XML
UN METALINGUAGGIO PER STRUTTURARE LA CONOSCENZA

Il computer non è una macchina intelligente
che aiuta le persone stupide, anzi è una
macchina stupida che funziona solo nelle
mani delle persone intelligenti.
Umberto Eco

INTRODUZIONE

Il linguaggio marcatore XML (e**X**tensible **M**arkup **L**anguage), la cui prima bozza di lavoro risale al 1996 e il cui protocollo in versione 1.0 è stato ufficializzato nel febbraio 1998 e la cui versione 1.1 è in elaborazione per diventare ufficiale nel corso dell'anno 2007, è il linguaggio più recente ed innovativo nato per il Web.

XML risponde all'esigenza di disporre un linguaggio che permetta lo sviluppo di applicazioni distribuite in rete, mantenendo una compatibilità completa con HTML.

Studiato da un gruppo di lavoro composto da esperti di reti (Internet e Intranet), esperti di editoria e comunicazione ed esperti di standardizzazione di linguaggi marcatori presso il W3C (***W**orld **W**ide **W**eb **C**onsortium*), l'XML, come l'HTML, deriva la struttura dall'**SGML** (**S**tandard **G**eneralized **M**arkup **L**anguage) lo standard internazionale di linguaggio marcatore strutturato.

È netta la distinzione architetturale tra HTML e XML: mentre il primo specifica al proprio interno struttura e rappresentazione dei documenti senza una separazione logica delle due parti, in XML il documento viene descritto attraverso una definizione rigorosa della struttura del contenuto.

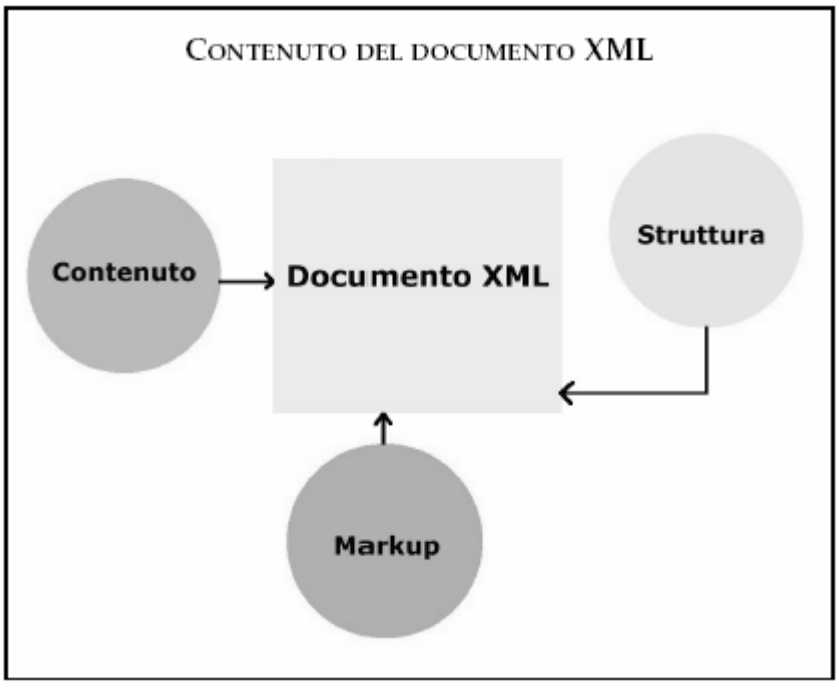

Figura 1 Componenti di un documento XML

L'obiettivo principale che orienta le linee di sviluppo e di applicazione di XML è la gestione di una tipologia di documentazione che può rapportarsi direttamente sia ai diversi formati di input e output, sia ai diversi formati di archiviazione strutturata delle informazioni.

Il documento XML è rappresentabile, così come è rappresentato graficamente in figura 1, con una equazione diversa dal documento HTML o descritto con altri linguaggi marcatori:

DOCUMENTO XML = CONTENUTO + STRUTTURA DELLE INFORMAZIONI + STILE DI RAPPRESENTAZIONE

Per meglio contestualizzare le ragioni di un approcciosistematico allo studio del linguaggio XML e delle sue possibili applicazioni citiamo e analizziamo l'incipt del documento di presentazione del gruppo di lavoro del W3C che si è occupato di XML[1] che così presenta il linguaggio: "*e**X**tensible **M**arkup **L**anguage (XML) è una formato testuale semplice e molto flessibile derivato da SGML (ISO 8879). Progettato in origine per venire incontro alla sfida ha venire a di pubblicazione elettroniche su vasta scala, XML sta svolgendo un ruolo sempre più importante nello scambio di un'ampia tipologia di dati sul WEB ed altrove*".

