

Quali sono gli standard e le raccomandazioni del World Wide Web Consortium per l'XML

Oreste Signore

Ufficio Italiano W3C presso il C.N.R. - Istituto CNUCE
Area della Ricerca di Pisa San Cataldo - Via G. Moruzzi, 1 - 56124 Pisa
Email:oreste@w3.org

Sommario

Le tecnologie Web sono in rapido e costante sviluppo, e il World Wide Web Consortium definisce regole e strategie di evoluzione del Web. In questo lavoro vengono inizialmente descritti il ruolo e l'organizzazione del World Wide Web Consortium. Successivamente, vengono presentate le caratteristiche essenziali del linguaggio XML, illustrando lo stato delle famiglie di tecnologie correlate. Dopo una sommaria descrizione delle linee di sviluppo future, in particolare del Semantic Web, che dovrebbe costituire la sfida tecnologica dei prossimi anni, vengono brevemente discussi potenzialità e limiti di XML.

Il World Wide Web Consortium

Il World Wide Web Consortium (W3C), è un consorzio che sviluppa tecnologie interoperabili (specifiche, linee guida, software, e strumenti) per portare il Web al massimo del suo potenziale come forum per l'accesso all'informazione, il commercio, le comunicazioni e la creazione di una cultura comune.

Il W3C (<http://www.w3.org>) è stato costituito nell'ottobre 1994 con lo scopo di sviluppare al massimo il potenziale del World Wide Web, definendo protocolli comuni che ne favoriscano l'evoluzione e assicurino l'interoperabilità.

È un consorzio internazionale di imprese, ospitato congiuntamente da Massachusetts Institute of Technology Laboratory for Computer Science (MIT/LCS) negli Stati Uniti, Institut National de Recherche en Informatique et en Automatique (INRIA) in Europa, Keio University Shonan Fujisawa Campus in Giappone.

Il consorzio è guidato da Tim Berners-Lee, Direttore e creatore del World Wide Web, e Jean-François Abramatic, Presidente.

Il W3C è finanziato dai membri, ed è *neutrale* rispetto ai venditori, lavorando con tutta la comunità per produrre specifiche e software di riferimento reso disponibile gratuitamente in tutto il mondo.

Il funzionamento del consorzio è regolato da un insieme di regole, contenute nel *Process Document* ([W3CPD]), che viene periodicamente verificato e adeguato, a seguito di accettazione da parte dei membri, alle esigenze emergenti. Un aspetto essenziale è che le decisioni vengono prese a seguito di un processo che prevede il raggiungimento del consenso dei partner. Questo significa che, anche se non sempre è possibile raggiungere l'unanimità, si ha comunque cura di non prendere decisioni su cui non ci sia accordo da parte di una vasta maggioranza. Tutte le osservazioni vengono valutate dal punto di vista tecnico.

Gli *obiettivi a lungo termine* del W3C sono coerenti con le motivazioni iniziali che hanno portato alla nascita del web. Essi possono essere espressi sinteticamente come:

- *Universal Access*: Rendere il Web accessibile a tutti, promuovendo tecnologie che tengono conto delle notevoli differenze in termini di cultura, formazione, capacità, risorse materiali, e limitazioni fisiche degli utenti in tutti i continenti
- *Semantic Web*: Sviluppare un ambiente software che consenta ad ogni utente di fare il miglior uso possibile delle risorse disponibili sul Web
- *Web of Trust*: guidare lo sviluppo del Web tenendo in attenta considerazione gli aspetti innovativi che questa tecnologia solleva in campo legale, commerciale e sociale

In un contesto in cui un gran numero di imprese innovative applica le specifiche W3C, si assiste a una rapidissima evoluzione delle tecnologie Web, e ad un accorciamento del "time to market", i compiti

principali del W3C sono:

- *Vision*: W3C promuove e sviluppa la sua visione del futuro del World Wide Web. I contributi provenienti da molte centinaia di ricercatori che operano nelle organizzazioni membro del W3C e nell'intera comunità Web permettono al W3C di identificare i requisiti tecnici che devono essere soddisfatti perché il Web diventi effettivamente un "*universal information space*".
- *Design*: W3C progetta le tecnologie Web per realizzare questa visione, tenendo conto delle tecnologie presenti e future.
- *Standardization*: W3C contribuisce alla standardizzazione delle tecnologie Web producendo specifiche (dette "Recommendation") che descrivono i componenti fondamentali del Web. Tutte le Recommendation e i Rapporti Tecnici sono *disponibili a tutti*.

Le W3C Recommendation

Le W3C Recommendation sono il risultato di un processo lungo e cooperativo, regolato dal Process Document, che prevede una serie di passi e di documenti prodotti:

- La proposta da parte di uno dei membri
- Lo svolgimento di un Workshop
- La definizione di una Activity
- La costituzione di un Working Group (aperto solo ai membri ed eventualmente a esperti qualificati)
- Working Draft (primo documento tecnico, sottoposto a revisioni e raffinamenti nell'ambito del Working Group)
- Last Call Working Draft (rapporto tecnico pubblico)
- Candidate Recommendation (prevede una "call for implementation")
- Proposed Recommendation (dopo aver maturato una adeguata esperienza implementativa)
- W3C Recommendation (documento definitivo, distribuito al vasto pubblico)

Anche se *Network Computing* cita il W3C nella sua lista dei "*Ten Most Significant Standards Groups*", dal punto di vista formale il W3C non è un organo di standardizzazione. Tuttavia, va riconosciuto che il W3C è una comunità di membri che cooperano spontaneamente per definire le linee guida e le specifiche, verificando che esse siano realmente implementabili, e mantiene stretti contatti con gli organi di standardizzazione e con gli User Forum.

La partecipazione al W3C

I membri del W3C, ad oggi, sono più di 500 ([W3CML]), e comprendono organizzazioni di vario tipo, che collaborano allo sviluppo del Web. I vantaggi dell'associazione al W3C sono riconducibili a un guadagno in termini di maggiore informazione, miglioramento di immagine, coinvolgimento attivo e possibile influenza nell'evoluzione del Web, maggiore prontezza nel seguire l'evoluzione del mercato. Il punto essenziale della associazione al W3C consiste proprio nella partecipazione attiva allo sviluppo del Web. Infatti, le ragioni principali per aderire al consorzio sono:

- Maggiore conoscenza dei processi evolutivi e possibilità di partecipare ai processi decisionali
- Conoscenza anticipata dell'evoluzione delle raccomandazioni, e quindi opportunità per svolgere attività tecnologica di punta, sviluppando prototipi in fase con lo sviluppo tecnologico e realizzando prima degli altri applicazioni e prodotti conformi alle raccomandazioni

Al momento, i membri italiani sono sei:

- C.N.R.
- CSP (Centro di Eccellenza per la Ricerca, Sviluppo e Sperimentazione di Tecnologie Avanzate Informatiche e Telematiche) di Torino;
- HiT Software s.r.l.
- Presidenza del Consiglio
- Telecom
- Università di Bologna

Il C.N.R. è membro del W3C fin dal 1995, e la sua partecipazione al W3C è supportata nel contesto delle attività internazionali del C.N.R.

