linea vuota.
- Soltanto una parte minore dell'informazione contenuta nell'intestazione del messaggio è mostrata dai clienti di posta
- I client di posta in genere consentono agli utenti di ispezionare l'intestazione completa, se richiedono di indagare l'origine del messaggio e il processo di consegna
- Ci sono quattro tipi di intestazioni; intestazione di identità, di consegna, di sicurezza e di formattazione.

Identità

Header	Description
• Date	• *Date/time sent*
• From	• *Address of sender*
• Sender	• *Address of sender's assistant*
• Organization	• *Organization of author*
• To	• *Address of recipients (may be a list)*
• Cc	• *Address of recipients in carbon copy*
• Bcc	• *Address of recipients in blind carbon copy*
• Subject	• *Message summary*
• Message-ID	• *Unique identifier assigned by the sender*
• Return-Path:	• *Address for 'bounce messages'*

Transmissione

- 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, expressed in the standard format in which the GMT and the time shift are given.
- Return-Receipt-To/ Disposition-Notification-To: specify if the sender requested a receipt, and to which address it should be sent.

Sicurezza

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

Formato/Codifica

- Struttura del corpo del messaggio e versione MIME version (sempre 1.0)
- Content-Type: specifica se il messaggio contiene una o più parti; in quest'ultimo caso un separatore è specificato: una stringa che separa le parti multiple di un messaggio nel corpo del messaggio

Se, al contrario, il messaggio contiene una sola parte

Content-Type e Content-Transfer-Encoding sono specificate direttamente nell'intestazione

Corpo del messaggio: single part

- E' un solo messaggio di testo senza allegati. Il tipo di contenuto (Content-Type) corrispondente nell'intestazione è semplice testo e specifica la codifica di carattere.
- Nel caso di caratteri plain ASCII la codifica di trasferimento del contenuto (Content-Transfer-Encoding) è di 7-bit.
- Se il set di caratteri è diverso da plain ASCII, si utilizza la codifica Unicode.
- Uno schema di codifica simile, chiamato *Encoded-Word*, è utilizzato per le informazioni di carattere testuale in set di caratteri diverso dal plain ASCII.

Struttura di un messaggio single part

```
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
```

Corpo del messaggio: Multiplepart

- Un messaggio MIME multipart è costituito da più parti separate da un *separatore* (*boundary)* ad esempio dalla stringa definita nell'intestazione a livello superiore dell'intestazione Content-Type inserita tra due parti.
- Si tratta di messaggi di vari tipi, specificati come *subtypes* nell'intestazione Content-Type:
  - multipart/mixed
  - multipart/alternative
  - multipart/digest
  - multipart/related
  - multipart/report
  - multipart/signed
  - multipart/encrypted

Corpo del messaggio: Multipleparts/mixed

Impacchetta in un singolo messaggio molteplici file con diversi tipi di dati, specificati dalle intestazioni Content-Type .

Il content type è di default text/plain. Questo *subtype* è generalmente utilizzato per inviare messaggi con allegati

L'ordine delle parti è significativo ed è utilizzabile dai client di posta in fase di visualizzazione.

Corpo del messaggio: Multiplepart/ alternative)

- Utilizzato per spedire versioni “alternative” dello stesso contenuto, il cui formato per ciascuna versione viene specificato dall’intestazione relativa al Content-Type
- Le parti alternative appaiono secondo un ordine di crescente affidabilità rispetto al contenuto originale, con la scelta migliore in ultima posizione
- I client di posta dovrebbero riconoscere che il contenuto delle varie parti è intercambiabile e visualizzare il tipo “migliore”
- Una tipica istanza è costituita dai messaggi che sono spediti sia in plain text (Content-Type: text/plain) che in HTML (Content-Type: text/html).
- Il plain text fornisce compatibilità retroattiva, mentre la parte in html consente formattazione e hyperlink. Pertanto le due parti non contengono esattamente la medesima informazione, poiché la parte in html è in qualche modo più ricca.

Corpo del messaggio: Multiplepart/digest

- E' una tipologia sintatticamente identica, ma si distingue dal punto di vista semantico
- Il valore di default del Content-Type passa dal formato plain/text a rfc822, che indica che il contenuto include un messaggio incapsulato con la sintassi di un messaggio RFC 822.
- E' utilizzato per spedire gruppi di messaggi in un singolo messaggio e, molto spesso, per inoltrare (*forward*) messaggi

Corpo del messaggio: Multiplepart/related

- Fornisce un meccanismo per rappresentare oggetti composti costituiti da molteplici parti interrelate, ciascuna delle quali è inviata come parte del messaggio.
- Un caso tipico è costituito da messaggi che inviano una pagina web, completa delle sue immagini. La root part contiene il documento HTML e utilizza i marcatori di immagine come collegamento alle immagini memorizzate nelle ultime parti del messaggio

Corpo del messaggio: Multiplepart/related (testo e HTML)

```
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---
```

Corpo del messaggio: Multiplepart/report)

- E' utilizzato per qualunque tipo di rapporto di posta elettronica. In genere è ustato per i rapporti di consegna (delivery report).
- E' costituito di due parti obbligatorie e una opzionale.
  - La prima contiene un messaggio comprensibile (*human readable*) con una descrizione della ragione che ha determinato la produzione del rapporto
  - La seconda è machine-parsable (ovvero un programma può scinderla in bit di dati facilmente memorizzabili o manipolabili) in relazione all'evento relativo al messaggio originario cui il rapporto si riferisce.
  - La terza parte, opzionale, contiene il messaggio cui il rapporto si riferisce o una sua parte, al fine di contribuire alla diagnosi del problema.

