Introduzione ad XML

Mario Arrigoni Neri

XML

- XML è contemporaneamente:
  - Linguaggio di annotazione (Markup) che permette di creare gruppi di marcatori (tag set) personalizzati
    (MathML, XHTML, chemicalML, ecc..)
  - Formato standard per lo scambio dei dati
  - Metalinguaggio per creare documenti arricchiti da informazioni addizionali
  - Un supporto per la costruzione di formati specifici per gli usi più disparati
- XML non è:
  - Un sostituto di HTML, le pagine web continueranno ad essere scritte in HTML. XML è un metalinguaggio, HTML è un linguaggio
  - Un linguaggio di programmazione: ogni documento XML contiene dati ed informazioni sui dati. Questi vengono poi estratti ed elaborati dalle varie applicazioni.

Il caso HTML

- HTML (HyperText Markup Language) nasce come DTD di **SGML** (Standard Generalized Markup Language) per la pubblicazione di semplici documenti testuali con qualche immagine e collegamento ipertestuale
- L'elemento fondamentale è il **tag**, testo racchiuso tra **'<'** e **'>'** che contiene informazioni circa il testo, costituisce quindi un meta-dato circa il dato vero e proprio che è nel testo
- Con il successo del Web HTML viene utilizzato per **scopi diversi** da quelli per cui era stato progettato
- Vengono implementate molte **estensioni proprietarie** che creano barriere all'interoperatività degli strumenti
- I parser (browser) **rilassano le regole sintattiche** ed interpretano anche documenti HTML "scorretti" (in maniera differente l'uno dall'altro)

Problemi con SGML

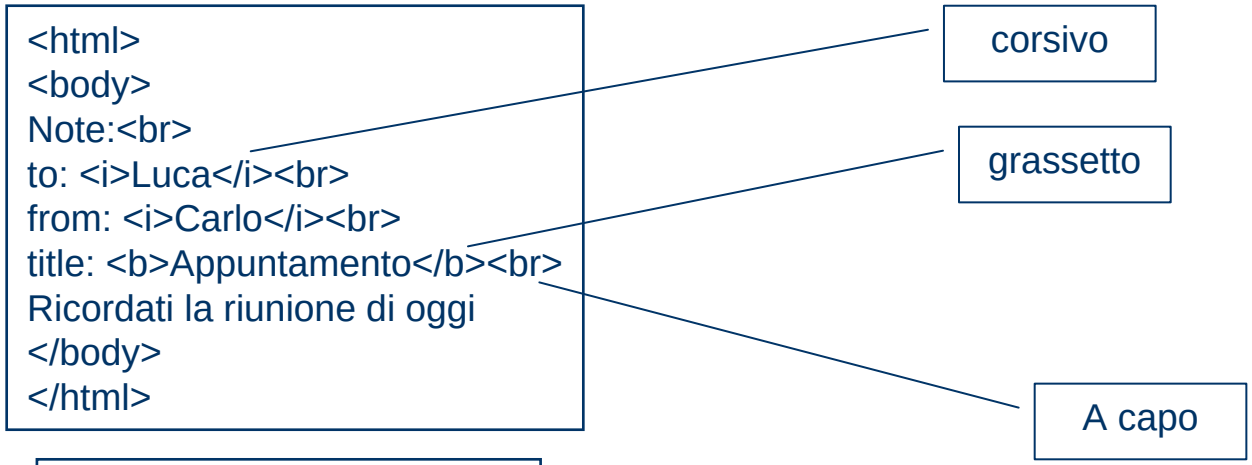
- Complesso da comprendere ed utilizzare
- Non è pensato per la rete: mancano link ipertestuali e specifiche grafiche
- Il successo di HTML ha fatto capire che:
  - Il mondo degli sviluppatori è pronto ad accogliere il modello basato sui TAG
  - La semplicità del linguaggio HTML è stato il suo principale punto di forza

Da HTML ad XML

- XML (eXtensible Markup Language) nasce dall'intento di applicare il paradigma dei tag in campi diversi dalla presentazione di ipertesti
- Si basa sul markup in modo simile ad HTML
- XML è pensato per descrivere dati
- I tag XML non sono predefiniti
- XML non è un linguaggio, ma un insieme di regole per costruire particolari linguaggi (metalinguaggio)

I tag in HTML

- I tag di HTML contengono informazioni per la visualizzazione dei dati



Note:
to: *Luca*
from: *Carlo*
title: **Appuntamento**
Ricordati la riunione di oggi

I tag in XML (1)

- In prima battuta, un documento XML è simile ad un HTML, in cui però possiamo “inventare” i tag

```
<?xml version="1.0"?>
<note>
        <to>Luca</to>
        <from>Carlo</from>
        <title>Appuntamento</title>
        <message>Ricordati la riunione di oggi</message>
</note>
```

- La scelta dei tag può essere effettuata a seconda delle informazioni che interessa rappresentare e che la specifica applicazione dovrà riconoscere

I tag in XML (2)

- La prima linea del documento (opzionale) identifica lo stesso come un XML ed indica anche la versione
- Il primo tag <*note*> identifica la radice del documento. "*questo documento è una nota*"
- I restanti tag specificano il contenuto della nota in termini di titolo, mittente, destinatario e messaggio
- L'ultimo tag conclude la descrizione della nota
- I tag si dividono in:
  - Tag di apertura: <nometag> es: <note>
  - Tag di chiusura: </nometag> es: </note>
  - Tag vuoti: <nometag/> es: <note/>

Elementi XML (1)