Gli Uffici W3C

Per migliorare la sua presenza e i rapporti con le singole comunità nazionali, il W3C ha creato un certo

numero di W3C Offices (<http://www.w3.org/Consortium/Offices/>). Gli uffici al momento sono 10, e si trovano in Australia, Germania, Grecia, Hong Kong, Inghilterra, Israele (di prossima apertura), Italia, Olanda, Marocco, Svezia.

I compiti degli Uffici sono:

- organizzare il supporto per gli eventi locali
- pianificare la traduzione dei documenti W3C
- installare e gestire un mirror site
- costituire un forum per i membri nazionali
- riportare al W3C le posizioni locali
- produrre materiale promozionale
- offrire supporto per la diffusione delle comunicazioni del W3C e dei comunicati stampa

L'Ufficio Italiano W3C (<http://w3c.it>) ha iniziato la sua attività nel 1999. Sul suo sito ufficiale si trova anche una sezione specifica per le *Risorse W3C Italia* in cui compaiono i puntatori a:

- La *Directory delle risorse italiane* alla quale tutti possono aggiungere i riferimenti, che vengono successivamente controllati e autorizzati
- La *mailing list di W3C Italia*
- I *comunicati stampa* di W3C Italia
- Le *traduzioni* dei documenti tecnici
- Gli *eventi W3C Italia*

L'Ufficio Italiano sta dando vita ad una nuova iniziativa, denominata *WebLab*, che intende rafforzarne il ruolo tecnico scientifico, agendo da punto di incontro tra le attività di ricerca e l'integrazione e sviluppo di tecnologie Web.

I quattro "Domain"

Per una sua migliore organizzazione interna, il W3C ha individuato quattro domini di attività, che hanno comunque una significativa interazione tra di loro.

I quattro domini sono:

- *Architecture*. L'Architecture Domain sviluppa le tecnologie di base del Web, potenziando la sua infrastruttura e migliorando la sua automazione.
- *User Interface*. Mira a migliorare l'interazione degli utenti con il Web. Questo comporta sviluppi sui formati e i linguaggi per presentare l'informazione con maggiore accuratezza e più elevato livello di controllo.
- *Technology and Society*. Cerca di sviluppare l'infrastruttura del Web per affrontare i problemi di tipo sociale e politico. La crescita esponenziale del Web ha costretto la comunità del Web a considerare gli aspetti etici e legali in una prospettiva internazionale. Il Technology and Society Domain cerca di comprendere questi problemi alla luce della nuova tecnologia, sia modificandola, sia formando gli utenti su benefici, costi e limiti della tecnologia.
- *Web Accessibility*. L'impegno del W3C per portare il Web al massimo del suo potenziale prevede la promozione di un elevato grado di usabilità per le persone portatrici di handicap. La Web Accessibility Initiative (WAI) persegue l'obiettivo di una maggiore accessibilità operando in cinque settori: tecnologia, linee guida, strumenti, formazione e promozione, ricerca e sviluppo.

I principi informatori del Web

Il Web è una applicazione costruita su Internet, e quindi ne ha ereditato i principi fondamentali, che guidano il lavoro svolto nelle varie Activities.

Questi principi informatori sono:

- *Interoperabilità*: Le specifiche dei linguaggi e dei protocolli del Web devono essere compatibili tra di loro, e consentire a qualunque hardware e software di operare tra di loro.
- *Evoluzione*: Il Web deve essere in grado di accogliere le nuove tecnologie. Principi di progettazione quali la semplicità, la modularità e l'estensibilità aumentano le possibilità che il Web sia in grado di funzionare con le tecnologie emergenti, quali i device mobili e la televisione digitale, o con altre tecnologie che compariranno.
- *Decentralizzazione*: La decentralizzazione è senza dubbio il principio più nuovo e difficile da applicare. Per consentire che il Web si diffonda realmente su scala mondiale senza rischiare errori o interruzioni, l'architettura (e Internet) devono limitare o eliminare le dipendenze da nodi centrali.

XML: la pietra angolare?

Perché XML?

L'HTML era stato ottimizzato per un facile apprendimento, per cui aveva un unico tag set, con una semantica predefinita per ogni tag. Essendo le strutture dati predefinite, e mancando una validazione formale, offriva un supporto inadeguato per gestire strutture dati complesse, articolate su più livelli. Ovviamente, l'HTML era il risultato di un compromesso tra le potenzialità di utilizzo e la facilità di uso, e si è dimostrato molto utile e potente per la realizzazione di applicazioni ragionevolmente semplici. I limiti di HTML emergono quando si hanno di fronte applicazioni che presentano requisiti più stringenti, determinati dalla complessità delle strutture dati da gestire, dalla necessità di utilizzare gli stessi dati in maniera diversa e di gestire in modo adeguato dati con un ciclo di vita significativo.

Extensible Markup Language (XML) è nato per far fronte alle nuove applicazioni Web, in cui i dati costituiscono un elemento essenziale (*data-centric Web applications*).

XML è stato quindi il primo passo per assegnare una semantica ai tag e supportare le transazioni sul Web, permettendo lo scambio di informazioni tra database diversi. Ulteriori e significativi vantaggi sono costituiti dalla possibilità di avere viste diverse degli stessi dati, e la possibilità di personalizzare le informazioni mediante opportuni agenti. L'adozione di XML agevola la gestione di collezioni di documenti, e costituisce un supporto fondamentale per la pubblicazione di informazioni a livello internazionale, con il non piccolo vantaggio di essere indipendente dalla piattaforma e dal linguaggio.

XML è un formato testo semplice e molto flessibile, derivato da SGML (ISO 8879). Progettato inizialmente per far fronte ai problemi dell'editoria elettronica a grande scala, XML sta ora giocando un ruolo sempre più significativo nello scambio di una grande varietà di dati sul Web. In sintesi, XML:

- Consente l'editoria elettronica internazionale in maniera indipendente dai supporti.
- Permette alle imprese di definire protocolli indipendenti dalla piattaforma per lo scambio di dati, in particolare nel settore del commercio elettronico.
- Consente di trasmettere l'informazione in modo che opportuni user agent possano elaborarla.
- Semplifica lo sviluppo di software per gestire informazione specializzata sul Web.
- Semplifica l'elaborazione dei dati utilizzando software a basso costo.
- Permette di visualizzare le informazioni nel modo più adeguato alla specifica esigenza, sotto il controllo dei fogli di stile.
- Agevola la specifica di metadati, consentendo un reperimento dell'informazione più soddisfacente e fornendo una piattaforma comune di incontro tra i produttori e gli utilizzatori dei dati.

