



Semantic Web: il futuro è già qui?

W3C OFFICE IN ITALY
Oreste Signore

KM CROSS

INTERNET MANAGEMENT

Sommario

Nel Semantic Web la conoscenza è rappresentata in maniera elaborabile dalla macchina, e quindi può essere utilizzata da componenti automatizzati, denominati agenti software. Il Semantic Web, naturale evoluzione del web attuale, potrà funzionare solo se le macchine potranno accedere ad un insieme strutturato di informazioni e a un insieme di regole di inferenza da utilizzare per il ragionamento automatico. Il Semantic Web non è un obiettivo irraggiungibile o puramente accademico: anche se molto lavoro è ancora da compiere, esistono già molte realizzazioni interessanti che si basano sull'utilizzo delle sue tecnologie di base, e i grandi attori del panorama ICT hanno investito e stanno investendo molte risorse nel settore.

1.1 Il Semantic Web: principi e tecnologie

1.1.1 Generalità

Nel contesto del Semantic Web, che è un' estensione del web attuale, il termine semantico assume sostanzialmente la valenza di "*elaborabile dalla macchina*". La semantica dei dati consiste nelle informazioni utili perché la macchina possa utilizzarli nel modo corretto [SemWeb], [TBL1998], [TBL1999], [TBL2001].

La filosofia di base del web è quella di uno spazio informativo universale, navigabile, con un mapping da URI alle risorse. Il Semantic Web potrà funzionare solo se le macchine potranno accedere ad un *insieme strutturato di informazioni* e a un *insieme di regole di inferenza* da utilizzare per il ragionamento automatico. La sfida del Semantic Web, quindi, è fornire un linguaggio per esprimere *dati* e *regole* per ragionare sui dati, che consenta l' *esportazione* sul web delle regole da qualunque sistema di rappresentazione della conoscenza.

1.1.2 I metadati

Nel navigare sul web si seguono dei link, che portano a quella che formalmente viene detta *risorsa* (resource) identificata univocamente da un URI. Le informazioni sulla risorsa vengono generalmente dette "*metadati*". Si può quindi dire che "*i metadati sono informazioni, comprensibili dalla macchina, relative a una risorsa web o a qualche altra cosa*". Si noti quindi che secondo questa definizione i metadati possono descrivere non solo risorse web, ma anche qualsiasi altra entità che possa

essere identificata in maniera univoca con un meccanismo analogo a quello utilizzato per individuare le risorse web (quindi mediante un URI). Di conseguenza, i metadati possono descrivere, per esempio, un dipinto, un reperto archeologico, un sito, un prodotto, una persona, etc. Il punto chiave è il fatto che i metadati sono comprensibili dalla macchina (*machine understandable*), e quindi costituiscono un tipo di informazione che può essere utilizzata in maniera opportuna dai software agent.

1.1.3 Resource Description Framework

L'uso efficace dei metadati richiede che vengano stabilite delle convenzioni per la *semantica*, la *sintassi* e la *struttura*. Coerentemente con l'assunzione di fondo di una architettura *peer-to-peer*, le *single comunità* interessate alla descrizione delle loro risorse specifiche definiscono la *semantica* dei metadati pertinenti alle loro esigenze. La *sintassi*, cioè l'organizzazione sistematica dei data element per l'elaborazione automatica, facilita lo scambio e l'utilizzo dei metadati tra applicazioni diverse. La *struttura* può essere vista come un vincolo formale sulla sintassi, per una rappresentazione consistente della semantica.

Resource Description Framework (RDF) è lo strumento 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 permette di definire la semantica dei tag XML, e fornisce un modello per descrivere le risorse, che hanno delle proprietà [RDF]. Il data model RDF, che consente di rappresentare statement RDF in modo sintatticamente neutro, è molto semplice ed è basato su tre tipi di oggetti: *Resources* (sempre individuate da un URI), *Properties* (un aspetto specifico, identificato da un "nome", che assume un "valore"), *Statements* (una tripla composta da un "soggetto", un "predicato" e un "oggetto", ovvero una tripla composta da una "risorsa", una "proprietà" e un "valore").

