Verso il semantic web

La tappa italiana del "W3C Semantic Tour"

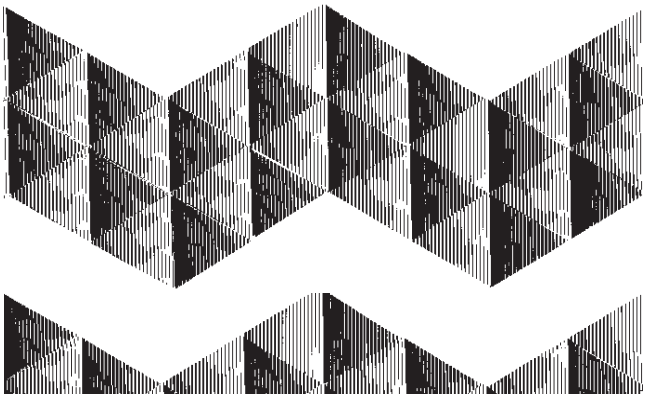
Lo scorso 10 giugno si è svolta a Roma, presso l'Università Gregoriana, la tappa italiana del "W3C Semantic Tour", un'iniziativa del World Wide Web Consortium (W3C) che nel giugno 2003 ha organizzato una serie di eventi della durata di un giorno in tutta Europa allo scopo di promuovere le tecnologie W3C, sottolineando il fatto che il web ha ormai dati ben definiti e linkati in modo da permettere una discovery più efficace, l'automazione, l'integrazione e il riutilizzo in diverse applicazioni.

Stando alla definizione di Tim Berners-Lee, inventore del World Wide Web e direttore del consorzio, il *semantic web* "*is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation.*"[1]

L'organizzazione a livello locale è stata curata dall'Ufficio W3C italiano, situato presso l'Istituto di scienza e tecnologie dell'informazione "A. Faedo" (ISTI) del CNR a Pisa.

Gli interventi più interessanti sono stati presentati nell'arco della mattinata, a cominciare da Daniel Dardailler (W3C associate chairman for Europe) che dopo aver presentato il consorzio, creato nell'ottobre del 1994 per portare il World Wide Web al pieno potenziale sviluppando protocolli comuni che ne promuovessero l'evoluzione e assicurassero l'interoperabilità,[2] ha illustrato la *mission*, le attività e il modello organizzativo, in particolare la distribuzione di uffici e host nel mondo. Massimo Marchiori (W3C-MIT) ha quindi introdotto il concetto di *semantic web,* partendo dal fatto che il passaggio all'XML non rappresenta la risoluzione del problema di un recupero efficace dell'informazione. Il *semantic web* è orientato all'accesso e a una migliore semantica: l'attuale modello di ricerca (diretta o locale) non funziona più dal momento che l'utente vuole accedere ai dati e avere il documento. I *web services*,[3] architetture di base SOA (Service Oriented Architecture)[4] rappresentano una soluzione allettante ma la necessità attuale è introdurre più semantica. La grande flessibilità dell'XML può rivelarsi in realtà un'arma a doppio taglio poiché la proliferazione di dialetti XML può portare a problemi di interoperabilità, causando l'effetto cosiddetto *lost in translation*.[5] In uno schema che descrive l'architettura del web possiamo definire RDF (Resource Description Framework) come il mattone principale, il linguaggio unico, che precisa la semantica delle risorse descritte nelle diverse comunità di utilizzatori di metadati. Come XML, RDF è un linguaggio estensibile, un metalinguaggio, è una cornice (*framework*) di descrizione delle risorse usabile indipendentemente dal dominio di applicazione. Un'applicazione interessante è Annotea,[6] un progetto del W3C che ha prodotto un sistema di annotazioni condivise delle pagine web, ossia commenti, note, spiegazioni o altri tipi di osservazioni esterne che possono essere "attaccate" (possiamo assimilarlo all'uso che facciamo normalmente dei foglietti post-it ma immaginandoli come oggetti dinamici) a un qualsiasi documento web senza alcuna necessità di toccare il documento. Per descrivere annotazioni come i metadati, Annotea usa uno schema di annotazioni basato su RDF che gestisce sia l'annotazione sia i dati ad essa relativi.[7] Gli equivoci che scaturiscono dalle attuali applicazioni semantiche sono innumerevoli se ci basiamo sulla natura monotonica di RDF:[8] partendo dal principio che se si crede all'insieme, si crede anche alle sottocomponenti, possono scaturire problemi interpretativi di una certa consistenza.