Message Body (Multipleoart/signed)

- E’ utilizzato per spedire messaggio provvisti di firma digitale
- E’ costituito di due parti, il testo del messaggio e la firma
- La firma è utilizzata per validare l’intero contenuto della prima parte. E’ possibile utilizzare tipi di firma diversi e I processi di normalizzazione in questo ambito sono ancora insufficienti
- I messaggi firmati possono essere inviati anche utilizzando lo schema multipart-mixed

Corpo del messaggio:Multipleparts/encrypted

- E’ utilizzato per inviare messaggi cifrati. Ha due parti.
- La prima contiene l’informazione necessaria per decifrare la seconda parte.
- Come nel caso dei messaggi firmati, ci sono implementazioni diverse che sono specificate nel Content-Type della prima parte.
- Anche in questo caso i processi di normalizzazione in questo ambito sono ancora insufficienti

MIME Media Type

- E' un elemento di identificazione utilizzato nella intestazione di *content type* per specificare la natura dei dati nel corpo di un'entità MIME, ad esempio nel corpo di un messaggio *single part* o in una parte di un messaggio *multipart.* Sono utilizzati in altri protocolli internet con lo scopo di consentire la corretta interpretazione del contenuto del messaggio specificando il formato del corpo del messaggio e dei suoi allegati
- E' definito in RFC 2046 in modo da assicurare la sua estensibilità (è previsto un processo di registrazione)
- Gli identificatori sono di due livelli: t*op-level type* e *subtype*, con parametri aggiuntivi opzionali
- Ci sono sette tipi *top-level media*, di cui cinque *discreet data type* (ad esempio per specificare il formato di un solo file: testo, immagine, audio, video, applicazioni) e due sono *composite data type*, (ad esempio per specificare la struttura di un corpo MIME composto di parti multiple)

Media Type: il mantenimento nel tempo

Nel processo di mantenimento nel tempo, è necessario garantire la capacità di restituire ogni parte di un messaggio in qualunque momento. E' in particolare indispensabile assicurare che:

- I media type che appaiono in un messaggio siano stati oggetto di memorizzazione insieme alla informazione necessaria a gestirli
- Un'applicazione sia disponibile per ciascun media type memorizzato; o
- Una copia di conversione dell'allegato sia conservata in un formato che garantisca la possibilità di restituzione futura

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.
- 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.
- 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 maintenance, since it affects the selection of the messages to be kept

SPAM

- Spam volume exceeded legitimate e-mails in 2007
- Common anti-spam products may be set according to one the following policies:
  - presumed spam messages are simply marked as spam and grouped in special folders;
  - presumed spam messages are discarded by the filter.
- The choice between these polices is affected by the relevance of so called ‘false positives’, i.e. messages that are tagged by the filter as spam but are not, and the consequent potential loss of legitimate messages.

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
- 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
- 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.

Certified E-Mail

- Certified e-mail messages can be sent among users registered with certified providers, who have to comply with security and interoperability requirements and are supervised by a national agency.
- When a message is sent, in addition to the standard delivery service, the provider authenticates the originator, and issues two electronically signed receipts: one proving that the message has been sent by the person in question, and the other one proving that the message has been delivered to the destination mailbox.
- Electronic receipts have legal force and may be used in litigations. Moreover the receipts may contain a 'fingerprint', i.e. a digest of the content of the message signed by the certified provider that can be used to avoid repudiation of the message content by the recipient.
- Certified e-mail, thanks to the presence of a trusted third party, guarantees the authenticity (identity and integrity) of the message, and provides formal evidence that the message has been delivered.

Privacy Issues

- Privacy concerns should focus more on mailbox unauthorized access than on disclosure of message content during its transmission.
- In countries with constitutional guarantee of the secrecy of correspondence, e-mail is equated to traditional mail, and only the owner of the mailbox is allowed to access it, even in an office environment. Privacy authorities protect e-mail confidentiality and grant the administrators the right to access users e-mail message only in particular situations and with due care.
- Employees should use company mailboxes only for business purposes and use a personal mailbox for their private messages. In practice it may be difficult to distinguish between personal and business communication. Also, there may be business messages that, since they are meant to be read by a specific person, may also contain personal information that should not be disclosed.
- When the distinction is not obvious, employers and legal authorities can ask for full disclosure of all e-mail.

