Sistemi di Elaborazione dell'informazione II

Corso di Laurea Specialistica in Ingegneria Telematica
II anno – 4 CFU
Università Kore – Enna – A.A. 2008-2009

Alessandro Longheu
http://www.diit.unict.it/users/alongheu
alessandro.longheu@diit.unict.it

Dati Semistrutturati: Il Web Semantico

Semantic Web

- "Ho fatto un sogno riguardante il Web [...] ed è un sogno diviso in due parti.
- Nella prima parte, il Web diventa un mezzo di gran lunga piú potente per favorire la **collaborazione tra i popoli**. Ho sempre immaginato lo spazio dell'informazione come una cosa a cui tutti abbiano accesso immediato e intuitivo, non solo per navigare ma anche per creare.
- Nella seconda parte del sogno, **la collaborazione si allarga ai computer**. Le macchine diventano capaci di analizzare tutti i dati sul Web, il contenuto, i link e le transazioni tra persone e computer. [...] i meccanismi quotidiani di commercio, burocrazia e vita saranno gestiti da macchine che parleranno a macchine, lasciando che gli uomini pensino soltanto a fornire **l'ispirazione e l'intuito**.
- ... il Web sarà un luogo in cui l'improvvisazione dell'essere umano e il ragionamento della macchina coesisteranno in **una miscela ideale e potente**"
- Con queste parole **Tim Berners Lee** presentava la sua visione del Web

Definizione

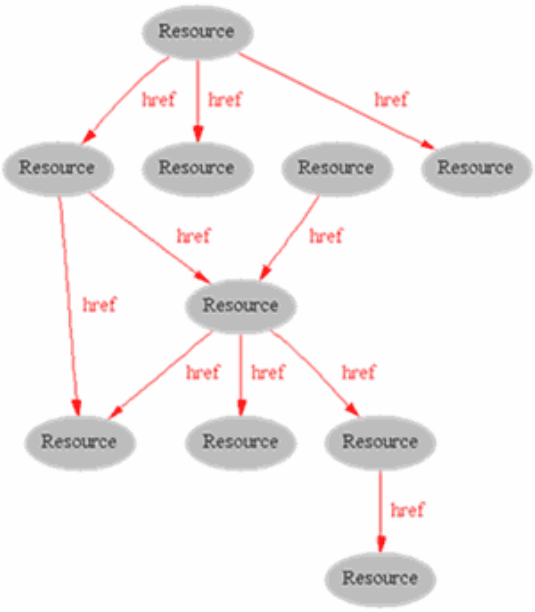
- Con il termine Web Semantico si intende la **trasformazione** del World Wide Web in un ambiente dove i documenti pubblicati (pagine HTML, file, immagini, e così via) siano associati ad informazioni e dati (metadati) che ne specifichino il contesto semantico in un formato adatto all'interrogazione, all'interpretazione e, più in generale, all'elaborazione automatica.
- Con l'interpretazione del contenuto dei documenti che il Web Semantico propugna, saranno possibili ricerche molto più evolute delle attuali, basate sulla presenza nel documento di parole chiave, ed altre operazioni specialistiche come la costruzione di reti di relazioni e connessioni tra documenti secondo logiche più elaborate del semplice link ipertestuale, permettendo un approccio simile a quello presente nei *sistemi esperti*

Lo scenario odierno

- Sono ormai passati diversi anni dalla comparsa della prima pagina web (fine anni '80, primi anni '90) e per quanto siano nette le differenze esistenti tra il web attuale e quello dei primi anni, tuttavia **l'infrastruttura di base è fondamentalmente la stessa**: "una rete di risorse di informazioni, basata sull'infrastruttura di Internet che si basa su tre meccanismi per rendere queste risorse prontamente disponibili al più vasto insieme possibile di utenti:
  - uno schema di denominazione uniforme per localizzare le risorse sul Web (ad es., gli URL);
  - protocolli per accedere alle risorse denominate sul Web (ad es., HTTP);
  - ipertesto, per una facile navigazione tra le risorse (ad es., HTML).

