5.2. Il linguaggio XML

Insegnamento di Informatica

Elisabetta Ronchieri

Corso di Laurea di Economia, Universitá di Ferrara

I semestre, anno 2014-2015

Argomenti

Cosa si intende per XML

Struttura di documenti SGML e XML

DTD

URI

Argomenti

Cosa si intende per XML

Struttura di documenti SGML e XML

DTD

URI

Cenni di SGML

- SGML (Standard Generalized Markup Language) é uno standard formale, approvato dall'ISO[1] nel 1986[2], ampiamente usato nel campo della editoria.
- SGML inizialmente doveva servire a definire documenti strutturati nel campo legale.
- Puó operare in differenti modi su diffierenti piattaforme o sistemi operativi, garantendo una certa flessibiliá.
- Non é un linguaggio di marcatura, ma un linguaggio con cui é possibile definire linguaggi di marcatura.
- SGML non sa cosa é un paragrafo, una lista, un titolo, ma fornisce una grammatica che ci permette di definirli.
- É all'origine della definizione di XML.

[1] ISO sta per International Standards Organization

[2] ISO 8879, Information Processing-Text and Office Systems-Standard Generalized Markup Language

Cosa si intende per XML

- XML (eXstensible Markup Language) venne introdotto nel 1996.
- Nel 1998 XML é diventato una raccomandazione (ossia standard) per il World Wide Web Consortium (W3C).
- Tra i suoi obiettivi troviamo:
  - deve essere usato in modo semplice su internet;
  - deve supportare una gran numero di applicazioni;
  - deve essere compatibile con SGML.

Cosa si intende pre XML

- XML assicura che i dati strutturati siano uniformi e indipendenti dalle applicazioni.
- Le sue caratteristiche fanno sí che si presti ad una larga varietá di applicazioni (come per il commercio elettronico o il network management).
- XML definisce in modo non ambiguo la struttura dei dati contenuti in un documento.

Caratteristiche di SGML e XML

- Sono linguaggi di marcatura dichiarativa e non procedurale:
  - le marche o etichette rappresentano solo dei nomi che permettono di identificare le parti logiche del documento;
  - le istruzioni su come processare il documento sono distinte e fisicamente separate dalle etichette;
  - le istruzioni sono in genere integrate in procedure o programmi che processano documenti con una determinata marcatura.

Caratteristiche di SGML e XML

- Consentono di associare ad una risorsa dei metadati (descrittivi, strutturali e amministrativi o gestionali).
- Sono linguaggi di marcatura referenziale.
- Hanno come obiettivo quello di definire metodi di rappresentazione di testi in forma elettronica, indipendenti dalle caratteristiche dei dispositivi e dei sistemi utilizzati.

Caratteristiche di SGML e XML

- Sono anche dei metalinguaggi, offrendo dei meccanismi per definire in modo formale un linguaggio di marcatura, attraverso un determinato tipo di grammatica, che ne controlla il vocabolario.
- Il meccanismo principale é la definizione di un tipo di documento, detto Document Type Definition (DTD) che permette di imporre che determinati documenti abbiano un determinato tipo, ovvero una determinata struttura.
- Un analizzatore sintattico puó sfruttare il DTD per verificare che la struttura del documento sia corretta, ovvero sia consistente con il linguaggio di marcatura definito con il DTD.
- I linguaggi di marcatura derivati da SGML, come XML, sono detti applicazioni di SGML.

Caratteristiche di SGML e XML

- Applicativi diversi possono usare l'informazione codificata in SGML/XML in modo diverso, a seconda delle esigenze.

  Esempi:

  - un applicativo per la formattazione puó scegliere di associare dei particolari caratteri di stampa a determinati elementi;
  - un applicativo di ricerca puó migliorare la precisione cercando solo elementi di un certo tipo, o all'interno di un particolare contesto.

Argomenti

Cosa si intende per XML

Struttura di documenti SGML e XML

DTD

URI

Struttura di documenti SGML e XML

- Un documento SGML e XML si compone di due parti:
  1. un prologo a sua volta composto da:
     - insieme di dichiarazioni SGML e XML;
     - un Document Type Definition (DTD).
  2. un'istanza di documento.

Dichiarazioni SGML e XML

- SGML definisce dei valori di default per un insieme di proprietá, quali il set di caratteri, i codici per i delimitatori di SGML, la lunghezza massima degli identificatori.
- Uno specifico documento puó modificare tali valori attraverso opportune dichiarazioni nel prologo, specificando cosí di fatto il dialetto di SGML usato nel documento.
- XML non lascia la stessa flessibilitá di SGML, e fissa a priori determinati aspetti dei documenti.
- La parte di dichiarazioni di un documento XML é opzionale.

