



Facoltà di Scienze della
Comunicazione

TECNOLOGIE DELLA COMUNICAZIONE APPLICATE ALL'IMPRESA

LAUREA SPECIALISTICA IN COMUNICAZIONE D'IMPRESA

PROF. STEFANO EPIFANI

stefano.epifani@uniroma1.it

DISPENSA N. 5

IL WEB SEMANTICO¹

¹ La presente Dispensa è tratta e adattata dal lavoro di tesi di Natalino Fiacco - *Applicazioni e tecnologie semantiche: prospettive per lo sviluppo di una rete intelligente* - laureato nel 2009 con la Cattedra di Organizzazione e Gestione della Comunicazione Interattiva. L'adattamento è a cura di Mauro Gallinaro, con la supervisione del Prof. Stefano Epifani.



INDICE

INTRODUZIONE	3
1. IL WEB SEMANTICO	6
1.1.INTRODUZIONE AL WEB SEMANTICO	6
1.2.L'ARCHITETTURA DEL WEB SEMANTICO	10
1.2.1. <i>URI - Uniform Resource Identifiers</i>	13
1.2.2. <i>Unicode</i>	13
1.2.3. <i>XML, RDF e Ontologie</i>	14
1.3.I PRINCIPI DEL WEB SEMANTICO.....	17
2. I LINGUAGGI DEL WEB SEMANTICO	22
2.1.IL CONCETTO DI MARKUP	22
2.2.XML - EXTENSIBLE MARKUP LANGUAGE	23
2.3.RDF E METADATI.....	29
2.3.1. <i>I Metadati</i>	29
2.3.2. <i>RDF - Resource Description Framework</i>	32
<i>Rappresentazione Grafica di RDF</i>	34
<i>RDF Schema</i>	36
2.4.LE ONTOLOGIE	37
2.4.1. <i>Introduzione</i>	37
2.4.2. <i>Caratteristiche delle Ontologie</i>	39
2.4.3. <i>Ontologie e Web Semantic</i>	43
2.4.4. <i>I Linguaggi delle Ontologie</i>	44

INTRODUZIONE

La maggior parte dei contenuti del Web, attualmente, sono espressamente rappresentati affinché possano essere letti da utenti umani e non sono progettati, viceversa, per essere elaborati da agenti software. Un browser o un motore di ricerca sono in grado, secondo ben precise istruzioni, di disporre e visualizzare le informazioni in una certa maniera, ma non hanno capacità di riconoscerle o attribuire loro un particolare significato per elaborarle. Ad esempio, quando un motore di ricerca come Google salva le pagine Web, non distingue se la parola *espresso* si riferisce a un treno o a un tipo di caffè, oppure se *verdi* sia un colore o il nome di un compositore (lo stesso vale, a maggior ragione, per i contenuti non testuali come le immagini o gli elementi multimediali). Un motore di ricerca può tentare di percepire i significati ricavandoli dal contesto della pagina, utilizzare sinonimi per espandere le ricerche, ma in ultima analisi il suo archivio è un elenco, certamente molto ben organizzato, di parole chiave. Se invece fosse in grado di comprendere il significato dei contenuti di una pagina, potrebbe accettare domande più complesse e fornire un'unica risposta altrettanto complessa e soprattutto efficace.

Il "Semantic Web" o "Web Semantico" nasce proprio in questo contesto e permette di introdurre una struttura semantica nei contenuti del Web. Non separato dal Web "tradizionale", ma introducendo il significato dei dati, ne è un'estensione che aggiunge una nuova funzionalità alle macchine, che diventano in grado di comprendere ed elaborare i dati, che fino ad ora semplicemente visualizzavano. Quindi, se fino ad oggi il Web è stato un supporto per la trasmissione di documenti tra le persone, in questo modo diventa una fonte di dati e informazioni che possono essere elaborate automaticamente.

Aggiungere la semantica ai contenuti del Web richiede la creazione di linguaggi e tecnologie in grado di estrarre significati da informazioni. In particolare, per lo sviluppo di un Web in versione semantica si configura come necessario lo sviluppo di un linguaggio che permetta di esprimere dati e regole per i ragionamenti: tale linguaggio dovrebbe consentire che le regole proprie di un sistema di rappresentazione della



conoscenza possano essere utilizzate sul Web. Naturalmente, per la natura stessa del Web, è possibile che dati concettualmente diversi siano rappresentati con lo stesso nome: un programma dovrebbe essere in grado di risolvere tali ambiguità e quindi elaborare le informazioni in maniera appropriata. Perché questo possa avvenire, si devono creare delle "ontologie". Il termine ontologia, preso in prestito dal linguaggio filosofico, indica un documento condiviso, che contiene la descrizione formale dei concetti di un dato dominio; identifica le classi più importanti, le organizza in una gerarchia, specifica le loro proprietà (che caratterizzano anche gli oggetti appartenenti alla classe) e descrive anche le relazioni più significative, che legano queste classi.

La struttura dei dati e la semantica introdotta dalle ontologie sono in grado di migliorare le potenzialità del Web: i programmi di ricerca, basandosi su un preciso concetto, cercano e trovano le pagine che effettivamente si riferiscono a quel concetto, anziché quelle che contengono parola-chiave ambigue o generiche; in questo modo la ricerca è più accurata: i programmi di ricerca, che nel Web tradizionale riportano una serie di pagine tra le quali l'utente deve ulteriormente cercare quelle d'interesse, con queste nuove funzionalità danno le pagine che si riferiscono a un preciso concetto; inoltre le interrogazioni possono riguardare informazioni che non risiedono sulla stessa pagina Web: il programma di ricerca, inferendo sulle regole specificate, può individuare il dato richiesto e rispondere all'interrogazione.

È chiaro che HTML, il linguaggio standard per la creazione di pagine Web, non ha le caratteristiche per strutturare i dati in base alla loro semantica, poiché specifica come devono apparire le informazioni, e pertanto deve essere affiancato da un altro linguaggio per la strutturazione semantica: tutto questo comporta la separazione tra il contenuto e il layout delle pagine e permette l'indipendenza tra i due.

In linea generale, applicazioni e prospettive del Web Semantico non rappresentano possibilità soltanto teoriche: nell'ultimo anno sono cominciate a comparire applicazioni Web intelligenti in grado di comprendere i dati pubblicati on-line. Questo nuovo modo di gestire le informazioni sta provocando una trasformazione del Web, facendolo evolvere, a vent'anni dalla sua nascita, in qualcosa di molto più avanzato rispetto a quello che siamo abituati a navigare ogni giorno. Nella prima decade della sua vita il Web è stato un



medium di sola lettura: una grande libreria dalla quale ricevere informazioni. La seconda decade ha visto invece la diffusione di servizi che lo hanno trasformato in un medium di scrittura. Attraverso Blog e Social Network i navigatori si sono convertiti da spettatori in attori, iniziando a pubblicare informazioni e opinioni: è il cosiddetto Web 2.0, il Web della partecipazione. Nella terza decade, iniziata da poco, le informazioni disponibili on line stanno diventando comprensibili anche ai calcolatori, in modo da poter essere collegate tra loro e riutilizzate sotto altre forme (e dando vita a nuove informazioni) tramite strumenti automatici. Si tratta di prospettive che stanno coinvolgendo già da qualche anno non soltanto il Web: applicazioni e strumenti definibili come "semantici" sono infatti già utilizzati in infrastrutture aziendali, istituzioni e altre realtà che hanno bisogno di catalogare e gestire grandi quantità di informazioni.

1. IL WEB SEMANTICO

1.1. INTRODUZIONE AL WEB SEMANTICO

Il termine Web Semantico è stato proposto da Tim Berners Lee nel 2001, nell'ambito di un articolo pubblicato dalla rivista *Scientific American*, in cui l'ideatore del WWW avanzava l'ambizione di creare una sorta di Internet “pensante”, la più avanzata ipotesi di intelligenza collettiva. L'idea di fondo di Tim Berners Lee e del W3C (World Wide Web Consortium) è quella di evolvere il Web attuale, non di sostituirlo, traslando il concetto di Web dal “**machine readable**” al “**machine understandable**”, ovvero rendere le informazioni processabili e comprensibili direttamente alle macchine. La creazione di una Rete Semantica, così come immaginata da Lee, è finalizzata a far sì che i computer diventino in grado di gestire automaticamente le informazioni e che “imparino” a realizzare una serie di processi in modo preciso, continuo e ripetuto. L'obiettivo è quello di creare un ambiente in cui l'informazione sia creata in modo tale da «sviluppare una cooperazione efficace tra computer e persone».

Nella concezione di Tim Berners Lee, il termine Web Semantico è associato all'idea di un Web nel quale agiscano agenti intelligenti, ovvero applicazioni in grado di comprendere il significato dei contenuti presenti on line e dunque in grado di guidare l'utente direttamente verso l'informazione ricercata, così come di sostituirsi all'utente stesso nello svolgimento di alcune operazioni. Un agente intelligente dovrebbe essere in grado di:

- comprendere il significato di documenti e risorse informative presenti sulla rete;
- creare percorsi in base alle informazioni richieste dall'utente, guidandolo poi verso di esse (in alcuni casi si può anche sostituire all'utente);
- spostarsi di sito in sito collegando elementi diversi dell'informazione richiesta in base a relazioni logiche e semantiche.



Nella definizione di Berners Lee, il concetto di tecnologia semantica assume una rilevanza centrale, dal momento che consente di comprendere il significato reale dei contenuti presenti sulla rete e considera il contesto in cui sono inseriti. L'espressione informazione semantica, infatti, indica quell' informazione che ha un significato, ovvero che è inserita e che si riferita ad un contesto ben determinato. Lo sviluppo di informazione semantica on line, dunque, si configura come un passo fondamentale nella direzione del Web Semantico, affinché le macchine siano in grado di interagire, comprendersi e interpretarsi. Il significato che le macchine saranno in grado di processare è ovviamente qualcosa di diverso da quello che viene messo in gioco nella comunicazione fra esseri umani; le macchine non sono certo capaci di comprendere realmente quello che sarà chiesto loro, ma dovranno essere capaci di riconoscere quelle informazioni che sono importanti per gli utenti, grazie ai software di cui dovrebbero essere dotate a questo scopo. A riguardo, sarà centrale che l'informazione venga strutturata secondo degli standard che consentano al computer di sapere quali bit di informazione rappresentano il significato di un documento on line.

Aggiungere il significato al Web non significherà rimpiazzare il vecchio Web con qualcosa di nuovo, ma creare una infrastruttura che si situi in cima al Web come una struttura descrittiva dell'informazione, un ambiente che sia una piattaforma per i motori di ricerca, per i mediatori di informazione e, infine, per gli agenti intelligenti.

In linea generale, il Web si configura come ambiente per l'interscambio di informazioni destinate ad essere utilizzate da esseri umani; tutti i documenti pubblicati in rete sono progettati per avere come utente finale l'essere umano. Non a caso, se si pensa a come sono scritti i documenti pubblicati on line, è possibile rilevare che le informazioni disponibili, che siano testi o immagini, assumono un valore semantico nel momento in cui vengono interpretate dagli utenti. Le attività di selezione, decontestualizzazione e interpretazione di contenuti non strutturati non possono essere realizzate dai computer. In altri termini le informazioni sono *machine-readable*, ma non *machine-understandable*.