1.1.4 Web Ontology Language

RDF e RDF Schema sono due componenti importanti, ma non risolvono tutti i problemi. Infatti applicazioni sofisticate richiedono di poter "ragionare" sui dati. Il Semantic Web deve quindi essere supportato da ontologie, che definiscono i concetti e le relazioni utilizzate per descrivere e rappresentare uno specifico dominio di conoscenza. In altri termini, occorre disporre di un linguaggio che consenta di definire la terminologia usata, le caratteristiche logiche e i vincoli delle proprietà, l'equivalenza dei termini, le cardinalità delle associazioni, etc. Un' ulteriore complessità deriva dal fatto che il web è intrinsecamente distribuito, e di conseguenza applicazioni diverse possono usare ontologie diverse, o le stesse ontologie, ma espresse in lingue diverse. Il W3C, sfruttando anche i risultati di altri progetti, quali DAML e OIL, ha definito un linguaggio, denominato OWL, che permette di esportare le ontologie in modo interoperabile [OWL]. OWL offre tre sottolinguaggi, di crescente potere espressivo: *OWL Lite* (per rappresentare classificazioni gerarchiche e vincoli semplici), *OW DL* (per una maggiore potenza espressiva, garantendo comunque che tutte le conclusioni siano computabili e concluse in un tempo finito), *OWL Full* (che offre la massima potenza espressiva, ma non fornisce garanzie sui tempi di computazione, e difficilmente sarà supportato nella sua interezza da software che implementano il ragionamento). Ognuno di questi

linguaggi è un' estensione del precedente, sia in termini di ciò che può essere espresso che in termini della validità delle conclusioni.

1.2 Alcune considerazioni

1.2.1 Ontologie

Nell'organizzare le informazioni si ricorre spesso al meccanismo della classificazione, che può però presentare diversi problemi, derivanti essenzialmente dalle differenze di dominio, terminologia, scelta e significato delle varie caratteristiche ritenute importanti, e dalle differenze tra le relazioni più significative.

È importante distinguere le ontologie dai meccanismi di classificazione. Mentre questi ultimi prestano attenzione alle esigenze di accesso all'informazione, basandosi su criteri predefiniti codificati mediante elementi *sintattici*, le ontologie si concentrano sul *significato* dei termini e sulla *natura* e *struttura* di un dominio. Ne deriva che il problema essenziale è sostanzialmente quello della corrispondenza semantica (*semantic matching*) e dell'integrazione semantica.

Anche sottili differenze tra le diverse terminologie risultano importanti, e la definizione di vocabolari standard non è la soluzione al problema: essi sono spesso il risultato di anni di lavoro da parte di esperti, ma non sempre si adattano bene, soprattutto in presenza di domini di conoscenza eterogenei, e non sempre vengono utilizzati in maniera corretta. Un'ontologia, almeno nel senso in cui questo termine viene impiegato nel settore informatico, è un oggetto specifico, progettato per esprimere il "senso inteso" (*intended meaning*) di un vocabolario. Una caratteristica importante di un'ontologia è il suo livello di precisione, cioè il modo in cui è in grado di esprimere i concetti in maniera non ambigua e con un buon grado di copertura e di precisione rispetto al modello inteso.

In un contesto ampio e decentralizzato come quello del web, assume una particolare importanza l'integrazione dell'informazione. In questo processo risulta essenziale il ruolo giocato da una *core ontology*, il cui obiettivo è fornire un modello globale ed estensibile in cui possono essere messi in corrispondenza ed integrati i dati provenienti da fonti eterogenee.

[Doerr 2003] sottolinea la distinzione sottile, ma importante, tra una *core ontology* e la definizione di *core metadata* (come Dublin Core). Entrambe sono finalizzate all'integrazione dell'informazione, ma differiscono per l'importanza attribuita alla comprensibilità da parte di un lettore umano. I metadati sono compilati e utilizzati principalmente da esseri umani, mentre una *core ontology* è un modello formale utilizzato da strumenti che provvedono all'integrazione di varie fonti di dati e svolgono varie altre funzioni. Di conseguenza, mentre i fattori umani, e principalmente la leggibilità, costituiscono un elemento cardine nella definizione dei *core metadata*, una *core ontology* può accettare un livello di complessità maggiore, privilegiando la completezza e la correttezza logica rispetto alla comprensibilità umana.