Lo scenario odierno

- Le pagine web sono collegate sintatticamente mediante indici che localizzano la URL della pagina e tali collegamenti consentono di identificare le pagine in modo univoco.
- Uno dei principali limiti di tale impostazione risiede nell'assenza di significato dei collegamenti, in altre parole questo sistema manca di una qualche capacità semantica: i collegamenti dovrebbero non solo condurci in un determinato luogo (la pagina web) ma anche descriverci il luogo in cui saremmo condotti.



Lo scenario odierno

- Il **funzionamento di un motore di ricerca** può essere descritto nel seguente modo:
  - l'interazione fra l'utente e il motore di ricerca inizia con l'invio di un'interrogazione, tramite form HTML;
  - il motore di ricerca utilizza le parole dell'interrogazione per cercare nei file indice che si è precedentemente costruito scaricando e analizzando le pagine web, quali pagine contengono quelle parole;
  - tali pagine vengono quindi ordinate per pertinenza utilizzando vari criteri, che essenzialmente si basano sul contenuto testuale delle pagine stesse e sulle informazioni rappresentate dai link sul web che puntano ad esse;
  - il risultato viene mostrato all'utente utilizzando una pagina HTML che contiene rappresentazioni condensate delle pagine più pertinenti.

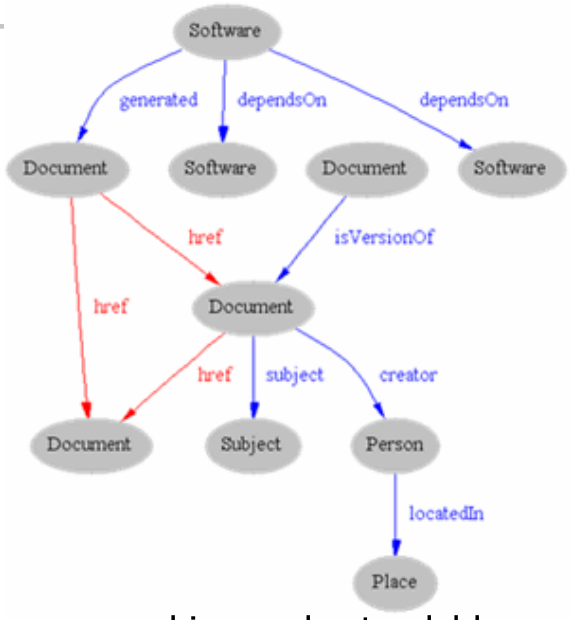
Lo scenario odierno

I motori di ricerca soffrono di **evidenti limiti**:

- Il primo dato è l'esistenza del cosiddetto **web nascosto**, ovvero una quantità di risorse informative disponibili sul web ma non rintracciabili dai motori di ricerca per varie cause quali contenuti non indicizzati, pagine periferiche, immagini, files audio, files video, file flash, archivi zippati, informazioni contenute in basi di dati, contenuti dinamici che cambiano in tempo reale ecc., stimato essere pari all'80% delle risorse disponibili
- visualizzazione dei risultati poco intuitiva ed esplicativa;
- limitata pertinenza con la richiesta inviata.
- **problemi di vocabolario**, ad esempio casi di sinonimia e polisemia che rendono praticamente impossibile per i motori di ricerca restituire esclusivamente i risultati attesi, questo a causa della notevole ricchezza (ma anche ambiguità) del linguaggio naturale, di fronte a cui anche i sistemi di ricerca più evoluti soffrono di enormi limiti di interpretazione, ad esempio la parola albero riguarda informatica, botanica, nautica? e ancora, un documento che parla di finanziamento del governo alle società calcistiche in pericolo di fallimento in che ambito ricade? Sport, politica, finanza?