Caratteristiche della struttura del linguaggio

Le caratteristiche di XML possono essere illustrate con un esempio semplice, relativo alla gestione degli Ordini. In Figura 1 viene riportato un esempio di documento XML che descrive un Ordine. La sintassi XML usa tag di inizio e fine, come per esempio `<importo>` e `</importo>`, per marcare i campi informativi. Un campo informativo racchiuso tra due marcatori viene detto elemento (*element*) e può essere ulteriormente arricchito dalla presenza di coppie nome/valore (nell'esempio, `id="ord001"`) dette attributi (*attribute*). Come si può vedere, si tratta di una sintassi semplice, la cui elaborazione automatica è poco complessa, senza codifiche particolarmente criptiche, per cui resta comprensibile alla lettura diretta. XML, basato su SGML, riesce familiare a chi conosce HTML.

In XML i tag devono essere inseriti correttamente uno dentro l'altro, deve esistere una corrispondenza tra il tag di apertura e quello di chiusura, sono previsti elementi a campo informativo nullo e gli attributi dei tag devono essere racchiusi tra doppi apici.

```
01 <?xml version="1.0"?>
11 <ordine id="ord001">
12   <cliente db="codcli123"/>
13   <prodotto db="prod345">
14     <importo>23.45</importo>
15   </prodotto>
16 </ordine>
```

Figura 1 - Un documento XML per descrivere un Ordine

La presenza di una struttura formale del documento, espressa nel DTD (Document Type Definition), non

ha un impatto diretto sul modello strutturale implicito: nell' esempio di Figura 2, in cui il DTD è incluso nel documento (ma potrebbe anche essere referenziato come risorsa esterna) le righe 4 e 6 specificano che l' attributo *db* è obbligatorio. Un documento XML si dice "well formed" quando rispetta le regole di scrittura; viene detto "validato" quando è coerente con la struttura definita nel DTD (Document Type Definition).

```

01 <?xml version="1.0"?>
02 <!DOCTYPE ordine [
03 <!ELEMENT ordine ( cliente, prodotto+ )>
04 <!ATTLIST ordine id ID #REQUIRED>
05 <!ELEMENT cliente EMPTY>
06 <!ATTLIST cliente db CDATA #REQUIRED>
07 <!ELEMENT prodotto ( importo )>
08 <!ATTLIST prodotto db CDATA #REQUIRED>
09 <!ELEMENT importo ( #PCDATA )>
10 ]>
11 <ordine id="ord001">
12   <cliente db="codcli123"/>
13   <prodotto db="prod345">
14     <importo>23.45</importo>
15   </prodotto>
16 </ordine>

```

Figura 2 - Un documento XML con il suo DTD (in grassetto)

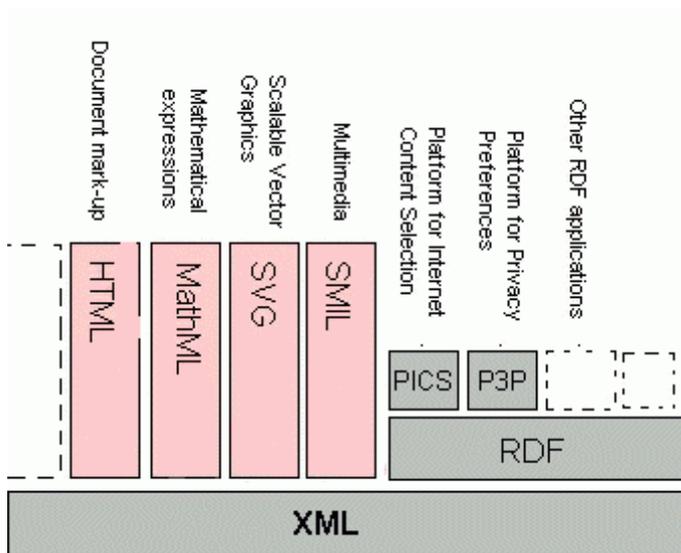


Figura 3 - XML come supporto ad applicazioni e linguaggi di markup

Per cosa?

XML costituisce attualmente la tecnologia chiave di W3C, è infatti un mezzo espandibile e flessibile per modellare il Web, e ormai si è affermato come la "lingua franca" del Web. Va comunque tenuto presente che in realtà W3C considera XML come una famiglia di tecnologie, su cui vengono investite le risorse del Consorzio, e non intende centralizzare il controllo di XML, preferendo lasciare agli utenti, coerentemente con la filosofia del Web, il compito di sviluppare applicazioni particolari. Il W3C vede l' XML come una Activity([XML]), che ha molte interazioni con tutte le altre. Vari gruppi di lavoro concorrono allo sviluppo della famiglia di tecnologie XML, definendo il suo minimo modello informativo (*infoset*), le sue proprietà strutturali (*schema*), la sua struttura ipertestuale (*link*), il modo di manipolare le strutture dati (*query*). La centralità di XML nell' architettura del Web è provata anche dal fatto che ogni nuovo linguaggio utilizzato per definire un nuovo standard deve essere descritto in XML.

XML è una sintassi a basso livello per rappresentare dati strutturati, e può essere utilizzato per una larga varietà di applicazioni. Questa idea è rappresentata in maniera molto schematica in Figura 3, che descrive

come già ora XML sia il supporto di un gran numero di applicazioni e di linguaggi di markup. Sia nel W3C che in altri contesti, molti gruppi stanno definendo nuovi formati per lo scambio delle informazioni, e questo sta portando ad una autentica esplosione di applicazioni XML, con un processo che appare inarrestabile. In molti contesti le applicazioni XML vengono utilizzate per rappresentare e organizzare i dati, con l' evidente convinzione che questo approccio possa rendere più facilmente gestibili i problemi derivanti dalla incompatibilità dei dati, riducendo il lavoro di reimmissione di informazioni che già sono in forma elettronica.

XML: storia e stato attuale di sviluppo

La *prima fase* dell' Activity XML iniziò nel giugno 1996, e si concluse nel febbraio 1998 con l' emanazione della *XML 1.0 Recommendation*, successivamente rivista nell' ottobre 2000 ([XML1.0]).

Nella *seconda fase*, il lavoro è stato svolto in parallelo da un certo numero di Working Group, che hanno prodotto le *Recommendation Namespaces in XML* nel gennaio 1999 e *Style Sheet linking* nel giugno 1999 (XMLns), [XMLstylesheet]).