1.2.2 Semantic Web e Intelligenza Artificiale

Nel settore dell'Intelligenza Artificiale (IA) già da molti anni esistono sistemi per la gestione della conoscenza che sono in grado di operare ragionamenti e deduzioni. Una caratteristica importante del Semantic Web è, coerentemente con i principi del

web, la sua totale *decentralizzazione*, e l' *interoperabilità* tra applicazioni, macchine, ontologie diverse. In questa visione, l'opera di armonizzazione delle ontologie e degli schemi descrittivi viene affidata ad agenti software che, disponendo di una rappresentazione della conoscenza e di regole di deduzione espresse con un linguaggio interoperabile, operano per armonizzare conoscenze diverse.

Il Semantic Web, comunque, non è Intelligenza Artificiale, anche se esistono molti punti di contatto tra le due aree, va piuttosto visto come uno strumento utile per l' Intelligenza Artificiale. In confronto a quest' ultima, RDF e OWL sono strumenti relativamente semplici, che forniscono un modo per esprimere e memorizzare i metadati, un modo di strutturare e caratterizzare i termini, e alcuni metodi per operare processi di inferenza in un contesto relativamente ristretto. L' obiettivo principale resta la semplicità, sacrificando eventualmente la completezza.

1.3 Alcune applicazioni delle tecnologie del Semantic Web

Nel corso di questi ultimi anni sono comparse molte applicazioni basate su tecnologie del Semantic Web. Mentre le prime applicazioni erano basate unicamente su RDF, le più recenti utilizzano in maniera più significativa le ontologie. Il processo non è semplice, perché, anche se esiste già un gran numero di ontologie, molte applicazioni sono ancora intrinsecamente centralizzate e le ontologie sono codificate in formati proprietari, e non è banale trasformarle in RDF/OWL, al fine di renderle interoperabili. Va tuttavia tenuto presente che esistono alcuni strumenti software per effettuare le conversioni.

Molti riferimenti si trovano sul sito WSIndex, portale su "Web Services and Semantic Web Resources" con una specifica sezione dedicata alle applicazioni SW [WSIndex].

Un altro riferimento interessante è la SW Technology Conference [SWTC], in cui molti attori del mondo web si sono incontrati per parlare non di metafisica del semantic web, ma di come l' uso di queste tecnologie possa migliorare il loro assetto commerciale.

Oltre al portale [SemWeb] va segnalato [Beckett], ricco di riferimenti a documenti, pubblicazioni, strumenti.

1.3.1 RSS

RSS (*RDF Site Summary* o *Rich Site Summary* o *Really Simple Syndication*) è un formato XML molto noto e utilizzato per la distribuzione di contenuti Web. Gli utenti, per mantenersi aggiornati sui contenuti di loro interesse pubblicati sul Web hanno tre opzioni:

1. *Visitare il sito Web*. Possono aggiungere il sito alla loro lista di *bookmark* e visitarlo periodicamente. Il rischio, ovvio, è di non ricordarsi sempre di farlo.
2. *Registrarsi per comunicazioni via e-mail*. La diffusione di filtri anti spam può determinare la perdita di queste comunicazioni, e comunque può essere oneroso per l' utente tener traccia delle comunicazioni ricevute.
3. *RSS feed*. Gli utenti possono registrarsi a particolari "canali" in cui l' informazione viene aggiornata automaticamente.

RSS fu lanciato per la prima volta da Netscape. Il formato inizialmente adottato era una semplificazione del primo *draft* proposto dal W3C, era derivato da RDF per la

gestione dei contenuti del portale My Netscape Network [MyNetscape], e permetteva la visualizzazione sul portale di *headline* e link relativi a notizie pubblicate su altri siti e rese disponibili attenendosi a specifiche ben precise. L' iniziativa riscosse subito un grande successo. La comunità dei blogger ha successivamente adottato lo stesso formato e le sue varianti. Il successo dei blog è stato probabilmente uno dei fattori trainanti del successo di RSS. Oggi RSS è lo standard *de facto* per l'esportazione di contenuti Web. Gli utenti possono oggi accedere a migliaia di feed RSS: alcuni siti (directory) raccolgono i riferimenti agli innumerevoli feed RSS disponibili sul Web.