Breve storia dell'XML

XML tra linguaggio e metalinguaggio

CONTENUTO DEL DOCUMENTO XML

Prime note e definizioni

I documenti XML, come i documenti HTML, possono contenere informazioni relative a dati codificati ed espressi in diverse modalità come testi, audio, immagini o filmati; i supporti di comunicazione che non trovano una codifica diretta nel linguaggio XML vengono inclusi, utilizzati, distribuiti e visionati nel loro formato originale.

In un linguaggio marcatore la definizione del set di caratteri e la codifica che si intende utilizzare nella loro rappresentazione digitale è essenziale: il linguaggio XML utilizza il set di caratteri UNICODE, che comprende diverse migliaia di caratteri appartenenti alle lingue di tutto il mondo, ed è corredato di diverse codifiche; in particolare la codifica UNICODE UTF-8 comporta la compatibilità tra i primi 128 caratteri e il codice ASCII. Questa compatibilità di codifica permette l'utilizzo dei normali editor di testi per la creazione di documenti XML. L'unità di informazione, in XML, si definisce ***entità***.

L'entità identifica e rappresenta una sezione di documento; può essere correlata a un nome, attraverso il quale viene inserita e utilizzata in un qualsiasi punto del documento XML, e a ciascuna entità è associata una notazione che permette l'identificazione del tipo (ad esempio GIF o JPEG per le immagini in tali formati).

L'aggiornamento di una entità esterna ad un documento che la richiama, l'entità esterna può essere un file, una parte di file o un'informazione di archivio, determina l'aggiornamento automatico del documento.

Le entità che compongono un documento XML possono essere memorizzate ovunque, sia sul computer locale che in differenti punti di una rete locale o della rete Internet, per questo la loro ricerca per l'inserimento nel documento, durante le fasi di parsing[2] e processing[3], viene indirizzata dalle informazioni formalizzate con la notazione XML.

Struttura delle informazioni nel documento XML

Il fondamento della struttura del documento XML è il concetto di elemento: l'elemento è l'unità di contenuto di un documento, caratterizzato da un **tag iniziale** e un **tag finale** che lo racchiudono.

Ad esempio, considerando un testo nell'ottica del documento XML, è possibile osservarne la struttura ad albero suddivisa progressivamente in capitoli, composti da titolo e paragrafi, a loro volta divisi in sottotitolo e capoversi, questi ultimi composti da testo, eventuale immagine ed eventuali note.

La struttura ad albero che descrive il testo avrò la seguente struttura:

Testo
- **Titolo del testo: «XML e applicazioni multimediali»**
- **Autore del testo: …**
- **Capitolo**

[1] Il documento è reperibile all'indirizzo ***http://www.w3.org/XML***
[2] Analisi sintattica del documento
[3] Elaborazione del documento

Titolo del capitolo: …
Paragrafo
Titolo del paragrafo: …
Capoverso
Testo: …
Nota: …
Figura
Didascalia: …
Capoverso
Testo: …
Nota: …
Figura
Didascalia: …
Immagine: …
Paragrafo
Titolo del paragrafo: …
Capoverso
Testo: …
Nota: …
Figura
Didascalia: …
Immagine: …
Capitolo
Titolo del capitolo: …
Paragrafo
Titolo del paragrafo: …
Capoverso
Testo: …
Nota: …
Figura
Didascalia: …
Immagine: …

Sulla base di tale esempio, osserviamo che gli elementi sono le parti della struttura rappresentata come i capitoli, i titoli, i paragrafi e i capoversi; in particolare in questa struttura si individuano:

- **elemento radice o elemento principale** (nell'esempio l'elemento Testo), che è il padre tutti gli altri elementi e che li racchiude, ed è sempre uno ed uno solo;
- **sottoelemento**, un elemento contenuto in un elemento genitore (nel nostro esempio gli elementi quali Capitolo, Paragrafo, Titolo, Figura e Didascalia); con la terminologia specifica della teoria dei grafi si definisce ramo ogni sottoelemento che ne contiene a sua volta altri (nel nostro esempio l'elemento Capitolo è un ramo) e si definisce foglia ogni sottoelemento che non contiene ulteriori sottoelementi (nel nostro esempio l'elemento Immagine è una foglia).