Nel settembre 1999 è iniziata una *terza fase*, con l' obiettivo di completare il lavoro della seconda e di introdurre un nuovo Working Group su XML Query.

Al momento sono attivi diversi Working Group:

- *XML Query Working Group*
La missione dell' *XML Query working group* è fornire dei meccanismi flessibili per formulare query per estrarre informazioni da documenti reali o virtuali sul Web. Il Working Group ha pubblicato diversi documenti allo stadio di *Public Working Draft* ([Xquery], [XMLqa], [XMLqdm], [XMLqf], [XMLqgc]).
- *XML Schema Working Group*
Anche se XML 1.0 fornisce il meccanismo del Document Type Definition (DTD) per dichiarare i vincoli nell' utilizzo del markup, l' elaborazione automatica dei documenti XML richiede funzionalità più rigorose e ampie, per specificare come si combinano le parti componenti una applicazione, la struttura del documento, gli attributi, i tipi di dati, etc. Il Working Group sta considerando i metodi per definire struttura, contenuto e semantica dei documenti XML. Tre specifiche ([XMLschema0], [XMLschema1], [XMLschema2]) sono passate, nell' ottobre 2000, nello stadio di Candidate Recommendation.
- *XML Linking Working Group*
Il Working Group sta progettando i link ipertestuali per XML, operando la distinzione tra link *esterni* (tra oggetti) e *interni* (verso parti di documenti XML). Entrambe le tipologie di link sono trattate in dettaglio, con l' obiettivo di progettare una funzionalità di hyperlink e indirizzamento avanzata, scalabile e mantenibile. *XML Linking Language* ([Xlink]) e *XML Base* ([XMLBase]) sono attualmente allo stadio di Proposed Recommendation, mentre *XML Pointer Language* ([Xpointer]) è nello stadio di Last Call Working Draft.
- *XML Core Working Group*
Il Working Group elabora la XML 1.0 Recommendation, tenendo conto degli errori riscontrati e di eventuali altri commenti. Il Working Group continua il lavoro dei Working Group: *XML Syntax*, *XML Fragment*, and *XML Information Set Working Groups*, e costituisce una sede di discussione per le attività relative agli *XML Namespaces*. Sono allo stadio di Working Draft i documenti su *XML Information Set* e *XML Inclusions* (*XInclude*) ([XInfoSet], [XInclude]), mentre il documento su *XML Fragment Interchange* ([XMLFragment]) è già nello stato di Candidate Recommendation.
- *XML Coordination Group*
Provvede a coordinare i lavori dei vari Working Group della XML Activity, e tra questa e le altre componenti del W3C o altre organizzazioni.

Il Document Object Model (DOM)

Il Document Object Model (DOM) è una API (Application Programmer Interface) standard verso la struttura del documento, il cui obiettivo è semplificare l' accesso alle componenti dei documenti, per cancellare, aggiungere o modificare il loro contenuto, stile e attributi. Il DOM rende possibile lo sviluppo di applicazioni che funzionino correttamente su tutti i browser, server, e piattaforme.

Il W3C sta attualmente completando il *DOM Level 3*, che presenta alcune caratteristiche nuove. Tra i vari moduli del DOM, il DOM-XML si basa su una rappresentazione interna ad albero del documento, nella quale l' utente può navigare. Il modo standard di vedere il documento è una gerarchia di tag,

XML nell' architettura di applicazioni

Al di là delle molte e spesso interessanti applicazioni specifiche di XML, un aspetto rilevante è costituito dalla possibilità di applicazioni di tipo più generale, che utilizzano XML per realizzare architetture applicative innovative. Tra queste, va segnalata la possibilità di *distribuire le funzioni tra client e server*, o realizzare un *ipertesto adattivo*. Molte di queste applicazioni si basano sulla comprensione di un principio chiave dell' XML, cioè la *separazione tra contenuto e forma*. Mediante le trasformazioni di stile è possibile presentare l' informazione in modo adeguato alle caratteristiche dell' utilizzatore. Considerato che frequentemente la medesima informazione può essere indirizzata a utenti diversi, è spesso necessario adeguare la presentazione alle specifiche esigenze.

Va sempre tenuto presente ([Holman2000]) che ogni volta che viene modificato lo stile della presentazione dei dati, si ha in realtà un doppio processo:

- in prima battuta, va *trasformata* l' informazione, dalla forma utilizzata quando è stata creata nell' ambito dell' organizzazione, a quella indicata per la distribuzione delle informazioni;
- successivamente questa informazione, ulteriormente ristrutturata, deve essere *presentata* nella maniera più efficace per il fruitore.

Nel flusso delle informazioni dall' origine alla fase di presentazione (*rendering*), è conveniente legare l' informazione alla sua forma finale il più tardi possibile, mantenendo così la possibilità di utilizzare l' informazione anche per altri obiettivi durante l' intero processo. La pratica di modellare l' informazione sulla base di quello che sarà il suo utilizzo finale, per quanto molto diffusa, è da considerare poco accorta. In generale, nella restituzione di un documento XML si hanno due passi distinti di trasformazione di stile:

- la *trasformazione* dell' istanza del vocabolario XML in una nuova istanza coerente con il vocabolario della semantica di restituzione
- la *formattazione* dell' istanza del vocabolario di restituzione nello user agent.

Per far fronte alle esigenze di questi due processi in modo indipendente, il Working Group del W3C responsabile dell' Extensible Stylesheet Language (XSL) ha messo a punto due Recommendation, una per la trasformazione dell' informazione, e l' altra per la restituzione (*rendering*).

La Recommendation *XSL Transformations* ([XSLT]) descrive un vocabolario riconosciuto da un XSLT processor per trasformare la struttura dell' informazione prodotta nell' ambito dell' organizzazione in un' altra struttura, adeguata per i successivi passi elaborativi.

La Candidate Recommendation *Extensible Stylesheet Language* ([XSL]) descrive un vocabolario riconosciuto dall' agente di rendering per trasformare i formati espressi in forma astratta in particolari mezzi di presentazione.

È importante tener presente che XSLT è stato progettato essenzialmente per *trasformare i vocabolari XML nel vocabolario XSL di formattazione*, per cui sono assenti alcune funzionalità che lo rendano uno strumento di uso affatto generale, anche se è utilizzabile per la maggior parte delle esigenze di trasformazione. Per questo motivo, i fogli di stile XSLT vengono spesso utilizzati in contesti diversi da quelli legati alla restituzione (*stylesheet rendering*) e sono detti *transformation script*. Un caso tipico è quello dell' interazione con sistemi preesistenti (*legacy system*), molto frequente anche nel settore dell' e-commerce, in cui un XSLT processor può trasformare una particolare istanza di transazione, basata sul suo vocabolario, nel vocabolario su cui è basato l' input previsto dall' applicazione preesistente. In altri termini, XSLT può rivelarsi utile in tutti quei casi in cui è rilevante non tanto la sintassi con cui viene rappresentato il contenuto, ma il risultato dell' analisi sintattica.

