Il linguaggio HTML

Mario Bravetti

Linguaggio HTML

- Hyper Text Markup Language è un linguaggio per la marcatura di testi che deriva dal più generale SGML (Standard Generalized Markup Language).
- Definisce insieme di elementi (marcatori o tag):
  - strutturazione e definizioni diverse parti di una pagina web
  - corredarlo di immagini, suoni e animazioni (elem. multimediali)
- I marcatori interpretati da: programmi per visualizzazione dei contenuti di questo tipo di documenti (browser).

Linguaggio HTML

- Documento HTML file di testo contenente tag (struttura di marcatura) che determinano il rendering della pagina:
  - struttura del contenuto (in HTML5 enfasi su semantica!) della pagina ma non aspetti di presentazione definiti con CSS
  - eventuali elementi esterni da integrare con il testo scritto.
- Browser: interpreta i marcatori, reperisce tutti gli elementi necessari per completare i contenuti e presenta all'utente il risultato finale del processo.

Storia di HTML

- HTML in varie versioni dal 1989 ad oggi:
  - Prima versione effettivamente utilizzata è la 2.0 (1994), che venne implementata su Mosaic, il primo browser a larga diffusione da cui derivò Netscape.
  - Seconda versione importante: 3.2 (1997) che definiva tabelle, applet, script e altre migliorie, ma non i frame (già implementati dal 1995 da Netscape e Microsoft)
  - Prima di HTML5 l'ultima versione di HTML è stata per diversi anni la 4.01 (1999), una correzione della 4.0 (1997-1998) che ha introdotto alcune novità come il supporto all'internazionalizzazione, ai fogli di stile e ai frame.

Storia di HTML

- HTML in varie versioni dal 1989 ad oggi (cont.):
  - XHTML 1.0 che definisce HTML in formato di specifica XML (linguaggio di marcatura generale per lo scambio dei dati) nasce nel 2000 (in seguito XHTML 1.1) e viene considerato il futuro di HTML
  - Mentre il W3C stava lavorando alle specifiche del nuovo XHTML 2.0 si è formato il gruppo WHATWG (Web Hypertext Application Technology WG), Apple-Mozilla-Opera-Google, che contrastava lo sviluppo (non retrocompatibilità, non in linea con sviluppo web)
  - Dal 2007 il WHATWG ha collaborato con il W3C in alla creazione di HTML5, prima versione a fine 2014

Osservazioni

- Fino ad oggi: browser progettati senza preoccuparsi della necessità di verificare la correttezza sintattica o strutturale dei documenti HTML
  - un browser può visualizzare (talvolta con risultati imprevedibili) anche documenti HTML non corretti.
- Con XHTML (versione di HTML aderente alle specifiche XML) il browser rifiuta di visualizzare pagine non corrette
- HTML5: specifica di come visualizzare pagine errate in modo che sia uniforme sui tutti i browser

Specifiche HTML

- Le specifiche sono pubblicate sul sito del W3C:
  - www.w3.org/TR/html401
- Per HTML working draft :
  - www.w3.org/TR/html5
- Si può verificare se un documento è conforme alle specifiche usando strumenti di convalida:
  - validator.w3.org

Proliferazione dei tag

- Proliferazione dei tag tra le successive versioni.
- La versione 4.01 consente di utilizzarli praticamente tutti, ma molti deprecati (sconsigliati):
  - per produrre gli stessi effetti sono stati previsti meccanismi alternativi più corretti ed efficaci e
  - non è detto che nelle versioni future del linguaggio ne sia ancora garantito il supporto.
- HTML5 definisce più direttamente quali sono supportati e quali non lo sono più

Elementi di HTML

- Un documento HTML può essere formato dai seguenti elementi:
  - marcatori (o tag)
  - attributi
  - testo ordinario
  - commenti
  - riferimenti a entità

Marcatori (tag)

- Nel caso più semplice, la struttura/significato di una sezione di testo.
- Nome di comando dentro parentesi angolari, es.
  - <table>
- Nomi dei tag usabili (e dei relativi attributi) e loro significato specificati dallo standard W3C
  - non si può "inventare" tag (a differenza di XML)

