Web semantico:
Il recupero dell'informazione affidato alla macchina

Università degli Studi di Siena
Corso di Laurea in Scienze della Comunicazione

Anno Accademico 2002-2003
Corso di Linguistica Computazionale

Professor Amedeo Cappelli

Seminario di Claudia Maccari.

L'attuale struttura del Web

•I dati sul Web sono organizzati in formati molto lontani dall'essere interpretati e compresi in modo automatico dal punto di vista semantico: è necessaria sempre l'interpretazione dell'uomo.

•Le pagine HTML non distinguono il contenuto dalla sua rappresentazione (l'informazione _ link, formattazione, metadati_ viene espressa ad un unico livello e questo rende semplice l'utilizzo di HTML, ma ne compromette la potenza, creando difficoltà nella trasmissione e nell'interscambio di dati).

•I dati non hanno una struttura evidente e significativa per una macchina che deve estrarre informazioni dal Web.

Recupero dei dati nell'attuale Web:

- Ricercando per parole chiave nell'intero contenuto di un sito o di una pagina
- Utilizzando algoritmi di codificazione che operano individuando indizi in grado di permettere la classificazione
- Attraverso la ricerca in data base strutturati

Problemi che caratterizzano l'uso di parole chiave e di algoritmi:

•Questi metodi non tengono conto del modo in cui è strutturato il linguaggio: una parola può trovarsi per caso in un ipertesto, ma non caratterizzarne il contenuto, oppure riferirsi solo ad una parte di esso.

•Può succedere che la ricerca per parola chiave vada a vuoto perché il termine che abbiamo digitato non compare nella descrizione del prodotto, oppure che otteniamo molti risultati non pertinenti perché abbiamo scelto un termine troppo comune.

Ricerca in data base strutturati

•Questo metodo può essere impiegato con successo per la ricerca di informazioni per le attività o i prodotti di un'azienda.

•In testi in cui le parole possono essere ricondotte a tassonomie.
Una *tassonomia* di oggetti o termini è una classificazione per categorie: per esempio le parti di un'automobile o l'insieme dei componenti elettrici.

Problemi: la costruzione di tassonomie accurate diventa difficile con l'aumentare del vocabolario, e richiede notevoli investimenti nella codificazione dell'informazione e nell'addestramento dei programmatori e degli utenti.

Problemi dell'attuale Web

- Ricerca delle informazioni: sistemi di ricerca attuali basati su parole chiave (rumore o silenzio)
- Estrazione dell'informazione: ad esclusivo carico dell'utente umano che deve ricercare e selezionare i documenti
- Manutenzione dell'informazione: l'aggiornamento dell'informazione effettuata manualmente, e senza il supporto di vincoli semantici è un'operazione che richiede tempo ed investimenti
- No generazione automatica dell'informazione: non è possibile creare siti web che si adattino alle specifiche esigenze di ogni utente.

L'innovazione:

•Con il Web semantico si intende sviluppare un linguaggio per esprimere le informazioni in una forma comprensibile e processabile dalla macchina.

•Il Web semantico non è separato dall'attuale, ma costituisce un livello di estensioni in cui alle informazioni viene attribuito un significato ben preciso.

•Obiettivo: far diventare la rete in grado di capire le nostre richieste. I documenti non dovranno più risultare delle isole di dati, ma dei database aperti nei quali un agente possa distinguere le informazioni contenute, ricavandone solo quelle richieste.

L'elaborazione automatica

•Il Web semantico si basa sui metadati, grazie ad essi un programma può elaborare un documento come se ne conoscesse il significato

•Si aprono tre scenari:

- ✓Applicazione che gestisce dati che conosce
- ✓Applicazioni che si scambiano metadati concordati
- ✓Applicazioni che si scambiano metadati senza accordo preventivo (è il modello di Internet)

•Per Internet occorre sviluppare dei modelli che permettano alle applicazioni di interagire anche in assenza di un accordo preventivo. Per ciò servono degli standard per:

- ✓Rappresentare i metadati
- ✓Scambiare i metadati
- ✓Estrarre significati dai metadati attraverso motori inferenziali