Affinché il "significato" dei dati veicolati sul Web sia "compreso" dai computer, è necessario che ad essi siano associate *meta-informationi* che ne descrivano il contenuto



“semantico” in una forma comprensibile ai computer. In questo senso i *metadati* sono informazioni, strutturate per essere elaborate da una macchina, che descrivono una risorsa Web. In particolare, i metadati sono scritti con una sintassi che li rende comprensibili alla macchina. In questo contesto, dunque, per contenuto semantico si intende quel contenuto elaborabile dal computer, “*machine-understandable*”: tale definizione non fa riferimento alla semantica del linguaggio naturale o alle tecniche di intelligenza artificiale, ma alla semantica dei dati, che consiste nelle informazioni utili affinché la macchina, ricevendo dei dati, li utilizzi nella maniera corretta.

Nel famoso articolo apparso nel 2001 sulla rivista *Scientific American*, in cui lo stesso Lee, insieme a James Hendler e Ora Lassila, dava questa definizione:

«The Semantic Web is not a separate web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation»²

La definizione di Berners Lee esemplifica l’idea sottesa al Web Semantico, evidenziandone tre elementi chiave. Innanzitutto, il Web Semantico non rappresenta un nuovo tipo di WWW, ma si configura come un’estensione del Web corrente e delle sue potenzialità. Alla struttura del Web tradizionalmente inteso, dunque, si affianca una rete dei contenuti “semantici” in grado potenzialmente di consentire la creazione di significati condivisi, utilizzabili da agenti software per attivare operazioni automatiche, a partire da semplici istruzioni dell’utente.

In secondo luogo, obiettivo centrale dello sviluppo del Web Semantico è la cooperazione tra computer e persone, nel senso che le macchine possano essere di supporto concreto agli esseri umani nell’esecuzione e nell’automazione di operazioni, a partire dalla comprensione semantica delle informazioni e dei contenuti on line. L’idea di Tim Berners Lee è quella di consentire ai computer, attraverso l’utilizzo di metadati, di operare in maniera automatica, a partire dai dati presenti sul Web e dalla loro

² T. Berners-Lee, J. Hendler, O. Lassila, *The Semantic Web*, 2001, <http://www.scientificamerican.com>.

comprendere e interpretare. In questo modo agenti software potrebbero essere in grado di eseguire molte delle operazioni di ricerca e selezione delle informazioni solitamente realizzate dagli utenti in una versione del Web ancora di tipo *machine readable*.

Infine, con la sua definizione Berners Lee sottolinea come il potenziamento effettivo del WWW nella direzione del Web Semantico sia possibile soltanto attraverso l'attribuzione di un significato ben definito all'informazione, in modo tale che le macchine possano raggiungere quel livello di comprensione dei contenuti *on the Web* che attiva funzionalità avanzate di ragionamento, capacità di risposta a domande complesse e funzioni di ricerca di tipo semantico.

Con riferimento a tali principi, nell'articolo che definisce il Web Semantico gli autori propongono un esempio emblematico che mette in evidenza peculiarità e caratteristiche della nuova visione del Web: lo scenario di partenza è quello di due ragazzi che si affidano ad agenti software per la ricerca on line di un bravo fisioterapista. Le richieste dei due giovani però sono complesse: il terapista deve essere conosciuto per la sua professionalità e affidabilità, il suo ambulatorio deve trovarsi nei pressi della loro abitazione, le sue prestazioni devono rientrare tra quelle coperte dall'assicurazione medica.

Gli agenti software che utilizzano le tecnologie del Web Semantico sono in grado non solo di portare a termine l'operazione di ricerca di un fisioterapista in base ai criteri indicati, ma sono capaci anche di raffinare la ricerca stessa per mettere a disposizione dei due utenti differenti alternative, come ad esempio di confrontare la disponibilità ad un appuntamento dello specialista con la propria agenda personale o di motivare la selezione dei risultati prodotti dall'operazione di ricerca.

Attraverso l'attuale versione del Web, ovvero *machine readable*, per ottenere la stessa efficienza i due utenti dell'esempio di Berners Lee dovrebbero attivare manualmente numerose operazioni di ricerca, con il rischio di non riuscire a trovare tutte le possibili alternative. Il Web attuale, infatti, è uno straordinario universo di informazioni, la cui selezione, manuale o supportata dai motori di ricerca, implica un notevole sforzo da parte dell'utente. Come evidenziato dall'esempio, dunque, il Web



Semantico ha l'obiettivo di consentire alle macchine l'elaborazione delle informazioni attraverso l'impiego della logica e della semantica. Obiettivo raggiungibile soltanto se i computer, ovvero le macchine, sono potenzialmente in grado di accedere a collezioni strutturate di informazioni e di portare a termine operazioni di ragionamento e "comprensione" automatici attraverso l'utilizzo di regole di inferenza logica e semantica.

La sfida del Web Semantico, in definitiva, è quella di trovare un linguaggio logico adatto ad esprimere sia i dati che le regole per il ragionamento automatico sui dati. Questo linguaggio deve ovviamente poter coesistere con i diversi sistemi di rappresentazione della conoscenza esistenti e deve essere sufficientemente espressivo da abilitare il ragionamento su scala Web.

1.2. L'ARCHITETTURA DEL WEB SEMANTICO

Il Web Semantico è caratterizzato da un'architettura a livelli. La figura rappresenta questi layer o livelli, individuati da Berners-Lee nel 2001³:

- alla base dell'architettura del Web Semantico c'è lo **standard URI (Uniform Resource Identifiers)** - che viene utilizzato per la definizione univoca di risorse on line, come documenti, file e indirizzi Internet - così come lo **standard Unicode**, un sistema di codifica che associa un numero (o meglio, una combinazione di bit) a ciascun simbolo, segno o carattere in maniera indipendente dal programma, dalla piattaforma e dalla lingua utilizzata;
- al livello superiore, si trova il metalinguaggio **XML (eXtensible Markup Language)**, che riveste un ruolo di primo piano nell'architettura, unitamente ai **NS (namespaces)** e all'**XML Schema**. XML è un linguaggio che porta con sé alcune informazioni sulla semantica degli oggetti, mentre i NS - identificati tramite gli URI - garantiscono l'interoperabilità tra i dizionari di metadati;

³ T. Berners Lee, L'Architettura del nuovo Web, Feltrinelli, Roma, 2001.



- **RDF (Resource Description Framework)** e **RDF Schema** costituiscono, invece, il linguaggio per descrivere le risorse on line e i loro tipi, per l'interoperabilità semantica di dati e informazioni e consentono agli agenti intelligenti di effettuare inferenze logiche;
- al livello superiore, si pone il **Vocabolario delle Ontologie**. Le ontologie sono sistemi di metadati, riferiti a specifici vocabolari, che permettono di descrivere le relazioni tra le risorse del Web, consentendo agli agenti intelligenti di interpretare e comprendere i dati.

Infine un accenno alle componenti del Web Semantico ancora in via di sviluppo e che nei prossimi anni potrebbero indicare la via da seguire per la rappresentazione della conoscenza on line, in modo ordinato ed efficiente, attraverso:

- **Linguaggi logici universali**, che fanno riferimento al livello logico dell'architettura individuata da Berners Lee. In questa direzione, attraverso un linguaggio universale fatto di euristiche - vale a dire di procedimenti che consentono di prevedere dei risultati - le asserzioni esistenti sul Web potranno essere utilizzate per derivare nuova conoscenza. Tuttavia, si tratta di una componente ancora in via di sviluppo, dal momento che tale linguaggio universale ancora non esiste;
- **Firma Digitale**, il sistema di autenticazione dei documenti digitali, componente di significativa importanza nell'architettura del Web Semantico. La Firma Digitale potrebbe consentire di stabilire la provenienza delle ontologie e delle deduzioni, oltre che dei dati presenti sul Web. La Firma Digitale, in pratica, attesta che una determinata persona ha scritto un determinato documento o un'istruzione; in questo modo, gli utenti che reperiranno tali documenti o istruzioni potranno essere sicuri della loro autenticità;
- **Trust (Fiducia) e Proof (Dimostrazioni)**. Il Web Semantico può essere inteso come un enorme database di informazioni a cui l'utente può accedere; la cosa



fondamentale è che l'utente stesso sia in grado di distinguere i dati affidabili da quelli inattendibili. L'informazione sul livello di fiducia o Trust dei dati può essere in qualche modo esplicito (*questo dato viene dalla fonte x e x fa parte dell'elenco di fonti che io considero attendibili*), oppure può essere derivata in modo automatico attraverso agenti intelligenti basati sulle tecnologie del Semantic Web (*questo dato viene dalla fonte x; x non fa parte dell'elenco di fonti che io considero attendibili, però si trova nell'elenco delle fonti attendibili dalla persona y, che io considero fidata*). Questa deduzione o inferenza può essere più o meno complessa e può non essere prevedibile a priori. In ogni caso l'utente potrebbe anche comprendere come il livello di Trust di una data risorsa informativa on line sia stato calcolato dagli agenti intelligenti. Questo potrà avvenire soltanto attraverso il livello Proof, ovvero la dimostrazione della logica inferenziale che sottende al ragionamento dell'agente. In particolare, le informazioni ottenute dal processo di deduzione o inferenza del livello Trust degli agenti saranno validati dal livello di Proof o dimostrazione che attraverso motori di validazione verificheranno e dimostreranno la veridicità di ciascuna deduzione.

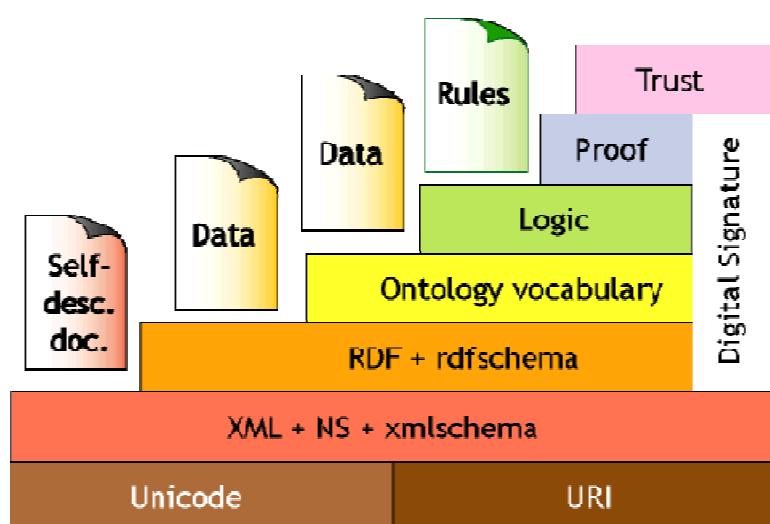


Figura 1 - Architettura del Web Semantico: il “layer cake” (Koivunen, Miller, 2001)



1.2.1. URI - Uniform Resource Identifiers

Nel Web Semantico è stato definito un sistema di identificatori unificato, gli Uniform Resource Identifiers (URI), per la definizione univoca del tema e dell'oggetto di una determinata risorsa on line. È possibile definire un URI per un qualsiasi documento, programma o contenuto on line; in particolare un URI indica al computer il meccanismo da attivare per accedere ad una determinata risorsa, in quale computer trovare la risorsa e attraverso quale percorso interno.