Lo scenario futuro

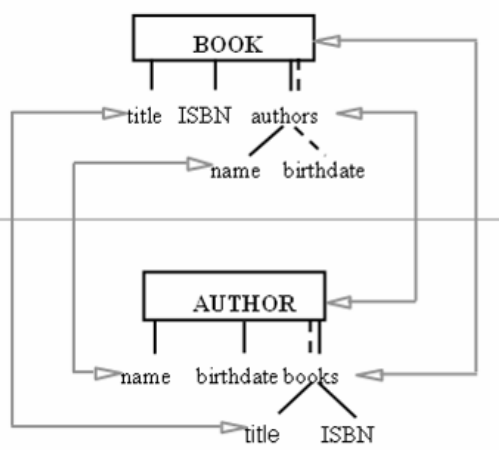
- Il WS non implica una qualche forma di intelligenza, paragonabile a quella di cui è dotata la mente umana da parte delle macchine, esso implica solo un'abilità delle macchine a risolvere problemi ben definiti realizzando operazioni ben definite su dati ben definiti esistenti.
  - Invece di richiedere ai computer di comprendere il linguaggio umano e la sua logica, si richiede all'uomo di fare uno sforzo in più in fase di progettazione web.



- Il web attuale è machine-readable ma non machine-understandable
- a tal fine potrebbero aiutare i **collegamenti semantici** piuttosto che i semplici hyperlink.

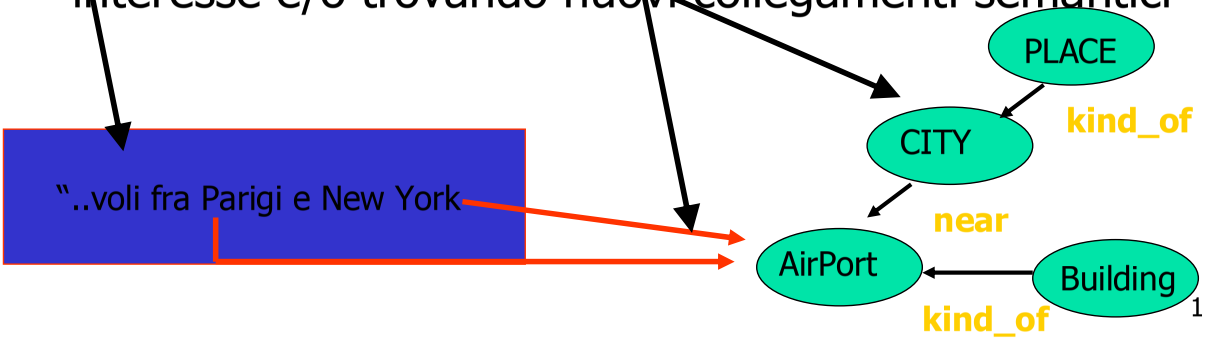
Lo scenario futuro

- Lo scenario futuro cerca di **riprodurre sul Web quello che già in parte esiste nel mondo dei database**: quando interroghiamo una base di dati, possiamo infatti fare ricerche piuttosto raffinate, ad esempio, chiedere "quali autori hanno scritto almeno due libri sull'IR"
- L'utente può formulare una richiesta che imponga precise relazioni ("almeno due libri sull'IR"), e tali relazioni sono stabilite fra concetti ("autore" e "libro") non fra parole chiave (non si ricerca la stringa "autore" o "libro"). Questo è possibile perché esiste uno **schema del DB**, cioè un modello ed un insieme di regole che stabiliscono coe debbano essere organizzati i dati



Lo scenario futuro

- Nel web, invece, le informazioni sono in genere NON strutturate; è quindi necessario fornire tale struttura ai dati (le pagine web) tramite:
  - I metadati (HTML) o annotazioni (XML, RDF) per indicare i collegamenti semantici
  - Lo schema (o ontologia) del dominio per ragionare su tali collegamenti, estraendo le informazioni di interesse e/o trovando nuovi collegamenti semantici