Il Web semantico:

1)Livello logico: ragionamento automatico, assiomi …

2)Livello ontologico: concetti, relazioni, definizioni

DAML+OIL

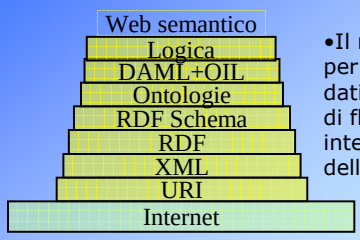
3)Livello di marcatura. Etichettatura semantica dei documenti ed identificazione di strutture.

RDF, RDFS

4)Livello dei dati: documenti, dati, immagini, testi…

XML, XMLS

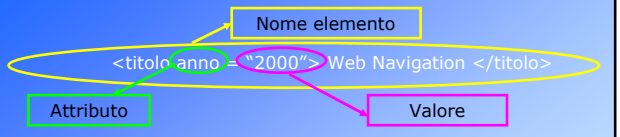
I molti livelli del Web semantico



•Il Web semantico si basa su una struttura formata da molti livelli di significazione.

•Il ricorso a più livelli per la trattazione dei dati si deve alla ricerca di flessibilità ed interscambiabilità dell'informazione.

1.Livello dei dati



•Impiega XML: eXtensible Markup Language e XML Schema.

•Fornisce al Web semantico l'interoperabilità sintattica.

•Un documento XML è un insieme annidato di tag aperti e chiusi, che prendono il nome di elementi in cui ognuno può avere un numero arbitrario di coppie attributo-valore. Il vocabolario degli elementi è definito per ogni applicazione specifica

Limiti di XML

XML è troppo flessibile, mentre per gestire il Web semantico è necessario un modello per esprimere delle conoscenze processabili automaticamente.

```
<?xml version = "1.0"?>
<libro>
  <autore> Jennifer Fleming </autore>
  <titolo> Moll Flanders </titolo>
</libro>
```

```
<?xml version = "1.0"?>
<libro autore = "Jennifer Fleming ">
<titolo> Web Navigation </titolo>
</libro>
```

```
<xml version = "1.0"?>
<libro>
  <titolo> Web Navigation </titolo>
  <autore> Jennifer Fleming </autore>
</libro>
```

•I tre documenti hanno lo stesso contenuto informativo, seppur presentato con strutture diverse

•Per noi umani i nomi degli elementi sono di grande aiuto per interpretarne il contenuto.

```
<?xml version = "1.0">
<v> <x> pppp</x>
<y> qqqq </y> </v>
```

Cosa può dire un calcolatore di questo esempio? Che pppp è y di qqqq o che qqqq è x di pppp? Dal punto di vista della macchina i nomi degli elementi non aiutano, e non aiuta l'indentazione.

- La semantica di un documento XML non è specificata formalmente, ma è incorporata nei nomi dei tag quindi può risultare comprensibile solo all'uomo e non alla macchina.
- Il computer non può confrontare i dati di due documenti decidendo se appartengono ad uno stesso concetto, ad uno stesso tipo di relazione.
- Il W3C ha cercato di ovviare inventando una struttura rigida: RDF (Resource Description Framework).
- Prima della standardizzazione di RDF ci si doveva assicurare che i dati appartenessero ad una stessa DTD o ad uno stesso schema per non creare confusioni.

DTD: prima di XML Schema

```
<! ELEMENT Biblioteca >
<! ELEMENT Libro (Autore | Titolo | Editore | anno) >
<! ELEMENT Autore (Nome | Cognome)>
<! ELEMENT Nome (#PCDATA) >
```

- DTD (Document Type Definition): dichiarazione delle regole gerarchiche che i vari elementi del documento devono seguire
- La DTD definisce la struttura del documento
- Una DTD deve essere inserita all'interno della dichiarazione di documento di un file, la DTD può essere interna o esterna al file

Difficoltà di DTD

- Il principale limite è di essere costruita con una sintassi diversa rispetto a XML. Questo comporta la necessità di due programmi diversi per analizzare un documento DTD, inoltre, non si può eseguire una descrizione nodo per nodo.
- Il W3C per ovviare a queste difficoltà ha dato alcune definizioni di schema. Una definizione di schema è un documento XML, scritto con sintassi XML, che vincola un altro documento ad assumere una certa forma e gerarchia.
- Il W3C ha designato XML Schema come sostituto di DTD.

