

Rdf e microformati: vocabolari condivisi per il web

di Luca Benazzi

Diamo uno sguardo alle tecnologie che consentono di connettere i contenuti della rete attraverso associazioni di significato per offrire informazioni sempre più elaborate, risposte sempre più precise e una maggiore interoperabilità.

Nel numero 45 abbiamo parlato di web semantico e di come potrebbe evolvere internet se i contenuti fossero interpretabili dalle macchine oltre che dagli esseri umani. I linguaggi utilizzati per pubblicare informazioni sulla rete, Html e Xhtml, specificano una serie di metadati, cioè informazioni che aggiungono dettaglio ai dati visibili, e che non vengono scritte con l'obiettivo di essere visualizzate dagli utenti, ma che servono a strutturare i contenuti e in una certa misura, assegnare a essi un valore semantico, cioè un significato.

Luca Benazzi vive a Londra e lavora da libero professionista come progettista di interfacce. Ha seguito progetti in vari settori e in particolare per eBay, PayPal, T-Mobile, Unilever, Dubai Holding, Royal Mail. Si occupa di varie cose tra cui scrittura, arte, fisica e cosmologia, scienze naturali, linguaggi, comunicazione, interazione uomo-computer, web semantico e cucina.
www.lucabenazzi.eu

Dai metadati interni ai metadati condivisi

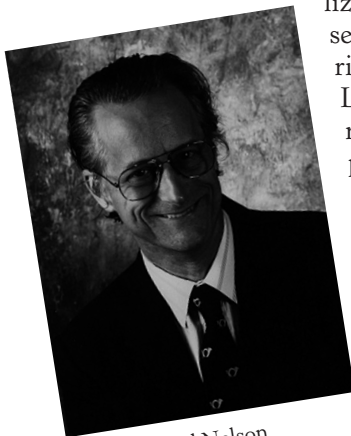
Se analizziamo il codice (X)Html con cui è scritta una pagina web, per esempio da un notiziario online, troviamo all'interno contenuti scritti per essere visibili misti a metainformazioni che sono visibili nel codice, ma non nella pagina visualizzata sul browser. In versione semplificata, una parte del codice potrebbe essere costituita come segue:

```
[...]
<h1>Durante la visita in Francia, il premier si esibisce raccontando barzellette</h1>
<img src ="Premier_primo_piano.jpg">
<p>Il premier in visita per due giorni a Parigi, ospite della serata, al termine dei rinfreschi ha scherzato con i commensali e dopo aver raccontato alcune barzellette ha improvvisato un lungo monologo [...]</p>
[...]
```

Dove si vede chiaramente che i contenuti scritti per essere visualizzati sono intervallati da identificatori contenenti le informazioni, e racchiusi tra parentesi (< >). All'interno del blocco di codice qui considerato, sappiamo ad esempio che la prima riga (introdotta mediante <h1>) definisce il titolo principale della pagina web considerata, la seconda riga (<img src= [...]) fa riferimento a un'immagine che deve essere visualizzata e che ha come nome ed estensione Premier_primo_piano.jpg, quindi segue un paragrafo di testo definito mediante la dicitura <p>, e che è stato qui riportato solo in parte per semplicità.

L'utilità di tale convenzione è che i metadati vengono utilizzati per definire l'organizzazione interna del documento in modo coerente, ma anche per poter applicare regole di formattazione in modo più elegante rispetto a un normale editor di testo quale Word.

Applicando i cosiddetti fogli di stile (Css) ai documenti creati, possiamo ad esempio specificare che tutti i titoli di pagina presenti sul sito e identificati dalla dicitura <h1> devono essere visualizzati in grassetto, e che il testo deve avere carattere, dimensione e colore determinati. Stessa cosa per il testo del paragrafo (racchiuso tra gli indicatori <p> in apertura e </p> in chiusura), che potrà avere un carattere di dimensioni minori e un colore differente rispetto al titolo.