Dichiarazioni SGML e XML

- La parte di dichiarazioni di un documento XML serve a dichiarare:
  - la versione di XML alla quale il documento si riferisce, che é la 1.0;
  - opzionalmente, il set di caratteri usato nel documento.
    - Tutti i processori di documenti XML devono essere in grado di riconoscere almento il set di caratteri UTF-8 (di cui i caratteri ASCII sono un sottoinsieme) o UTF-16 (tipo di codifica per Unicode).
  - opzionalmente, se il documento contiene una parte di dichiarazioni esterna al documento stesso, o se l'unica parte di dichiarazioni é quella contenuta nel documento stesso.

Esempio:

```
<?xml version="1.0" encoding="UTF-8" standlone="yes"
?>
```

Elementi SGML e XML

- Un elemento é un'entiá testuale considerata come una componente strutturale.
- Ogni elemento ha un nome.
- Tipi diversi di elementi hanno nome diversi:
  - SGML e XML non forniscono alcun modo per dichiarare il significato di un particolare tipo di elemento, se non la sua relazione con gli altri elementi;
  - la semantica degli elementi testuali dipende dall'applicazione.
- Un elemento é costituito da una coppia di etichette che si corrispondono rispetto all'annidamento di etichette:

  `<nome-elemento>` etichetta iniziale
  `</nome-elemento>` etichetta finale
- La parte compresa tra etichetta iniziale e quella finale viene detta contenuto dell'elemento.

Esempio di elemento

Supponendo che stiamo marcando una porzione di un testo per indicare che forma un capitolo.

- Possiamo usare un elemento di nome capitolo il cui contenuto é il testo del capitolo:

  `<capitolo> testo del capitolo </capitolo>`

Attributi SGML e XML

- Ogni elemento puó avere degli attributi che servono a specificare delle caratteristiche proprie dell'elemento che non corrispondono direttamente al testo che costituisce il documento.
- Un attributo ha un nome o un valore.
- I valori degli attributi di un elemento sono specificati tramite coppie contenute nell'etichetta iniziale dell'elemento.

  `nome_attributo=valore`

  Esempio:
  - L'elemento capitolo puó avere un attributo numero che specifica il numero d'ordine del capitolo che stiamo marcando:

    ```
    <capitolo numero="5"> testo del capitolo
    </capitolo>
    ```

Entitá SGML e XML

- Sono il meccanismo di marcatura referenziale proprio di SGML e XML.
- Ogni entitá rappresenta un'abbreviazione di una sequenza di caratteri.
- Ogni entitá é una parte di documento a cui si é dato un nome.
- Il riferimento ad un'entiá puó essere usato nell'istanza di documento al posto della sequenza di caratteri che l'entitá rappresenta.

Entitá SGML e XML

- Le entitá a cui ci si puó referire all'interno di un documento possono essere:
  - definite nel DTD di un documento:
    - le entiá esterne rappresentano una parte di documento memorizzata su un file esterno per modularizzare il documento;
  - predefinite:
    - il meccanismo usato per specificare caratteri riservati o non standard e per ovviare alle differenze di codifica su sistemi diversi.
- Il riferimento ad un'entitá avviene inserendo il nome dell'entitá preceduto dal carattere "&" e seguito dal carattere ";".

Esempi di entitá SGML

- Entitá:
  - Si puó definire *inc* che rappresenta la stringa:

    ```
    <Sezione> Ancora da scrivere.
    </Sezione>
    ```
  - Ogni volta che si vorrá indicare nel documento la presenza di una sezione ancora da scrivre, si potrá usare il riferimento:

    ```
    &inc;
    ```
- Entitá predefinite:
  - in XML, i caratteri "&", "<", ">", possono essere inclusi nel testo libero attraverso riferimenti alle entitá predefinite *amp*, *lt*, *gt*.

Istanza di documento

- L'istanza di documento, che segue il DTD in un documento SGML o XML, contiene testo libero, etichette e riferimenti ad entitá.
- É un unico elemento che viene detto elemento radice, tipicamente composto da altri elementi.

  Esempio:

```
<TechRepDip nome_dip="CNAF">
    <Intestazione> ... </Intestazione>
    <Sezione>
        <Sottosezione> ... </Sottosezione>
        <Sottosezione> ... </Sottosezione>
    </Sezione>
    &inc;
</TechRepDip>
```

Argomenti

Cosa si intende per XML

Struttura di documenti SGML e XML

DTD

URI

Document Type Definition (DTD)

- Serve a specificare quali sono le strutture ammesse per l'istanza di documento che segue il prologo.
- Definisce il nome e la struttura degli elementi, gli attributi e le entitá che possono essere usati da un'intera classe di documenti marcati.
- Da un punto di vista astratto puó essere considerato una grammatica che genera documenti marcati con delle etichette.
- DTD definisce uno specifico linguaggio di marcatura.
- DTD non dice nulla sulla semantica della marcatura, sulla rappresentazione di un documento:
  - l'applicazione SGML e XML necessita di una componente aggiuntiva alla DTD, ossia al foglio di stile.