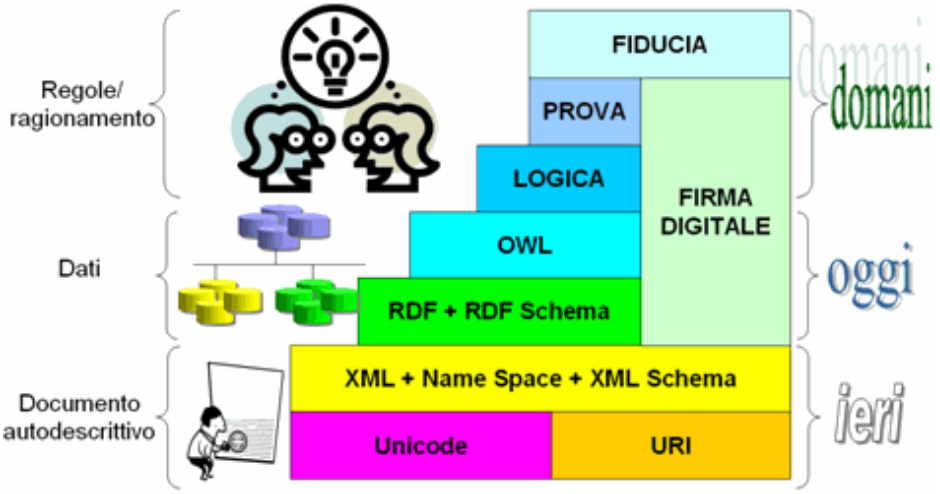
Lo scenario futuro

- **Benefici del Semantic Web**? Berners Lee ipotizza tale scenario con un esempio divenuto ormai un classico della letteratura sul WS:
  - "Lucy ha la necessità di prenotare una visita medica specialistica per sua madre. Dopo aver istruito il suo agente (che possiamo definire in modo semplicistico un programma capace di eseguire compiti definiti da un utente in modo autonomo, ovvero senza il controllo diretto dell'utente stesso) circa le proprie esigenze (tipo di visita specialistica, massima tariffa consentita, distanza dalla casa della madre, date disponibili, ecc.), delega ad esso il compito di ricercare sul web in modo del tutto autonomo quali soluzioni sono disponibili. Una volta che l'agente avrà individuato le possibili alternative, Lucy avrà l'unico compito di scegliere la più adatta e darà il comando all'agente di prenotare in sua vece".
- per rendere possibile tutto questo **NON basta l'XML**, occorre una nuova architettura…

Lo scenario futuro

L'architettura del semantic web presenta diversi livelli:

- i dati: definiti in modo strutturato tramite XML;
- i metadati: "informazioni sui dati" gestite tramite RDF;
- le ontologie: semantica di dati e metadati tramite specifici linguaggi.



Lo scenario futuro

- **Diverse risorse tecnologiche** sono implicate in quest'architettura. Alcune di queste sono già oggi pienamente disponibili, altre rappresentano il futuro (quelle fondanti il livello della logica, prova e fiducia). Trasversale rispetto a più livelli risultano le tecnologie legate alla firma digitale.
- **Esaminiamo ora i singoli componenti della piramide:**
- **Unicode**: sistema di codifica che assegna una combinazione di bit a ogni carattere in maniera indipendente dal programma, piattaforma e dalla lingua. Tramite Unicode è possibile rappresentare i caratteri usati in quasi tutte le lingue vive e in alcune lingue morte, nonché simboli matematici e chimici, cartografici, l'alfabeto Braille, ideogrammi etc…

