The European Online Magazine for the IT Professional
http://www.upgrade-cepis.org
Edizione italiana a cura di ALSI e Tecnoteca
http://upgrade.tecnoteca.it

Interrogare il Semantic Web

di
José-Francisco Aldana-Montes, Antonio-César Gómez-Lora, Nathalie Moreno-Vergara, and María del Mar Roldán-García

(Traduzione italiana a cura di Roberto Fresco (ALSI – www.alsi.it) dell'articolo
Querying the Semantic Web:Feasibility Issues
pubblicato sul Vol. III,No. 4, Agosto 2002
della rivista online UPGrade, a cura del CEPIS)

Riassunto italiano: **Interrogazioni del Web: Quali possibilità?** Al momento, una intera nuova tecnologia è in corso di sviluppo sulla base degli standard XML per il processamento delle interrogazioni su sorgenti eterogenee di dati. Questo include, fra l'altro, la specifica di linguaggi di interrogazione e di tecniche per la mediazione ed integrazione di sorgenti di dati. La nuova grande sfida consiste nello sviluppare ciò che è già noto come il Semantic Web, prendendo questa tecnologia come punto di partenza. E' il trattamento dell'informazione molto efficiente, basato su livelli logici ed ontologici, che costituisce uno dei punti di forza che determineranno il suo successo pratico.

Parole chiave: Integrazione di dati, query di elaborazione basate sulle ontologie, Web, semantica.

1. Introduzione

Dal momento in cui Internet ha assunto il ruolo di rete globale di comunicazione e di strumento per lo sviluppo dei sistemi informativi, è stato evidente che qualcosa di più delle pagine HTML statiche fosse necessario.
Dopo l'arrivo dell'XML, considerato come la modalità di descrizione dei dati strutturati e semi-strutturati, il World Wide Web Consortium (W3C) ha iniziato a sviluppare tecnologie attorno a questo nuovo standard. Possiamo citare tra queste XMLS (Extensible Mark-up Language Schemas), RDF (Resource Description Framework), RDF-Schema, XSL/XSLT (Extensible Style sheet Language/Transformation) e SOAP (Simple Object Access Protocol).
Le raccomandazioni del W3C, insieme all'uso dell'XML inteso come formato per lo scambio elettronico dei dati, aprono la possibilità di integrazione di sorgenti eterogenee di dati nell'ambiente del World Wide Web. Come risultato di questo, le più importanti aziende nello sviluppo software, come Oracle, Microsoft e IBM, hanno introdotto il supporto all'XML nelle loro basi di dati e prodotti.
Tutta questa tecnologia ha contribuito allo sviluppo di applicazioni innovative che non richiedono interazione umana, rendendo il Web più "comprensibile" alle macchine.
Berners-Lee, il creatore del Web, descrive il Semantic Web come il successivo passo nell'evoluzione delle applicazioni Web, in cui "l'informazione ha un preciso significato, rendendo possibile per i calcolatori e le persone lavorare assieme" [Berners-Lee et al. 2001].
Lo sviluppo del semantic Web coinvolge l'adozione di svariate tecnologie (Figura 1), che danno la possibilità di aggiungere significato alla struttura dei documenti XML e di descrivere le necessarie informazioni semantiche per abilitare l'elaborazione automatica da parte delle macchine.

© Novática

Da un po' di tempo questo problema viene affrontato per mezzo delle "ontologie" descritte come documenti o file che "definiscono in maniera assoluta le relazioni tra i termini". Le ontologie consentono di lavorare con i concetti, invece che con parole chiave, all'interno dei sistemi di recupero dell'informazione.