Altra tecnologia di recente sviluppo legata a RDF è Web Ontology Language (OWL),[9] usato per descrivere ontologie (classificazioni), concetti (classi e sottoclassi), equivalenze. Un'ontologia definisce i termini usati per descrivere e rappresentare un'area della conoscenza: includono definizioni di concetti base di un dominio (ossia un'area di conoscenza specifica) usabili dalle macchine e le relazioni tra essi, codificando la conoscenza in uno o più domini e rendendola pertanto riutilizzabile. Si tratta di rappresentazioni tipo le tassonomie (ad es. la gerarchia di Yahoo), gli schemi di metadati (come il Dublin Core) o teorie logiche. Il *semantic web* ha bisogno di ontologie con un livello strutturale significativo. OWL presenta una struttura a tre livelli: le *Classi* (cose generiche) nei molti domini di interesse; le *Relazioni* che possono esistere tra esse e le *Proprietà* (o attributi) che tali cose possono avere. Le ontologie sono dunque uno strumento per accordarsi in modo preciso sul significato dei termini utilizzati in un certo contesto e sulle relazioni che intercorrono tra questi termini. Lo schema RDF fornisce lo strumento per definire il vocabolario, la struttura e i vincoli per esprimere metadati relativi alle risorse web. Uno stesso concetto può avere diverse URI di riferimento e l'utilità dell'URI decresce in base al numero delle varianti: OWL riduce questo problema grazie alla possibilità di stabilire relazioni tra ontologie, agendo come "traduttore" a livello superiore RDF. Senza un'appropriata informazione semantica le forme di aggiornamento automatico fal-



liscono e potrebbero trarre notevole vantaggio dal *semantic web*. Un esempio di applicazione avanzata è Closed World Machine (CWM) sviluppato qualche tempo fa da Tim Berners-Lee, un data processor che utilizza il linguaggio RDF/XML e può essere usato per richiedere, controllare, trasformare e filtrare l'informazione. RDF rappresenta comunque solo il primo passo verso quello che proprio Tim Berners-Lee chiama *semantic web*, un World Wide Web dove i dati sono strutturati e gli utenti possono pienamente beneficiare di questa struttura quando accedono all'informazione sul web. In realtà RDF fornisce il vocabolario base in cui i dati possono essere espressi e strutturati: il problema dell'accesso e della gestione di questi dati resta. Metalog[10] fornisce una vista "logica" dei metadati presenti sul web. L'approccio è composto da tre livelli (modello, logica, sintassi).

Al vertice del modello del web presentato da Marchiori c'è la fiducia: ci si può fidare dell'informazione recuperata tramite queste applicazioni innovative? Troppo web semantico può trarre in inganno, se se ne fa un uso non corretto. L'informazione sul web non è trasparente, piuttosto è soggetta a diversi tipi di cosiddette fratture quali i criteri per influenzare il ranking dei motori di ricerca, le *pop-up windows* o i *web-bugs*.[11] Un breve cenno è stato fatto dal relatore allo standard Topic Maps[12] (ISO 13250 pubblicato nel gennaio 2000) definito in effetti uno sforzo parallelo al *semantic web*, frutto di due attività diverse ma con obiettivi molto simili: anche le *topic maps* sono infatti una rappresentazione della struttura della conoscenza volta a risolvere i problemi legati alla gestione dell'informazione nel creare, mantenere ed elaborare indici di documentazione complessa. Secondo la specifica XML Topic Maps (XTM) 1.0 lo scopo di una *topic map* è far convergere la conoscenza delle risorse attraverso uno strato sovrapposto, o mappa, delle risorse: una *topic map* cattura i soggetti delle risorse di cui parla e le relazioni tra i soggetti in modo indipendente dall'implementazione.

