ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Content encoding

Fabio Vitali

Corsi di laurea in Informatica e
Informatica per il Management
Alma Mater – Università di Bologna

Cos'è il content encoding

- Molti ambienti informatici forniscono restrizioni sulla varietà di caratteri usabili. I più noti sono
  - Modelli di rappresentazioni dei dati (ad es.: stringhe nei linguaggi di programmazione, formati dati, linguaggi di markup, etc.). Spesso alcuni caratteri hanno scopi tecnici interni al linguaggio, e non è possibile utilizzarli semplicemente come contenuto
  - Canali di trasmissione (ad es: protocolli Internet): molti di questi canali sono stati creati quando ASCII 7 bit imperava, e non sono trasparenti all'uso di flussi di dati a 8 bit (8bit clean).

Termini frequenti

- Escaping: il carattere proibito viene preceduto o sostituito da una sequenza di caratteri speciali.
  - String c = "Questa stringa \"contiene\" caratteri speciali" ;
  - Questa stringa "contiene" caratteri speciali

- Encoding: il carattere proibito viene rappresentato numericamente con il suo codice naturale secondo una sintassi speciale
  - "felicit\u00E0" ;
  - felicità
  - felicità

L'origine dei problemi: SMTP

Simple Mail Transfer Protocol

- È uno dei protocolli di VII livello più importanti di TCP/IP, sicuramente il più antico tra quelli ancora in uso oggi (1982).
- SMTP è un protocollo text-based, per lo scambio di messaggi di posta e la verifica dei destinatari dei messaggi.
- Una connessione SMTP è composta da una apertura, uno o più sequenze di comandi, ed una chiusura.
- Ad ogni comando corrisponde una risposta composta da un codice numerico ed una stringa leggibile.
  - MAIL FROM:<Smith@alpha.com>
    250 OK
  - RCPT TO:<Green@beta.com>
    550 No such user here

Limiti di SMTP

- Questi sono i limiti fondamentali di SMTP:
  - La lunghezza massima del messaggio è di 1 Mb
  - I caratteri accettati sono solo ASCII a 7 bit
  - Ogni messaggio deve contenere una sequenza CRLF ogni 1000 caratteri o meno (alcune antiche implementazioni lo aggiungevano automaticamente se non lo trovavano).
- Questi limiti impediscono la trasmissione di documenti binari:
  - Un file binario usa tutti i 256 tipi di byte
  - Un file binario può facilmente essere più lungo di 1 Mb
  - In un file binario la sequenza CRLF è una sequenza come tutte le altre, e può esserci o mancare senza vincoli. Introdurla artificialmente può corrompere il file.
- MIME permette di bypassare questi limiti all'interno di SMTP