Apertura e chiusura di un tag

- I tag appaiono in coppie apertura/chiusura:
  - text
- Serve per individuare la parte di testo (text sopra) a cui viene applicata la marcatura del tag
- Vi sono tag singoli (chiusura non necessaria):
  -
 (o <hr>, <img>)
- Inteso come
 </br> (no testo di riferimento)
- Meglio utilizzare l'abbreviazione di
 </br> :
  -
 (o <hr />, <img />)

Attributi

- Ogni tag può avere attributi che conferiscono alla parte di testo ulteriori proprietà di visualizzazione:
  - <img width="50px" height="20px"
    src="./images/garden.jpg“ />
- Per alcuni tag determinati attributi sono obbligatori (altrimenti non avrebbero significato):
  - <img />
- Normalmente usare doppi apici per scrivere il valore degli attributi (apici singoli sono consentiti)
  - HTML5 permette l'indicazione di attributi senza apici

Attributi

- Per ciascun elemento, numerosi attributi:
  - l'impostazione di dimensioni, caratteristiche di presentazione, ecc.
- Quando è possibile, impostare caratteristiche di presentazione del testo tramite gli stili (CSS), evitando il ricorso agli attributi per i singoli tag
  - comunque sono, in genere, consentiti

Particolarità

- Nomi dei tag e degli attributi **non case-sensitive:**
  - <head>, <HEAD>, <Head>, <hEaD> tutti equivalenti
- Errori in documenti HTML (browser li corregge), per esempio:
  - tag di apertura senza il corrispondente tag di chiusura
  - tag annidati scorrettamente:
    - <table><tr>…</table></tr>

Commenti

- Porzioni di testo inserite allo scopo di chiarire il significato di determinate parti del documento, che non vengono visualizzate dal browser
  - <!-- commento -->

Testo e codice HTML

- Per elementi di markup e testo da pubblicare non sono considerati: spazi bianchi multipli tra le parole, ritorni a capo, righe vuote o tabulazioni.
- Questo significa che:
  - usare spazi, ritorni a capo, tabulazioni, ecc.. serve solo ad aumentare la leggibilità del codice sorgente
  - non ha alcun effetto sulla visualizzazione da parte del browser

Testo ordinario

- Contenuto vero e proprio del documento (informazione che si vuole visualizzare).
- Comprende le parole, gli spazi e la punteggiatura che costituiscono il testo.
- Non devono essere utilizzati direttamente:
  - i caratteri speciali (€, ©, ®, ™, ecc.),
  - quelli "localizzati" (è, ù, ö, ñ, ecc.) e
  - i caratteri riservati di HTML (<, >, &);
- per inserire questi caratteri in una pagina Web ricorrere ai riferimenti a entità (vedi nel seguito).

Caratteri speciali

- L'insieme di caratteri ASCII non è sufficiente per un sistema informativo globale quale è il Web
- Si utilizzano quindi sistemi di codifica estesi per i caratteri (Unicode, UCS, UTF-8,…)
  - Contengono i caratteri di moltissime lingue (16 bit)
- Nella pratica UTF-8 molto usato, però la visualizzazione di caratteri speciali dipende dalla
  - configuraz. del browser (se non si indica in HTML)
  - supporto da parte del sistema operativo su cui gira
- Meglio non usarli nel file di testo HTML

Riferimenti a entità

- Parole chiave racchiuse tra i delimitatori "&" e ";" che causano la visualizzazione di un carattere
  - © → ©
  - ® → ®
  - " → "
  - & → &
  - < → <
  - > → >
  - …. (lettere accentate ecc…)

Riferimenti a entità

- Particolarmente utile può risultare il carattere di spazio vuoto:
  -    (non-breaking space)
- Spazio di non interruzione (di riga, cioè non va a capo)
- Serve per introdurre spazi multipli.

Struttura di un file HTML

- Caratterizzato da una struttura ad albero:
  - gerarchia di contenimento delle sezioni di marcatura
- L'elemento radice dell'albero definisce i limiti del documento. Al suo interno, divisone in due parti:
  - una intestazione (header)
  - un corpo (body)
- Prima di radice: dichiarazione tipo di documento (DTD) opzionale (non fa parte del ling. HTML).
- In generale un file HTML deve essere costruito rispettando questa struttura di base

Struttura di un file HTML

```
<!DOCTYPE HTML >
<HTML>
    <HEAD>
        <TITLE>Titolo del documento</TITLE>
    </HEAD>
    <BODY>
        <P>Testo di un paragrafo</P>
    </BODY>
</HTML>
```