Il comportamento di uno stylesheet è guidato dalla presenza dei tag di markup, che definiscono un modello *implicito*, non dalla struttura *esplicita* del documento, descritta dal DTD o da altro schema esterno al documento. Questo comportamento, indipendente dal DTD, risulta molto utile nell' esaminare documenti XML *well formed*, che non fanno esplicito riferimento a un modello, anche se, ovviamente, la presenza di un DTD consente di operare sul documento con un livello di conoscenza più elevato.

È probabilmente superfluo rammentare che un singolo file può essere elaborato da più stylesheet, in quanto non c'è nulla che leghi una istanza di file ad uno o più specifici stylesheet.

Un XSLT processor prende in input uno o più stylesheet e uno o più file sorgente. L' input iniziale è costituito da un singolo stylesheet e un singolo file sorgente, gli altri stylesheet vengono incorporati prima dell' elaborazione completa del primo file sorgente. L' XML processor legge poi gli altri file sorgente sulla base del contenuto del primo file XML, quindi l' XSLT processor può accedere altri file sorgente in qualunque momento, sotto il controllo dello stylesheet.

Tutti gli input devono essere documenti XML well formed. L' ovvia conseguenza è che non è possibile

elaborare file HTML che seguono convenzioni lessicali non XML, ma è possibile elaborare file XHTML (*Extensible Hypertext Markup Language*)¹

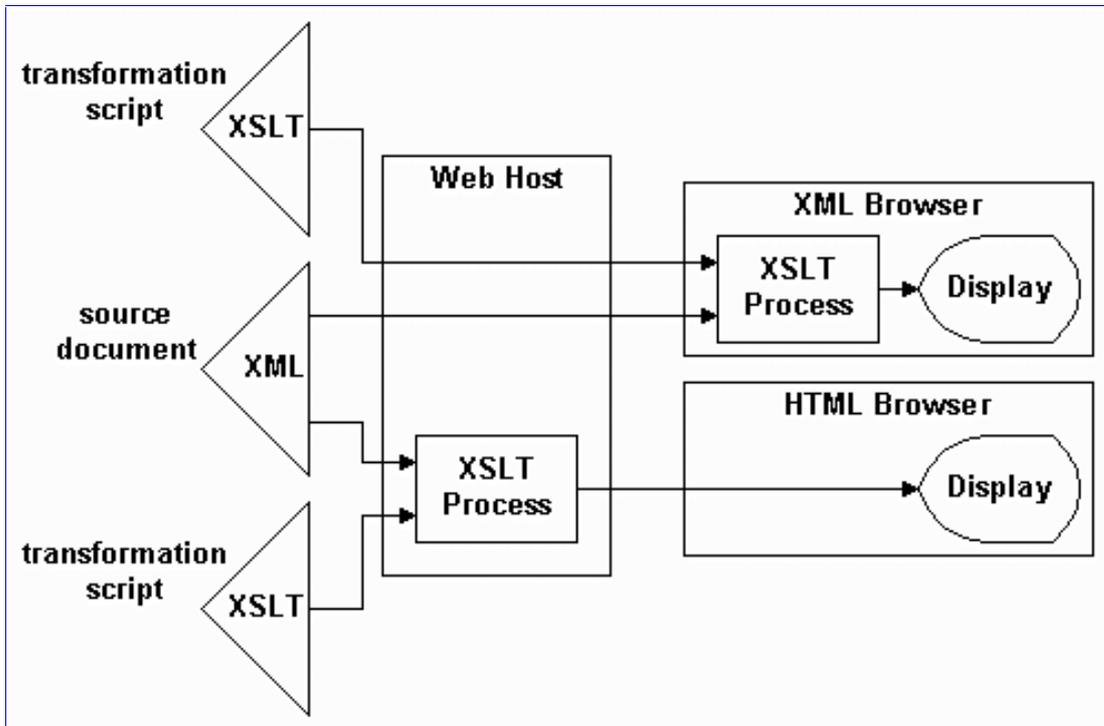


Figura 4 - Esempio di trasformazione mediante XSLT (da [Holman2000])

In una architettura a tre livelli, è possibile inviare informazione strutturata al web browser utilizzando XSLT sull' *host*, sullo *user agent*, o su *entrambi* (Figura 4). Da un punto di vista tecnico, il server può distribuire il carico elaborativo ad agenti XML/XSLT inviando una combinazione di stylesheet e informazione sorgente da trasformare sul lato client; oppure può effettuare tutte le trasformazioni direttamente sul server, in modo da poter supportare anche user agent che supportano solo vocabolari HTML o HTML/CSS, o WML ([Lee2000]).

Indipendentemente dalle caratteristiche del browser, la scelta se trasferire documenti con markup XML o trasformarli in documenti HTML (semanticamente meno ricchi) può dipendere da considerazioni legate all' opportunità di rendere disponibili a tutti le informazioni contenute nel documento XML completo. Una scelta potrebbe essere quella di trasformare il contenuto del documento in base alle caratteristiche del potenziale fruitore. Infine, per diminuire l' utilizzo della banda di trasmissione e incrementare l' elaborazione distribuita, è anche possibile utilizzare due processi XSLT, uno per trasformare il markup delle informazioni in un generico formato di distribuzione, e un altro sul client agent, per la personalizzazione,

Va comunque tenuto presente che esistono due linee di pensiero: secondo alcuni è necessario porre un "semantic firewall" tra i dati aziendali e gli utilizzatori, che quindi potrebbero anche essere dei semplici browser HTML, destinati solo alla presentazione di informazioni spogliate di qualunque semantica. Altri invece ritengono il Web debba comunque evolvere verso un *semantic web* (di cui si parla nel seguito) in cui l' informazione trasmessa debba sempre essere semanticamente ricca, ed essere poi elaborata da user agent intelligenti e dal comportamento rispettoso delle regole.

¹ File XHTML possono essere generati da file HTML utilizzando un tool disponibile gratuitamente sul sito W3C (<http://www.w3.org/People/Raggett/tidy/>), che corregge le codifiche HTML errate e segnala quelle che non riesce a correggere.