Non è sempre possibile definire una struttura ad albero in grado di rappresentare in modo completo ed esaustivo il contenuto di un documento, alcuni documenti digitali, che consentono la navigazione ipertestuale, sono strutturati a grafo e nel ridurre la struttura da grafo ad albero si rischia una perdita di informazione. Il linguaggio XML offre una efficace e flessibile soluzione per la codifica e l'attivazione di collegamenti ipertestuali tra documenti e all'interno dello stesso documento, migliore del protocollo di collegamento compreso in HTML, fondata sull'utilizzo delle entità.

In questo contesto si può definire la seguente categorizzazione:

- ***elementi***, le unità di base per la descrizione logica della struttura del documento, che ne consentono l'organizzazione del contenuto;
- ***entità***, le unità di base per la descrizione della struttura fisica del documento, che

ne consentono la composizione del contenuto.

Definito il concetto di elemento, poniamo l'accento sulla strutturazione del contenuto nei documenti XML, ovvero quella proprietà che differenzia tra loro XML e HTML, ma che soprattutto consente il superamento della soluzione di continuità tra documento e dato organizzato.

La strutturazione del contenuto si fonda sul concetto di "tipo di documento" a cui corrisponde un protocollo di definizione del documento in funzione dell'organizzazione delle informazioni contenute.

Il tipo di documento è l'insieme delle specifiche di descrizione della struttura logica delle informazioni contenute in un documento.

Il linguaggio XML utilizza le specifiche del tipo di documento per identificare le caratteristiche degli elementi di contenuto e la loro posizione logica nella struttura del documento.

Il protocollo di descrizione del tipo di documento in XML è stato, nella prima strutturazione del linguaggio, il DTD (Document Type Definition), richiamato all'interno del documento XML dal tag DOCTYPE, che permette di descrivere, in codifica rigorosa, strutture complesse e flessibili.

Sotto l'aspetto teorico, la definizione del tipo di documento comporta uno sforzo organizzativo insolito ma proficuo per la successiva redazione del documento, in quanto razionalizza la composizione dei contenuti e ne semplifica il controllo e l'aggiornamento, evitando una confusione tra queste attività di tipo strutturale con la formattazione e gestione dell'output.

Seguendo l'esempio del testo, osserviamo il DTD limitatamente alla parte, subordinata all'elemento paragrafo, che descrive la struttura di capoverso e nota a lato:

```
<!ELEMENT capno (capoverso,nota)+>
<!ELEMENT capoverso(#PCDATA)+>
<!ELEMENT nota (#PCDATA)+>
```

In questo stralcio di listato notiamo che la struttura viene caratterizzata da un nome, capno, di cui vengono indicate le componenti, capoverso e nota, ciascuna delle quali è poi definita in modo congruente con il tipo di dati che conterrà, in questo caso testo ASCII.

Si noti che la struttura prevede la possibilità di essere reiterata (il simbolo "+"), dando origine a paragrafi composti da più capoversi e più note.

Il documento XML individua il DTD di riferimento utilizzando il tag:

```
<!DOCTYPE nome_tipo "http://www.dominio. com/file.dtd">
```

in cui viene dichiarata la struttura di documento a cui il documento XML si uniforma (nell'esempio, nome_tipo) e l'indirizzo al quale è reperibile il file in cui la struttura è resa disponibile come risorsa (nell'esempio, "http://www.dominio.com/file.dtd").

La struttura del documento può essere descritta anche all'interno tag DOCTYPE).

È possibile e formalmente corretto non dichiarare un DOCTYPE all'interno di un documento XML, in questo caso il documento non dispone di una struttura di riferimento.

Le regole che determinano il linguaggio XML e la relazione esistente tra documento XML e il concetto di DOCTYPE determinano tre possibili casi quando il documento XML è sottoposto alle funzioni di analisi (parsing):

1. **documento non fruibile**, il documento contiene errori grammaticali (ad esempio un tag aperto e poi non chiuso correttamente) per cui non viene processato e reso disponibile all'utente;

2. **documento well-formed**, il documento è formalmente corretto ma non si fa riferimento esplicito ad alcun tipo di documento (DTD); in questo caso il documento è fruibile (può essere sottoposto alla fase di processing, quindi visualizzato) ma non è strutturato, le informazioni contenute sono disorganizzate come all'interno di un documento HTML;
3. **documento valid**, il documento è formalmente corretto e fa riferimento esplicito ad un tipo di documento; il documento è fruibile e strutturato, quindi può essere sottoposto a complesse elaborazioni del contenuto e dialogare con archivi di dati parimenti strutturati.