Nicola Guarino del Laboratory for Applied Ontology (LOA[13]) ha presentato una lunga relazione dedicata alle ontologie e al loro ruolo nel *semantic web*. Concordando sul fatto che XML è solo il primo passo, dal punto di vista sintattico, e che i vocabolari standard rappresentano un ulteriore passo avanti (nonostante le difficoltà che comportano nella costruzione, nell'aggiornamento e nella comprensione da parte degli utenti) ha ribadito la necessità delle ontologie che vanno ben oltre le liste di vocabolari standard. Distanziandosi innanzitutto dal concetto filosofico, egli definisce l'ontologia nell'ambito dell'Information Technology come un artefatto, un oggetto progettato per esprimere il significato *inteso* di un vocabolario condiviso. La posizione di Guarino si rivela ancora molto legata al concetto di database e infatti delinea il ruolo delle ontologie nel *semantic web* come finalizzate all'interoperabilità semantica e all'information retrieval. L'ontologia è una concettualizzazione, una struttura formale della realtà così come è interpretata da un agente, indipendentemente dal vocabolario usato e dalla attuale occorrenza di una specifica situazione. La funzione dell'ontologia è dunque principalmente quella di catturare il significato *inteso* di un vocabolario secondo una concettualizzazione: si tratta di una teoria logica i cui modelli catturano il più possibile i modelli intesi. Qual è dunque la differenza tra le ontologie e i vecchi modelli concettuali? In effetti, secondo Guarino, non è tanta, poiché le ontologie rappresentano modelli concettuali più raffinati e con alcune attenzioni particolari (i modelli non sono accessibili *run time* e non hanno una semantica formale). Nel modello vengono specificati solo i vincoli importanti per le *queries,* mentre nelle ontologie si tiene conto dei vincoli necessari per il significato inteso. Tra ontologie e base di conoscenza più che differenze esiste un rapporto di inclusione, poiché la base di conoscenza consta di due componenti, asserzionale e terminologica, e quest'ultima di fatto è un'ontologia. Le formule ontologiche devono essere assunte sempre come valide, in tutte le situazioni. Per verificare la qualità di un'ontologia ci rifacciamo ai concetti di precisione e di copertura (rispetto a tutti i modelli presenti). Nell'insieme dei modelli che più si avvicinano alle ontologie: i glossari, i thesauri, gli schemi di Database Object Oriented e le tassonomie, queste ultime ritenute le più somiglianti come livello di precisione, comunque inferiore rispetto a quello delle ontologie. Un'ulteriore dimensione di valutazione è l'accuratezza, basata sull'esclusione di situazioni non intese.[14] La qualità dell'ontologia è dunque legata ai valori di consistenza (almeno locale), copertura, precisione e accuratezza.

Per valutare, comparare e certificare le ontologie è necessario un *framework* rigoroso. Un altro aspetto importante è la trasparenza cognitiva, poiché bisogna che le ontologie risultino comprensibili. Delle ontologie vengono fatti diversi usi:

– *Semplice accesso semantico.* Il significato dei termini è conosciuto a priori dalla comunità di utilizzo, l'espressività è limitata e la navigazione all'interno dell'ontologia veloce.

– *Negoziazione.* Siamo al di fuori di una comunità ristretta che condivide i significati, come nel caso del web: un'immensa comunità eterogenea che richiede linguaggi più sofisticati. In questo caso il ruolo delle ontologie fondazionali (o generali) è determinante perché esse pongono una struttura tassonomica robusta, aiutano a capire gli *agreements* e i falsi *agreements* migliorando la fiducia nei *web services*.

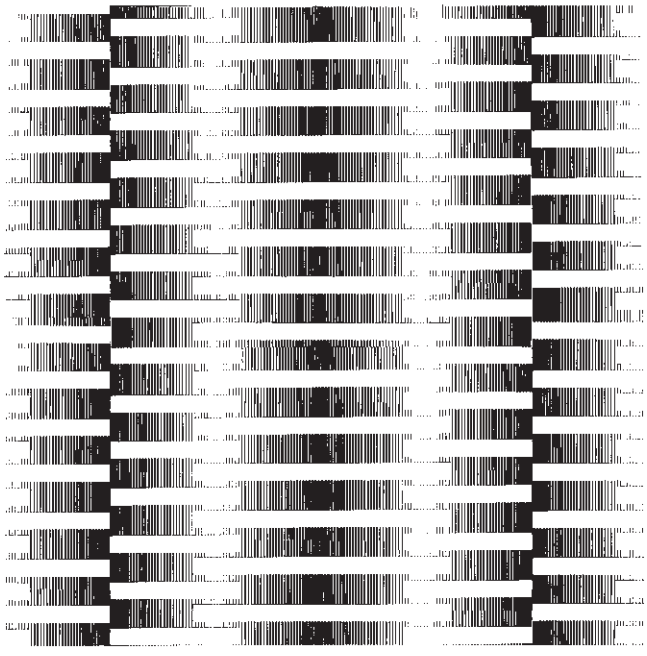
Alcuni esempi di queste realtà sono il progetto WonderWeb Library of Fundational Ontologies (WFO) che ha un approccio interdisciplinare a un'ontologia unificata di modellazione, nessun singolo livello superiore ma piuttosto un piccolo set di ontologie fondazionali e il progetto Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE) primo modulo di riferimento per la Fundational Ontology Library che vede le categorie come contenitori concettuali piuttosto che come profonde implicazioni metafisiche. È legato a Word Net,[15] un database lessicale online per la lingua inglese sviluppato presso il Cognitive Science Laboratory