XML: uno sguardo al futuro

Se già oggi è quindi possibile utilizzare la tecnologia *per distribuire le funzioni tra client e server*, o realizzare un *ipertesto adattivo*, la sfida del futuro, illustrata compiutamente da Tim Berners-Lee, Direttore del Wide Web Consortium, in varie sedi, e più recentemente in una sessione di XML 2000, è il *Semantic Web*. Nella visione di Berners-Lee, il Web ha una architettura a livelli (Figura 5), che verrà sviluppata completamente nel giro di una decina di anni.

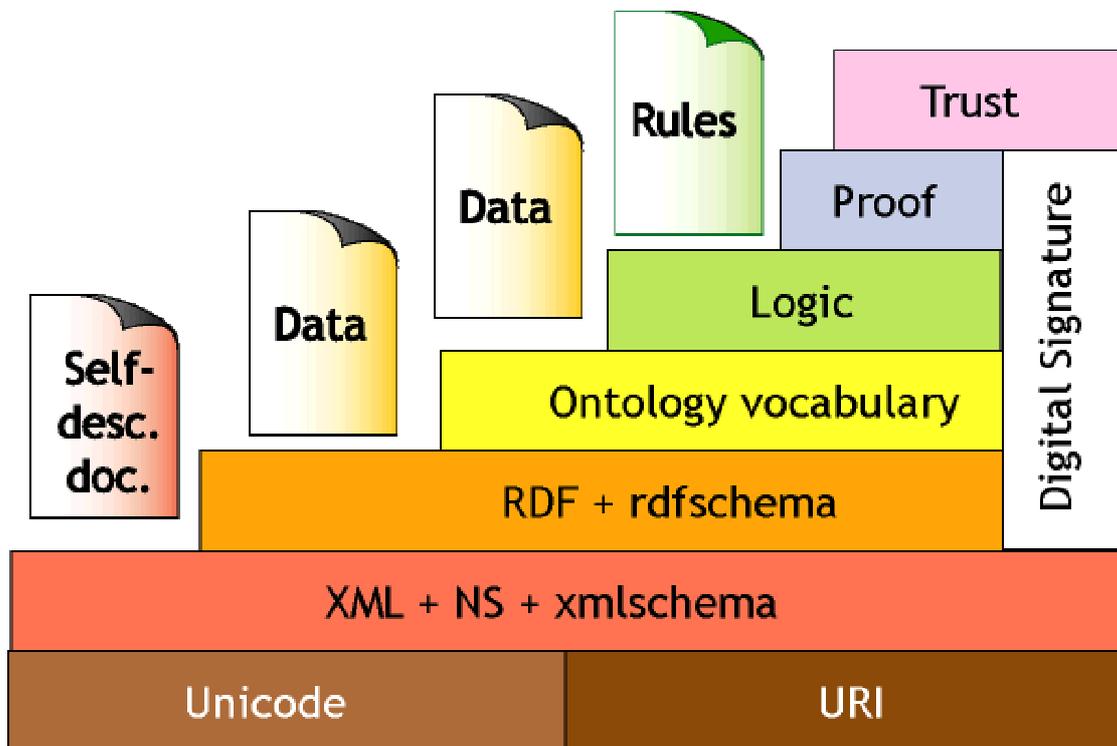


Figura 5 - L'architettura a strati del Semantic Web

Per chiarezza di terminologia, va ricordato che la filosofia di base del Web è quella di uno spazio informativo universale, navigabile, con un mapping da *URI (Uniform Resource Identifier)* alle risorse. Nel contesto del Semantic Web, il termine semantico assume la valenza di "*elaborabile dalla macchina*" e non intende fare riferimento alla semantica del linguaggio naturale e alle tecniche di intelligenza artificiale. Il Semantic Web è, come l'XML, un ambiente dichiarativo, in cui si specifica il significato dei dati, e non il modo in cui si intende utilizzarli. La semantica dei dati consiste nelle informazioni utili perché la macchina possa utilizzarli nel modo corretto, eventualmente convertendoli. In questo senso, si potrebbe parlare del superamento di un "*semantic test*" quando una macchina, ricevendo dei dati, li utilizza nella maniera corretta.

Si riproduce quindi la situazione di una decina di anni fa, quando vi furono le prime dimostrazioni del Web: l'aspetto essenziale è il fattore di scala. Su piccola scala, un collegamento tra due informazioni semanticamente equivalenti, ma rappresentate in maniera diversa su due database, non rappresenta qualcosa di significativo, ma ampliare queste possibilità alla comunità mondiale, senza imporre vincoli stringenti, assume tutta un'altra rilevanza².

Esaminando più in dettaglio gli elementi del Semantic Web, appare evidente, oltre al ruolo di base giocato da XML (con *Name Space* e *xmlschema*), la centralità di RDF e RDF Schema, che costituiscono il linguaggio per descrivere le risorse e i loro tipi.

² A puro titolo di esempio, per sottolineare l'importanza dell'interoperabilità semantica, e le difficoltà che si incontrano anche quando si fa riferimento ad un contesto specifico, ancorché ricco e complesso, si rimanda allo sforzo compiuto nel progetto *Aquarelle*, in cui, tra le altre cose, erano stati definiti dei DTD generali per la descrizione del patrimonio artistico-culturale, coerenti con le tradizioni culturali dei vari paesi partecipanti (<http://aquarelle.inria.fr/aquarelle/welcome.html>).

Al di sopra di questo strato si pone il *livello ontologico*. Una ontologia permette di descrivere le relazioni tra i tipi di elementi (per es. "questa è una proprietà transitiva") senza però fornire informazioni su come utilizzare queste relazioni dal punto di vista computazionale.

La firma digitale (*digital signature*) è di significativa importanza in diversi strati nel modello astratto del Semantic Web. La crittografia a chiave pubblica è una tecnica nota da qualche anno, ma non si è ancora diffusa su larga scala come ci si poteva attendere. Nella visione di Berners-Lee, un elemento che potrebbe aver giocato contro la diffusione di questa tecnica è che essa è a "grana grossa", imponendo una scelta binaria tra fiducia o non fiducia (trusted/not trusted), mentre sarebbe necessaria una infrastruttura in cui le parti possano essere riconosciute e accettate come credibili in specifici domini. Con una granularità più fine come questa, la firma digitale potrebbe essere utilizzata per stabilire la provenienza delle ontologie e delle deduzioni, oltre che dei dati.

Il *livello logico* è il livello immediatamente superiore al livello ontologico. A questo livello le asserzioni esistenti sul Web possono essere utilizzate per derivare nuova conoscenza. Tuttavia, i sistemi deduttivi non sono normalmente interoperabili, per cui, secondo Berners-Lee, invece di progettare un unico sistema onnicomprensivo per supportare il ragionamento (*reasoning system*), si potrebbe pensare di definire un linguaggio universale per rappresentare le dimostrazioni. I sistemi potrebbero quindi autenticare con la firma digitale queste dimostrazioni ed esportarle ad altri sistemi che le potrebbero incorporare nel Semantic Web.

