Appunti su Content Encoding

Cos'è il Content Encoding

Il content encoding è un meccanismo utilizzato per superare le restrizioni imposte da vari ambienti informatici sulla varietà di caratteri utilizzabili. Queste restrizioni possono derivare da:

- **Modelli di rappresentazione dei dati:** Alcuni caratteri hanno scopi tecnici interni e non possono essere usati come contenuto (es. virgolette in stringhe di codice).
- **Canali di trasmissione:** Molti protocolli (es. SMTP) sono stati creati quando ASCII a 7 bit era lo standard, e non supportano flussi di dati a 8 bit.

Termini frequenti:

- **Escaping:** Il carattere proibito è preceduto o sostituito da una sequenza speciale (es. `\"` in C, `&quot;` in HTML).
- **Encoding:** Il carattere proibito è rappresentato numericamente con il suo codice (es. `\u00E0` in JavaScript, `&#x00E0;` o `&#224;` in HTML).

L'origine dei problemi: SMTP

Simple Mail Transfer Protocol (SMTP):

- Protocollo di livello VII di TCP/IP, usato per lo scambio di email (1982).
- Text-based: comandi e risposte testuali.
- Connessione: apertura, sequenze di comandi, chiusura.

Limiti di SMTP:

- Lunghezza massima messaggio: 1 MB.
- Caratteri accettati: solo ASCII a 7 bit.
- Sequenza CRLF ogni 1000 caratteri o meno.

Questi limiti impediscono la trasmissione di documenti binari.

MIME (Multipurpose Internet Mail Extensions)

MIME ridefinisce il formato del corpo dei messaggi SMTP (definito in RFC 822) per permettere:

- Messaggi di testo in altri set di caratteri.
- Formati per messaggi non testuali.
- Messaggi multi-parte.
- Header con set di caratteri diversi da US-ASCII.

Funzionamento:

1. Messaggio non-SMTP trasformato in messaggi SMTP da un preprocessore.
2. All'arrivo, i messaggi SMTP vengono decodificati e riassemblati.

MIME su canali SMTP:

Ciò che viaggia su un canale SMTP è *sempre* un messaggio SMTP, con i suoi limiti. MIME aggira i limiti:

- **Codifica caratteri:** Il messaggio con caratteri non ASCII viene codificato per essere trasmesso in ASCII a 7 bit. Diverso encoding per testo e binari.
- **Sequenze CRLF:** Meccanismi per permettere CRLF nel flusso di dati, a volte inserendoli forzatamente.

- **Lunghezza Messaggi:** Un messaggio MIME può essere diviso in vari messaggi SMTP.

I servizi MIME

- **Dichiarazione di tipo:** `Content-Type` identifica il tipo di dati (es. `text/plain; charset=ISO-8859-1`). Aiuta il ricevente a scegliere l'applicazione giusta. *Non* si basa sull'estensione del file.
- **Messaggi multi-tipo:** Un messaggio MIME può contenere parti di tipo diverso (es. testo e allegato binario). Si creano sottomessaggi MIME per ciascuna parte, e il messaggio complessivo diventa "multi-parte".

Header specifici MIME

- **`Content-Type`:** Tipo MIME del contenuto (tipo, sottotipo, parametri).
- **`Content-Transfer-Encoding`:** Tipo di codifica per la trasmissione. Valori: `7bit` (default), `8bit`, `binary`, `quoted-printable`, `base64`, altri definiti in IANA.

MIME - Quoted Printable

- Per dati con molte parti US-ASCII e poche eccezioni (es. testi in lingue europee).
- Codifica solo i byte non conformi:
  - Codice > 127 o < 32: `"= " + codice esadecimale`. Es: "Hello’99" -> "Hello=B499".
  - Righe > 76 caratteri: interrotte con "soft breaks" (`=` alla fine della riga).

MIME - Base 64

- Per dati binari o multi-byte.
- Usa un sottoinsieme di 64 caratteri US-ASCII "sicuri":
  - Lettere maiuscole ('A' => 0).
  - Lettere minuscole ('a' => 26).
  - Numeri ('0' => 52).
  - '+' e '/' (62 e 63).
- **Codifica:**
  1. Flusso dati diviso in blocchi di 24 bit (3 byte).
  2. 24 bit suddivisi in 4 blocchi di 6 bit.
  3. Ogni blocco di 6 bit codificato in uno dei 64 caratteri.
- **Formattazione:**
  - Stringa risultante divisa in righe di 76 caratteri (tranne l'ultima) con CR-LF.
- **Decodifica:**
  - CR e LF sono ignorati.
  - Algoritmica, senza chiavi o calcoli complessi.
- **Base64 NON è crittografia!**

Conclusioni

In breve:

- Problemi di encoding e ordinamento di byte sono stati affrontati.
- Sono stati affrontati meccanismi di encoding, per poter affrontare questi problemi.