alla Princeton University. A questo punto è lecito chiedersi se il *semantic web* non sia piuttosto una trovata pubblicitaria per catturare l'interesse: di fatto un vocabolario ontologico da solo non basta, sono necessarie primitive ontologi tipo quelle di OWL ma anche queste non sono sufficienti. In questo campo l'Europa è piuttosto avanti, più degli USA, a detta di Guarino, e la stessa Italia vanta l'organizzazione della prima conferenza sul ruolo delle ontologie negli Information Systems nel 1998.[16]

L'intervento di Michele Missikoff, previsto dal programma, è stato tenuto da Francesco Taglino, sul web semantico nelle applicazioni di impresa, focalizzato sull'interoperabilità semantica e le ontologie di impresa. Nella visione di un approccio ontologico per la cooperazione tra imprese finalizzata alla cooperazione tra sistemi informativi eterogenei, il progetto europeo Harmonise[17] mira a costruire un'infrastruttura basata su un'ontologia condivisa nel settore turistico. In particolare il progetto ha investigato su tre aree principali, classificando i problemi di interoperabilità causati dalle differenze degli schemi concettuali di due applicazioni che tentano di cooperare in due gruppi:

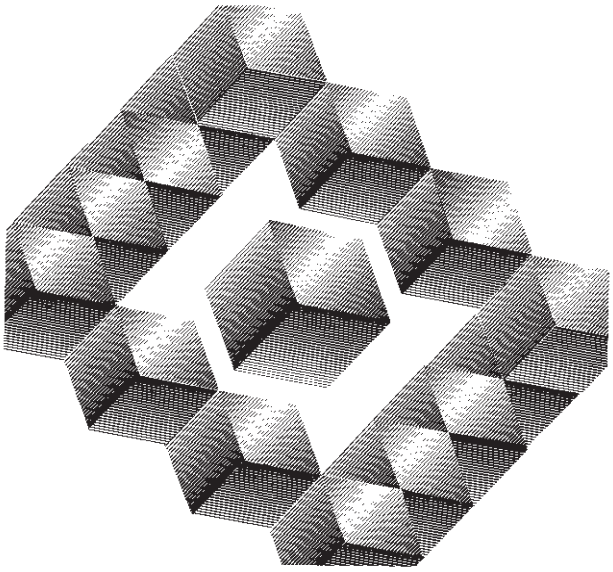
1) *lossless clashes* che possono essere risolte senza perdita di informazione (per es. relative al *naming* quando la stessa informazione è rappresentata da due etichette diverse, o differenze strutturali quando gli elementi dell'informazione sono raggruppati in modi diversi, e infine differenze di unità quando un valore scalare, tipo una quantità di denaro, è espresso con unità di misura diverse);

2) *lossy clashes* che comportano perdita di informazione, come nel caso di diversi livelli di granularità, raffinamento e precisione (per esempio nell'esprimere una distanza si può utilizzare *near* o un valore in miglia o nella precisazione relativa, nel caso di un hotel, al numero delle camere e al numero dei letti in un altro).