Gli identificatori più comuni sono quelli utilizzati per le pagine sul Web, ovvero gli indirizzi URL (Uniform Resource Locator). Ad esempio, nel Web il Blog dedicato alla Cattedra di Tecnologie della Comunicazione applicate all'impresa viene identificato univocamente in rete attraverso il seguente URI, in questo caso un indirizzo URL: <http://www.tci09.wordpress.com>. Più nel dettaglio, nel caso si intenda cercare sul Blog la slide dedicata al Web Semantico, la URI o URL <http://tci09.wordpress.com/2009/12/08/slide-dellottava-lezione/> indicherebbe al computer: con http://, il protocollo da utilizzare per accedere alla risorsa; con tci09.wordpress.com il computer sul quale la risorsa è stata salvata; /2009/12/08/slide-dellottava-lezione/ il percorso da seguire all'interno del computer e il nome della risorsa. In pratica, un indirizzo URL comunica ai computer dove trovare una risorsa specifica sul Web, identifica e allo stesso tempo localizza le risorse on line.

1.2.2. Unicode

Per immagazzinare nella memoria di un computer lettere, simboli o segni è necessario che a ciascun carattere venga assegnato un numero che lo identifica. Esistono differenti sistemi di codifica che associano numeri identificativi ai caratteri, attraverso tecniche e processi differenti. Questi sistemi di codifica, oltre a non comprendere un numero di caratteri sufficiente per tutte le circostanze, sono in contraddizione l'uno con l'altro, nel senso che due codifiche possono utilizzare lo stesso numero per identificare due caratteri diversi o che, viceversa, associno identificatori diversi per lo stesso



carattere. Inoltre, qualsiasi elaboratore necessita di utilizzare codifiche diverse. Il problema è che, quando i dati passano da una codifica a un'altra, o da una piattaforma all'altra, si corre il rischio di perdere informazioni. All'interno di questo contesto si inserisce lo standard Unicode, che attribuisce un numero univoco a ciascun carattere, indipendentemente dalla piattaforma, dall'applicativo e dalla lingua utilizzata. Lo standard Unicode è alla base di molti linguaggi come XML e Java ed è supportato da molti sistemi operativi, oltre che da quasi tutti i Web browser.

L'affermazione dello standard Unicode, insieme alla recente disponibilità di tecnologie e piattaforme che lo supportano, è fra i più significativi sviluppi della tecnologia della globalizzazione del software. L'adozione di Unicode, infatti, consente di sviluppare e di utilizzare un'unica versione di un software o di un sito Web anche attraverso piattaforme e lingue diverse, evitando ai colossi del software di reingegnerizzare il prodotto. Principi del Web Semantico.

1.2.3. XML, RDF e Ontologie

XML è un linguaggio pensato per progettare file di testo che descrivono dati strutturati, in un modo che produca file che siano facili da generare e da leggere per una macchina, che non siano ambigui e non dipendenti dalla piattaforma. L'acronimo significa Extensible Markup Language e indica un linguaggio di marcatura che permette agli individui di creare, definire e utilizzare i propri tag, quindi rappresenta una sorta di metalinguaggio per definire altri linguaggi per specifiche applicazioni. Costituisce il linguaggio per raggiungere l'interoperabilità sintattica, ovvero la capacità dei sistemi di riuscire ad interpretare la sintassi e la struttura dei documenti scambiati, ma anche l'interoperabilità strutturale, seppur in via parziale, cioè la possibilità di interpretare le strutture dei documenti di schemi logici differenti, conoscendo le regole di traduzione tra tali schemi. XML, però, non consente di soddisfare la cosiddetta interoperabilità semantica, ovvero la capacità di più sistemi di scambiarsi dati e informazioni in modo che il loro significato sia accettato e interpretato automaticamente dal sistema ricevente. Per questo, vengono in soccorso altre tecnologie, prima fra tutte RDF.



RDF è il linguaggio base per la codifica, lo scambio e il riutilizzo di metadati strutturati e consente l'interoperabilità tra applicazioni che si scambiano sul Web informazioni *machine-understandable*. RDF è il modello relazionale dei dati definito e promosso dal W3C per l'interscambio di dati sul Web. In RDF il significato dell'informazione è codificato in un insieme di triple che corrispondono a frasi elementari costituite da tre elementi, soggetto, predicato e complemento, identificabili attraverso le URI. Le triple poi consentono di esprimere asserzioni sulle risorse, dicendo che tali entità (soggetto) hanno proprietà (predicato) con certi valori (attributi). Con RDF è possibile definire i termini o i concetti che fanno parte dello schema logico dell'informazione da elaborare.

Rispetto a XML, RDF presenta alcuni vantaggi sostanziali, come per esempio la possibilità di definire agevolmente il tipo e le proprietà degli oggetti (si tratta sempre di un insieme definito di risorse, proprietà e valori). Tuttavia, anche RDF è caratterizzato da una serie di limiti che necessitano dell'adozione di ulteriori tecnologie in grado di rendere effettivamente concreta l'interoperabilità tra le varie applicazioni. Due sistemi, per esempio, potrebbero utilizzare URI diverse per identificare la stessa risorsa e pertanto le asserzioni fatte in un sistema non sarebbero confrontabili o integrabili con le asserzioni della stessa risorsa di un altro sistema. Per ottenere ciò, invece, occorre un modo per affermare che due identificatori abbiano lo stesso significato e quindi rappresentino due identificatori per riferirsi alla stessa risorsa.

Ecco che entrano in gioco le ontologie. Le ontologie costituiscono la tecnologia più importante alla base di tutta la concezione del Web Semantico, ma probabilmente anche la più complessa. Nel corso degli anni sono state proposte diverse definizioni, sia da parte dalla comunità dell'intelligenza artificiale che della comunità stessa del Web Semantico. Probabilmente la definizione più accreditata è quella data da Gruber nel 1993, secondo cui «*un'ontologia è la specificazione di una concettualizzazione*»⁴. Tale definizione richiama due punti chiave:

⁴ T. Gruber, *A translation approach to portable ontology specifications*, 1993, su <http://www.tomgruber.org>.



- **Un'ontologia fa riferimento a una concettualizzazione:** in un'ontologia saranno definiti i concetti rilevanti alla descrizione del dominio in esame. Una concettualizzazione, infatti, è una vista astratta e semplificata del mondo o di quella parte di realtà che ci proponiamo di rappresentare; ogni sistema di rappresentazione o gestione della conoscenza avrà sempre una qualche concettualizzazione esplicita o implicita del suo dominio;
- **Un'ontologia consiste in una specificazione:** tale concettualizzazione è espressa attraverso un'appropriata definizione, un'indicazione particolareggiata dei concetti e delle relazioni che fra essi sussistono.

Un semplice caso di ontologia può essere dato da una tassonomia o da un tesauro; ad esempio, le diverse categorie e sotto categorie utilizzate dai bibliotecari per catalogare libri formano una semplice ma efficace ontologia. Le ontologie e l'OWL (Ontology Web Language) sono utilizzate per definire dei metadati che descrivano dei testi. Il loro utilizzo permette di recuperare documenti esprimendo query complesse. Si parte da concetti semplici, ma si può raffinare la ricerca esprimendo vere e proprie asserzioni composte da un soggetto, un predicato e un oggetto. Le ontologie si configurano come supporti per l'organizzazione dei domini cognitivi e dei loro dati.

In sintesi, dunque, Berners-Lee disegna l'architettura del Web Semantico con quattro livelli essenziali. Il livello dei dati deve provvedere a stabilire un modello dei dati e la sintassi dei metadati (XML, RDF); il livello schema stabilisce una base per la definizione di un vocabolario (RDF Schema); il livello ontologico è deputato alla definizione delle ontologie (OIL, DAML, OWL); infine, il livello logico, che invece dovrebbe assicurare le basi per il ragionamento e per la realizzazione di inferenze⁵. Come già visto in precedenza, questi livelli principali dell'architettura del Web Semantico sono collegati anche ad altri, uno di base (quello delle URI) e gli altri superiori (Fiducia e Dimostrazioni), che sono strettamente legati alla realizzazione di quello logico, a sua volta dipendente dalla piena realizzazione del livello ontologico.

⁵ Cfr., T. Berners Lee, *What the Semantic Web can represent*, 1998, su <http://www.w3.org>.

Gli sforzi principali nella realizzazione del Web Semantico sono andati verso il livello dei dati e quello di schema. Il linguaggio XML mira alla definizione della struttura dei documenti e non prevede la possibilità di riconoscere esplicitamente delle proprietà semantiche. RDF invece mostra una certa interoperabilità semantica, perché secondo molti esperti «*contiene tanta tecnologia per la Rappresentazione della Conoscenza da poter essere condivisa largamente con i più svariati linguaggi per i meta-dati*»⁶. A partire da questo livello, che è ormai standardizzato, gli sforzi attuali per la realizzazione del Web Semantico si stanno concentrando sul livello della costruzione delle ontologie. Grazie a RDF e RDF Schema sarà possibile stabilire il livello superiore, proprio perché i linguaggi ontologici attualmente messi a punto si basano su questi standard. Attualmente i maggiori candidati a formare il livello ontologico sono i linguaggi OIL e DAML (probabilmente nella forma del DAML+OIL) e l'OWL messo a punto dal W3C.

1.3. I PRINCIPI DEL WEB SEMANTICO

Prima di analizzare le tecnologie e i linguaggi alla base del Web Semantico, è opportuno focalizzare l'attenzione sui principi che caratterizzano questa rinnovata versione del Web, prendendo in considerazione lo studio realizzato da Marja Ritta Koivunen ed Eric Miller, pubblicato in un articolo dedicato al seminario mondiale sul Web Semantico tenutosi in Finlandia nel 2002⁷. In particolare, i due studiosi hanno rilevato **sei principi** che sottendono all'architettura del Web Semantico e ne descrivono le caratteristiche essenziali:

- **Ogni cosa può essere identificata attraverso una URI.** Come già evidenziato in precedenza, si tratta del concetto di partenza del Web Semantico, secondo il quale qualsiasi oggetto, proprietà e stato presente in rete dovrebbe essere

⁶ *Ibidem*.