Anche se i livelli più alti dell'architettura potranno richiedere diversi anni prima di poter raggiungere uno stadio in cui siano effettivamente implementati e affidabili, già oggi esiste una quantità significativa di lavoro nel settore delle ontologie. Le soluzioni pratiche includono l'uso di XSLT per derivare RDF da documenti XML, la comparsa di database e motori di ricerca RDF di tipo generalizzato, interfacce grafiche generalizzate e specifiche per RDF. La riorganizzazione delle attività sui metadati in ambito W3C dovrebbe dare un impulso al lavoro sul Semantic Web.

Non sarebbe corretto concludere questa illustrazione del Semantic Web senza segnalare che su di esso, che pure si presenta come un'idea affascinante, è più che lecito mostrare qualche perplessità o scetticismo, specialmente riandando con la memoria alle illusioni create dall'Intelligenza Artificiale negli anni '60 e '70. Inoltre, questa linea di sviluppo non sembra avere molti punti in comune con i "web services" che molti venditori XML propongono come il prossimo futuro, investendovi anche significative risorse.

Su queste tematiche nella prossima World Wide Web Conference [WWW10] è previsto un Workshop ([SemWeb2001]).

XML: luci e ombre

L'e-business, supportato dalla tecnologia Internet, può a buona ragione essere considerato oggi la forza trainante nell'evoluzione del modello di business. A breve termine, questa rivoluzione avrà un impatto massiccio sulle transazioni business-to-business. L'avvento dell'e-business potrebbe essere segnalato dall'intenzione di Ford e General Motors di utilizzare l'e-commerce per tutti i loro acquisti, per un giro d'affari che solo per la Ford è dell'ordine di 80 miliardi di dollari l'anno ([Worden2000]).

Per supportare le transazioni e-business, è necessario un linguaggio comune per scambiarsi le informazioni. Abbiamo già visto come HTML non possedesse le caratteristiche necessarie, mentre XML, con le sue potenzialità di strutturazione, permette di definire significato e struttura delle informazioni. Quasi ovvio che esso sia apparso come la panacea, il nuovo esperanto per il mondo dell'e-business, e sia stato accolto con enorme favore dai principali attori del mondo IT. Imprese concorrenti tra di loro, come IBM, Microsoft, Sun, Oracle, supportano lo standard XML e stanno effettuando investimenti su prodotti coerenti con gli standard che essi stessi, in modo cooperativo, aiutano a definire. Tuttavia, è evidente che XML costituisce solo la base per definire uno standard per le transazioni e-business, mentre mancano ancora tasselli importanti per raggiungere una reale interoperabilità semantica. La libertà di markup di XML ne costituisce infatti sia un punto di forza che di debolezza. Inoltre, la spinta verso una strutturazione estremamente elevata può portare, in alcuni ambiti, in cui il testo libero ha una valenza considerevole, a una autentica incomunicabilità delle informazioni, che fanno riferimento ad organizzazioni concettuali e vocabolari diversi e non riconciliabili.

La situazione che si è venuta a creare richiama in qualche modo alla memoria quella determinata agli inizi degli anni '80 dall'affermarsi dei DBMS relazionali. Anche in quel caso, il modello si limitava a definire come rappresentare e memorizzare le informazioni, senza imporre nessun vincolo semantico (anche perché le implementazioni commerciali gestivano un concetto molto debole di dominio). Esattamente

come in XML ognuno può definire a suo piacimento strutture di documenti e nomi di tag, così nel mondo relazionale ognuno poteva definire a suo piacimento i nomi delle tabelle e delle colonne. I risultati sono evidenti a tutti: anche se tuttora accade di sentir dire che due sistemi informativi che utilizzano lo stesso DBMS sono "compatibili senza problema", tutti sanno che in realtà informazioni semanticamente equivalenti sono memorizzate e gestite in maniera difforme sia tra imprese diverse, che talvolta anche nell'ambito della stessa organizzazione.

Qualcuno stima i costi di integrazione delle applicazioni anche a un 40% dei costi IT totali. Anche se questa stima fosse eccessiva, andrebbero comunque considerati i costi indotti, determinati dalla difficoltà a sviluppare nuove applicazioni e i conseguenti ritardi nella loro realizzazione.

È quindi evidente il rischio implicito nell'adozione di XML come soluzione dei problemi di interscambio dell'informazione: esso costituisce unicamente uno standard a livello fisico, e parzialmente a livello logico, ma non offre nessuna garanzia a livello concettuale, l'unico che potrebbe garantire una reale interoperabilità.

In apparenza, il problema parrebbe risolto, almeno nell'ambiente e-business, dall'adozione di uno degli standard emergenti, come C-XML per le transazioni commerciali o FIN-XML per le transazioni finanziarie, anche se va comunque tenuto presente che questi standard sono abbastanza complessi, e mantenersi coerenti con essi potrebbe rappresentare un lavoro di non trascurabile peso.

Purtroppo, al momento esistono diversi standard emergenti, sia specifici dei vari settori, sia talvolta per lo stesso settore. Ne consegue una potenziale Babele, con grossi costi di interfacciamento tra i vari standard. L'ipotesi di realizzare un repository delle varie descrizioni dei messaggi XML (BizTalk, ebXML, altre iniziative sostenute dal gruppo XML-EDI) non sembrano in realtà essere davvero in grado di risolvere l'enorme problema di avere comunque un numero di procedure di conversione che è dell'ordine di N^2 (se esistono N standard di messaggi si devono implementare $N \cdot (N-1)$ traduzioni). Per riportare la complessità del problema all'ordine N , occorrerebbe implementare uno schema generale onnicomprensivo, valido per tutti i settori, che riscuota l'accordo di tutti. Questo tipo di approccio, caldeggiato in passato anche in altri contesti, non ha mai avuto successo, ed è improbabile che lo abbia in questo caso.

Conclusioni

Le linee di sviluppo del Web sono coordinate dal W3C, che le definisce con il supporto e l'accordo dei suoi membri. La tumultuosa crescita delle applicazioni Web e delle tecnologie relative impone ritmi quasi impensabili, come provato dalla quantità di documentazione tecnica che si sussegue con cadenza incredibile. In un clima di apparente confusione, esistono tuttavia alcune linee di sviluppo ben definite.

L'utilizzo di XML si va diffondendo con una rapidità e una pervasività che testimoniano come molti abbiano compreso che si tratta di una tecnologia, o meglio di una famiglia di tecnologie, che sono centrali nello sviluppo del Web.