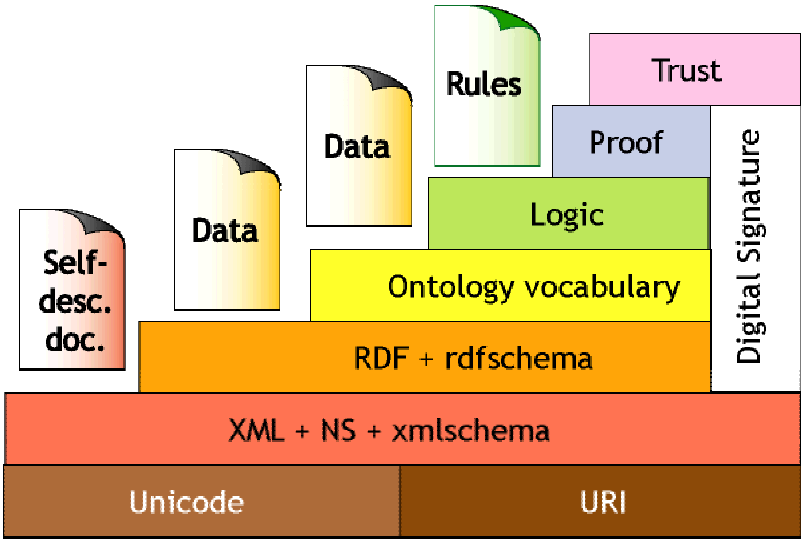


Figura 1: Visione del Semantic Web presentata da Tim Berners-Lee
http://www.w3.org/2000/Talks/1206-xml2k-tbl/slide10-0.html

In relazione alle sorgenti dell'informazione, le ontologie descrivono il contenuto di repository di dati al di là della loro rappresentazione sintattica, dando possibilità di integrazione semantica [Mena et al. 2000]. Un problema che scaturisce nel prossimo futuro, dovuto al rapido aumento di specifiche ontologie (e standard per i metadati), è l'integrazione stessa delle ontologie.

2. Linguaggi per il Semantic Web

Allo scopo di definire una concettualizzazione ovvero un'ontologia, è necessario disporre di un linguaggio per rappresentare quella conoscenza.
Tradizionalmente questi linguaggi sono stati sviluppati nell'area dell'Intelligenza artificiale e focalizzano la loro base formale sui paradigmi del calcolo dei predicati del primo e secondo ordine, la logica descrittiva o anche i paradigmi object oriented.
Sono le diverse proprietà espressive e computazionali che differenziano poi questi linguaggi. Tra quelli più rilevanti citiamo Ontolingua, Loom, OCML e Flogic.
Poiché XML è diventato "de facto" lo standard per lo scambio di dati tra le applicazioni, sono stati sviluppati linguaggi di rappresentazione della conoscenza per il Web che rendono possibile la pubblicazione delle ontologie usando la sintassi di questo standard. In questo modo, per quanto è possibile, si evita l'arduo compito di definire dei parser appositi. I linguaggi basati sull'XML sono candidati ideali per il Semantic Web. Alcuni di questi sono il Simple HTML Ontology Extensions (SHOE), l'Ontology Exchange Language (OML) e il Resource Description Framework Schema (RDFS). Essi ereditano le caratteristiche dell'XML e incorporano nuove caratteristiche che migliorano l'espressività del modello dati iniziale.
Ulteriori proposte di linguaggi estendono RDF e RDFS come: Ontology Interchange Language (OIL) e il suo successore (DAML+OIL).
Visto che l'XML impone la necessità di restrizioni strutturali (una base comune), nell'idea di fornire metodi esenti da errore per le espressioni semantiche, RDF fornisce l'infrastruttura per permettere la codifica, il riuso e lo scambio di metadati strutturati. In questo modo, l'RDF è il modello più promettente per associare informazioni al contenuto delle risorse Web.
E' improbabile che un singolo linguaggio possa coprire tutte le necessità e i requisiti presentati nel Semantic Web. Un'analisi delle caratteristiche maggiormente auspicate, riguardanti l'espressività e

la potenza dei meccanismi di ragionamento, ci fornirà il profilo per un efficiente linguaggio per il Semantic Web. Viene fatta una chiara distinzione tra i termini: rappresentazione della conoscenza e il ragionamento.
La formalizzazione della conoscenza è portata avanti, in molti casi, attraverso l'uso di concetti, relazioni n-arie, funzioni, procedure, istanze, assiomi, regole di produzione ed attraverso la semantica formale. Questi fattori determinano l'espressività del linguaggio.
La tabella 1 mostra un confronto tra questi elementi; il simbolo '+' indica un aspetto supportato dal linguaggio, il simbolo '–' indica invece un aspetto non supportato, '+/–' indica un aspetto non direttamente supportato (ma che può essere modellato con le risorse fornite dal linguaggio) e 'NA' si riferisce ad aspetti non supportati di piccola rilevanza.
E' possibile notare come i linguaggi basati sull'XML non forniscano di solito la possibilità di definire funzioni, procedure e assiomi, eccetto qualche piccola assiomatizzazione come le regole deduttive nel caso di SHOE. Essi perdono anche la semantica formale inerente al linguaggio stesso. Tutto ciò rende difficile l'implementazione di meccanismi efficienti di ragionamento. In questo senso Ontolingua è forse il più espressivo di tutti i formalismi presentati, sebbene attualmente non esista nessun motore inferenziale che li implementi.

