



La traduzione dei dati nel linguaggio del web semantico

Tom Baker

I linked data sono dati esposti sul web in modo tale da poter essere collegati ad altri dati – a dati provenienti da fonti eterogenee – per formare un tessuto di dati più ricco ed espressivo della somma delle sue parti: un web semantico. L'idea di web semantico può essere affrontata anzitutto come un fenomeno linguistico. Il metodo che permette l'integrazione di dati diversi in un modo coerente si presenta come una "lingua" per i dati. Non si tratta di una lingua parlata. Si tratta tuttavia di una lingua inventata dagli umani, una lingua inventata per comunicare informazioni e pensieri fundamentalmente umani e per l'uso umano, una lingua fatta però per essere letta ed elaborata non da lettori o ascoltatori umani bensì da computer secondo algoritmi ben definiti.

La lingua del web semantico ha una grammatica. Questa grammatica, chiamata Resource Description Framework (RDF), definisce una struttura per esprimere frasi – nel linguaggio tecnico: "statements" – secondo un semplice modello tripartito. Ogni frase RDF esprime una singola asserzione secondo un modello composto da un soggetto (per esempio "Questo libro"), un predicato ("ha per titolo") e un oggetto ("L'inferno"). Secondo la grammatica di RDF queste frasi tripartite (triple) possono descrivere o un rapporto tra due entità – per esempio, tra un libro ed un autore – o un rappor-

to tra un'entità e delle annotazioni testuali che la descrivono – ad esempio le parole "L'inferno" o la data "1308". Con l'eccezione delle annotazioni testuali, una frase RDF non è composta da parole della lingua parlata bensì da Uniform Resource Identifiers (Uniform Resource Identifier (URI)) – sequenze di caratteri contraddistinte dal prefisso "http://". Gli indirizzi URI assomigliano a indirizzi web (URL), però nella lingua del web semantico vengono usati come parole per nominare persone, luoghi, epoche storiche, servizi di ogni tipo, persino concetti astratti – in linea di principio tutte le "cose" che possono essere nominate o pensate.

Nella grammatica del web semantico, gli indirizzi URI – le "parole" – appartengono a categorie grammaticali ben definite. Un predicato (o "proprietà") col significato "ha per titolo," per esempio, ha la funzione di verbo, o più precisamente di frase verbale. Un predicato può esprimere la natura della relazione tra un oggetto e un altro; per esempio, questo libro "è stato tradotto da" quella persona. I soggetti e gli oggetti delle frasi possono appartenere a "classi" (per esempio "libro"), le quali fungono da sostantivi. Oltre a servire come parole – parole appartenenti a categorie grammaticali ben precise – questi indirizzi URI servono anche come "note a piè di pagina". Nel contesto del Domain Name System (DNS) – il sistema che controlla la risoluzione degli indirizzi web a livello globale – un indirizzo URI è controllato dall'ente che ha registrato un determinato dominio. Un indirizzo URI – grazie al modo in cui è formato – porta dunque in sé, almeno in modo implicito, una citazione. Gli indirizzi di tipo URI usati per le proprietà e per le classi portano il lettore a definizioni documentate in maniera ufficiale. Così è il web stesso che fornisce alla lingua dei dati il suo dizionario. Con la sua funzione di nome globalmente unico per una determinata "cosa", un indirizzo URI serve ad ancorare i fili delle reti del significato (web semantico). Ogni filo della rete è una frase grammaticale. Ogni frase esprime

un'idea semplice – "Dante ha scritto L'Inferno" oppure "Dante è nato a Firenze". Un insieme di frasi collegate ha la funzione di paragrafo – un paragrafo che graficamente si rappresenta come un complesso di relazioni. La scelta di esprimere i dati (le informazioni) in forma di frasi ricombinabile rende più semplice l'integrazione di dati provenienti da fonti differenti. Si aprono nuove opportunità per chi si occupa di patrimonio culturale. L'assegnazione di indirizzi URI agli oggetti della descrizione bibliografica – ad opere, pubblicazioni, persone, eccetera, ma anche alle proprietà, alle intestazioni di soggetto e alle classificazioni che vengono usati per descriverli – agevola la traduzione del lavoro delle biblioteche nel linguaggio del web semantico, mettendolo a disposizione del mondo dei linked data. L'affidabilità delle biblioteche come istituzioni stabili si trasferisce agli indirizzi URI da loro creati. Le biblioteche possono perciò diventare nodi centrali nella crescente rete dei linked data.