La fruizione di un documento RSS è un processo molto semplice, e avviene generalmente in due modalità: o attraverso appositi software che interpretano un feed permettendo agli utenti di visualizzarne i contenuti, o integrando i contenuti del feed all'interno di un sito Web. Un aspetto particolarmente interessante è che è possibile accedere a notizie e informazioni da qualunque luogo, grazie alla disponibilità di applicazioni per la lettura di feed RSS su dispositivi mobili (telefoni cellulari, palmari, ecc.). Ciò consente l'accesso a notizie e informazioni da qualunque luogo.

1.3.2 Data integration

Un problema che si presenta spesso nella realtà è l' *integrazione semantica* di risorse aziendali memorizzate su database diversi. I vocabolari basati su RDF, RDFS e OWL costituiscono una sorta di "esperanto" tra i vari componenti. Approcci simili sono stati adottati in altri casi e in altri contesti ([Sculpteur], MITRE Corp., [MuseoSuomi]), e alcune imprese stanno investendo in questo settore.

1.3.3 Portali

Il Live Mobile Portal di Vodaphone utilizza RDF per le applicazioni di ricerca delle risorse (suonerie, giochi, immagini). La maggior efficacia della ricerca di informazioni ha fatto diminuire del 50% le visite delle pagine per il download, e ha determinato un incremento dei profitti relativi alle suonerie del 20% in due mesi. In questo caso, RDF è stato un fattore chiave per diminuire il carico e aumentare i profitti.

Il portale della Sun (SwordFish) gestisce le richieste per supporto, manuali, etc. utilizzando un search engine basato su RDF per cui le query vengono soddisfatte interrogando un database RDF interno [SunWP], [SunSH].

Nokia ha portale di supporto simile a questo [NokiaPortal].

1.3.4 Ricerca più efficace grazie alle ontologie: il caso GoPubMed

La ricerca su [GoPubMed], importantissima fonte di citazioni nel settore biomedico, viene migliorata grazie all' utilizzo di ontologie. In particolare, i risultati delle query vengono ordinati in base ad una ontologia (per esempio, la Gene Ontology), e vengono generati termini aggiuntivi per la ricerca, evidenziati nella presentazione degli abstract. Questa applicazione dimostra l' efficacia di *ontologie di dominio* per migliorare la ricerca di informazioni.



1.3.5 Standardizzazione dei metadati: il caso Dublin Core

La *Dublin Core Metadata Initiative* [DC] ha lo scopo di sviluppare standard interoperabili per i dati online, per supportare un ampio spettro di obiettivi e di modelli di business. Dublin Core è uno dei primi vocabolari in RDF per metadati, e costituisce la base per i vocabolari per le Digital Libraries distribuite.

Il *Dublin Core Metadata Element Set* [DCMES] può essere visto come l' elemento semantico di base per i metadati sul Web. Le sue 15 categorie generali (elements) permettono di creare descrizioni semplici e facilmente comprensibili per la maggior parte delle risorse informative. DCMES è tuttavia solo l' elemento di base, perché le singole comunità hanno spesso bisogno di una semantica più ricca per poter descrivere compiutamente le loro risorse. A questo scopo, altri metadati possono essere combinati con DCMES, al fine di creare descrizioni più ricche.

Le applicazioni dovrebbero essere in grado di mescolare i metadati provenienti da altri standard, anch' essi espressi in RDF. La comunità RSS è un esempio di adozione con successo dell' approccio di combinare la semantica di Dublin Core con altri vocabolari, per consentire la produzione di sommari e recensioni più efficaci dei siti web.

Il Dublin Core Metadata Element Set è il componente di base per supportare la ricerca di informazioni su domini applicativi diversi. Il suo successo è testimoniato dall' adozione da parte di moltissime organizzazioni nei più disparati settori, e cominciano a comparire applicazioni anche nel settore commerciale, con imprese appartenenti all' area salute e finanziaria che lo utilizzano come base per lo scambio di informazioni.