Struttura di un DTD

- É costituito da:
  1. una dichiarazione del tipo di documento che corrisponde alla specifica del simbolo iniziale della grammatica;
  2. un insieme di dichiarazioni di tipi di elemento:
     - ogni tipo di elemento corrisponde ad un non termine della grammatica;
     - ogni dichiarazione di un tipo di elemento corrisponde ad una produzione della grammatica;
  3. un insieme di dichiarazioni di attributi:
     - ogni attributo é associato ad un determinato tipo di elmento;
  4. un insieme di dichiarazioni di entitá.

Struttura di un DTD

- La dichiarazione del tipo di documento é sempre inclusa nel documento stesso.
- Le altre dichiarazioni del DTD sono in generale distribuite in due parti:
  1. dichiarazioni interne, incluse direttamente nel documento;
  2. dichiarazioni esterne, specificate in un file esterno che viene invocato per riferimento.

DTD: caso generale

- Nel caso piú generale in cui il DTD contiene sia dichiarazioni interne che esterne, il DTD ha la seguente forma:

```
<!DOCTYPE tipo-doc
SYSTEM "file-DTD-subset-esterno" [
<!-- inizio DTD subset interno -->
...
<!-- fine DTD subset interno -->
]>
```

- tipo-doc é un nome che indica il tipo di documento;
- file-DTD-subset-esterno é il nome del file esterno che contiene l'external DTD subset;

DTD: caso generale

- Nel caso piú generale in cui il DTD contiene sia dichiarazioni interne che esterne, il DTD ha la seguente forma:

```
<!DOCTYPE tipo-doc SYSTEM "file-DTD-subset-esterno" [
<!-- inizio DTD subset interno -->
...
<!-- fine DTD subset interno -->
]>
```

- la parola chiave SYSTEM serve a specificare che l'external DTD subset si trova in un oggetto di sistema che é appunto il file esterno;
- le parti racchiuse tra "<!- -" e "- ->" rappresentano commenti.

In alcuni casi la parola chiave SYSTEM seguita dal nome del file puó essere sostituita dalla parola chiave PUBLIC seguita da un identificatore pubblico e poi dal nome del file.

DTD con sole dichiarazioni esterne

- Nel caso in cui il DTD non abbia dichiarazini interne, la parte tra parentesi quadre rimane vuota.

  Esempio:
  - Consideriamo il DTD dei rapporti tecnici di un dipartimento
  - Il DTD completo potrebbe essere specificato nel file esterno TechRepDip.dtd
  - ogni rapporto tecnico deve contenere la seguente dichiarazione di tipo di documento

    ```
    <!DOCTYPE TechRepDip SYSTEM "TEchRepDip.dtd." []>
    ```

DTD con sole dichiarazioni interne

- Nel caso in cui il DTD sia specificato completamente nel documento stesso devono essere omessi la parola SYSTEM ed il nome del file esterno.

  Esempio:

  ```
  <!DOCTYPE docmio [...]>
  ```

Struttura di un DTD

- Un DTD viene condiviso da un gran numero di documenti tutti di un certo tipo.
- Il caso comune prevede un DTD con un file esterno contenente:
  - la dichiarazione di tutti gli elementi;
  - le entitá;
  - gli attributi

  che sono in comune ai documenti di un certo tipo.
- Le dichiaraioni del DTD interne al documento stesso, se presenti, riguardano solo la dichiarazione di entitá specifiche per il documento.

Uso del DTD negli analizzatori sintattici

- SGML e XML si differenziano per quanto riguarda la necessitá di specificare un DTD per un documento.
- SGML richiede che ogni documento abbia un DTD associato in un documento.
- XML puó non avere il DTD.
- Per XML s considerano due tipi di analizzatori sintattici:
  - analizzatori sintattici non validanti, che verificano se un documento é ben formato;
  - analizzatori sintattici validanti, che devono verificare se un documento é valido.