	Ontolingua	OCML	LOOM	FLOGIC	XOL	SHOE	RDF(S)	OIL
Concepts	+	+	+	+	+	+	+	+
n-ary Relations	+	+	+	+/–	–	+	+	+
Functions	+	+	+	+/–	–	–	–	+
Procedures	+	+	+	–	–	–	–	–
Instances	+	+	+	+	+	+	+	NA
Axioms	+	+	+	+	–	–	–	NA
Rules of Production	+	+	+	–	–	–	–	NA
Formal Semantics	+	+	+	+	+	–	–	–

Tabella 1: Confronto tra linguaggi relativamente agli elementi per la rappresentazione della conoscenza, presentato in [Corcho/Gómez 2000

Altri interessanti confronti possono essere affrontati in relazione ai meccanismi di ragionamento che i linguaggi permettono. La tabella 2 considera questi aspetti. OIL ha un automatico sistema di classificazione (desiderabile nel caso delle ontologie), Flogic implementa il trattamento di eccezioni ed entrambi questi linguaggi presentano tipici sistemi di inferenza. Confrontati ai linguaggi basati sull'XML, i linguaggi tradizionali supportano l'implementazione di procedure, il mantenimento di restrizioni, ed entrambi le valutazioni top-down e bottom-up.

	ONTOLINGUA	OCML	LOOM	FLOGIC	XOL	SHOE	RDF(S)	OIL
Inference Mechanisms								
Correct	–	+	+	+	–	–	–	+
Complete	–	–	–	+	–	–	–	+
Classification								
Automatic Classification.	–	–	+	–	–	–	–	+
Exceptions								
Use of Exceptions	–	–	–	+	–	–	–	–
In heritance								
Monotonic	+	+	+	+	NA	+	NA	+
Non-Monotonic	+/–	+/–	+	+	NA	–	NA	–
Simple Inheritance	+	+	+	+	NA	+	+	+
Multiple Inheritance	+	+	+	+	NA	+	+	+
Procedures								
Implementation of Procedures	+	+	+	–	–	–	–	–
Restrictions								
Examination of Restrictions	+	+	+	+	–	–	–	–
Evaluation Model								
Top-Down	–	+	+	+	–	NA	–	–
Bottom-Up	–	+	+	+	–	NA	–	–

Tabella 2: Meccanismi di ragionamento nei linguaggi, presentati in Corcho/Gómez 2000

In tal modo esiste un importante compromesso tra completezza ed espressività dei meccanismi di inferenza usati e la loro efficienza. Tutto ciò rende molto interessante lo studio e lo sviluppo di tecniche per la valutazione distribuita di programmi logici in questo contesto e le tecniche che forniscono supporto all'elaborazione delle interrogazioni basate sulle ontologie nel semantic web.

3. Un framework per il Semantic Web: implicazioni pratiche

Heflin propone un framework formale per la rappresentazione della conoscenza nel Semantic Web, basato sulle ontologie [Heflin 2001]. Siamo interessati nell'applicazione di quel generico framework all'integrazione e interrogazione dei Web services. Sebbene Heflin estenda la definizione formale dell'ontologia durante lo sviluppo, per semplificare la spiegazione noi useremo soltanto una delle sue prime definizioni.
Un'ontologia è definita come una tripla O=<V,A,E>, dove V corrisponde al vocabolario, A è l'insieme di assiomi e E rappresenta l'insieme delle ontologie che sono di fatto estese per mezzo di O.
Ci sono allora due funzioni associate alle risorse: C(r) che determina l'insieme di ontologie con le quali r è in corrispondenza e K(r) che traduce l'informazione della risorsa r, introducendola nella teoria formale in termini del vocabolario dell'ontologia (modellando in un certo senso quello che è il concetto di wrapping).
In figura 2 possiamo vedere, in uno schema grafico, un esempio nel quale sia O_U e sia O_F estendono O_G. In aggiunta si ha la visione di O_U (un concetto base in questo framework) in Figura 2.
Nel modello presentato da Heflin, le interrogazioni sono eseguite con prospettive ontologiche.
La prospettiva PO_U definita nell'ontologia O_U è l'unione degli assiomi dell'ontologia O_U, le funzioni della conoscenza di tutte le risorse che corrispondono all'ontologia O_U e le funzioni di conoscenza di tutte le risorse che corrispondono con tutte le ontologie estese da O_U.
Questa costruzione formale può essere sviluppata in diversi modi in relazione ai differenti modelli di implementazione, ognuno dei quali ha importanti implicazioni pratiche (dal punto di vista della realizzazione ed efficienza di implementazione) e teoriche (anche esprimere la necessità di estendere, "specificare" il precedente modello).
L'impostazione più semplice per il Semantic Web dovrebbe essere quella in cui tutte le fonti di dati corrispondano ad un insieme limitato, ma ben conosciuto ed largamente accettato, di ontologie.