XML Schema

- eXtensible Markup Language Schema
- Stabilisce la struttura del documento XML

È formato da:

- ✓Documento istanza: il documento che contiene le informazioni che interessano realmente
- ✓Documento schema: il documento che descrive la struttura dei dati contenuti nel documento istanza

2.Il livello schema o di marcatura

•RDF (Resource Description Framework).
Mentre XML aiuta ad associare alle risorse (pagine html o altri oggetti accessibili attraverso il Web) dei metadati descrittivi, RDF ne esprime il significato, e la struttura concettuale (in termini di relazioni con altre risorse o di caratterizzazioni)

•RDF fornirà al Web semantico le informazioni sui dati ed esprimerà le relazioni che intercorrono tra i dati: strumento per aggiungere semantica ad un documento senza dover fare alcuna assunzione sulla sua struttura

•RDF intende supportare l'interoperabilità semantica tra applicazioni che si scambiano informazioni in forma accessibile alla macchina

Caratteristiche di RDF

- Gli elementi costitutivi e la sintassi di RDF sono XML
- RDF consente di costruire delle asserzioni sui contenuti di una pagina web
- Le asserzioni si definiscono in base a delle terne formate da: soggetto, predicato ed oggetto
- Le asserzioni individuano delle relazioni tra i dati di cui trattano, ma non esplicitano ancora il significato di tali relazioni (se dico che un soggetto è predicato di un oggetto ho individuato una relazione, ho indicato gli elementi che entrano in questa relazione, ma non ne so ancora il significato, potrei voler dire che un certo nome corrisponde all'autore di una pagina, ma anche che una pagina è indicata da un certo link, etc.)
- Per definire il significato delle relazioni occorrono le ontologie.

RDF

Il modello base di RDF si basa su tre concetti:

1)Il SOGGETTO è la *risorsa*, cioè qualunque cosa che si desideri descrivere (es: un documento HTML)

<rdf:Description about="http://www.hopslibri.com">

2)Il PREDICATO è la *proprietà*, un aspetto o una caratteristica specifica usata per descrivere il soggetto

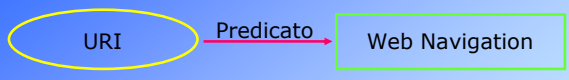
Vende

3)L'OGGETTO è l'affermazione che attribuisce un *valore* alla caratteristica della risorsa

Web navigation

Rappresentazione grafica di RDF

Un'asserzione RDF può anche essere rappresentata graficamente tramite grafi orientati ed etichettati:



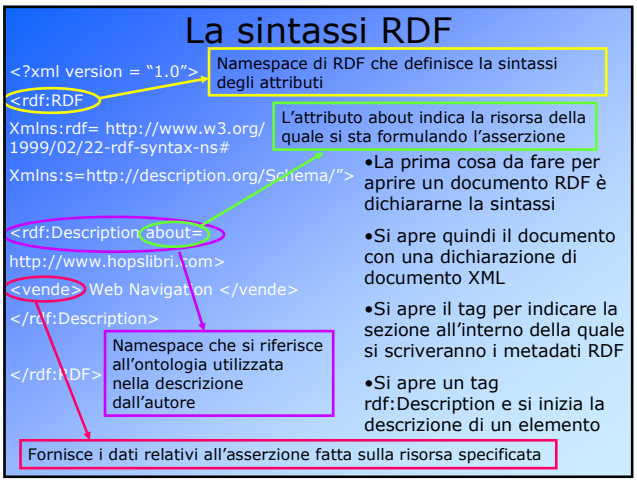
•I nodi a forma ovale rappresentano le risorse

•Gli archi rappresentano le proprietà

•I nodi rettangolari rappresentano i valori

RDF supporta solo relazioni binarie, cioè relazioni fra due risorse, per instaurare relazioni di livello superiore è necessario servirsi di una risorsa intermedia in grado di fornire le rimanenti relazioni che si intendono esprimere.