⁷ M. Koivunen, E. Miller, *W3C Semantic Web Activity*, presentato al Seminario “Semantic Web kick-off” tenutosi in Finlandia nel novembre 2001, disponibile all’indirizzo <http://www.w3.org>.



identificato in maniera univoca e soprattutto con una denominazione non ambigua nel Web Semantico, attraverso un identificatore ben definito. Tale identificatore può essere l'URL di una pagina Web, così come l'indirizzo e-mail. Le URI rappresentano i componenti di partenza dell'infrastruttura del Web Semantico per consentire alle macchine e ai software di descrivere la semantica delle risorse disponibili on line e dunque per comprendere il risultato di contenuti e informazioni presenti sul Web;

- **Le risorse e i link possono avere un “tipo”.** Nel Web attuale, quello in versione *machine readable*, le risorse (ovvero contenuti, informazioni e documenti presenti in rete) e i relativi link rappresentano l'ossatura dello spazio dell'informazione on line. Tuttavia, risorse e link sono solitamente realizzati per essere compresi dagli utenti e normalmente non contengono metadati che spieghino il loro contenuto e le loro relazioni con altre pagine Web. Gli utenti sono in grado di interpretare, comprendere e associare un significato ben preciso alle relazioni alle risorse tra loro linkate; senza metadati, la natura di una risorsa e le relazioni logiche e semantiche con altre risorse non sono comprensibili dalle macchine. I computer, dunque, richiedono che alla risorsa sia aggiunta una *meta-information*, che consenta di stabilire delle relazioni “più ricche”: «*Una relazione più informativa potrebbe essere, per esempio, “dipenda da”, “è una versione di”, “ha soggetto”, “autori”*»⁸. Con il Web Semantico le risorse e i link acquistano un “tipo” (*type*), ovvero una meta-information che associa un concetto a ciascuna risorsa e a ciascun link; concetto che può consentire ai computer di comprendere le relazioni logiche e semantiche tra le risorse, esplicitate attraverso link. Nella figura 2, alcuni link tra risorse evidenziano che un documento è una versione di un altro, che è scritto in riferimento ad altri documenti circa persone o luoghi, o che la risorsa contiene un software dipendente da altri software;

⁸ *Ibidem*.

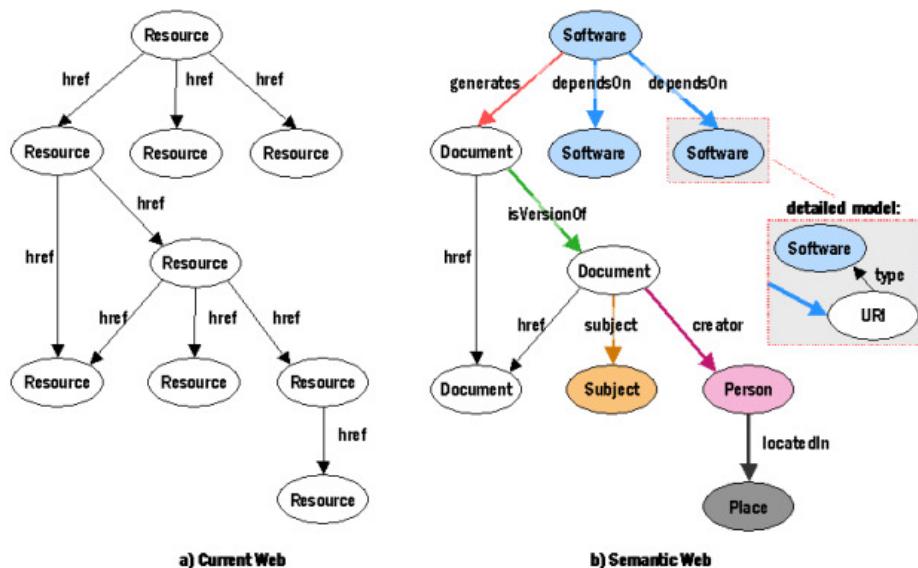


Figura 2 - I principi del Web Semantico - Koivunen, Miller 2001

- **L'informazione parziale è tollerata.** Il Web attuale è democratico e decentralizzato, nel senso che qualsiasi utente può facilmente definire tutti i link che ritiene opportuni tra le proprie risorse Web, dunque, contenuti, informazioni e documenti. Tuttavia, spesso si tratta di collegamenti ipertestuali soltanto “parziali” e incompleti, nel senso che risultano inefficaci e poco validi dal punto di vista semantico (ad esempio, un link che collega un documento dedicato al Web Semantico con un'immagine che rappresenta il Web 1.0), oltre che in alcuni casi non più attivi. Si tratta di link, dunque, che rischiano di creare ambiguità e difficoltà semantiche rilevanti per gli utenti. Anche il Web Semantico sarà democratico e decentrato, dal momento che tutti gli *Internet user* saranno in grado di mettere in rete documenti e creare link fra le risorse. La struttura dell'informazione sarà ancora frammentata e “parziale”; tuttavia, gli strumenti e le tecnologie che saranno messi a punto per la realizzazione effettiva del Web Semantico saranno in grado di tollerare parzialità e incompletezza delle informazioni on line e capaci, quindi, di garantire alle



macchine di comprendere il significato delle relazioni, anche se incomplete e inesatte dal punto di vista logico e semantico;

- **Non c'è bisogno della verità assoluta.** La veridicità delle informazioni disponibili sul Web Semantico non sarà dimostrabile in assoluto, come già accade nel Web attuale. Tuttavia, il Web Semantico sarà basato su applicazioni potenzialmente in grado di creare una «*macchina della fiducia*» - così come definita da Berners-Lee - concetto in base al quale i computer e i software saranno capaci di verificare in automatico - senza l'intervento degli utenti - l'affidabilità dei documenti e dei contenuti presenti on line attraverso interpretazioni e inferenze;
- **L'evoluzione è tollerata.** Come accade per il Web in versione *machine readable*, anche nel Web Semantico documenti, informazioni e contenuti saranno sempre soggetti ad evoluzioni e mutamenti di significato, dal momento che il Web è inserito all'interno di un contesto sociale (ad esempio, un termine come "rete sociale" può assumere differenti significati a seconda del contesto e dell'epoca in cui è inserita). Inoltre, dal momento che il Web è decentralizzato - e lo sarà anche nella versione *machine understandable* - per una stessa risorsa possono esistere diversi significati; alcuni potrebbero prevalere su altri ma, in assenza di un'autorità centrale che standardizzi definizioni e significati per ciascuna risorsa, il Web Semantico dovrà essere in grado di gestire tutti i differenti vocabolari messo in rete dalle diverse comunità che ne possano fare uso. Ambiguità e incertezza circa i significati descritti per le varie risorse dovranno essere gestite nel migliore dei modi dalle tecnologie e applicazioni del Web Semantico. L'infrastruttura del Web definito 3.0, dunque, dovrà essere in grado di gestire il processo di evoluzione e aggiornamento delle informazioni in rete e dei relativi significati, non sostituendo quelli nuovi a quelli vecchi, ma integrandoli e coordinandoli;
- **Design minimalista.** Nella definizione tecnica del Web Semantico, il W3C ha inteso proporre una infrastruttura meno standardizzata possibile del Semantic Web, con l'obiettivo di facilitare lo sviluppo di uno spazio aperto alle



innovazioni e alle nuove applicazioni on line. Come già evidenziato, tale infrastruttura fa riferimento ai quattro livelli (dati, schema, ontologico e logico), che disegnano un'architettura minimalista del Web Semantico.



2. I LINGUAGGI DEL WEB SEMANTICO

2.1. IL CONCETTO DI MARKUP

Il termine **markup** (o marcatura) indica un insieme di informazioni riguardanti un testo o i singoli elementi che lo compongono. In linea generale, la tecnica di composizione del testo utilizzando marcatori o etichette (*i tag*) richiede la definizione di una serie di convenzioni, ovvero di un **linguaggio di markup** o linguaggio di marcatura di documenti. Si tratta di un linguaggio che descrive i meccanismi di rappresentazione del testo (ad esempio il colore, il font e la struttura) che, attraverso convenzioni standardizzate, sono utilizzabili su più supporti e rendono esplicite particolari interpretazioni del testo stesso. Il markup o etichetta è separato dal contenuto, ed eventualmente può lui stesso contenerlo.

I linguaggi di markup possono essere diversi e si distinguono per differenti criteri, come la rappresentazione del testo, la struttura di base, il significato degli elementi che lo compongono e la visualizzazione o formattazione. Nell'ambito dello sviluppo del Web Semantico, quelli più utilizzati sono i **linguaggi di markup di tipo descrittivo**, come ad esempio SGML, HTML e XML. Questi linguaggi di marcatura consentono di separare il testo (formattazione) dal contenuto (marcatura); infatti, generalmente quello che si produce con un linguaggio di markup descrittivo non è il documento stesso, ma un file di codice e testo che deve essere poi interpretato da un'applicazione (nel caso dell'HTML il browser).

SGML (Standard Generalized Markup Language) è stato il primo metalinguaggio di markup descrittivo standardizzato a livello internazionale, che ha definito dei metodi di rappresentazione del testo in forma elettronica in modo indipendente dall'hardware e dall'applicazione utilizzati. La potenza di SGML, dunque, è la flessibilità: è uno standard che può essere applicato ad ogni tipo dato, è non proprietario e soprattutto indipendente dalla piattaforma utilizzata, caratteristica che consente di strutturare e marcire i documenti in maniera indipendente dall'applicazione utilizzata.

2.2. XML - EXTENSIBLE MARKUP LANGUAGE

Attualmente il Web è un enorme contenitore di informazioni e l'utilizzo del linguaggio HTML (Hypertext Markup Language) ne ha facilitato sicuramente lo sviluppo. Si tratta tuttavia di un linguaggio che ha come unico obiettivo quello di curare la formattazione grafica dei documenti del Web e non fornisce nessun supporto per strutturare e descrivere le informazioni in essi contenute, dal punto di vista semantico. Questo linguaggio - considerato la base del World Wide Web - consente infatti di creare in maniera standardizzata pagine di informazioni formattate in grado di raggiungere, tramite Internet, un numero di utenti in costante aumento. Insieme al protocollo HTTP (Hypertext Transport Protocol), HTML ha rivoluzionato il modo in cui le persone inviano e ricevono informazioni, ma lo scopo principale per cui è stato realizzato è la visualizzazione dei dati. Per questo motivo, l'HTML fa riferimento al modo in cui le informazioni vengono presentate e non al tipo o alla struttura di tali informazioni.

Il linguaggio HTML era nato per permettere agli utenti del World Wide Web di scrivere con semplicità documenti e condividerli su sistemi differenti: il presupposto era che quelle informazioni fossero costituite da un testo con immagini e collegamenti ipertestuali. Attualmente, tuttavia, le informazioni sul Web sono database di testo, immagini, suoni, video, audio; nel contesto di un rinnovato e più complesso Web, occorre un linguaggio che sia in grado di descrivere tipi di informazioni differenti e specifiche, definire relazioni complesse di collegamenti fra documenti, trasmettere informazioni in diversi formati. Inoltre, negli ultimi anni il notevole sviluppo del Web ha portato la necessità di avere dei linguaggi e degli strumenti che permettano il recupero intelligente dei dati e di organizzarli efficientemente, dal punto di vista logico e semantico.