- Un elemento XML è tutto ciò che è compreso tra un tag di apertura (incluso) ed il corrispettivo tag di chiusura (incluso)

<NOMETAG> testo </NOMETAG>

- Tra i due tag si trova il contenuto dell'elemento, che può essere:
  - Element content: se il contenuto è costituito da altri elementi, ad esempio l'elemento <note>
  - Simple content: se il contenuto è un semplice testo. Es: <message>
  - Mixed content: se contiene testo inframezzato da altri elementi. Ad esempio <message> se si permettono tag di formattazione come <b>,<i>
  - Empty content: se il contenuto è vuoto, es: il tag <img> dell'HTML
- Per un tag vuoto la coppia apertura/chiusura possono essere sostituiti da un tag vuoto

Elementi XML (2)

- Gli elementi in XML sono estendibili
- In questo modo è possibile mantenere compatibilità con versioni precedenti del software (backward compatibility)
- Es:

```
<?xml version="1.0"?>
<note>
        <to>Luca</to>
        <from>Carlo</from>
        <title>Appuntamento</title>
        <message>Ricordati la
riunione di oggi</message>
</note>
```

```
<?xml version="1.0"?>
<note>
        <to>Luca</to>
        <from>Carlo</from>
        <title>Appuntamento</title>
        <message>Ricordati la
riunione di oggi</message>
        <date>2003-01-10</date>
</note>
```

Elementi XML (3)

- Gli elementi in XML sono in relazione tra di loro e queste relazioni determinano il modello del documento
- Il documento è organizzato come un albero, in cui la relazione di contenimento tra tag è equivalente alla relazione nodo-sottonodo
- Es: <to>, <from>, <title> e <message> sono sottoelementi di <note>

```
<?xml version="1.0"?>
<note>
        <to>Luca</to>
        <from>Carlo</from>
        <title>Appuntamento</title>
        <message>Ricordati la riunione di oggi</message>
</note>
```

- L'entità che non è sottoentità di nessuno (es: <note>) è l'entità radice

Attributi XML

- Gli attributi sono informazioni aggiuntive che possono essere inserite negli elementi XML per completarne o arricchirne l'informazione, in maniera simile a quanto accade in HTML

HTML	XML
<img src="logo.gif" > <a href="login.jsp">	<message language="IT">Ricordati la riunione di oggi</message>

- Vengono inseriti solo nei tag di apertura (o nei tag vuoti)
- Possono essere racchiusi sia tra apici singoli che doppi

Attributi o elementi ? (1)

- Spesso le stesse informazioni possono essere rappresentate sia tramite attributi che tramite (sotto)elementi. Es:

Sottoelementi	Attributi
<note> <to>Luca</to> <from>Carlo</from> <title>Appuntamento</title> <message>…</message> </note>	<note title="Appuntamento"> <to>Luca</to> <from>Carlo</from> <message>…</message> </note>

Attributi o elementi ? (2)

- La scelta tra attributi o elementi è soggettiva, tuttavia le due soluzioni non sono in genere equivalenti.
- Problemi con gli attributi:
  - Non possono contenere **valori multipli**
    <parent name="Luca"><child>Marco</child> <child>Mario</child></parent>
  - Sono difficilmente **espandibili** (aggiunta di sottoelementi)
  - Non possono descrivere **strutture**
    <book><author><name>..</name><surname>..</surname></author></book>
  - Non hanno un supporto standard per la gestione nei programmi
  - Sono difficili da controllare rispetto ad un formato di documento DTD

- E' opportuno usare gli attributi per informazioni essenziali per l'elemento, come ad esempio gli identificativi (ID)

Regole sintattiche (1)

- Tutti i tag devono essere chiusi

HTML	XML
paragrafo1 paragrafo2	paragrafo1 paragrafo2

- I tag devono essere correttamente innestati

HTML	XML
corsivo e grassetto	corsivo e grassetto

Regole sintattiche (2)

- Ogni documento XML deve avere uno ed un solo elemento radice

Corretto	Scorretti
<note> <to>Luca</to> <from>Carlo</from> <title>Appuntamento</title> <message>…</message> </note>	<to>Luca</to> <from>Carlo</from> <title>Appuntamento</title> <message>…</message>
	<note>…</note> <note>..</note>

Regole sintattiche (3)

- Gli attributi devono sempre essere inclusi tra apici singoli o doppi

Corretto	Scorretto
<note date="12/11/2002">	<note date=12/11/2002>

- XML è case sensitive

Corretto	Scorretto
<to>Luca</to>	<to>Luca</To>

- Differentemente da quanto accade in HTML, in XML gli spazi vengono preservati
- I commenti possono essere inseriti tra i segni “<!- -” e “- ->”

<! – - questo è un commento XML - - >

CDATA

- E’ possibile introdurre del testo in modo che questo non venga elaborato dal parser XML, ma venga semplicemente restituito all’utente

```
<![CDATA [
Questo testo non viene elaborato e <questo> non è un tag
]]>
```

- Ciò è utile per evitare errori di parsing anche quando il “contenuto” potrebbe essere interpretato come codice XML

XML ed applicazioni XML

- Dato che XML è un (meta)linguaggio per specificare altri linguaggi costituisce un livello comune per il dialogo in ambienti differenti
- XML non dice nulla su che tag utilizzare, ma fissa solo delle regole comuni per eseguire correttamente il parsing del file
- E' possibile usare XML per gli scopi più disparati, a seconda delle operazioni che verranno eseguite dalla specifica applicazione di fronte agli specifici tag. Es: XHTML