La sintassi RDF



- La prima cosa da fare per aprire un documento RDF è dichiararne la sintassi
- Si apre quindi il documento con una dichiarazione di documento XML
- Si apre il tag per indicare la sezione all'interno della quale si scriveranno i metadati RDF
- Si apre un tag rdf:Description e si inizia la descrizione di un elemento

La reificazione

- RDF impiega la reificazione per riportare un concetto, l'elemento che funge da predicato, a due localizzazioni, due nodi pagina, dei quali si esprime una relazione. La particolarità e la forza di RDF sta nel fatto di indicare delle URI (Uniform Resource Identifier).

Esempio:

Il professore del corso Z dice che il libro Web Navigation è venduto al sito http://www.hops.com.

Scriviamo prima la tripla http://www.hops.com vende Web Navigation

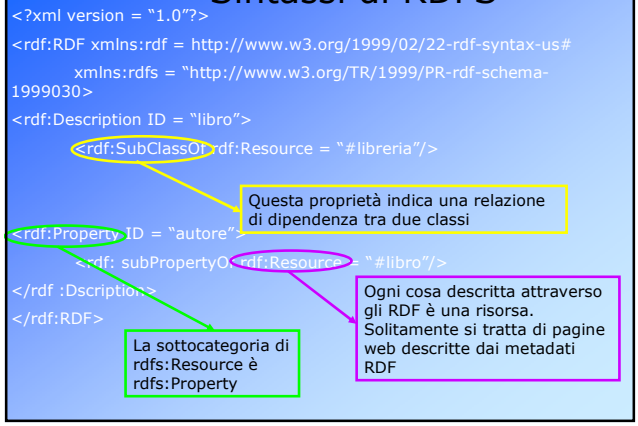
Poi la facciamo diventare una risorsa #Sentence

Facciamo l'asserzione. #Sentence è asserita dal professore del corso Z.

RDF Schema

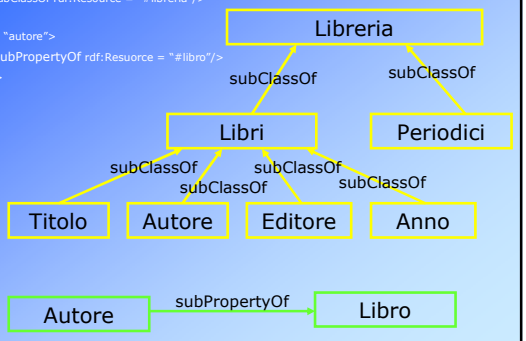
- Resource Description Framework Schema
- RDFS in sintassi XML e attraverso l'uso di namespace, definisce relazioni tra gli elementi RDF usati nella descrizione di metadati
- Il vocabolario RDFS è richiamato tramite il riferimento all'URI http://www.w3.org/1999/02/22-rdf-syntax-us#
- Nel documento RDFS si dovranno specificare le gerarchie tra gli elementi, in modo da fornire ad un agente informatico la capacità di applicare ad essi regole di deduzione

Sintassi di RDFS



```
<?xml version = "1.0"?>
<rdf:RDF xmlns:rdf = http://www.w3.org/1999/02/22-rdf-syntax-us#
        xmlns:rdfs = "http://www.w3.org/TR/1999/PR-rdf-schema-1999030>
<rdf:Description ID = "libro">
        <rdf:SubClassOf rdf:Resource = "#libreria"/>

<rdf:Property ID = "autore">
        <rdf: subPropertyOf rdf:Resuorce = "#libro"/>
</rdf :Description>
</rdf:RDF>
```



3.Livello ontologico.