Il parsing del documento XML, in virtù delle sue caratteristiche, è necessariamente più rigido del parsing di un documento HTML (analogamente, l'organizzazione delle informazioni è di regola più rigida in un archivio che in un documento destrutturato); una tale organizzazione del documento evidenzia l'importanza della correttezza formale del documento stesso; ad esempio, i tag XML, a differenza dei tag HTML, sono *case-sensitive*, ovvero sensibili alla differenza tra caratteri maiuscoli e minuscoli, per cui l'elemento ***titolo*** e l'elemento ***Titolo*** risultano essere formalmente e logicamente diversi .

Per comprendere e dare unitarietà all'analisi della strutturazione dei documenti XML è interessante esaminare il breve documento, riportato di seguito, che utilizza il DTD ***nome_tipo***, e riporta il primo capoverso del quarto capitolo de "*I promessi sposi"* di Alessandro Manzoni:

```
<?XML version="1 0"?>
<!DOCTYPE nome_tipo «http://www dominio com/ file dtd»>
<capno>
        <capoverso>
                Il cielo era tutto sereno. A mano a mano che il sole si
                alzava dietro il monte, si vedeva la sua luce, dalla sommità
                dei monti opposti, scendere come spiegandosi rapidamente
                giù per i pendii e nella valle.
        </capoverso>
        <nota> Compare Frà Cristoforo </nota>
</capno>
```

Una possibile visualizzazione di questo documento, come risultato di una procedura di processing, può essere:

Il cielo era tutto sereno. Amano a mano che il sole si alzava dietro il monte, si vedeva la sua luce, dalla sommità dei monti opposti, scendere come spiegandosi rapidamente giù per i pendii e nella valle.	*Compare Frà Cristoforo*

I DTD sono dei tipi di "schema" che erano presenti nel linguaggio SGML ma che utilizzati con XML perdono alcune delle loro potenzialità di modellazione della struttura di un documento.

I DTD utilizzati con il linguaggio XML sono uno strumento efficiente per definire:

- **elementi**
- **attributi**
- **entità**
- **notazioni**

oppure per determinare la possibile molteplicità con cui gli elementi possono comparire all'interno di un documento XML.

Laddove la necessità è quella di modellare dati e relazione con un discreto grado di

complessità si deve fornire uno strumento altrettanto efficace, e questo bisogno ha prodotto lo sviluppo del linguaggio XML Schema.

La storia di questo linguaggio è molto breve, ed inizia ufficialmente nel gennaio del 2000 con la presentazione del prima nota "Datatypes for DTDs" del W3C dove si afferma che c'era la necessità di tipizzazione dei dati ed inoltre delineava la strada per la migrazione dei documenti degli utenti scritti in XML e DTD verso XML Schema.

L'evolversi del linguaggio XML Schema non è stato tale da rendere obsoleti i DTD ne questo era lo scopo; la tecnologia dei DTD seppur semplice è consolidata e dove velocità e compatibilità sono una discriminante importante i DTD permangono ancora strumenti idonei.

Se si volesse fare un esempio legato ai mezzi di trasporto si potrebbe dire che anche bicicletta e auto sono due mezzi di trasporto, così come DTD e XML strumenti per la strutturazione dei documenti, ma se devo andare in centro ad acquistare il giornale la bicicletta è sicuramente la scelta migliore, diversamente l'auto se devo percorrere molti chilometri o spostarmi insieme ad altre persone.

Gli XML Schema sono in ogni caso un grosso passo evolutivo in funzione della definizione di strutture di documenti, in particolare si possono ricordare i punti che già in una nota del 1999 (il documento è reso disponibile all'indirizzo *http://www.w3.org/TR/NOTE-xml-schema-req*) riporta:

1. ***more expressive than XML DTDs*** – XML è più espressivo, in particolare la tipizzazione permette ulteriori funzionalità rispetto al DTD;
2. ***expressed in XML*** – scritto in XML, gli XML Schema sono a loro volta documenti XML e questo in termini di flessibilità e potenza è estremamente rilevante;
3. ***self-describing*** – è autodescrittivo ossia la struttura degli schemi è identico al documento XML che codificano;
4. ***usable by a wide variety of applications that employ XML*** – utilizzabile da un'ampia gamma di applicazioni che integrano XML;
5. ***straight for wardly usable on the Internet*** – utilizzabile direttamente in Internet, la sua definizione come linguaggio tiene ben presente l'obiettivo di utilizzare XML Schema per la rete e le applicazioni distribuite;
6. ***optimized for interoperability*** – ottimizzato per l'interoperabilità, XML Schema correlato alle applicazioni distribuite in rete può basarsi anche su tutte le architetture che hanno come scopo l'interoperabilità di cui la rete necessità; in particolare è dotato di meccanismi interni che permettono agli schema di richiamarsi ed innestarsi l'uno nell'altro;
7. ***simple enough to implement with modest design and runtime resources –*** abbastanza semplice da aver bisogno di risorse modeste sia per la progettazione che per l'esecuzione;
8. ***coordinated with relevant W3C specs*** (XML Information Set, Links, Namespaces, Pointers, Style and Syntax, as well as DOM, HTML, and RDF Schema) – coordinato con le specifiche relative di W3C.