Se la lingua del web semantico è stata inventata dagli umani per la comunicazione umana per mezzo dell'elaborazione informatica, è tuttavia importante riconoscere che la mera esistenza di una tale lingua franca – intesa come uso combinato di grammatica e vocabolari disponibili in RDF – non può da sola risolvere le difficoltà della comunicazione umana. Per analogia, una certa prevalenza dell'inglese non garantisce l'efficacia della comunicazione a livello mondiale e i ricercatori nel settore della fisica non parleranno mai la stessa lingua degli avvocati. La diversità linguistica è inevitabile ed anche auspicabile. La lingua del web semantico favorisce la creazione di nuove conoscenze "unendo i puntini", in quanto rende il contenuto dei dati inequivocabilmente e globalmente citabile, grammaticalmente uniforme e quindi universalmente elaborabile. Come l'inglese serve spesso come seconda lingua parlata, RDF fornisce una seconda lingua per i dati – una lingua che può essere usata per esporre dati locali in una forma utilizzabile da tutti. Come l'inglese è utile senza

essere una grammatica perfetta, allo stesso modo è utile anche RDF – ed è l'unico linguaggio di questo tipo ad uso generale che possediamo. Ma come l'inglese, analogamente a ogni altra lingua parlata, si basa su profonde strutture grammaticali derivanti dalla capacità umana di comunicare attraverso un linguaggio, è probabile che RDF, se reinventato, finirebbe per assomigliare in maniera molto marcata a quello che attualmente abbiamo. Oltre a permettere lo scambio di dati nel presente, RDF fornisce una risposta alla progressiva e inevitabile obsolescenza di applicazioni informatiche e di formati particolari di dati: una risposta basata sul fatto che la conoscenza è espressa attraverso una grammatica capita da tutti, usando vocabolari e indirizzi URI di risorse pubblicamente documentati. Data la disponibilità della grammatica e dei vocabolari, i dati RDF non richiedono in linea di principio informazioni aggiuntive per la loro interpretazione: sono dati "autoparlanti." Siamo nel bel mezzo di un rapido cambiamento da un mondo basato prevalentemente sulla stampa a un mondo basato prevalentemente sul digitale. L'ordine di grandezza e la velocità del cambiamento garantiscono in pratica che tutte le applicazioni e le interfacce utente che usiamo oggi saranno ad un certo punto, probabilmente presto, sostituite. I dati che non sono "autoparlanti" saranno così vulnerabili da diventare irrilevanti. Non dobbiamo accontentarci soltanto che i dati possano essere correttamente interpretati nel presente. Dobbiamo fare in modo che essi rimangano comprensibili anche nel futuro. In un tempo in cui giovani e meno giovani tendono sempre di più a credere che informazioni non trovabili tramite Google semplicemente non esistano, la lingua del web semantico fornisce alle istituzioni della memoria culturale un mezzo per esprimere il loro patrimonio in una forma ancorata nel web, globalmente citabile e, purché i loro vocabolari vengono conservati, citabile anche nel futuro.

Ai fini di una corretta indicizzazione, si invitano i lettori a citare esclusivamente il testo in lingua inglese; l'unico, infatti, che presenta l'indicazione del numero di pagina, l'abstract, le keywords e le date del processo redazionale.

Baker, T. "Designing data for the open world of the Web". *JLIS.it*. Vol. 4, n. 1 (Gennaio/January 2013): Art: #6308. DOI: [10.4403/jlis.it-6308](https://doi.org/10.4403/jlis.it-6308). Web.



TRADUZIONI