In questa prospettiva si inserisce **XML - Extensible Markup Language**, metalinguaggio che permette di creare e di definire sintatticamente dei linguaggi personalizzati di markup. Si tratta di un linguaggio aperto e basato su testo, che in termini pratici consente di esplicitare la struttura di un documento dal punto di vista

sintattico, attraverso marcatori all'interno del testo; in questa prospettiva, XML fornisce informazioni di tipo strutturale relative ai dati veri e propri (in particolare, si fa riferimento a dati strutturati, come quelli presenti in archivi e tabelle). Queste informazioni rappresentano “dati sui dati”, cioè metadati, che offrono un contesto aggiuntivo all'applicazione che utilizza i dati e consente un nuovo livello di gestione e manipolazione delle informazioni sul Web. XML, dunque, permette di descrivere i documenti della rete dal punto di vista sintattico, rendendoli comprensibili alle macchine oltre che all'uomo. Un documento XML è organizzato secondo una struttura logica ad albero e non presenta un insieme predefinito di elementi, ma permette di crearne dei nuovi a seconda delle esigenze. Esiste un elemento radice che deve essere necessariamente presente. Ogni elemento può avere dei figli e si possono specificare degli attributi che ne definiscono particolari caratteristiche. La sintassi del metalinguaggio XML è uguale a quella dell'HTML: ogni elemento è composto da un tag di apertura e da un tag di chiusura, gli attributi vengono elencati all'interno del tag di apertura, sono opzionali e non si ha un limite massimo nella loro quantità.

Inoltre, il metalinguaggio XML è indipendente dalla piattaforma di utilizzo, completamente testuale, autodescrittivo e strutturato. In considerazione di tali caratteristiche, XML può essere utilizzato anche come mezzo di trasmissione dei dati tra differenti applicazioni che supportano l'accesso ad un database. Le informazioni contenute nel database di un'applicazione possono essere tradotte in un documento XML e inviate ad altre applicazioni, che procedono ad elaborarle.

In linea generale, dunque, il metalinguaggio di markup XML si differenzia dal linguaggio HTML per quattro aspetti fondamentali:

- in primo luogo, XML è stato concepito per descrivere sintatticamente le informazioni di un documento o di un database, mentre HTML è stato ideato soltanto per visualizzarle;



- con XML la struttura di un documento può essere descritta in modo gerarchico, attraverso una struttura ad albero formata da tag nidificati in differenti livelli di complessità;
- ogni documento XML può contenere una descrizione opzionale della sua grammatica, in modo che possa essere utilizzata da applicazioni che richiedono una validazione della struttura del documento;
- con XML gli utenti possono definire nuovi tag e attributi all'interno di ciascun documento;
- il linguaggio XML consente ai singoli utenti, alle aziende e alle organizzazioni di creare il proprio linguaggio di markup, specifico per il tipo di informazione che si intende trattare. Per molte applicazioni e per diversi settori di business, ad esempio, sono stati già sviluppati linguaggi di markup specifici, come ad esempio il Channel Definition Format e il Mathematical Markup Language.

XML risponde all'esigenza del W3C di portare nel World Wide Web lo Standard Generalized Markup Language (SGML), lo standard internazionale per la descrizione della struttura e del contenuto di documenti elettronici di qualsiasi tipo. XML sfrutta tutta la potenza espressiva di SGML, ma non contiene tutte le complesse funzioni che raramente vengono utilizzate. Si caratterizza quindi per la semplicità con cui è possibile scrivere documenti, condividerli e trasmetterli nel Web. Nella fattispecie, lo sviluppo dell'XML è stato portato avanti dal W3C con i seguenti obiettivi progettuali:

- **XML deve essere utilizzabile in modo semplice su Internet.** Il metalinguaggio XML deve operare in maniera efficiente sul Web per soddisfare le esigenze delle applicazioni eseguite in un ambiente di rete distribuito;
- **XML deve supportare un gran numero di applicazioni.** Tale metalinguaggio dovrebbe essere utilizzato attraverso un'ampia gamma di applicazioni, tra cui



strumenti di creazione do dati e informazioni, motori per la visualizzazione di contenuti, strumenti di traduzione e applicazioni di database;

- **XML deve essere compatibile con SGML.** Tale obiettivo è stato definito sulla base del presupposto che un documento XML valido debba anche essere un documento SGML valido, in modo tale che gli strumenti SGML esistenti possano essere utilizzati con l'XML e siano in grado di analizzare il codice XML;
- **Lo sviluppo di programmi che elaborino documenti XML deve risultare semplice.** L'adozione del linguaggio XML è ovviamente strettamente legato alla disponibilità di strumenti; negli anni la proliferazione di tali supporti è la dimostrazione che tale obiettivo sia stato raggiunto;
- **I linguaggi di markup devono essere personalizzabili.** La straordinaria potenza di XML è rappresentata dalla possibilità di creare linguaggi di markup personalizzati. Quando molti utenti concordano su un linguaggio di markup, è possibile creare browser o applicazioni personalizzate per gestire questi linguaggi;
- **Il numero di caratteristiche opzionali dell'XML deve essere mantenuto al minimo possibile.** Al contrario dell'SGML, l'XML elimina le opzioni, in tal modo qualsiasi elaboratore potrà pertanto analizzare qualunque documento XML, indipendentemente dai dati e dalla struttura contenuti nel documento stesso;
- **I documenti XML dovrebbero essere leggibili da ogni utente e ragionevolmente chiari.** Dal momento che utilizza il testo per descrivere i dati e le relazioni tra i dati stessi, l'XML è più semplice da utilizzare e da leggere del formato binario che esegue la stessa operazione; inoltre, poiché il codice è formattato in modo diretto, è necessario che l'XML sia facilmente leggibile da parte sia degli utenti che dei computer. In questo modo si facilita lo scambio informativo tra le varie applicazioni e si superano gli ostacoli imposti dai nuovi formati proprietari dei dati;
- **La progettazione di XML dovrebbe essere rapida.** L'XML è stato sviluppato per soddisfare l'esigenza di un linguaggio estensibile per il Web. Questo



obiettivo è stato definito dopo aver considerato l'eventualità che se l'XML non fosse stato reso disponibile rapidamente come metodo per estendere l'HTML, altre organizzazioni avrebbero potuto provvedere a fornire una soluzione proprietaria, binaria o entrambe;

- **I documenti XML devono essere facili da creare.** Dal momento che si tratta di un linguaggio testuale, i documenti XML possono essere creati facendo ricorso a strumenti di semplice utilizzo, quali editor di testo normale.

Attualmente XML si configura come il linguaggio padre di molti linguaggi di Markup, come ad esempio RDF e OWL, le due componenti centrali dell'architettura del Web Semantico, insieme all'XML stesso. Nella fattispecie, tali linguaggi ereditano le caratteristiche dell'XML, ma risultano progettati ad hoc per determinate aree di lavoro.

Come già rilevato, XML consente di descrivere la sintassi di un documento mediante marcatori. In linea generale, si tratta di una sintassi molto simile a quella dell'HTML, anche se molto più rigida. In particolare la definizione della sintassi di un documento XML può essere espressa mediante una DTD o un XML Schema, ai quali si aggiunge un namespaces XML, di seguito brevemente descritti:

- **DTD - Document Type Definition.** Si tratta di un modello sintattico, che descrive la struttura e il contenuto di un documento XML. Contiene le definizioni dei tag e degli elementi, l'ordine nel quale devono apparire all'interno del documento e formalizza la struttura ad albero. Inoltre una DTD contiene la definizione degli attributi e del loro tipo, l'obbligatorietà e il valore predefinito e include la definizione delle entità e delle notazioni. Le DTD impongono dei vincoli al documento XML affinché esso possa essere analizzato da software e applicazioni; in altri termini, stabiliscono le regole formali che governano la struttura di un documento, specificando quali tag devono essere presenti nel documento stesso, quali annidamenti sono possibili, le combinazionimesse di elementi e

attributi. Ad esempio, una DTD potrebbe stabilire che per ogni tag <libro> devono essere annidati i tag <titolo>, <autore>, <informazioni_editoriali>, senza i quali il documento non è formalmente corretto. Ciascun utente o azienda può creare un DTD personalizzato: è in questa direzione che l'XML ci configura come metalinguaggio, vale a dire un linguaggio di livello superiore attraverso il quale è possibile creare altri linguaggi;

- **XML Schema.** Un XML Schema è un'alternativa al DTD e permette di definire la struttura, il contenuto e la semantica di documenti XML. Mentre le DTD consentono una semplice tipizzazione dei dati per elementi ed attributi all'interno di un documento, gli schemi XML consentono una tipizzazione dei dati sistematica, ovvero permettono di definire un vocabolario XML che può essere utilizzato per descrivere documenti XML facendo ricorso sempre alla stessa sintassi. Un documento collegato ad uno schema XML deve obbligatoriamente rispettare le regole stabilite dallo schema per poter essere convalidato come schema valido;
- **Namespaces XML.** Spesso si può avere la necessità di utilizzare elementi definiti diversamente da DTD. È possibile che il nome di un elemento presente in una DTD sia lo stesso di un elemento di un'altra DTD. Questa è la classica situazione di collisione fra nomi di elementi e di attributi. È possibile risolvere il problema tramite l'utilizzo dei *namespace*. Un *namespace* è una raccolta di nomi identificati da una URI, utilizzata nel documento XML, nel momento in cui ci si riferisce ad un certo elemento, eliminando il rischio di ambiguità.

Infine, un accenno ai **fogli di stile**, necessari per la visualizzazione di un documento XML su un qualsiasi browser. Il foglio di stile è un documento esterno, che può essere anche collocato sul Web, generalmente utilizzato per applicare in modo coerente stili o formattazione ai documenti. Nello specifico il foglio di stile consente di modificare l'aspetto di ciascun elemento di un documento o di una serie di documenti, per esempio il colore, le dimensioni e il tipo del carattere, lo sfondo, i bordi, le immagini, i file audio e video, separando la presentazione del documento dal contenuto del documento stesso. Il



tipo di foglio di stile più utilizzato sul Web è basato sulla specifica dei fogli di stile CSS (Cascading Style Sheets), ma lo standard più recente è XLS (Extensible Stylesheet Language). L'XLS è un sottoinsieme del Document Style Semantics and Specification Language (DSSSL), il linguaggio di stile utilizzato in ambiente SGML: gode delle proprietà di essere estensibile, potente ma nello stesso tempo di facile utilizzo. Con l'XSL è possibile creare fogli di stile che permettono la visualizzazione di un documento XML in un qualsiasi formato.

2.3. RDF E METADATI

2.3.1. I Metadati

Tutta la famiglia di linguaggi XML ha lo scopo di generare dati che non siano soltanto destinati alla lettura, ma che possano essere utilizzati da software e applicazioni. XML risponde alle esigenze di standardizzazione sintattica per cui è utilizzabile da diverse applicazioni e piattaforme come linguaggio di interscambio. Tuttavia, come più volte rilevato nel corso della dispensa, affinché i dati vengano elaborati da computer e software è necessario che siano accompagnati da informazioni riguardanti i dati stessi, che ne specifichino il contesto e ne descrivano i contenuti. Tali informazioni prendono il nome di "metadati". Le meta-informationi, infatti, permettono agli utenti di specificare informazioni sui documenti creati o fruiti che siano non soltanto leggibili, ma anche interpretabili dalle applicazioni di rielaborazione e dai motori di ricerca.