Come esemplificazione di seguito viene riportato un prototipo di listato che definisce gli stessi elementi definiti attraverso DTD, nell'esempio del testo, affrontato in precedenza:

```
<?xml version='1.0' encoding='ISO-8853-1' ?>
<xs:schema xmlns:xs='http://www.w3.org/2000/10/XMLSchema'>
        <xs:element name='capoverso' type='xs:string' />
        <xs:element name='nota' type='xs:string' />
</xs:schema>
```

Stile di rappresentazione del documento XML

Il processo di rappresentazione del documento XML (processing) è gestita attraverso strumenti software che permettono di rappresentare il contenuto in funzione delle caratteristiche stilistiche proprie del documento e del supporto di output, alcuni strumenti finalizzano la rappresentazione alla stampa, altri alla visualizzazione su computer attraverso browser. In particolare, i browser Web stanno integrando funzioni di processing di codice XML, nella prospettiva di una progressiva adozione di questo protocollo nella produzione di siti Web ad alta interattività, modularità e dinamicità.

Il processo di rappresentazione dei documenti non è sempre necessario, in molti sistemi di comunicazione i documenti XML vengono costruiti e utilizzati in virtù delle loro caratteristiche di strutturazione interna al solo scopo di gestire flussi di dati tra archivi altrimenti incompatibili, e per questo non richiedono fasi di processing.

Le informazioni sulle caratteristiche stilistiche dei documenti (formattazione del testo, posizione dell'immagine, tipi, dimensioni e stili di font) sono separate dai contenuti, come le informazioni sulla struttura, e sono supportate da file esterni, denominati ***"fogli di stile"***, descritti secondo i linguaggi appartenenti alla famiglia **XSL** (e***X****stensible* ***S****tyle* ***L****anguage*).

XSL è un linguaggio di descrizione dello stile di rappresentazione di un documento o di una classe di documenti, progettato e sviluppato presso il 3WC (*World Wide Web Consortium*, il consorzio di determinazione degli standard dei protocolli del Web), per la standardizzazione dei documenti di stile sulla rete.

XSL rispetta le specifiche ***DSSSL*** *(****D****ocument* ***S****tyle* ***S****emantics and* ***S****pecification* ***L****anguage)*, protocollo di descrizione stilistica sviluppato per il protocollo SGML e rappresenta il collegamento tra la codifica dei dati e le operazioni di formattazione che rendono fruibile il contenuto del documento.

Elaborando l'esempio precedente di redazione del documento XML, è possibile definire la formattazione del capoverso (ad esempio: va scritto utilizzando un font di grandezza 10) e della nota (font di grandezza 8 in stile corsivo), attraverso il protocollo XSL, e un foglio di stile che formalizza queste regole è il seguente:

```
<xsl>
    <rule>
        <target-element type= "capno"/>
        <paragraph font-size="10pt">
            <children/>
        </paragraph>
    </rule>
    <rule>
        <target-element type= "nota"/>
            <paragraph font-size="8pt" font-style="italic">
                <children/>
            </paragraph>
    </rule>
</xsl>
```

XSL è quindi un linguaggio marcatore in cui vengono descritte le specifiche desiderate; in particolare nel listato precedente è di interesse il tag <rule>, che contiene un elemento e un'azione da compiere su di esso ogni volta che viene rilevato.

Osservando nei dettagli l'esempio, per l'elemento "capno" (<target-element type="capno"/>) viene specificato che il software di elaborazione dell'XML, ogni qualvolta rileva tale elemento, deve creare un paragrafo (<paragraph…> </paragraph>) nel visualizzatore testi utilizzato (browser o elaboratore testi), dando alla sequenza di caratteri che lo compongono una grandezza di 10 punti (font-size="10pt"), mentre per l'elemento "nota" è richiesto un carattere grande 8 punti e uno stile italico (font-style="italic").

Nella definizione dei fogli di stile e quindi nella descrizione dei formati da applicare, il livello

di difficoltà può essere tale per cui il linguaggio XML necessita della potenza e flessibilità di un vero linguaggio di programmazione.