Un'ontologia è una specificazione formale ed esplicita di una concettualizzazione condivisa. Following Gruber

1) Un'ontologia è un vocabolario che contiene collezioni di asserzioni, che definiscono le relazioni tra i concetti e specificano le regole logiche per ragionare su di essi

- L'ontologia non influenza il grado di comprensione del calcolatore, ma ne accresce il patrimonio di informazioni su cui operare deduzioni
- WordNet è una ontologia per la lingua inglese che distingue due tipi di relazioni:
  - ✓Relazioni lessicali: sinonimia, antinomia, polisemia
  - ✓Relazioni concettuali: iponimia, meronimia

- La relazione più importante in WordNet è la sinonimia. Due espressioni sono sinonime in un contesto linguistico C se la sostituzione di una con l'altra in C non cambia il valore di verità.
- Una stessa parola può comparire in più synset quando è polisemica
- L'ipo/iperonimia, IS A KIND OF, mette in relazione significati subordinati e sovraordinati
- La meronimia, IS A PART OF, stabilisce una gerarchia delle parti all'interno di un insieme di significati
- WordNet consente: retrival delle relazioni polisemiche di un termine, retrival dei sinonimi di una parola, retrival dei percorsi iperonimici e meronimici.

Metodi costruire ontologie

L'acquisizione è un passo critico nella costruzione di una base di conoscenza, infatti se l'ontologia non è sufficientemente ricca, chiara e consistente non ha utilità pratica.

Approcci per costruire le ontologie :

- Ispirazione: si basa sull'immaginazione e sulla creatività personale dello sviluppatore riguardo alle strutture ed alle proprietà del dominio
- Induttivo: tale approccio parte dall'analisi di un buon numero di esempi tratti dalla letteratura per estrarre i dati e le loro relazioni
- Deduttivo: tale approccio si basa sull'adozione di principi generali e sulla loro applicazione in modo da costruire un'ontologia che si adatti ad un caso specifico

DAML+OIL

•Dal punto di vista degli standard, il livello ontologico impiega DAML (DARPA Agent Markup Language) + OIL (Ontology Inference Layer)

•DAML+OIL è un linguaggio ontologico che permette di rappresentare le informazioni del Web in modo che il loro significato sia comprensibile alle macchine

•Originariamente erano due linguaggi separati:

- ✓DAML permetteva di inserire il contenuto semantico dei dati basandosi sulle ontologie definite con RDFS
- ✓OIL è nato come un linguaggio per la creazione di ontologie, basato anch'esso su RDFS

DAML+OIL si basa su:

•Linguaggi che modellano la realtà in classi, ognuna delle quali ha delle proprietà. Tali classi sono legate tra di loro da relazioni di superclasse o sottoclasse

•Linguaggi che descrivono le informazioni in forma matematica in modo che si possano fare dei ragionamenti basati sulla descrizione dei concetti e su classificazioni automatiche.

•La sintassi DAML+OIL è quella di XML e di RDF. Lo schema RDF fornisce un insieme di primitive (come la relazione di sottoclasse) e le regole sintattiche per definire le gerarchie (che non sono previste in XML).

La sintassi di DAML+OIL

•DAML+OIL struttura il dominio in termini di classi e proprietà conformemente alle specifiche memorizzate in un'ontologia

```
•<daml:Class rdf:ID= "libri">
    <rdfs:subClassOf rdf:resource="#libreria"/>
    <daml:disjointWith rdf:resource="#periodici"/>
</daml:Class>

•<daml:Class rdf:about ="#libri">
    <rdfs:subClassOf>
        <daml:Restriction daml:cardinality="1"
            <daml:onProperty rdf:resource="#è scritto"/>
    </rdfs:subClassOf>
</daml:Class>
```

Indica che è sottoclasse di un'altra classe rappresentata come "rdf resource".