In questo quadro si inserisce l'ambizioso progetto del Semantic Web, che, considerato da un certo punto di vista, tende a rendere reale il grande sogno di Ted Nelson ([NelsonTH], [LitMach1993]). Va però sottolineato che un reale sviluppo di queste idee si potrà avere solo se si condividerà la filosofia ispiratrice del Web, mondo libero, che rifiuta i vincoli della centralizzazione, delle soluzioni proprietarie, della costituzione e prevalenza di posizioni di potere.

La realtà aziendale nell'implementazione di progetti in questo contesto non è omogenea, accanto ad imprese o gruppi che sposano appieno la filosofia del Web, e realizzano applicazioni di assoluta eccellenza, permangono alcune posizioni molto conservative, legate ad approcci tradizionali e sistemi sostanzialmente chiusi, che privilegiano la stabilità e affidabilità a breve termine, perseguendo strade che difficilmente garantiscono un recupero degli investimenti in tempi meno ravvicinati, ma non remoti, in un contesto evolutivo così rapido come il Web.

Riferimenti

- [DOM] *Document Object Model (DOM)*, <http://www.w3.org/DOM/>
- [Holman2000] Holman G. Ken: *What is XSLT? (I)*
<http://www.xml.com/lpt/a/2000/08/holman/s1.html>
- [Jigsaw] *"Jigsaw - The W3C Web Server"* <http://www.w3.org/Jigsaw>

- [Lee2000] Lee Wei Meng: *Transforming XML into WML*,
http://www.wirelessdevnet.com/channels/wap/training/xslt_wml.html
- [LitMach1993] Nelson Theodor Holm: *Literary Machines 93.1*,
<http://www.sfc.keio.ac.jp/~ted/TN/PUBS/LM/LMpage.html>
- [NelsonTH] <http://www.sfc.keio.ac.jp/~ted/>
- [RDFMSS] O.Lassila, R.Swick: *Resource Description Framework (RDF) Model and Syntax Specification*, W3C Recommendation 22 February 1999, <http://www.w3.org/TR/REC-rdf-syntax>
- [RDFSS] *Resource Description Framework (RDF) Schema Specification*, W3C Recommendation 03 March 1999, <http://www.w3.org/TR/1999/PR-rdf-schema-19990303>
- [SemWeb2001] <http://semanticweb2001.aifb.uni-karlsruhe.de/index.htm>
- [SGML] ISO 8879:1986: *Information processing -- Text and office systems -- Standard Generalized Markup Language (SGML)*; <http://www.iso.ch/cate/d16387.html>
- [SW2000] *Jigsaw - An open Source server from W3C that maps out one vision of the Web's future*,. <http://serverwatch.internet.com/reviews/web-jigsaw.html>
- [TBL1998] Tim Berners-Lee: *Semantic Web Road Map*, (1998),
<http://www.w3.org/DesignIssues/Semantic.html>
- [TBL1999] Tim Berners-Lee: *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by Its Inventor*, HarperSanFrancisco (1999), ISBN 0-06-251587-X
- [TBL2000] <http://www.w3.org/2000/Talks/1206-xml2k-tbl/slide1-0.html>
- [W3C] <http://www.w3.org>
- [W3CML] <http://www.w3.org/Consortium/Member/List>
- [W3CPD] World Wide Web Consortium Process Document
(<http://www.w3.org/Consortium/Process/Process-19991111/>)
- [Worden2000] Worden, R.: *XML E-Business Standards: Promises and Pitfalls*
<http://www.xml.com/lpt/a/2000/01/ebusiness/index.html>
- [WWW10] <http://www10.org/>
- [XInclude] XML Inclusions (XInclude) Version 1.0 - W3C Working Draft 26 October 2000
<http://www.w3.org/TR/xinclude>
- [XLink] XML Linking Language (XLink) Version 1.0 - W3C Proposed Recommendation 20 December 2000, <http://www.w3.org/TR/xlink/>
- [XLL] *XML Linking Language (XLink)*, W3C Working Draft 21 February 2000
<http://www.w3.org/TR/xlink/>
- [XML] *Extensible Markup Language (XML)* , <http://www.w3.org/XML>
- [XML1.0] *Extensible Markup Language (XML) 1.0 (Second Edition) W3C Recommendation 6 October 2000*, <http://www.w3.org/TR/2000/REC-xml-20001006>
- [XMLBase] XML Base - W3C Proposed Recommendation 20 December 2000
<http://www.w3.org/TR/xmlbase>
- [XMLFragment] XML Fragment Interchange - W3C Candidate Recommendation 12 February 2001
<http://www.w3.org/TR/xml-fragment>
- [XMLInfoSet] XML Information Set - W3C Working Draft 2 February 2001
<http://www.w3.org/TR/xml-infoSet>
- [XMLNs] Namespaces in XML World Wide Web Consortium 14-January-1999 -
<http://www.w3.org/TR/REC-xml-names>

| | |
|-----------------|--|
| [XMLqa] | The XML Query Algebra - W3C Working Draft - 15 February 2001 http://www.w3.org/TR/query-algebra/ |
| [XMLqdm] | XML Query Data Model - W3C Working Draft - 15 February 2001 http://www.w3.org/TR/query-datamodel |
| [XMLqr] | XML Query Requirements - W3C Working Draft - 15 February 2001 http://www.w3.org/TR/xmlquery-req |
| [XMLquc] | XML Query Use Cases - W3C Working Draft - 15 February 2001 http://www.w3.org/TR/xmlquery-use-cases |
| [XMLschema0] | XML Schema Part 0: Primer - W3C Candidate Recommendation 24 October 2000 http://www.w3.org/TR/xmlschema-0/ |
| [XMLschema1] | XML Schema Part 1: Structures - W3C Candidate Recommendation 24 October 2000 http://www.w3.org/TR/xmlschema-1/ |
| [XMLschema2] | XML Schema Part 2: Datatypes - W3C Candidate Recommendation 24 October 2000 http://www.w3.org/TR/xmlschema-2/ |
| [XMLstylesheet] | Associating Style Sheets with XML documents Version 1.0 - W3C Recommendation 29 June 1999, http://www.w3.org/TR/xml-stylesheet |
| [XPointer] | XML Pointer Language (XPointer) Version 1.0 - W3C Last Call Working Draft 8 January 2001 http://www.w3.org/TR/xptr |
| [Xquery] | XQuery: A Query Language for XML - W3C Working Draft - 15 February 2001 http://www.w3.org/TR/xquery/ |
| [XSL] | Extensible Stylesheet Language (XSL) Version 1.0 - W3C Candidate Recommendation - 21 November 2000, http://www.w3.org/TR/xsl/ |
| [XSLT] | XSL Transformations (XSLT) Version 1.0 - W3C Recommendation - 16 November 1999, http://www.w3.org/TR/xslt |