1.3.6 Il progetto SIMILE

Una delle difficoltà che l' utente incontra nel consultare le biblioteche digitali deriva dalle differenze di terminologia e di interfaccia. Il progetto SIMILE (acronimo per *Semantic Interoperability of Metadata In unLike Environments*) si propone di superare queste difficoltà e aiutare l' utente a navigare nel Semantic Web [SIMILE].

SIMILE nasce dal progetto DSpace, sviluppato da Hewlett-Packard Research Labs e MIT Libraries, che era un repository per indicizzare, conservare e ridistribuire risorse digitali. Il supporto ai metadati offerto da DSpace è limitato allo schema Dublin Core, che, come abbiamo già detto, è molto generale ma limitato. Se Dspace dovrà in futuro supportare altri schemi di metadati, dovrà necessariamente affrontare il problema dell' interoperabilità.

Le tecnologie di base del Semantic Web (RDF, RDFS e OWL) consentono alle singole comunità di creare i loro metadati specifici, eventualmente sulla base di ontologie più generali, e di renderli fruibili a tutti. Tuttavia, non tutti gli utenti sono esperti utilizzatori delle tecnologie del Semantic Web, e quindi è utile che possano disporre di strumenti che li aiutino nel verificare il loro lavoro. Questa è stata una delle motivazioni principali che hanno portato alla nascita del progetto SIMILE. L' obiettivo primario è l' estensione di Dspace, per consentire il supporto di qualunque schema di metadati, fornendo nel contempo un modo per distribuire le risorse digitali. SIMILE è quindi nato nel contesto bibliotecario, ma gli strumenti realizzati possono essere utilizzati in molti altri domini. Inoltre, per far fronte alle difficoltà nel definire



ontologie, creare RDF, e convertire in RDF i metadati espressi in XML, è stato necessario perseguire l'ulteriore obiettivo di creare strumenti con cui utenti non specialisti potessero creare un RDF di buona qualità.

SIMILE ha sviluppato una suite di applicazioni Web che permettono il browsing dei metadati RDF utilizzando un normale browser web. La suite è composta da due prodotti: Longwell e Knowle. Il primo [Longwell] è un "browser a faccette" che consente all'utente di visualizzare il contenuto del thesaurus nascondendo il modello RDF sottostante. Knowle è invece uno strumento per navigare sui grafi RDF, e il suo scopo è consentire la visione dettagliata e la correzione del modello RDF sottostante.

La potenza di RDF risiede nel supporto che questa tecnica fornisce per definire i modelli, e nella sua implicita natura distribuita. Per risparmiare agli utenti l'intrinseca complessità della sintassi XML/RDF, nell'ambito del progetto Simile viene mantenuto un catalogo degli strumenti esistenti per trasformare i dati da sintassi esistenti a RDF. Questi strumenti (detti *RDFizers*) sono utili per risolvere il problema fondamentale del semantic web: *"non è possibile sviluppare applicazioni significative per mancanza di una sufficiente mole di dati RDF, ma non si creeranno mai abbastanza dati RDF senza una applicazione che li utilizzi pesantemente"*. La disponibilità di *RDFizers* consente agli utenti di esplorare i dati esistenti utilizzando dei browser RDF, e quindi apprezzare i benefici derivanti dall'utilizzo di RDF come *lingua franca*,

Piggy Bank [Huynh2005] è un'estensione del browser Firefox che lo trasforma in un "Semantic Web Browser". Il principio guida su cui si basa è quello di consentire agli utenti Web di estrarre elementi informativi dalle pagine web che stanno visitando, e salvarli in un formato coerente con le tecnologie del Semantic Web (quindi basato su RDF), corredandoli di metadati. Piggy Bank consente poi di utilizzare queste informazioni all'interno del browser stesso. Infatti queste informazioni, estratte da vari siti, possono essere ricercate, ordinate, organizzate in sottoinsiemi, consultate, indipendentemente dalla loro originaria collocazione. Questo approccio è molto interessante, perché si presenta come un modo di migliorare l'utilizzo delle risorse web e l'esperienza dell'utente, contribuendo quindi alla diffusione delle tecnologie del Semantic Web. Un ulteriore punto di forza è "*Semantic Bank*", un repository condiviso di RDF, cui la comunità che utilizza Piggy Bank può contribuire, mettendo a comune le informazioni acquisite.