Inoltre, tutte le fonti di dati dovrebbero essere universali ed omogeneamente accessibili. In tal modo, ogni sistema collezionerebbe velocemente gli assiomi e le risorse della propria prospettiva e potrebbe elaborarli. E' evidente che questo scenario non è molto realistico in qualità di architettura del Semantic Web. Non ci si aspetta che ogni risorsa sia messa in corrispondenza ad un certo numero di ontologie e tanto meno che un sistema isolato sia in grado di elaborare un'intera prospettiva che investe un gran numero di risorse. E' più facile immaginare che ci saranno molte ontologie e molti sistemi in grado di valutare ontologie specifiche (SEO_i).
In questo modo, le prospettive dovrebbero essere realizzate in modo cooperativo, grazie all'interazione di questi agenti e sistemi. In questo contesto è essenziale ottenere l'interoperabilità ed autonomia del SEO_i. Questa richiede uno specifico isolamento di livelli differenti, che noi analizzeremo in tre differenti dimensioni. Ognuna di queste presenta un insieme di problemi che devono esser trattati in relazione ad aspetti quali la proprietà intellettuale, l'eterogeneità e la distribuzione.
Dal punto di vista della proprietà intellettuale, un'azienda potrebbe voler isolare o proteggere le fonti di dati e/o gli assiomi che definiscono gli elementi del vocabolario della propria ontologia. Prendendo i tre elementi fondamentali che costituiscono il modello di un'ontologia, ovvero il vocabolario, gli assiomi e le fonti di dati modellate dalle funzioni della conoscenza K(r), soltanto il primo di questi deve essere necessariamente pubblico (totalmente o parzialmente).
La seconda dimensione con la quale noi lavoriamo è l'eterogeneità, basata sul formalismo a due fattori fondamentali e il meccanismo di inferenza. L'estensione di un'ontologia ha bisogno di essere realizzabile persino quando si usano due formalismi distinti per rappresentare gli assiomi. Allo stesso modo, i due sistemi basati sulle ontologie possono usare diversi meccanismi di inferenza per calcolare i termini di tali assiomi.