Il termine metadato significa letteralmente "dato circa un altro dato". Questa definizione è la più basilare, ma è senz'altro la più esplicativa del concetto: i metadati infatti sono informazioni a supporto delle informazioni primarie, ovvero tutta quella serie di dati che il Web gestisce e mette a disposizione degli utenti. Un esempio tipico di metadati è costituito dalla scheda del catalogo di una biblioteca, la quale contiene informazioni circa il contenuto e la posizione di un libro: questi sono dati riguardanti gli elementi del libro cui si fa riferimento mediante la scheda e dunque sono alla base del processo di reperimento del libro in quel sistema informativo che è la biblioteca.

Tim Berners-Lee definisce formalmente i metadati come «*informazioni sulle risorse Web comprensibili ai calcolatori*»⁹. Essi quindi svolgono una funzione fondamentale nei processi di reperimento e di elaborazione delle informazioni. Essendo informazioni, inoltre, i metadati possono essere presenti e gestiti in tre modi. nei documenti stessi a cui sono associati; in documenti diversi, che vengono immediatamente associati all'oggetto a cui sono riferiti; in documenti diversi che possono essere trattati anche separatamente.

I metadati, dunque, sono potenzialmente in grado di migliorare e ottimizzare le ricerche di risorse e contenuti su Internet. In ottica di Web Semantico, l'inserimento di metadati consentirebbe di dare un senso logico alle risorse on line e offrirebbe alle macchine l'opportunità di ragionare sui dati e interpretarli. Il vantaggio più evidente sarebbe la possibilità di effettuare ricerche sempre più precise ed efficaci, non più basate semplicemente sulle keyword e una effettiva condivisione delle informazioni disponibili tra le diverse applicazioni Web.

Tuttavia, quella proposta rappresenta una definizione riduttiva delle finalità dei metadati, che viceversa possono essere applicati a vari contesti e soddisfare differenti esigenze. Infatti, oltre che per facilitare e rendere più mirata la ricerca, i metadati possono e devono essere utilizzati per fornire informazioni sulla risorsa, allo scopo di garantirne la fruibilità nel tempo. Si parla in questo caso di metadati conservativi e il campo d'azione si sposta dall'universo Web a quello del materiale acquisito in formato digitale con scanner, fotocamera o simili e temporaneamente registrato su un supporto (come un cd recordable). Sono soprattutto le preoccupazioni sulla durata di supporti, attrezzature e software di lettura a spingere i tecnici verso l'adozione di misure che riducano il pericolo di perdita di contenuti informativi importanti. Si rende necessario, pertanto, catalogare la risorsa digitale indicando non solo i suoi contenuti, ma anche le tecniche adottate per la sua creazione, dalle attrezzature utilizzate al formato, fino agli utenti che hanno eseguito materialmente i lavori.

⁹ T. Berners-Lee, D. Connolly, R. Swick, *Web Architecture: describing and exchanging data*, 1999, su <http://www.w3.org>



L'utilizzo di metadati assume un rilievo significato e in una prospettiva a trecentosessanta gradi anche all'interno delle aziende e delle organizzazioni. Tutte le aziende infatti sono dotate di un sistema Data Warehouse, la cui funzione principale è quella di rendere disponibili i dati per un sistema informativo a supporto dei processi decisionali di più alto livello di un'organizzazione, con l'obiettivo di consentire un miglioramento della pianificazione strategica e del controllo direzionale al fine di promuovere la competitività aziendale sul mercato. Nello specifico, i fattori che hanno determinato la necessità di integrare i metadati nei processi decisionali possono essere sintetizzati nei punti seguenti¹⁰:

- la mancanza di integrazione e flessibilità nei sistemi odierni. Il mercato globale, la competizione serrata e l'aumentare della complessità delle richieste dei clienti hanno indirizzato le aziende verso sistemi integrati e più flessibili in cui i cambiamenti siano all'ordine del giorno e quindi ben assimilabili. Un archivio di metadati (metadata repository) riduce significativamente sia il tempo che il costo di analisi e di sviluppo, documentando le regole di trasformazione di dati, le fonti, le strutture dei dati stessi ed il contesto del loro Data Warehouse;
- la necessità di miglioramento continuo e crescita dei Data Warehouse. I metadati sono diventati uno strumento di supporto fondamentale e le risorse ad essi connesse sono vertiginosamente aumentate negli ultimi anni, rappresentando per le aziende una soluzione nuova il cui sviluppo non è facilmente prevedibile;
- i bisogni degli utenti non sempre soddisfatti. I metadati sono divenuti la chiave del rapporto tra il lato tecnico dell'azienda e quello del business, traducendo il linguaggio dell'IT in termini comprensibili ai business men;
- la necessità per le compagnie di ridurre l'impatto della turnazione degli assunti. Un impiegato rappresenta un patrimonio di non sole nozioni tecniche, ma soprattutto di esperienza e conoscenza aziendale maturate sul campo che i

¹⁰ *Ibidem.*



metadati permettono di immagazzinare e mettere a disposizione di chi si affaccerà per la prima volta a questa nuova realtà;

- l'esigenza di far crescere la confidenza degli uomini verso i dati e le informazioni. I metadati creano un "layer semantico" tra dati e utenti che permette a questi ultimi di poter maneggiare le informazioni con molta più facilità e fiducia.

2.3.2. RDF - Resource Description Framework

I metadati, dunque, apportano un contributo fondamentale al miglioramento dell'accesso all'informazione. L'uso efficace dei metadati, tuttavia, richiede che vengano stabilite delle convenzioni per la semantica, la sintassi e la struttura dei documenti e delle risorse on line: la sintassi, vale a dire l'organizzazione sistematica dei dati per l'elaborazione automatica, facilita lo scambio e l'utilizzo dei metadati tra applicazioni diverse; la struttura, invece, può essere vista come un vincolo formale sulla sintassi, per una rappresentazione consistente della semantica.

A riguardo, come già analizzato, XML fornisce informazioni semantiche mediante un meccanismo per la definizione della struttura di un documento. Tale metalinguaggio definisce una struttura ad albero per i documenti, in cui ciascun nodo individua un tag ben definito, mediante il quale è possibile interpretare le informazioni che esso racchiude. La semantica di un documento XML non è però specificata in modo esplicito, ma è "incorporata" nei nomi dei tag.

Viceversa, lo strumento base per la codifica, lo scambio e il riutilizzo di metadati strutturati - e dunque per introdurre la semantica all'interno dei documenti del Web - è **RDF - Resorce Description Framework**, linguaggio figlio di XML progettato con l'obiettivo di aggiungere meta-information ai documenti Web e in generale alle risorse on line. In questa direzione, RDF permette l'aggiunta di semantica ai contenuti e ai documenti del Web, senza fare alcuna assunzione sulla sua struttura. La caratteristica principale dell'RDF è quella di consentire l'interoperabilità semantica tra applicazioni che si scambiano informazioni sul Web. RDF consente, dunque, la costruzione della

struttura semantica all'interno dei documenti, superando il limite maggiore di XML che, pur essendo uno strumento potentissimo per la strutturazione dei documenti e per la creazione di applicazioni e nuovi linguaggi, permette soltanto l'organizzazione delle informazioni di un documento e non dice nulla sul significato del relativo contenuto. In altri termini, mentre XML stabilisce una sintassi di base di un documento ma non fa riferimento alla semantica, RDF aggiunge meta-informationi ai documenti stessi, attraverso le quali fornisce informazioni sul contenuto e sul significato. In sintesi, XML supporta l'interoperabilità sintattica, mentre RDF mira all'interoperabilità semantica.

In linea generale, Resource Description Framework è una cornice, un framework, che stabilisce in che modo l'informazione possa essere rappresentata sul Web. «*RDF is a foundation for processing metadata; it provides interoperability between applications that exchange machine-processable information on the Web. Basically RDF defines a data model for describing machine-processible semantics in data*» (Broekstra et al., 2002).¹¹

Lo sviluppo di RDF nell'ambito del Web Semantico è stato condotto per consentire i seguenti utilizzi¹²:

- Web Metadata: fornire informazioni su risorse Web e i sistemi che utilizzano (restrizioni per la privacy, descrizioni di capacità, valutazione del contenuto);
- Applicazioni che richiedono modelli dell'informazione aperti invece che chiusi (annotazione di risorse Web, descrizione di processi organizzativi, attività di avvicendamento);
- Fare per l'informazione *machine processable* (le applicazioni di dati) quello che il Web ha fatto per l'ipertesto. Si tratterà di consentire che l'informazione sia processabile anche in ambienti diversi da quelli della sua creazione;
- Consentire che le applicazioni possano lavorare in interdipendenza, rendendo possibile la combinazione di nuovi dati per creare nuova informazione;

¹¹ J. Broekstra, *Enabling knowledge representation on the Web by extending RDF Schema*, 2002, su <http://www.sciencedirect.com>.

¹² Cfr., G. Klyne, J. Carroll, *Rdf concepts and abstract syntax*, 2004, su <http://www.w3.org>.

-
- Elaborazione automatica delle informazioni per gli agenti software; il Web dovrebbe trasformarsi da un ambiente caratterizzato esclusivamente da informazione *human-readable* ad un network di processi di cooperazione attraverso il linguaggio RDF. «*RDF is designed to represent information in a minimally constraining, flexible way. It can be used in isolated applications, where individually designed formats might be more direct and easily understood, but RDF's generality offers greater value from sharing. The value of information thus increases as it becomes accessible to more applications across the entire Internet*»¹³.

Quindi, la progettazione di RDF ha puntato a realizzare alcuni obiettivi quali lo stabilire un modello di dati molto semplice, facile da processare e manipolare per le applicazioni; stabilire una semantica formale ed inferenze verificabili; la formazione di un vocabolario basato sulle URI, che sono utilizzate per nominare qualsiasi cosa sul Web con RDF; l'uso di una sintassi basata su XML per codificare i dati per le interazioni fra le applicazioni; l'uso di valori basati sugli schemi XML, in modo da poter comunicare con ogni altra applicazione XML; infine, permettere a chiunque di compiere asserzioni su qualsiasi risorsa, asserzioni che potrebbero anche essere incomplete ma che le applicazioni che usano RDF dovrebbero essere in grado di gestire.

Rappresentazione Grafica di RDF

Una delle rappresentazioni di RDF è quella in sintassi XML. Dunque, ogni documento RDF è anche un documento XML. Tuttavia, mentre un documento XML rappresenta una struttura ad albero, RDF rappresenta un grafo i cui nodi rappresentano i concetti del

¹³ *Ibidem*. Di seguito si presenta la traduzione del testo: «*RDF è disegnato per rappresentare le informazioni in un modo flessibile col minimo di restrizioni. Può essere usato in applicazioni isolate, dove formati progettati individualmente possono essere compresi più direttamente e facilmente, ma le affermazioni generali di RDF offrono un valore più grande per la condivisione. Il valore dell'informazione quindi cresce col diventare accessibile a più applicazioni attraverso l'intera Internet*

dominio che si intende descrivere e gli archi rappresentano le relazioni tra i concetti. Appare dunque chiaro che XML si pone come base per la creazione di nuovi linguaggi, mentre RDF ha invece l'obiettivo fornire un valido supporto per modellare domini.