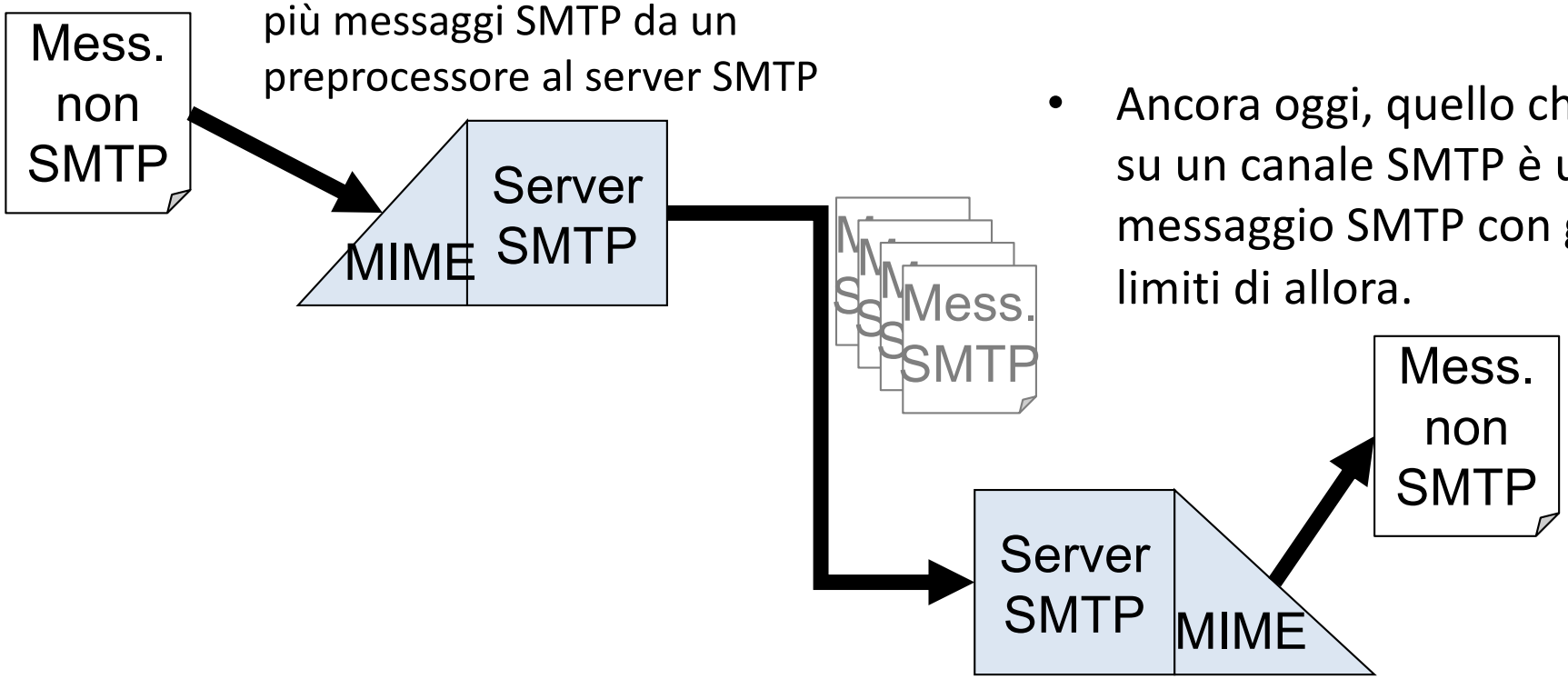
MIME

Multipurpose Internet Mail Extensions

- RFC 822 definisce con sufficiente dettaglio il formato degli header dei messaggi SMTP, ma specifica in modo molto generico che il corpo di un messaggio deve essere semplice testo US-ASCII.
- MIME ridefinisce il formato del corpo di RFC 822 per permettere:
  - Messaggi di testo in altri set di caratteri al posto di US-ASCII
  - Un insieme estensibile di formati per messaggi non testuali
  - Messaggi multi-parte
  - Header con set di caratteri diversi da US-ASCII.

Messaggi MIME su canali SMTP

- Il messaggio non compatibile con SMTP viene trasformato in uno o più messaggi SMTP da un preprocessore al server SMTP



- Ancora oggi, quello che viaggia su un canale SMTP è un puro messaggio SMTP con gli stessi limiti di allora.

- All'arrivo, il o i messaggi SMTP vengono decodificati e riaccorpati a formare il messaggio originale.

I limiti SMTP su MIME

- Codifica caratteri
  - il messaggio che contiene caratteri non ASCII viene codificato in maniera appropriata, cosicché ciò che viene effettivamente trasmesso sia veramente ASCII 7 bit. Il transfer encoding è diverso per messaggi di testo e messaggi binari (es. immagini).
- Sequenze CRLF
  - Tutti i sistemi di transfer encoding adottano un meccanismo per permettere la presenza di sequenze CRLF in mezzo al flusso di dati, alcuni anzi prevedendoli in maniera forzata ogni tot caratteri (76, per lo più).
- Lunghezza messaggi
  - Un processore MIME può generare vari messaggi SMTP da un singolo messaggio MIME, ciascuno inferiore per dimensione al limite SMTP. Il processore all'arrivo si occupa di verificare il corretto arrivo di tutti i singoli messaggi SMTP e ricostituisce il messaggio MIME originario

I servizi MIME

- Dichiarazione di tipo
  - Tutti i messaggi MIME vengono identificati da un Content Type, che definisce il tipo di dati del messaggio e aiuta l'applicazione ricevente a gestire il messaggio e a invocare l'applicazione più adatta.
  - N.B.: l'attribuzione dell'applicazione non viene fatta sulla base dell'estensione del nome del file.
- Messaggi multi-tipo
  - Un messaggio MIME può contenere parti di tipo diverso (es. un messaggio di tipo testo e un attachment binario).In questo caso si creano dei sottomessaggi MIME per ciascuna parte (con il suo bravo content-type) e il messaggio MIME complessivo diventa "multi-parte", qualificando e codificando in maniera diversa ciascuna sottoparte.