L'ontologia rappresenta una vista comune del dominio di applicazione, usata per dare significato alle strutture informative scambiate tra applicazioni. Il linguaggio utilizzato in Harmonise è OPAL (Object, Process, Actor Modelling Language) su cui è stato sviluppato l'*ontology management system* SymOntoX. Il concetto di ontologia di dominio viene rappresentato come una castagna divisa in senso verticale in tre aree: il livello superiore (*upper domain ontology*) contiene concetti generici (per es. evento), nella parte inferiore (*lower domain ontology*) concetti elementari (per es. prezzo, costo, indirizzo Internet). Il nucleo centrale (*application ontology*) è il più complesso perché i concetti e le definizioni dipendono dalla specifica applicazione, il problem solving messo in atto e la tecnologia sottostante, oltre agli aspetti culturali (per es. concetto di cliente, fattura, sconto o concetti più legati al settore specifico, nel caso di Harmonise il turismo). L'annotazione semantica nella rappresentazione del significato di uno schema concettuale locale viene espressa usando i concetti dell'ontologia che fornisce un unico riferimento semantico per ogni applicazione. Tali annotazioni semantiche vengono utilizzate per generare la mappatura tra schema concettuale locale e ontologia di riferimento.

Dall'intervento di Jeremy J. Carroll (HP Labs) è emerso un punto di vista piuttosto critico nei confronti del *semantic web*, sottolineandone la natura di puro investimento, allo stato attuale, in vista di ritorni futuri piuttosto incerti. Il gruppo di ricerca dei laboratori Hewlett Packard riconosce a questa tecnologia la potenzialità di rendere più flessibile l'approccio all'integrazione dei dati, ai *web services* e alla scoperta della conoscenza. Pertanto collabora con il W3C alla definizione di standard per supportare la crescita del web semantico, allo sviluppo di tool per sviluppatori che facilitino le implementazioni relative al *semantic web* e ne valorizzino il significato e le potenzialità. Jena rappresenta la piattaforma principale di HP per il web semantico: si tratta di un *toolkit* in Java per sviluppare applicazioni per il web semantico in base alle raccomandazioni del W3C per RDF e OWL (http://www.w3.org/TR/webont-req/), utile per la manipolazione di entità e modelli RDF.[18] Investire nel *semantic web* comporta dei rischi, rappresentati in primo luogo dalla *critical mass*: attualmente i costi sono più dei vantaggi, e in qualsiasi tipo di network la partecipazione al *semantic web* è un atto di fede poiché siamo ancora agli inizi mentre la crescita di questo fenomeno è esponenziale. Altro rischio o l'interoperabilità semantica: attraverso il *semantic web* è possibile realizzare cooperazione tra i modelli concettuali di applicazioni diverse, ma al momento non c'è abbastanza esperienza per poter affermare che il sistema *semantic web*-ontologie funziona. Infine consideriamo che esistono sistemi cosiddetti *neats* (formali e ben precisi) e *scruff* (sciatti): i primi, pur fornendo garanzie, hanno un'espressività limitata mentre il mondo reale non è così pulito e preciso. Il secondo tipo comporta maggiore difficoltà di composizione a vantaggio di una espressivi-



tà più estesa. Altro aspetto importante è costituito dalla globalizzazione: se c'è un'ontologia del web ci sarà un modello concettuale del mondo: prevarrà forse il modello americano? la traslazione delle ontologie porterà al prevalere di una sulle altre?

Con queste riflessioni si è passati all'ultimo intervento in programma sul tema del *semantic web* per la piccola e media impresa e sul progetto SEWASIE (Semantic Webs and AgentS in Integrated Economies)[19] presentato da Guido Vetere (IBM Software Group Rome Lab), progetto finanziato dalla Commissione europea nel 2001 allo scopo di creare un motore di ricerca avanzato che consenta l'accesso all'informazione attraverso una semantica dei dati elaborabile dalla macchina per fornire la base di una comunicazione web strutturata. L'iniziativa è nata in area modenese (Confartigianato e Università di Modena) sull'analisi delle barriere (B2B, web) identificate in problemi economici e nella carenza di semantica con relativa difficoltà a operare a livello di significato. SEWASIE rappresenta un tentativo di applicazione del *semantic web* nella piccola media impresa. Il sistema crea e mantiene ontologie multilingue con un livello di inferenza basato sugli standard W3C (XML, XML Schema, RDF(S)) che sono alla base dei meccanismo di ricerca avanzata e forniscono la terminologia per gli scambi comunicativi strutturati. I risultati delle ricerche possono essere personalizzati e visualizzati secondo le preferenze degli utenti. SEWASIE fornisce un'architettura distribuita basata su agenti intelligenti (*brokers*, *mediators* e *wrappers*) che supporta l'utente nella richiesta di fonti informative web eterogenee e realizza il recupero tramite i nodi informativi, ossia componenti indipendenti che arricchiscono semanticamente i dati collegandoli a ontologie o ad altri metadati. Un aspetto interessante del progetto è la valutazione dell'impatto economico e organizzativo di un web semanticamente ricco di informazione sui sistemi industriali delle piccole e medie imprese, in particolare i potenziali benefici economici e i fattori interni ed esterni necessari per ottenerli, tenendo conto dei cambiamenti organizzativi richiesti alle aziende dal nuovo sistema. Dal momento che diverse imprese presentano caratteristiche differenti relative al tipo di informazione richiesta, attualmente SEWASIE conduce l'analisi dei requisiti utente in due settori distinti, quello della modellatura industriale e il tessile.