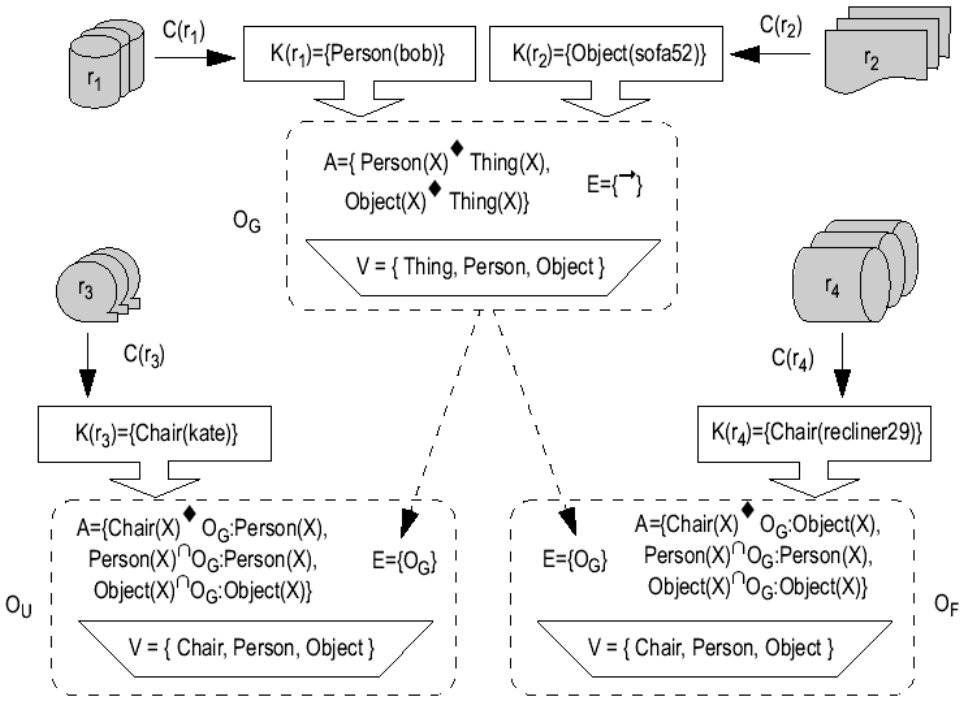
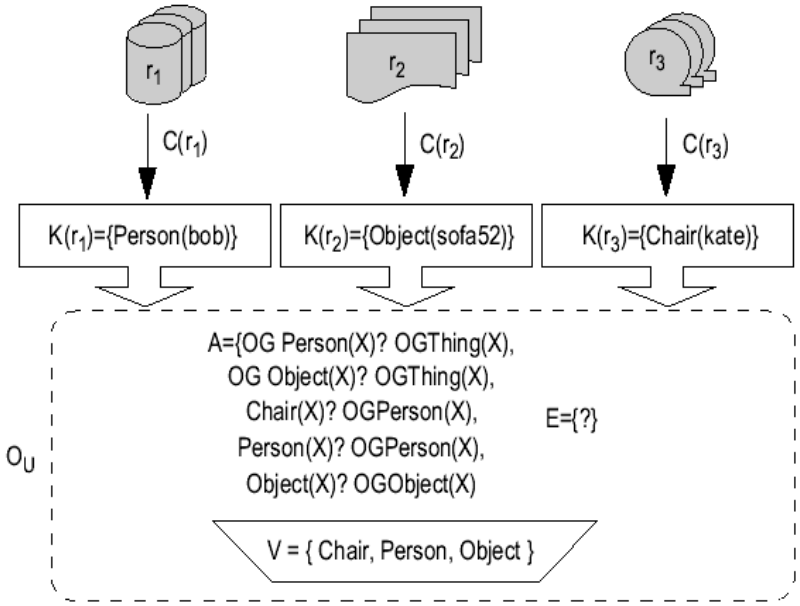


Figura 2: Esempi di ontologie, estratte da [Heflin 2001]

Nel semplice modello di implementazione del framework di Heflin, possiamo pensare, forse, che l'eterogeneità nel formalismo debba essere risolta introducendo un traduttore. Comunque, un caso estremo di eterogeneità (che è sia reale sia comune) corrisponde alla dichiarazione dei Servizi a Valore Aggiunto (Valued Added Service – VAS).

Un VAS è un'integrazione o un sistema di mediazione che può essere inteso come una funzionalità aggiunta ed incorporata nel sistema e la cui semantica non può essere espressa nel linguaggio usato per la sua integrazione. Il concetto di VAS appare in molti campi, come per esempio, le stored procedure e le procedure esterne in un sistema DBMS o i predicati esterni nei linguaggi logici. Nel lavoro di Heflin, tutto questo corrisponde ai predicati dell'ontologia le cui definizioni potrebbero non essere ottenute nel formato usato per esprimere l'insieme stesso degli assiomi.



L'analisi della terza dimensione, che noi abbiamo chiamato distribuizione, è sostenuta dalle considerazioni sugli aspetti che sono di notevole importanza per la reale realizzabilità del Semantic Web. Ci sono due aspetti fondamentali della distribuzione: il primo si riferisce alla valutazione dell'ontologia e il secondo all'ottimizzazione di questa valutazione. In relazione alla valutazione, si percorrono due aree importanti come la replicazione, il bilanciamento del carico e il livello di tolleranza

Figura 3: Teoria formale equivalente alla prospettiva di O_U nell'esempio di figura 2.

degli errori (errori attribuibili alla perdita e non accessibilità ai dati oppure alla perdita dei sistemi o degli agenti di valutazione). In aggiunta, le tecniche di ottimizzazione cambiano radicalmente quando si considerano gli aspetti di non trasparenza (opacità) delle risorse, degli assiomi ecc.
Riepilogando, ritornando all'esempio in figura 2, per mantenere l'eterogeneità e l'autonomia dovremmo ottenere lo stesso PO_U (Figura 3) in un contesto nel quale l'unica parte nota è quella mostrata nel diagramma di figura 4.