Mantenance Issues

- Distinction between the e-mail application and the recordkeeping system
- Most e-mails will only be kept in the application or repository and then discarded
- 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 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.

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 selection scheme adopted to decide if a message is going to be captured, the user should be involved in the classification of the records and in manually entering additional metadata.
- The InterPARES recommendation entrusts this function jointly to the user and the to the recordkeeping system, under the control of the system administrator (you should look at the InterPARES 2 COP model).

Maintenance and Preservation Issues

- 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 protected.
- 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.

Maintenance and Preservation of MIME

- 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.
- 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.

Long-term Preservation

- Preserving integrity: not a different issue for e-mail
- Preserving accessibility 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 whim, while in some other environments organizations may be able to strongly recommend, or even enforce, the use of formats suitable for long-term preservation.
- 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 are realistically possible to support on the long term.

Long-term Preservation (cont.)

- messages are maintained in RFC 2822/MIME format;
- 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 header lines for all converted objects; this provides some kind of assessment of the conversion procedure, and allows to understand to what extent the integrity of the record may have been compromised;
- when a supported format approaches obsolescence, all records in that format are converted in a new supported format.

Preservation by Individuals

- Users interested in preserving messages have to set up their own procedures, which may be based on three alternatives:
  - converting individual messages into text files and preserving these files;
  - converting individual messages or entire folders of messages into PDF documents using Adobe Acrobat Professional;
  - performing and preserving regular e-mail backups.
- Any attachments to an e-mail message form part of the message, so ensure that attachments are saved and remain associated with any e-mail that is of long-term importance.
- Single messages are usually stored in the original RFC 2822 format.
- Aggregations of e-mail messages are stored by some products in proprietary formats (e.g. Lotus notes, Outlook, Pegasus), and by other products in open formats (e.g. Alpine, Gnus, Eudora, Kmail, Mozilla Thunderbird, Novell evolution, Opera mail). Products following the latter approach often give the user choice between several different open formats.

Preservation by Individuals (cont.)

- Popular formats for aggregations of e-mails are Mbox and Maildir. In Mbox all messages are concatenated and stored as plain text in a single file, while Maildir uses a separate file for each message. A significant advantage of these formats is that, since they rely on standard files, the stored information may also be accessed using standard content management tools.
- As most of the preserving and filing actions have to be performed manually by the individual's and small office environment, some e-mail offer the possibility of automating a sequence of actions, by supporting scripts (e.g. Lotus notes, Mozilla Thunderbird, Outlook express) or java language (e.g Lotus notes).

E-Mail Archiving

- "E-mail archiving" products have been designed to
  - manage a huge number of stored e-mails without affecting e-mail server performances;
  - ensure regulatory compliance by capturing messages before they can be modified maliciously or deleted by the recipient;
  - allow organizations to implement structured policies for accessing stored e-mail;
  - include auditing capabilities to track access to archived records;
  - extend active mailboxes by providing user access to the repository via a web client and through the e-mail client;
  - enforce integration between e-mail management systems and record management systems;
  - provide advanced search and knowledge management capabilities.
- These functionalities allow bulk e-mail retention and access to stored messages according to security policies.
- "Archiving products" are mainly designed for large organizations managing the e-mail through integrated systems. Therefore, practically all vendors support Microsoft Exchange, and a large and increasing number of them support also Lotus Domino.

E-Mail Archiving (cont.)

- Symantec's product, Enterprise Vault, is an integrated content archiving platform supporting e-mail, instant messages, SharePoint and, through third parties add-on modules, popular proprietary repositories (e.g. Bloomberg, BlackBerry).
- Besides Symantec, products of several other vendors provide not only for the capture of all e-mail messages, but also allow the definition and the implementation of a retention policy, and give the end-user a view of stored messages similar to an extended mailbox.

E-Mail Archiving (cont.)

Typical functionalities found in these products, interesting for maintenance purposes, are:

- user options for message classification (e.g. Open Text, CommonStore, Message menager);
- automatic classification capabilities (e.g. CommonStore, EmailXtender, Enterprise Vault, Message Manager);
- cross-user "archive full-text searches" (e.g. Autonomy Zantaz, CommonStore);
- assurance of message authenticity via electronic signature (e.g. HP e-mail archiving, MailMeter);
- disaster recovery features (e.g. NearPoint, Enterprise Vault);
- storage optimization (e.g. Enterprise Vault).

Conclusions

- The market is still aimed essentially at medium and large organizations, and therefore small organizations and individuals still lack specific products to support their e-mail preservation policies, and may only rely on the functionalities offered by e-mail clients.
- This makes implementing an e-mail retention and maintenance policy a quite difficult task for the individual user and the small organization, since it requires a technical background and professional profiles they usually lack.
- Therefore, the best solution for them may be to rely on retention and maintenance services offered by e-mail providers or specialized companies, a kind of offer that is expected to increase both in volume and in quality in the next years.