La classe non ha elementi in comune con un'altra specificata

La classe è formata dagli elementi che hanno esattamente un certo numero di valori distinti della proprietà

Indica che la classe è ottenuta facendo una restrizione su un'altra, in base ad una specifica proprietà

Una proprietà può essere così specificata:

Es: la proprietà di una libreria di vendere i libri di un determinato autore, che è sottoproprietà di avere dei libri.

```
<daml: ObjectProperty rdf:ID= "ha i libri di ">
    <rdfs:SubPropertyOf rdf:resource = "#ha libri/>
    <rdfs:range rdf:resource= "#testi"/>
</daml:ObjectProperty>
```

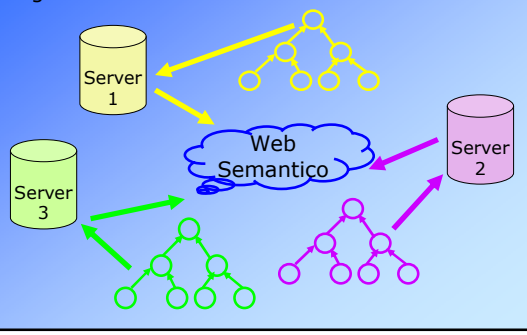
Indica che la proprietà lega un oggetto con un altro oggetto

Indica che la proprietà è una sottoproprietà di un'altra specificata

Indica l'insieme dei valori che la proprietà può assumere, che sono istanze di una classe indicata

Interazione tra ontologie

•Lo sviluppo di ontologie specifiche, costruite su settori diversi, comporta il problema dell'interoperabilità: ontologie differenti dovranno comunicare tra loro

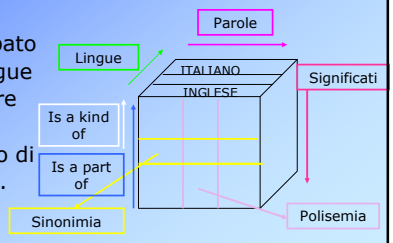


Interazione tra ontologie

•Il ciclo di vita di un'ontologia si basa sui seguenti passi: design, evaluation, validation e revision.

•L'ingegneria ontologica si è posta l'obiettivo di sviluppare delle ontologie che siano condivisibili e riutilizzabili da applicazioni differenti.

•È stato anche sviluppato un traduttore multilingue in modo da trasformare automaticamente informazioni all'interno di alcuni target codificati.



•Gli sviluppatori di ontologie cercano una metodologia, basata su nozioni abbastanza generali, da poter essere applicata indipendentemente dal particolare dominio per il quale è stata sviluppata.

•Si era pensato a dei filtri di conversione tra un database di metadati che contiene una particolare conoscenza ed un altro, ma tale possibilità non è praticabile:

- ✓Il numero dei filtri cresce all'aumentare delle fonti da mettere in relazione
- ✓I filtri dovrebbero essere aggiornati frequentemente per adattarsi al variare delle ontologie

•Un'altra soluzione ancora da studiare cerca di sviluppare un'algebra delle ontologie che rappresentano terminologie provenienti da domini distinti tra loro

4.Il livello logico

•Il Web semantico ha bisogno di modi per inserire la logica all'interno dei documenti

•Abbiamo già dei metadati strutturati, che nel livello logico vengono elaborati prima con gli operatori logici (not, and, or) e successivamente con i quantificatori (for all X, Y(X)).

•Le regole di inferenza potenziano il meccanismo deduttivo di RDF che consente di scrivere regole, ma non specifica in quale ordine esse dovranno essere applicate. Un'ontologia può esprimere la regola "Se un codice postale è associato a uno stato, e l'indirizzo di un'università usa quel codice postale, allora quell' università è in quello stato".

•Questo linguaggio logico unificato permetterà di collegare tutti i concetti tra loro in una unica rete di sapere universale che consentirà la comunicazione tra i gruppi umani che utilizzano "Ontologie" differenti, cioè che di norma non si comprendono a vicenda, ed aprirà la conoscenza all'analisi degli "agenti intelligenti" dotandoci di una nuova classe di strumenti con quali vivere e lavorare

•Le ontologie possono migliorare il funzionamento del web in molti modi: se le si usa ad un livello semplice rendono più efficaci ed accurate le operazioni dei motori di ricerca, ma ad un livello più avanzato agiscono come vere e proprie riorganizzatrici del sapere