Lo scenario futuro

- **Diversi codici per la rappresentazione dell'informazione**:
- il primo codice ASCII è a 7 bit (ISO 646)
- con il bit 8, si hanno altri 128 caratteri, scelti in base alla lingua a cui si vuole offrire il supporto, dando luogo agli standard ISO 8859, ad esempio 8859-1 (ISO Latin-1) per le lingue europee, 8859-15 cirillico, 8859-6 arabo…
- utilizzando 16 bit si possono rappresentare insiemi di caratteri fonetici e ideogrammi che rappresentano intere parole (indispensabile per cinese e giapponese); si ottiene il codice UNICODE a 16 bit; i primi 128 caratteri sono identici all'ISO 646 e i primi 256 sono gli stessi dell'ISO 8859-1
- utilizzando 32 bit, si ottiene lo standard ISO 10646, che ha l'obiettivo di raccogliere tutti i simboli utilizzati da tutte le lingue del mondo inclusi quelli matematici, valutari ecc.
- C++ usa 8 bits ~ 256 caratteri differenti
- Java usa 16 bit ~ 65,535 caratteri differenti

Lo scenario futuro

- URI: sta per **Uniform Resource Identifier** (Identificatori uniformi di risorse); un URI è una stringa che identifica una risorsa nel Web in maniera univoca: un documento, un'immagine, un file, un indirizzo email… (es. http://www.websemantico.org/index.php)
- L'URI richiama alla mente il concetto di URL, un po' diverso:
  - An Uniform Resource Locator (URL) is the term used to identify an Internet resource, and can be specified in a single line of text.
  - An Uniform Resource Name (URN) is the term used to identify an Internet resource, without the use of a scheme (protocol), and can be specified in a single line of text.
  - An Uniform Resource Identifier (URI) is used by a browser to identify a single document, and it too can be specified in a single line of text.

Lo scenario futuro

- URL vs. URN vs. URI
  - The difference between the three is subtle. An URL refers to a Web page, including the scheme, but without a name location. An URN may also include the location of a code fragment. An URI refers to a Web page including the location of the code fragment, if one exists, and the scheme.
  - URL http://www.cnn.org/iis/review1.htm
  - URN www.cnn.org/iis/review1.htm#one
  - URI http://www.cnn.org/iis/review1.htm#one
  - Because Web servers allow for default documents and do not require a scheme to retrieve a document, the subtle difference between an URL, URN and URI is hard to tell.

Lo scenario futuro

- Di fatto l'URI è una generalizzazione di URL ed URN:
  - Un URL (Uniform Resource Locator) è un URI che, oltre a identificare una risorsa, fornisce mezzi per agire su o per ottenere una rappresentazione della risorsa descrivendo il suo meccanismo di accesso primario o la sua "ubicazione" ("location") in una rete.
    - Per esempio, l'URL http://www.onu.org/ è un URI che identifica una risorsa e lascia intendere che una rappresentazione di tale risorsa (il codice HTML della versione corrente di tale home page) è ottenibile via HTTP da un host chiamato www.onu.org.
  - Un URN (Uniform Resource Name) è un URI che identifica una risorsa mediante un "nome" in un particolare dominio di nomi ("namespace"). Un URN può essere usato senza lasciar intendere l'ubicazione della risorsa.
    - Per esempio l'URN urn:isbn:0-395-36341-1 consente di individuare univocamente un libro mediante il suo nome 0-395-36341-1 nel namespace dei codici ISBN, ma non suggerisce dove e come possiamo ottenere una copia di tale libro.

Lo scenario futuro

XML, Name Space e XML Schema:

- XML (eXstensible Markup Language) è un meta-linguaggio di markup. In pratica fornisce un insieme di regole sintattiche per modellare la struttura di documenti e dati. Questo insieme di specifiche definiscono le modalità con cui crearsi un proprio linguaggio di markup. XML reca tra i suoi vantaggi fondamentali quello di garantire un'alta interoperabilità dei dati.
- La struttura e la grammatica soggiacenti ad un documento XML possono essere stabilite attraverso un DTD (Document Type Definition) o (meglio) attraverso XML Schema, che fornisce un metodo per comporre vocabolari XML.
- Un Namespace non è altro che un insieme di nomi di elementi e/o attributi identificati in modo univoco da un identificatore. La presenza di un identificatore univoco individua così un insieme di nomi distinguendoli da eventuali omonimie presenti in altri namespaces.