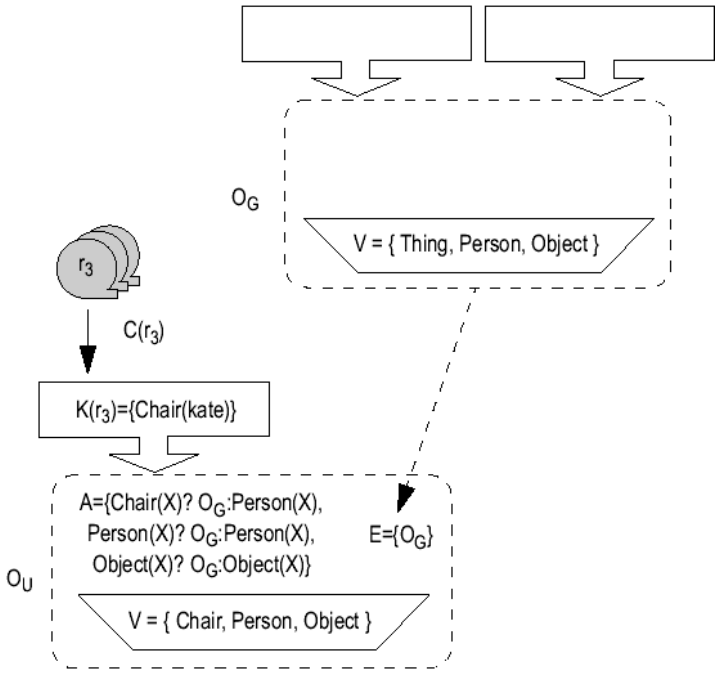


Figura 4: Opacità nella prospettiva di O_U

Mantenendo l'originale prospettiva questo diagramma introduce nuovi problemi, poiché richiede che i sistemi in grado di valutare O_U (SEO_U) e O_G (SEO_G) interagiscano per ottenere la soluzione. Questa interazione è molto più complessa del più semplice uso di SEO_G attraverso SEO_U. Entrambi devono cooperare per ottenere una soluzione valida. Per esempio, da PO_U può essere inferito OG:Person(Kate), che allo stesso tempo inferirà O_G:Thing(Kate) (vedi figura 3).
In tal modo SEO_U chiede a SEO_G se Person(Kate) è vero. In un ambiente distribuito (dove SEO_U e SEO_G non hanno tutte le informazioni) SEO_G risponderebbe nel caso ottimo, "no" o "Non so" (ciò dipende dal fatto se usa ipotesi di un mondo chiuso o no).
Dal punto di vista della costruzione di PO_U, senza violare l'autonomia delle ontologie o senza forzare una valutazione centrata, se SEO_U ha informato SEO_G del fatto che Person(Kate) è vero in PO_U, piuttosto che chiederlo, allora SEO_G non solo sapra' questo, ma risponderà che anche Thing(kate) in PO_U è vero.
Queste interazioni diventano più complesse se consideriamo il fatto che, all'interno di una interrogazione, le risposte appartengano ad un modello Markoviano. Nel precedente esempio, questo è equivalente a richiamare le affermazioni che Person(Kate) e Thing(kate) nelle successive interazioni.
Al momento stiamo lavorando allo sviluppo ulteriore di questo framework formale, adattando e sviluppando le tecniche dei database per ottenere una implementazione del modello che abbiamo descritto. Quest'ultimo è basato sul nostro precedente lavoro, relativo alla valutazione distribuita e ottimizzazione di programmi logici distribuiti per mezzo di un modello di flusso dei dati (D^3 – Distributed Dataflow Datalog) [Aldana 1998], [Aldana et al. 1996], e sullo sviluppo ulteriore di sistemi di mediazione basati su questo modello (D^3 – Web) [Aldana et al. 1997], così come il nuovo lavoro sviluppato alla Data Stream Systems [Babcock et al. 2002].

Conclusioni

Il semantic Web continuerà ad avere un modello di crescita simile a quello del Web e il successo dipenderà da noi se saremo in grado di sviluppare tecniche realistiche che renderanno questo modello di sviluppo realizzabile. Il compromesso adottato tra il concetti di potenza, completezza ed efficienza, per ognuno dei differenti meccanismi di inferenza, aprirà un'ampia gamma di studio per nuove tecniche di valutazione – basate sulle ontologie – per programmi logici distribuiti all'interno del contesto del Semantic Web. Comunque, l'interrogazione e la gestione efficiente di basi di conoscenza distribuite ha ancora molti aspetti non risolti, inclusi quelli relativi all'integrazione efficiente di informazione e lo sviluppo di meccanismi di inferenza distribuiti.