RDF fornisce un modello per descrivere le risorse, che possono avere delle proprietà (o anche attributi o caratteristiche). Per RDF una risorsa è un qualsiasi file accessibile sul Web mediante un URI o un URL. Nello specifico RDF è basato su tre tipi di oggetti o elementi costitutivi:

- **Risorsa (Resource):** ogni oggetto che può essere identificato attraverso un URI. Questo include le pagine e i siti, elementi specifici all'interno di pagine Web, immagini, server ed anche oggetti, come ad esempio libri stampati non direttamente fruibili sul Web;
- **Proprietà (Property):** è un aspetto specifico, caratteristica, attributo o relazione utilizzata per descrivere una risorsa. Ogni proprietà possiede un proprio significato, definisce i valori che possono essere attribuiti, il tipo di risorse che può descrivere e le relazioni che possono instaurarsi con altre proprietà. Tali definizioni vengono realizzate attraverso l'uso degli RDF Schema;
- **Asserzioni (Statements):** ogni risorsa specifica con una proprietà ed il valore che tale proprietà assume, costituisce un'asserzione RDF. Un'asserzione è basata su triple: soggetto (risorsa da descrivere), predicato (proprietà), oggetto (affermazione). In altre parole, un oggetto O ha proprietà P con valore V. Questo corrisponde alla relazione che può essere scritta come $P(O) = V$ oppure $P(O, V)$. Ad esempio: Autore (<http://www.pincopallino.com>) = PincoPallino P O V afferma che l'autore di <http://www.pincopallino.com> è Pinco Pallino. In un documento RDF si fanno asserzioni secondo cui determinati oggetti hanno proprietà con certi valori. Questa struttura è un modo naturale per descrivere la maggior parte dei dati processati dalle macchine.



Con particolare attenzione alle Asserzioni, graficamente ognuna delle triple costituenti la struttura d base di ogni espressione RDF viene definita “grafo RDF”, rappresentato da archi e nodi. Ogni tripla rappresenta una dichiarazione di relazione tra gli elementi denotati tra i nodi che sono collegati nella tripla. Quindi, la tripla contiene tre parti, rappresentate dai nodi (il Soggetto e l’Oggetto) e dagli archi (il Predicato), che denotano il tipo di relazione esistente. Soggetto e Oggetto possono essere una risorsa o una stringa di caratteri, mentre il predicato è la proprietà che definisce la relazione fra soggetto (chiamato appunto Risorsa) e oggetto dell’asserzione. La Risorsa deve sempre essere descritta attraverso una URI; la Proprietà stabilisce una caratteristica, un attributo, una relazione che può essere usata per descrivere la Risorsa; l’Oggetto è il Valore della proprietà e potrebbe essere un’altra risorsa o una stringa di caratteri. La relazione fra Soggetto, Predicato e Oggetto è la relazione esistente fra una Risorsa, una Proprietà e il Valore di questa stessa proprietà per la risorsa. Quando viene stabilita una relazione di questo tipo si è posta una “istruzione RDF”.

RDF Schema

Per essere applicato in concreto nella definizione e nell’interscambio di metadati, RDF necessita di una sintassi. Il W3C ha proposto XML come sintassi di scambio per RDF (RDF/XML Syntax Specification) e permettere attraverso “XML namespace” l’associazione delle proprietà con lo schema RDF che le definisce.

Scrivere un documento con RDF/XML implica la necessità di creare dei vocabolari specifici per definire classi di risorse e le proprietà che definiscono queste stesse risorse. In questa direzione, così come XML Schema permette di definire vincoli sulla struttura dei documenti XML, RDF Schema consente la definizione di un vocabolario che può essere usato per le proprietà RDF e la specifica dei tipi di oggetti ai quali queste proprietà possono essere applicate. In altre parole, RDF Schema non è altro che un insieme di risorse RDF che possono essere utilizzate per descrivere proprietà o altre risorse RDF. Una delle caratteristiche più interessanti di RDF Schema è la semplicità con cui può essere esteso mediante il meccanismo dei namespace. Non solo si può far uso di



diversi “vocabolari” RDF Schema, ma anche estendere una versione precedente senza dover ridefinire una gran mole di concetti. Con l’obiettivo di evitare problemi di ambiguità, ogni RDF Schema deve avere una propria URI, anche se si tratta dell’estensione di un altro schema.

2.4. LE ONTOLOGIE

2.4.1. *Introduzione*

Come già evidenziato in precedenza, allo stato attuale il livello finora più standardizzato nell’architettura del Web Semantico disegnata da Berners Lee è quello dei linguaggi di dati, rappresentato da XML e RDF come “core languages”. La creazione di XML, prima, e di RDF, poi, ha rappresentato soltanto il primo passo verso la creazione di una infrastruttura che si collochi in cima al Web per creare un ambiente di informazione processabile dalle macchine, in grado di sostenere l’evoluzione di tutte quelle attività che nel Web attuale sono limitate dalla natura intrinseca del WWW e dalle tecnologie principali che lo hanno caratterizzato. XML e RDF rappresentano il punto di partenza verso una rappresentazione della conoscenza attraverso il Web Semantico.

Lo sviluppo dell’XML, come visto in precedenza, ha rappresentato un passaggio fondamentale in direzione della interoperabilità sintattica, assicurando una sintassi di base per la rappresentazione di dati anche molto diversi fra loro in una maniera che le macchine, grazie ai fogli di stile, sono in grado di rappresentare indipendentemente dal software utilizzato. RDF, invece, presenta come obiettivo primario quello di definire un meccanismo per la descrizione di risorse, senza fare alcuna assunzione su un particolare dominio di applicazione e senza definire la semantica di alcun dominio. RDF in questo senso è un linguaggio che si pone come neutrale rispetto al campo in cui viene applicato, proprio perché deve potersi prestare a diversi usi e in ambiti differenti. Il linguaggio RDF rappresenta il gradino principale da cui far partire quella interoperabilità semantica che è richiesta alle tecnologie che formeranno il Web Semantico.



Rispetto a XML, RDF assicura una maggiore interoperabilità semantica perché le unità semantiche sono definite attraverso la struttura oggetto-attributo, cosicché con RDF un modello del contenuto, ai fini di una certa applicazione, viene rappresentato naturalmente senza dover compiere nessuna opera di traduzione. La differenza più importante fra XML e RDF, quindi, è che le espressioni scritte con XML non hanno una semantica interna e la sua semantica è determinata soltanto dal tipo di operazioni che è possibile fare con quel documento, mentre un documento RDF ha una semantica dichiarativa che fa sì che il significato di un documento RDF, processabile dalla macchina, sia indipendente dal software.

La definizione di RDF presenta dei vantaggi importanti nella rappresentazione delle risorse Web in termini semanticci, tra cui il più rilevante consiste nella possibilità di consentire la costruzione e l'impiego delle ontologie necessarie alla macchina per l'elaborazione delle informazioni. Il "livello ontologico" del Web Semantico è quello che raccoglie attualmente gli sforzi maggiori, perché centrale nell'opera di sviluppo di un "Knowledgeable Web", in cui la gestione della conoscenza e il suo trattamento da parte delle macchine siano predominanti. I linguaggi e le tecnologie principali sviluppati fino ad ora, che costituiscono già degli standard, arrivano a toccare lo strato ontologico del Web Semantico ma non lo hanno ancora costruito.

La nascita di questo livello dipenderà dalla possibilità concreta di creare delle ontologie (grazie ai linguaggi appositamente definiti), che possano essere utilizzate al fine di fornire un serbatoio di significati formali attraverso i quali i computer possano compiere inferenze e trattare i dati in modo da capirne un significato *machine processable*.

«For the semantic web to function, computers must have access to structured collections of informations and sets of inference rules that they can use to conduct automated reasoning. Artificial-intelligence researchers have studied such systems since long before the Web was developed. Knowledge Representation [...] is currently in a state comparable to that of hypertext before the advent of the Web: it is clearly a good idea, and some very nice



demonstrations exist, but it has not yet changed the world. It contains the seeds of important applications, but to realize its full potential it must be linked into a single global system»¹⁴

I sistemi di rappresentazione della conoscenza tradizionali, basati sul campo di studi dell’Intelligenza Artificiale, hanno tipicamente avuto una struttura centralizzata che richiede ad ognuno di condividere esattamente le stesse definizioni dei concetti. Questa centralizzazione comporta che un sistema di questo tipo limiti il tipo di domande a cui la macchina può rispondere soltanto a quelle nella forma e per i termini previsti dal sistema, non potendo quindi rendere conto delle domande che non rispettano queste condizioni. In questo senso il Web Semantico dovrebbe superare questo limite, proprio a partire dalla caratteristica principale del Web stesso, quella di essere “grande quanto il mondo”.

Al momento attuale, quindi, dato che alcuni linguaggi già sono standardizzati, lo sforzo maggiore sta nella creazione delle ontologie e del livello logico che dovrà sostenere le regole di inferenza dei computer.

2.4.2. Caratteristiche delle Ontologie

Il termine ontologia rimanda a campi di studi e a settori del sapere anche molto lontani fra loro. Con particolare attenzione all’ambito informatico, una ontologia è una descrizione dei concetti e delle relazioni che intercorrono fra di essi. Nella fattispecie, si tratta di un documento che indica in modo formale il significato e i legami fra termini distinti. Più precisamente, introdotta in una specifica realtà, individua gli aspetti che sono ritenuti rilevanti nel contesto di riferimento e quelli che possono essere ignorati: specifica i concetti e come sono legati tra loro, quali sono le proprietà di ciascuno e come

¹⁴ T. Berners Lee, Handler, Lassila, op. cit. Di seguito si presenta la traduzione del testo: «*per far funzionare il Web Semantico, i computer devono avere accesso a collezioni strutturate di informazioni e set di regole di inferenza che possano usare per condurre un ragionamento automatizzato. I ricercatori dell’Intelligenza Artificiale hanno studiato tali sistemi da molto tempo prima che il Web fosse sviluppato. La Rappresentazione della Conoscenza [...] è attualmente in uno stato paragonabile a quello dell’ipertesto prima dell’avvento del Web: è chiaramente una buona idea, e alcune dimostrazioni molto buone esistono, ma non ha ancora cambiato il mondo. Contiene i semi di importanti applicazioni, ma per realizzare il suo pieno potenziale deve essere collegata all’interno di un singolo sistema globale».*

queste proprietà sono connesse, mediante regole d'inferenza e logiche. In altri termini, l'ontologia permette di lavorare con un insieme strutturato di concetti dove le relazioni risultino chiare e, soprattutto, significative a livello semantico.

Una ontologia è formata da una raccolta di nomi per i concetti che si vuole definire e dalle relazioni di ordinamento tipo-sottotipo che si vogliono stabilire per un dato dominio o realtà che si intende descrivere e interpretare. Come già rilevato, una definizione rigorosa di ontologia è quella di Gruber per cui una ontologia è «*una specificazione formale ed esplicita della concettualizzazione condivisa di un dominio*»¹⁵.