Lo scenario futuro

- **RDF e RDF Schema**:
  - RDF (Resource Description Framework) fornisce un insieme di regole per definire informazioni descrittive sui dati, più precisamente sugli elementi costitutivi un documento web; queste asserzioni sono realizzate tramite triple che legano tra loro gli elementi in una relazione binaria. Le triple sono del tipo: Soggetto (la risorsa), Predicato (la proprietà) e Oggetto (il valore). Un modello RDF è rappresentabile da un grafo orientato sui cui nodi ci sono risorse o tipi primitivi e i cui archi rappresentano le proprietà.
  - RDF Schema fornisce, a sua volta, un metodo per combinare queste descrizioni in un singolo vocabolario. Il modo per sviluppare vocabolari specifici per un dato dominio di conoscenza è rappresentato dalle ontologie.

Lo scenario futuro

- Uno dei problemi principali di fronte a cui ci si trova davanti quando si parla di ontologie è quello della condivisione e della conciliazione di esigenze e punti di vista diversi, in sostanza delle **infinite visioni del mondo**.
- Per tale motivo la generazione di un'ontologia fondante e totale risulta essere un utopia e sempre più, anche nell'ambito del Web Semantico, si sta sviluppando un movimento di sviluppo di **ontologie provenienti dal basso**, ovvero emergenti dal senso comune e dai processi sociali di negoziazione dei significati.
- Sempre per lo stesso motivo si tende alla creazione di diverse ontologie, ciascuna riferita ad un preciso dominio e seguente un dato punto di vista. Nasce qui l'esigenza di **interoperabilità** dei diversi sistemi ontologici generati, problema a cui si può ovviare perseguendo processi di standardizzazione dei linguaggi descrittivi di tali sistemi.

Lo scenario futuro

- Nell'ambito del Web Semantico, il W3C ha sostenuto lo sviluppo di OWL (Web Ontology Language) quale linguaggio per la definizione di ontologie strutturate basate sul Web.
- OWL è un linguaggio di markup per rappresentare esplicitamente significato e semantica di termini con vocabolari e relazioni tra i termini. Tale rappresentazione dei termini e delle relative relazioni costituisce un'ontologia.
- L'obiettivo è permettere ad applicazioni software di elaborare il contenuto dei documenti scritti in OWL.

Lo scenario futuro

- OWL è composto da tre sottolinguaggi caratterizzati da una crescente espressività:
  - OWL Lite: utile per quanti necessitano soprattutto di una gerarchia di classificazione e semplici restrizioni;
  - OWL DL (Description Logics): utile per quanti ricercano il massimo dell'espressività mantenendo la completezza computazionale (tutte le conclusioni hanno la garanzia di essere calcolabili) e la decidibilità (tutte le computazioni finiscono in un tempo definito);
  - OWL Full: destinato agli utenti che vogliono la massima espressività e libertà sintattica di RDF senza le garanzie computazionali.
- Come indicato nei documenti ufficiali W3C "OWL Full può essere considerato come una estensione di RDF, mentre OWL Lite e OWL DL possono essere considerate come una estensione di una visione limitata di RDF".
- Ogni documento OWL è un documento RDF, ed ogni documento RDF è un documento OWL Full, ma solo alcuni documenti RDF saranno un documento OWL Lite oppure OWL DL".

Lo scenario futuro

- Logica, Prova e Fiducia:
  - **Logica:** Affinché il Web Semantico possa effettivamente aiutarci in una vasta gamma di situazioni, estraendo autonomamente informazioni utili dalla mole di documenti annotati semanticamente, è indispensabile costruire un potente linguaggio logico per realizzare le inferenze (ovvero procedimenti deduttivo mediante cui, a partire da una o più premesse, si ricava, per via logica, una conclusione).
  - **Prova:** Le conclusioni ottenute saranno validate a questo livello tramite motori di validazione costituiti da sequenze di formule derivate da assiomi.
  - **Trust:** Infine il sistema restituirà solo quelle informazioni che secondo il richiedente proverranno da utenti di indubbia attendibilità.