Bibliografia

[Aldana et al. 1996]
J. F. Aldana, J.M.Troya, DataFlow Evaluation of Datalog Queries, en: International Joint Conference and Symposium on Logic Programming, Bonn, Alemania (1996) 69–78.

[Aldana et al. 1997]
J. F. Aldana, M.I.Yagüe, WWW as a Distributed Deductive Database, en: 8th International Conference on Management of Data, Madras, La India (1997) 277–292.

[Aldana 1998]
J. F. Aldana, Un Modelo de Flujo de Datos para la Evaluación de Consultas en Bases de Datos Deductivas. Tesis Doctoral. Universidad de Málaga. 1998.

[Babcock et al. 2002]
B. Babcock, S. Babu, M. Datar, R. Motwani, J. Widom, Models and Issues in Data Stream Systems. 21st ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS), (2002).

[Berners-Lee et al. 2001]
T. Berners-Lee, J. Hendler, O. Lassila. The Semantic Web. Scienti
.c American. Mayo 2001.
<http://www.sciam.com/2001/0501issue/0501berners>.
[Corcho/Gómez 2000]
O. Corcho, A. Gómez. A Roadmap to Ontology speci.cation
Languages. Proceedings of Knowledge Acquisition, Modelling
and Managements, 12th International Conference, EKAW 2000.
Lecture Notes in Computer Science 1937 Springer 2000.
[He.in 2001]
J. D. He.in. Towards the Semantic Web: Knowledge Representation
in a Dynamic, Distributed Environment, PhD Thesis, 2001.
[Mena et al. 2000]
E. Mena, A. Illaramendi, V.Kashyap, A.P. Sheth "OBSERVER:
An Approach for Query Processing in Global Information systems
based on Interoperation Across Pre-Existing Ontologies".
Distributed and Parallel Databases 8(2), 2000.

Note biografiche sugli autori

José-Francisco Aldana-Montes
Ha ricevuto il suo grado Ph.D dall'Università di Málaga (Spagna) nel 1998. Egli attualmente ha il ruolo di professore associato nel dipartimento di Computer Science dell'Università di Málaga. E' stato membro del comitato per il JISBD dal 1999 al 2002. I suoi interessi di ricerca sono relativi allo studio dell'integrazione di database e tecnologie Web. Ciò include valutazione e ottimizzazione (distribuita) di interrogazioni ricorsive; valutazione distribuita di Xquery; ottimizzazione (semantica) di interrogazioni XML, integrazione semantica e modelli teorici per il Semantic Web, e l'elaborazione delle interrogazioni nel Semantic Web. jfam@lcc.uma.es.

Antonio-César Gómez-Lora
E' assistente nel dipartimento di Computer Science presso l'Università di Málaga (Spagna). Ha ottenuto la laurea M.S. in Computer Science presso l'Università di Málaga nel 1997 ed attualmente lavora per ottenere il titolo di Ph.D. presso questa Università. I suoi interessi di ricerca includono tecniche ibride per la valutazione di interrogazioni ricorsive e l'ottimizzazione in sistemi distribuiti e tecniche di ottimizzazione al tempo di valutazione della query. cesar@lcc.uma.es.

Nathalie Moreno-Vergara
È una studentessa di dottorato, con borsa di ricerca, al dipartimento di Computer Science dell'Università di Málaga (Spagna). Ha preso il titolo di M.S. in Computer Science nel 2000 e attualmente sta lavorando per ottenere il Ph.D. presso questa Università.
I suoi interessi di ricerca sono focalizzati sulle ontologie, integrazione semantica e modelli teorici per il Semantic Web (linguaggi di interrogazione e meccanismi di ragionamento per quell'ambiente) vergara@lcc.uma.es.

María del Mar Roldán-García
E' studentessa di dottorato, con borsa, al dipartimento di Computer Science di Málaga (Spagna). Ha preso il titolo di M.S. in Computer Science nel 2000 e attualmente sta lavorando per ottenere il Ph.D. presso questa Università. I suoi interessi di ricerca includono l'integrazione di sorgenti di dati eterogenee, ontologie e tecniche di indicizzazione. mmar@lcc.uma.es.