Documento XML ben formato

- Un documento XML é ben formato se é conforme alla grammatica che definisce XML.
- Le etichette iniziale e finale devono corrispondere ed essere innestate correttamente.

```
<x><y> ... </y></x> Ben Formato
<x><y> ... </x></y> Non Ben Formato
<x><y> ... </x> Non Ben Formato
```

- I valori degli attributi devono essere racchiusi tra virgolette

```
<x id="100"> ... </x> Ben Formato
<x id=100> ... </x> Non Ben Formato
<x id="100> ... </x> Non Ben Formato
```

- Deve esservi un unico elemento radice.
- Un analizzatore non validante deve effettuare un'analisi del documento rispetto alla grammatica XML e puó ignorare le dichiarazioni di elementi nel DTD.

Documento XML valido

- Un documento XML é valido se
  - é ben formato;
  - include un DTD;
  - l'istanza di documento é conforme al DTD e i valori degli attributi degli elementi rispettano il tipo dichiarato nel DTD.

Esempio di DTD XML

```
<!DOCTYPE TechRepDip [
<!ELEMENT TechRepDip (Intestazione, Sezione+,
    Bibliografia?)>
<!ELEMENT Intestazione (numero, Data, Titolo,
    Autore+, Sommario?)>
<!ELEMENT Sezione (TitoloSezione, Testo?, Sezione*)>
...
<!ATTLIST Sezione id ID #REQUIRED
num NMTOKEN #IMPLIED
stato (finale | provvisorio) "finale">
]>
```

Dichiarazione di elemento XML

- In XML, una dichiarazione di elemento ha la seguente forma:

  `<!ELEMENT elemento-dichiarato modello-contenuto>`
- In XML mancano le parti regole-di-minimizzazione ed eccezioni (presenti nell'elemento SGML), e non é possibile dichiarare piú elementi in una sola dichiarazione.
- In XML sia l'etichetta iniziale che quella finale di un elemento sono sempre obbligatorie.

Specifica del contenuto di un elemento

- Il contenuto puó essere specificato in termini di altri elementi, ovvero di un modello di contenuto o attraverso parole riservate.
- La piú comune delle parole riservate é PCDATA

Argomenti

Cosa si intende per XML

Struttura di documenti SGML e XML

DTD

URI

URI

- URI sta per Uniform Resource Identifier.
- É una stringa di caratteri che identifica univocamente una risorsa (locale, remota o di una rete quale quella di Internet) e che ha la seguente sintassi:

  `<schema>:<dettagli specifici dello schema>`
- La sintassi e la semantica della parte specifica dello schema dipendono dallo schema stesso.
- Le URI possono essere classificate in:
  - Universal Resource Locator (URL) che definiscono il metodo per trovare una risorsa;
  - Universal Resource Name (URN) che definiscono l'identitá della risorsa e non dipendono dall'indirizzo fisico della risorsa.

URL

- Le risorse su Internet sono identificate dal loro indirizzo URL che indica come localizzarle.
- Sono fragili a modifiche del meccanismo di accesso come il cambio del nome di una directory.
- Il nome dello schema:
  - puó coincidere con il protocollo che serve per accedere alla risorsa, quali http, ftp, mailto.
  - puó non essere associato a nessun protocollo quale file.

Esempio di URL

- Le risorse sul Web sono identificate dalla seguente forma di URL:
  `http://<hostname>:`
  `<porta>/<path-risorsa>?<query>#<frammento>`
  - hostname é il nome o indirizzo IP del server che ospita la risorsa;
  - porta é dove il server é in ascolto (é opzionale essendo la porta associata di default al protocollo http pari a 80);
  - path-risorsa é il percorso della risorsa all'interno del server;
  - query é la stringa passata alla risorsa per ulteriore interpretazione (opzionale);
  - frammento identifica una parte della risorsa stessa (opzionale)

Esempio di localizzazione di una risorsa a partire dall'URL

- La risorsa cercata é:
  `https://www.cnaf.infn.it/~elironc/index.html`
  - é accessibile tramite protocollo https
  - si trova su un certo server web il cui nome logico é *www.cnaf.infn.it* mappato sugli effettivi indirizzi IP;
  - si trova in una certa posizione nel file system del calcolatore server, identificata da un certo path:
    `~elironc/index.html`

URN

- Lo schema é per definizione un URN.
- Deve essere trasformato in un apposito servizio, nello URL attualmente associato alla risorsa sempre che esista.
- La mappa delle corrispondenze URN-URL deve essere aggiornata ogni volta che la risorsa viene spostata.

URL verso URN

- Il sistema ISBN é il tipico esempio dell'uso degli URN

  Esempio:
  `urn:isbn:0-486-27557-4`
  - identifica una specifica edizione dell'opera di Shakespeare "Romeo and Juliet"
- Per accedere alla risorsa e leggere l'opera si ha bisogno di conoscere l'indirizzo URL di un file.

  Esempio:
  `file:///home/username/books`
  - identifica univocamente il libro elettronico nella macchina locale.