Regole di base del linguaggio XML

Il linguaggio **XML** (*eXtensible Markup Language*) è utilizzato per rappresentare, trasmettere e memorizzare dati strutturati.
Un **dato strutturato**, chiamato *record*, può essere visto come un insieme di dati semplici organizzati secondo una struttura prefissata. Per esempio, un numero telefonico è un dato semplice, mentre un'anagrafica con cognome, nome, comune, provincia, è un dato strutturato.
Un elenco di dati aventi la stessa struttura, come per esempio una rubrica telefonica, forma un *archivio*. È importante sottolineare che una buona struttura, oltre che tenere in ordine i dati, facilita la ricerca e, soprattutto, fornisce un aspetto **semantico** ai dati, cioè aggiunge ai dati un significato.
XML stabilisce un insieme di semplici regole che permettono di rappresentare e trasmettere facilmente dati strutturati. In questo differisce dal linguaggio HTML che è un linguaggio di formazione delle pagine grafiche visualizzabili da un browser. Tuttavia, come HTML, non è un linguaggio di programmazione, cioè non viene utilizzato per creare programmi o script.
Le regole del linguaggio XML sono fissate, e aggiornate nel tempo, dal **W3C** (*World Wide Web Consortium*), l'ente che regola le tecnologie standard del Web (*www.w3.org* oppure *www.w3c.it* per la versione in lingua italiana).

XML fa uso di **tag** racchiusi dai simboli **<** e **>**, in analogia a HTML. Mentre HTML trasmette i dati insieme alle informazioni contenute nei fogli di stile per visualizzarli (il font, la grandezza, lo sfondo, il layout della pagina), XML trasmette i dati lasciando l'interpretazione degli stessi all'applicazione destinataria.
I tag di XML descrivono i dati e i loro attributi, organizzandoli in un normale file di testo.
Si osservi che questa caratteristica rende la rappresentazione dei dati indipendente dal dispositivo e dal sistema operativo utilizzati dall'utente finale.

Le seguenti sono alcune tra le regole di base di XML:

- in XML è necessario chiudere tutti i tag, con l'uso della barra, ad ogni tag deve corrispondere un **tag di chiusura**:

```
<nome>Mario</nome>
```

- l'XML è **case sensitive** per cui il tag *<contatto>* è diverso da *<Contatto>*, il quale a sua volta è diverso da *<CONTATTO>*

- i tag XML devono essere sempre **annidati** correttamente:

```
<lista>
<nome>Giorgio</nome>
</lista>
```

- tutti gli **attributi** vanno inseriti tra virgolette:

```
<lista genere="F">
<nome>Laura</nome>
</lista>
```

- nel documento XML possono essere inserite anche **righe di commento** per spiegare il significato dell'intero documento o di parti di esso; i commenti sono delimitati dai simboli <!–– e ––>:

```
<!-- informazioni sullo studente -->
```

La prima riga del documento XML ha il seguente formato:

```
<?xml version="1.0"?>
```

L'attributo **version** dichiara la versione del linguaggio XML utilizzata:

```
version="1.0"
```

Questo attributo è obbligatorio.

Si può aggiungere anche l'attributo **encoding** che specifica il codice utilizzato nel documento XML per rappresentare i caratteri. La codifica si riferisce in generale al codice **Unicode**: in particolare può essere **UTF-8** o **UTF-16** (*Unicode Transformation Format*) a 8 o 16 bit, oppure **ISO-8859-1** per i caratteri *ISO-8859 Latino 1*.
L'attributo *encoding* è facoltativo; se la codifica è UTF-8 oppure UTF-16, può essere omesso. Spesso però viene comunque inserito per una documentazione più precisa del file XML.

Esempi di dichiarazioni iniziali del documento XML sono quindi:

```
<?xml version="1.0"?>

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

<?xml version="1.0" encoding="ISO-8859-1"?>
```

Per illustrare l'uso pratico del linguaggio XML useremo come riferimento un semplice esempio che risolve il problema di organizzare una raccolta di siti Web preferiti (*bookmark*), in modo da poterli poi ritrovare e utilizzare facilmente.
Ogni record è composto da un nome (per esempio *Google*), un URL (*http://www.google.it*), una breve descrizione e una categoria di appartenenza (*motori di ricerca*).

Tutti questi dati vengono inseriti in un file XML: esso è un file di testo che viene salvato su disco con l'estensione **.xml**.

Il documento XML può essere scritto con un qualsiasi editor di testi, per esempio **Blocco note**. Si ricordi che, in *Blocco note*, per salvare il file con l'estensione XML, nella finestra *Salva con nome* occorre scegliere **Tutti i file (*.*)** nella casella **Salva come**, per impedire che il testo venga salvato con l'estensione *.txt*.

Inoltre, per utilizzare le lettere accentate, occorre salvare il testo del file XML con la codifica **Unicode**: selezionare la codifica con la casella combinata nella parte inferiore della finestra *Salva con nome*.

(siti.xml)

```
<?xml version="1.0" encoding="UTF-8"?>
<lista>
<sito>
<nome>Google</nome>
<URL>http://www.google.it</URL>
<descrizione>Il piu' famoso motore di ricerca</descrizione>
<categoria>motori di ricerca</categoria>
</sito>
</lista>
```

Il tag *<lista>* più esterno rappresenta l'intero archivio, il tag *<sito>* rappresenta il singolo record e gli altri tag più interni sono le informazioni relative a ciascun sito. In questo modo si crea una struttura gerarchica di tag, che risultano annidati uno all'interno del precedente.

È possibile definire anche gli **attributi** di un elemento.
Nell'esempio si potrebbe definire la *categoria* come attributo di *sito*.

```
<?xml version="1.0" encoding="UTF-8"?>
<lista>
<sito categoria="motori di ricerca">
<nome>Google</nome>
<URL>http://www.google.it</URL>
<descrizione>Il piu' famoso motore di ricerca</descrizione>
</sito>
</lista>
```

Non ci sono regole predefinite per decidere quali dati è opportuno mantenere a sè stanti e quali definire come attributi: la scelta può variare a seconda del tipo di dato trattato.
In generale gli attributi vengono utilizzati per i dati che possono identificare univocamente un record (come accade per le *chiavi* in un database).
Per esempio:

```
<?xml version="1.0" encoding="UTF-8"?>
<lista>
<sito IDsito="s001">
<nome>Google</nome>
<URL>http://www.google.it</URL>
<descrizione>Il piu' famoso motore di ricerca</descrizione>
<categoria>motori di ricerca</categoria>
</sito>
<sito IDsito="s002">
<nome>La Repubblica</nome>
<URL>http://www.repubblica.it</URL>
<descrizione>Versione on-line del quotidiano</descrizione>
<categoria>news</categoria>
</sito>
</lista>
```