Header specifici MIME

- MIME introduce alcuni nuovi header SMTP:
  - Content-Type: il tipo MIME del contenuto. Serve per permettere al ricevente di scegliere il meccanismo più adatto per presentare i dati. Specifica la natura del dato tramite la specificazione di tipo, sottotipo e ulteriori parametri utili.
    - Content-Type: text/plain; charset=ISO-8859-1
  - Content-Transfer-Encoding: il tipo di codifica utilizzata per trasmettere i dati. Serve per la trasmissione su canale SMTP di dati che non sono naturalmente corretti secondo le regole di SMTP: 7bit, sequenze CRLF ogni 1000 caratteri o meno. Sono valori accettabili “7bit” (default), “8bit”, “binary”, “quoted-printable”, “base64” o altre stringhe definite nel registro IANA
    - Content-Transfer-Encoding: base64

MIME - Quoted printable

- Uno dei due tipi di content transfer encoding definiti da MIME. Viene usata per la trasmissione di dati che contengono grosse quantità di byte nel set US-ASCII, e solo poche eccezioni
  - Ad esempio, documenti testuali in lingue europee.
- Codifica dunque solo quei pochi byte non conformi. Per esempio:
  - Un codice superiore al 127 o inferiore al 32 viene codificato con la sintassi “=“ + codice esadecimale. Ad esempio “Hello’99” diventa “Hello=B499”
  - Righe più lunghe di 76 caratteri vengono interrotte con “soft breaks”, cioè con un uguale come ultimo carattere della linea.

MIME - Base 64

- Base 64 è un tipo di transfer encoding MIME suggerito per dati binari o multi-byte.
- Viene identificato un sottoinsieme di 64 caratteri di US-ASCII sicuri (hanno la stessa codifica in tutte le versioni di ISO 646). Questi sono:
  - le lettere maiuscole (26, 'A' => 0),
  - Le lettere minuscole (26, 'a' => 26),
  - I numeri (10, '0' => 52)
  - I caratteri '+' e '/' (=> 62 e 63 rispettivamente).
- Ogni flusso di dati viene suddiviso in blocchi di 24 bit (3 byte). A loro volta questi 24 bit sono suddivisi in 4 blocchi di 6 bit ciascuno e codificati secondo una tabella prefissata in uno dei 64 caratteri già descritti.

MIME – Base 64 (2)

Input	M	a	n
Codice ASCII	77	97	110
mappa bit	0 1 0 0 1 1 0 1	0 1 1 0 0 0 0 1	0 1 1 0 1 1 1 0
Indice 6-bit	19	22	5	46
Output Base64	T	W	F	u
Codice ASCII	84	87	70	117

- La stringa risultante viene divisa in righe di 76 caratteri (tranne l'ultima, che è lunga quanto deve essere) con l'aggiunta di CR-LF.
- Nella decodifica i codici CR e LF sono da ignorare.
- La decodifica di Base64 è algoritmica, banale, non usa chiavi né calcoli di particolari complessità.
- Base64 NON È una tecnica crittografica!!!

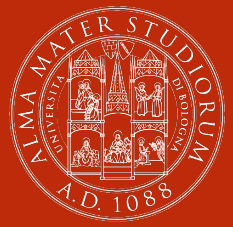
Conclusioni

Qui abbiamo parlato di set di caratteri

- I problemi di codifica e di ordinamento dei byte
- Meccanismi di encoding

Riferimenti

- N. Bradley, The XML companion, Addison Wesley, 1998, cap. 13.
- K. Simonsen, Character Mnemonics & Character Sets, RFC 1345, IETF, June 1992
- D. Goldsmith, M. Davis, UTF-7, A Mail-Safe Transformation Format of Unicode, RFC 2152, IETF, May 1997
- The Unicode consortium, Unicode® 8.0.0, Released: 2015 June 17, http://unicode.org/versions/Unicode8.0.0/



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Fabio Vitali

Corso di tecnologie web

Fabio.vitali@unibo.it

www.unibo.it