1.4 Conclusioni

L'idea del Semantic web, lanciata qualche anno fa, sembrava un sogno irrealizzabile, e molti non avevano nascosto il loro scetticismo. Oggi il quadro tecnologico appare più chiaro e confortante, anche se restano ancora aperti importanti problemi squisitamente di ricerca. Tuttavia, molte applicazioni utilizzate quotidianamente e su larga scala si basano sulle tecnologie del Semantic Web (in particolare RDF), e alcuni attori importanti del panorama ICT stanno cominciando a offrire prodotti che supportano queste tecnologie o si basano su di esse (a titolo di esempio, citiamo la piattaforma XMP di Adobe e Oracle Spatial 10g).

Un aspetto importante, in questo processo, è stato giocato dal ruolo guida del W3C, e soprattutto dal metodo di lavoro che lo caratterizza: un ambiente fortemente motivato e aperto, in cui tecnici e ricercatori di tutto il mondo portano il loro contributo alla crescita del Web. C'è da chiedersi se, in un mondo così fortemente competitivo e



in rapida evoluzione, le imprese che restano fuori da questo processo potranno poi giocare un ruolo importante e innovativo.

1.5 Ringraziamenti

I miei ringraziamenti vanno a Ivan Herman, Head of W3C Offices. Parte del materiale utilizzato proviene da sue presentazioni.

Riferimenti bibliografici

- [Beckett] Dave Beckett's Resource Description Framework (RDF) Resource Guide, <http://planetrdf.com/guide/>
- [DC] *Dublin Core Metadata Initiative*, <http://www.dublincore.org/>
- [DCMES] <http://purl.org/dc/documents/rec-dces-19990702.htm>
- [Doerr2003] Martin Doerr, Jane Hunter and Carl Lagoze: *Towards a Core Ontology for Information Integration*, Journal of Digital Information, Volume 4 Issue 1, Article No. 169, 2003-04-09, (April 2003), <http://jodi.ecs.soton.ac.uk/Articles/v04/i01/Doerr/>
- [GoPubMed] <http://www.gopubmed.org/>
- [Huynh2005] David Huynh, Stefano Mazzocchi, David Karger: *Piggy Bank: Experience the Semantic Web Inside Your Web Browser*, Lecture Notes in Computer Science, Volume 3729, Oct 2005, Pages 413 - 430 <http://hdl.handle.net/1721.1/29466> , <http://www.springerlink.com/media/56we0cguvp5jpgd7ta2w/contributions/p/8/7/1/p87114h811812n11.pdf>
- [Longwell] <http://simile.mit.edu/longwell/>
- [MuseoSuomi] <http://museosuomi.cs.helsinki.fi/>
- [MyNetscape] <http://my.netscape.com/>
- [NokiaPortal] <http://www.forum.nokia.com/main.html>
- [OWL] *Web Ontology Language (OWL)*, <http://www.w3.org/2004/OWL/>
- [RDF] *Resource Description Framework (RDF)*, <http://www.w3.org/RDF/>
- [Sculpteur] <http://www.sculpteurweb.org/>
- [SemWeb] <http://www.semanticweb.org/>
- [Signore2003] Signore, Oreste: *Strutturare la conoscenza: XML, RDF, Semantic Web - Clinical Knowledge 2003* (1st edition) - Udine, 20-21 September 2003 <http://www.w3c.it/papers/ck2003.pdf>, <http://www.w3c.it/talks/ck2003/>
- [SIMILE] <http://simile.mit.edu/>
- [SunSH] Sun System Handbooks collection, http://sunsolve.sun.com/handbook_pub/
- [SunWP] Sun White Papers collection, <http://www.sun.com/servers/wp.html/>,
- [SWTC} <http://www.semantic-conference.com/>
- [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
- [TBL2001] T. Berners-Lee, J. Hendler, O. Lassila: *The Semantic Web*, Scientific American, May 2001, <<http://www.scientificamerican.com/2001/0501issue/0501berners-lee.html>>
- [Weibel2000] Stuart Weibel, Eric Miller *An Introduction to Dublin Core* <http://www.xml.com/pub/a/2000/10/25/dublincore/>)
- [WSIndex] http://www.wsindex.org/Companies/Semantic_Web/index.html