Daniela Canali
ISPRI-CNR, Roma
daniela.canali@tin.it

Note

[1] Tim Berners-Lee – James Hendler – Ora Lassila, *The Semantic web,* "Scientific American", May 2001.

[2] "The World Wide Web Consortium (W3C) develops interoperable technologies (specifications, guidelines, software, and tools) to lead the web to its full potential", Fonte: <http://www.w3.org/>.

[3] I *web services* sono *middleware* (software che facilitano l'integrazione di applicazioni interconnesse) che semplificano la connettività tra applicazioni web. Sono basati su standard e specifiche XML (SOAP, UDDI, WSDL). L'architettura di un *web service* consiste di tre funzioni primarie: discovery, descrizione e trasporto. Per ognuna di esse c'è uno standard specifico, basato su XML: UDDI (Universal Description, Discovery and Integration of Web Services), per il discovery, WSDL (Web Service Description Language) per la descrizione e SOAP (Simple Object Access Protocol) per il trasporto. Le transazioni web service girano su HTTP and TCP/IP.

[4] Un'architettura *service oriented* è un modo di connettere le applicazioni attraverso una rete tramite un protocollo di comunicazione comune.

[5] Si tratta di un effetto relativo al cambiamento di significato nei successivi passaggi di traduzione, simile al vecchio gioco del telefono senza fili (un passaparola in cui accade che la parola finale non sia realmente quella pronunciata dal primo giocatore ma tutt'altro). Il concetto è discusso e applicato in <http://www.tashian.com/multibabe>l.

[6] <http://www.w3.org/2001/Annotea>.

[7] La prima implementazione di Annotea è Amaya editor/browser, un software open source per creare e aggiornare documenti direttamente sul web. La versione 8.0 (del 15 aprile 2003) è scaricabile all'URL: <http://www.w3.org/Amaya/User/BinDist.html>.

[8] Con l'aggettivo monotonico si intende soddisfare la condizione per cui se S richiede E allora (S+T) richiede E: aggiungendo informazione ad alcune supposizioni non si può invalidare una richiesta valida. Cfr.: *Glossary of technical terms* dell'Institute for the Interdisciplinary Study of Human & Machine Cognition (IHMC), <http://www.coginst.uwf.edu/~phayes/Glossary-RDF-draft.html>.

[9] <http://www.w3.org/TR/webont-req/>.

[10] <http://www.w3.org/RDF/Metalog/>:

[11] Il *web bug* è un elemento grafico su una pagina web o un messaggio e-mail che ha la funzione di monitorare ciò che sta leggendo la pagina o il messaggio. È spesso invisibile perché ha di solito la dimensione di un pixel ed è rappresentato come tag HTML IMG.

[12] <http://www.topicmaps.org/xtm/index.html>.

[13] Il laboratorio appartiene all'Institute of Cognitive Sciences and Technology (ISTC).

[14] I modelli non intesi sono esclusi ma ci sono situazioni nel mondo reale che hanno collassato l'ontologia che non è solo uno schema logico ma deve tener conto della realtà.

[15] <http://www.cogsci.princeton.edu/~wn/>.

[16] 1st International Conference on Formal Ontologies in Information Systems, FOIS '98.

[17] Michele Missikoff, Hannes Werthner, Wolfram Höpken, Mirella Dell'Erba, Oliver Fodor, Anna Formica, Francesco Taglino, *Harmonise – Towards interoperability in the tourism domain*, <http://ectrl.itc.it/home/publications/enter_paper_v7_revised1.pdf>.

[18] Nel marzo 2003 è stata rilasciata la Jena2 Preview Release 3.

[19] <http://www.sewasie.org/>.