Il termine “concettualizzazione” si riferisce ad un modello astratto di qualche fenomeno del mondo che identifica i concetti più rilevanti di quel fenomeno; con “esplicita” si intende che il tipo di concetti utilizzati e le restrizioni nel loro utilizzo sono definiti esplicitamente; per “formale” si intende il fatto che le ontologie debbano essere *machine understandable*; “condivisa” vuol dire che un’ontologia deve raccogliere un bagaglio di conoscenza consensuale, che non sia quindi limitata ad un qualche individuo specifico ma che sia accettata da una comunità.

L'introduzione delle ontologie nel Web consente la strutturazione delle informazioni e permette di superare alcuni aspetti critici del Web tradizionale. In primo luogo, le varie sorgenti “producono” informazioni in diversi formati e la creazione di indici per localizzare queste sorgenti è piuttosto complessa, poiché risulta molto difficile ottenere indicazioni da sorgenti audio o video: un'ontologia, viceversa, può facilitare questa operazione, descrivendo in modo formale i contenuti di ogni sorgente e supportando gli utenti nella ricerca di quelle sorgenti che generano un particolare tipo di informazione. Di conseguenza, l'introduzione del livello ontologico nel Web inteso in senso tradizionale consente di superare anche il problema della mancanza di struttura del Wwww, dovuta al fatto che HTML è un linguaggio di formattazione e non si occupa di gestire le informazioni per facilitare il loro reperimento: un'ontologia descrive il dominio dal punto di vista strutturale, definendone i componenti e i loro legami. Il terzo problema affrontato è quello della dipendenza dal contesto: definendo mediante

¹⁵ T. Gruber, *op. cit.*



un'ontologia il contesto in cui è presentato un certo documento, i termini utilizzati sono propri di quell'ambito e quindi il loro significato non è ambiguo.

Le ontologie configurano un settore di studio che coinvolge diverse aree della ricerca sull'intelligenza artificiale: la rappresentazione della conoscenza, lo studio dell'elaborazione del linguaggio naturale, il Knowledge Engineering, arrivando a toccare di recente campi quali il knowledge management, l'e-commerce, il recupero delle informazioni.

Lo studio sulle ontologie è stato prima orientato alla definizione di cosa fosse una ontologia e quali richieste una ontologia dovesse soddisfare. Il passo successivo nello sviluppo delle ontologie è stato quello della nascita di linguaggi di rappresentazione per la definizione e il cambiamento delle ontologie stesse. Allora si è trattato di guardare allo sviluppo del contenuto delle ontologie, alla creazione, quindi, di ontologie riutilizzabili nei sistemi di rappresentazione della conoscenza.

Il problema attualmente più urgente circa le ontologie (anche in riferimento alle esigenze del Web Semantico), per la creazione di domini di conoscenze condivise, è quello di assicurare l'evoluzione delle ontologie quando queste entrano a far parte di applicazioni concrete. In questa prospettiva, infatti, affinché le ontologie risultino efficienti nell'ottica del Web Semantico, devono per forza di cose subire un processo di cambiamento e aggiornamento continuo, visto che è continua la trasformazione della conoscenza condivisa su qualsiasi concetto, sia che cambino le categorie di interpretazione che il mondo di riferimento.

A partire dalla definizione di ontologia di Gruber si individuano anche i tre elementi che si sottopongono a cambiamenti e che causano le trasformazioni delle ontologie stesse. In particolare una ontologia può cambiare nel dominio, può cambiare nella concettualizzazione o, ancora, può cambiare nella specificazione esplicita. I cambiamenti di dominio dipendono dai cambiamenti del mondo reale, che sono una evoluzione del dominio, al quale devono corrispondere necessariamente gli aggiornamenti della descrizione di quel dominio. A differenti visioni del mondo, determinanti differenti scopi e compiti per la ontologia, corrispondono nuove concettualizzazioni che si rendono necessarie, perché quelle relative ad una diversa interpretazione del mondo, in base a



vecchi obiettivi, non sono più adatte. Infine, i cambiamenti nelle specificazioni esplicite avvengono quando una ontologia deve essere tradotta da un linguaggio di rappresentazione ad un altro, determinando un cambiamento non solo nella sintassi di una ontologia ma nella semantica stessa.

Nelle ontologie, accanto al concetto di evoluzione c'è anche quello di "versioning", ovvero la creazione di nuove versioni delle ontologie nel senso delle interfacce. Evoluzione e versioning fanno sì che si creino e si sostengano differenti versioni di una ontologia, ed è importante poter tracciare questo percorso di cambiamento nel tempo per far sì che a versioni diverse sia possibile comunque associare sempre tutte le informazioni, sia quelle vecchie che quelle nuove.

I cambiamenti delle ontologie comportano anche la necessità che differenti versioni di una ontologia siano tra loro compatibili, in modo che nessuna informazione vada perduta durante queste trasformazioni, neanche sulla ontologia stessa. Come affermano Noy e Klein, la compatibilità fra due ontologie va misurata rispetto alla conservazione dei casi che compongono la concettualizzazione del dominio (se c'è perdita di dati), alla conservazione dell'ontologia stessa (se la nuova ontologia costituisce un sovra-insieme della vecchia), alla conservazione delle conseguenze (se tutte le inferenze che potevano essere fatte con la vecchia ontologia possono essere fatte anche con la nuova), alla conservazione della consistenza (se viene mantenuta la consistenza logica di tutte le definizioni).

L'evoluzione delle ontologie può essere distinta in due modi: tracciata e non tracciata, a seconda se sono individuati o meno tutti i cambiamenti intercorsi fra una versione e l'altra. Seguendo l'evoluzione di una ontologia, tracciandone la trasformazione, considerando quindi l'evoluzione come una serie di cambiamenti, è possibile considerare gli effetti di questi cambiamenti sui casi dell'ontologia e sull'ontologia stessa. Nel caso, invece, non venga tracciata l'evoluzione quello che rimane sono due versioni di una ontologia e nessuna conoscenza sui passi che sono stati compiuti nel passaggio da una versione ad un'altra.

Infine, un'altra caratteristica delle ontologie è che queste si prestano ad essere riutilizzate ed estese ad altre ontologie, non venendo quindi confinate ad un sistema

specifico, ma un cambiamento in un vocabolario formale comporta delle trasformazioni in tutti gli altri cui esso è collegato. La relazione che si può stabilire fra ontologie diverse mostra che la natura delle ontologie è decentralizzata. Inoltre, il loro sviluppo, e quindi la loro evoluzione nel tempo, è anche collaborativo perché non viene stabilito un controllo centralizzato sui loro usi particolari. Le ontologie si prestano quindi ad usi di volta in volta diversi a seconda delle applicazioni che le utilizzano e delle altre ontologie a cui sono collegate.

2.4.3. Ontologie e Web Semantico

«Ontologies are an essential backbone technology because they provide an important feature: they interweave formal semantics understandable to a computer with real world semantic understandable to humans»¹⁶.

Le ontologie rappresentano probabilmente lo strato più importante verso la costruzione del Web Semantico perché soltanto con la creazione di ontologie, intese come insiemi strutturati di definizioni di vocabolario, potrà essere possibile la congiunzione fra il livello dei dati e quelli superiori. Le ontologie sono il tramite che permetteranno di collegare il Web attuale a strutture di rappresentazione della conoscenza basate sul significato, in modo da far entrare nel Web anche gli agenti intelligenti.

I linguaggi come XML e RDF non determineranno lo sviluppo del Web Semantico se non si creeranno le ontologie necessarie a supportare la distribuzione dell'informazione semantica e la sua elaborazione da parte delle macchine. Linguaggi come RDF costituiscono il punto di partenza per lo sviluppo delle ontologie. A partire da RDF Schema, linguaggi più evoluti sono stati messi a punto per la rappresentazione semantica delle informazioni, come l'OIL (Ontology Inference Layer) del progetto europeo On-toknowledge, e il DAML+OIL (DARPA Agent Markup Language+OIL), nato

¹⁶ Y. Ding, *The Semantic Web: yet another hip?*, 2002, su <http://www.sciencedirect.com>

dallo sforzo congiunto in America e in Europa di creare una base per la rappresentazione consensuale della conoscenza sul Web in base ad uno stesso standard. Entrambi i linguaggi sono basati su RDF Schema, di cui rappresentano delle estensioni. Il contributo di questi linguaggi a RDF Schema sta nell'aggiungere una semantica formale basata su descrizioni logiche e su primitivi più avanzati, come l'uso dei marcatori booleani (AND, OR, NOT, per esempio) e di alcuni assiomi logici.

Le ontologie sono fondamentali nell'ottica dello sviluppo del Web Semantico perché, come afferma Euzenat, «*esse permettono alle applicazioni di accordarsi sui termini che queste usano quando comunicano fra loro. Esse sono un fattore chiave per rendere possibile l'interoperabilità del Web Semantico*»¹⁷. Questa importanza assunta dalle ontologie rende il loro sviluppo indispensabile alla nascita del Web Semantico, visto che senza di esse manca proprio l'elemento chiave che dovrebbe consentire alle macchine di processare il significato formale delle informazioni. La creazione concreta delle ontologie, e la loro implementazione nelle applicazioni reali che dovranno essere messe a disposizione degli utenti del Web, richiederà di adottare delle metodologie di progettazione tali da tenere conto delle continue esigenze di evoluzione delle ontologie al cambiare del contesto e del loro uso nelle situazioni concrete, delle descrizioni del mondo.

2.4.4. I Linguaggi delle Ontologie

Le tecnologie per il Web Semantico sono linguaggi che hanno il vantaggio di una estrema semplicità nella rappresentazione dei dati, in modo che questi possano essere poi utilizzati da applicazioni in grado di processarli “semanticamente” in base a delle definizioni formali, che altro non sono che le ontologie.

Il livello dei dati, formato da XML e RDF, e il livello Schema, necessario alla definizione di un vocabolario, non sono però sufficienti. Un semantica formale per i primitivi espressi da RDF Schema non è fornita e l'espressività di questi primitivi non è

¹⁷ J. Euzenat, Comparison between Ontology Distance, 2002, su <http://www.informatik.uni-trier.de>.



sufficiente per un modellamento ed un ragionamento ontologico, ed è necessario uno strato di linguaggi ulteriore. Questo strato aggiuntivo di tecnologie è formato da quei linguaggi che consentono la creazione delle ontologie, e quindi del livello logico del Web Semantico. I linguaggi principali sviluppati a questo scopo, ovvero OIL, DAML+OIL E OWL mirano ad estendere RDF Schema per la definizione dei vocabolari formali necessari a compiere le inferenze.

In particolare OWL - Ontology Web Language è un linguaggio di markup semantico derivato da DAML+OIL e strutturato a partire da RDF. OWL consente la formalizzazione di uno specifico dominio della conoscenza, permettendo la definizione di classi e delle proprietà di tali classi. Permette la definizione di elementi singoli della classe e la definizione delle proprietà che questi assumono. Una ontologia OWL consiste di nessuna o più intestazioni, seguita da zero o più classi, elementi, proprietà degli elementi e istanze.