Lo scenario futuro

- Gli altri elementi fondamentali sono rappresentati da:
- **Agenti intelligenti**: programmi capaci di eseguire compiti definiti da un utente in modo autonomo, ovvero senza il controllo diretto dell'utente stesso: essi raccolgono, filtrano ed elaborano le informazioni che trovano sul web;
- **Firma digitale**: garantisce, basandosi su di un sistema crittografico, l'autenticità delle varie asserzioni e permette di scoprire la loro provenienza. Spetta poi all'utente istruire il software del proprio computer di quali firme digitali fidarsi. Essa può essere apposta come allegato dei documenti web. L'obiettivo finale è quello che viene comunemente definito **"Web of Trust"** (un web capace di offrire riservatezza, che ispiri gradualmente fiducia, e che faccia in modo che ci si prenda la responsabilità di ciò che viene pubblicato);

Lo scenario futuro

- **Metadati**: I metadati sono alla base di tutto il WS. I metadati sono dei "dati sui dati": informazioni relative ai dati, tramite le quali è possibile ricavare delle informazioni sulla risorsa a cui sono associate. Ad ogni risorsa disponibile sul web dovrebbe essere associata una precisa descrizione.
- Sono stati proposti diversi schemi di metadati; allo stato attuale uno dei più diffusi è il **Dublin Core**, un sistema di metadati costituito da un insieme minimale di elementi per descrivere materiale digitale accessibile via rete.
- Il set minimo è costituito da 15 elementi: Titolo, Creatore, Soggetto, Descrizione, Editore, co-autore (Contributor), Data, Tipo, Formato, Identificatore, Fonte, Lingua, Relazione Copertura (Coverage), Gestione dei diritti di autore.

Lo scenario futuro – in sintesi

- Ma in che modo queste tecnologie possono cooperare affinché il web possa fornirci i servizi ipotizzati da Berners Lee?
- Volendo semplificare il discorso, alla base occorre una **diversa filosofia di progettazione** delle risorse web - basate su XML -, le quali devono rispettare gli standard definiti e recare con se una descrizione delle proprie caratteristiche (tramite RDF e metadati).
- Ciascuna di queste risorse sarà identificabile in modo non ambiguo grazie all'uso degli URI (risolvendo così i problemi di ambiguità visti quando abbiamo parlato dei motori di ricerca).
- I **metadati** sono la base informativa su cui potranno operare gli agenti intelligenti per prendere le proprie decisioni.
- Gli **agenti**, a loro volta, potranno muoversi nello spazio-web sfruttando il sistema di rappresentazione della conoscenza disponibile (**ontologie**). Le decisioni degli agenti a questo punto saranno consentite grazie all'utilizzo di linguaggi di inferenza logica. Gli agenti, infine, nel prendere le proprie decisioni terranno conto del grado di fiducia attribuito alle risorse (ed ai loro autori identificati da sistemi di firma digitale) dagli utenti stessi.

Lo scenario futuro – in sintesi

- La piena realizzazione dei principi del Web Semantico è probabilmente ancora lontana da una sua realizzazione e gli ostacoli maggiori al suo sviluppo si incontrano proprio al livello ontologico dell'architettura precedentemente vista.
- L'onerosità della mappatura delle risorse, la piena interoperabilità tra i diversi linguaggi utilizzati per la descrizione dei dati e le relazioni tra essi, i cambiamenti, anche culturali, profondi che si richiedono soprattutto in fase di progettazione dei documenti destinati al web richiedono uno sforzo supplementare e quell'adeguamento sociale e tecnologico che fin dagli inizi Berners Lee aveva indicato come chiave del cambiamento.