Query language

•XQuery: Query Language for eXtensible Markup Language

•XQuery è un linguaggio funzionale che comprende alcuni tipi di espressioni che possono essere nidificate e composte con grande generalità

•XQuery si basa su XML Schema ed è disegnato per essere compatibile con gli altri linguaggi standard di XML

•Le interrogazioni tra applicazioni dovranno essere delle domande che esprimono nuovamente i dati in XML

Un esempio di XQuery

Trova il titolo del libro pubblicato da Jennifer Fleming nel 2000.

```
FOR $b IN document ("lib.xml")/lib/libro
WHERE $b /autore=" Jennifer Fleming"
      AND $b@anno= "2000"
RETURN $b/titolo
```

<titolo> Web Navigation </titolo>

Espressioni FLWR

•Elementi base di FLWR:

- ✓For: introduce una variabile
- ✓Let: introduce una variabile aggiuntiva
- ✓Where: agisce da filtro tra le condizioni
- ✓Return: stabilisce quale dovrà essere il risultato

Conclusioni

•Lo sviluppo del Web Semantico interessa molte aree del Web tradizionale:

✓Siti di commercio elettronico in cui le ontologie facilitano la comunicazione tra venditore ed acquirente, o tra mercati diversi e lontani, consentendo che le stesse descrizioni siano utilizzate in siti di vendita diversi

✓Motori di ricerca a cui le ontologie permettono di fare non solo ricerche basate su parole chiave, ma anche sui contenuti semantici inclusi nella pagina

•Il web Semantico facendo riferimento ad ogni entità con un URI permetterà a chi lo vorrà di descrivere nuovi concetti, basandosi eventualmente su altri già esistenti, e, con l'impiego di linguaggi standard consentirà di collegare le informazioni apportate da ognuno con il resto del Web

•La struttura che si verrà a formare fornirà una serie di strumenti nuovi in grado di comunicare, dialogare ed "imparare" insieme

Bibliografia

Amedeo Cappelli, Maria Novella Catarsi, Patrizia Michelassi, Lorenzo Moretti, *Conceptual and linguistic constraints for the construction of a knowledge base in archaeology*, Istituto di Linguistica Computazionale del CNR Pisa, Dipartimento di Informatica Università di Pisa

Mauro Iannucci, Massimo Imperiali, *Semantic Web,* Infomedia

Andrea Albini, *Il futuro della rete che capisce,* Internet News, Tecniche nuove S.p.a.

Ernesto Damiani, *www10 La tela di Penelope,* Internet News, Tecniche nuove S.p.a.

Michael Gruninger, Jintae Lee, *Ontology Application and Design,* Communication of the ACM, Febbraio 2002, vol. 45. No 2

Tim Berners-Lee, *Semantic Web Road Map,* Settembre 1998

Sean B. Palmer, *The semantic Web: an introduction*

Ora Lassila, Deborah McGiunness, *The role of Frame-Based Rappresentation on the Semantic Web,* Software Tecnology Laboratory, Nokia Research Center, & Knowlwdge System Laboratory, Stanford Univerity

Bernardo Magnini, Manuela Speranza, *Merging global and specialized linguistic ontologies,* ITC-irst Istituto per la ricerca scientifica e tecnologica, Trento, Italia

Bernardo Magini, *Costruzione di una base di conoscenza lessicale per l'iataliano basata su WordNet,* ITC-irst Istituto per la ricerca scientifica e tecnologica, Trento, Italia

Sean Bechhofer, jeen Broekstra, Stefan Decker, Michael Erdmann, Dieter Fensel, , Carole Goble, Frank van Harmelen, Ian Horrocks, Michel Klein, Deborah McGiunness, Enrico Motta, Peter Patel-Schneider, Steffen Staab, Rudi Studer, *An informal description of Standard OIL and Istance OIL,* Novembre 2000

Ian Horrocks, Deborah McGiunness, Christopher Welty, *Digital Libraries and Web-Based Information Systems*

D. Chamberlin, *XQuery: An XML query language,* IBM SYSTEM JOURNAL, Vol 41, No 4, 2002