Ted Nelson



Ora, immaginiamo che essendo stata pubblicata online, la pagina considerata sia rintracciabile dai motori di ricerca. Consideriamo tre scenari:

Scenario 1:

Un utente, volendo cercare informazioni sul premier francese, digita su Google le parole chiave “premier Francia”. La pagina web che abbiamo considerato, che proviene da un sito piuttosto popolare, viene visualizzata tra i primi risultati di ricerca. Dal contesto si capisce però che la pagina non ha niente a che vedere con il premier francese ma semmai con il premier di un altro paese non specificato nel titolo.

Scenario 2:

Insieme alla pagina sopra descritta, nei risultati di ricerca compaiono altre pagine web contenenti informazioni specifiche sul premier francese.

Chi garantisce però che le informazioni siano relative al premier in carica e non a quello uscente, se il sito a cui ci si connette è poco aggiornato? L'unico modo per saperlo è quello di passare in rassegna i contenuti uno a uno.

I METADATI VENGONO UTILIZZATI PER DEFINIRE L'ORGANIZZAZIONE INTERNA DEL DOCUMENTO IN MODO COERENTE...

Scenario 3:

Un altro utente potrebbe essere alla ricerca di quella bellissima barzelletta che parla dei premier europei riuniti al G8, letta su un sito di cui non ricorda bene il nome e ancora una volta imbattersi nella nostra pagina web che parla di tutt'altro...

Vediamo che in tutti e tre gli scenari descritti si presenta un problema di base: le macchine non sono ancora in grado di interpretare il significato della gran parte dei contenuti pubblicati su internet.

I metadati disponibili non sono sufficienti. Il codice Xhtml che abbiamo esaminato ha infatti un valore strutturale piuttosto che di significato.

Per chi ha una discreta conoscenza di Xhtml, possiamo anche aggiungere che quando anche lo si arricchisce di valore semantico usando degli attributi, si tratta pur sempre di convenzioni interne che non hanno valore universale. La situazione sta però evolvendosi e dato che internet è una tecnologia orientata alla condivisione, è facile immaginare che in un futuro non lontano le

pagine web verranno scritte in modo che gli esseri umani, con l'aiuto dei computer, possano trovare rapidamente le informazioni che cercano anziché dover passare in rassegna uno a uno i risultati e setacciare quelli potenzialmente utili da quelli inutili e/o fuorvianti.

Come i vocabolari condivisi possono migliorare internet

Alcune soluzioni hanno già cominciato a prendere piede in questi anni e tra le tecnologie più promettenti troviamo i *microformati* e *Rdf*, utilizzati da un crescente numero di sviluppatori di pagine web come standard de facto per trasmettere metadati ai motori di ricerca e agli applicativi in grado di interfacciarsi alla rete.

Sia i microformati, sia *Rdf* sono costituiti da porzioni di codice che è possibile includere all'interno delle pagine web per attribuire un significato specifico e inequivocabile ai contenuti, attraverso vocabolari condivisi definiti in modo universale.

Risultati di ricerca più accurati

Google e Yahoo hanno di recente cominciato a sperimentare modi di visualizzare le informazioni marcate mediante microformati e *Rdf*.

Immaginiamo ad esempio un sito di e-commerce dove si vendono libri. Normalmente, a ciascun libro viene assegnata una pagina contenente

la descrizione dell'opera e i vari attributi, tra cui autore, titolo, prezzo, casa editrice, numero di pagine. Possono inoltre essere presenti recensioni create dagli utenti e commenti alle recensioni.

Il codice (X)html convenzionale consente di identificare ciascuno di questi elementi mediante degli attributi (es. 'autore', 'titolo', 'anno di pubblicazione'), ma si tratta ancora una volta di dati ambigui e con scarso valore semantico che una macchina non è in grado di interpretare; la macchina, infatti, non è istruita per riconoscere il significato della parola 'autore', a meno che tale significato non sia chiaramente identificato da una linea di codice come l'occorrenza di un vocabolario condiviso. Si noti tra l'altro che, in assenza di tali convenzioni, anche un essere umano può essere indotto in errore, in quanto l'attributo generico 'autore' potrebbe indicare tante cose differenti: l'autore del libro, l'autore della pagina web, o l'autore del sito.

Tradotto in un linguaggio umano, il codice di *Rdf* e dei microformati suonerebbe come qualcosa del tipo: “Procediamo a descrivere un'opera di tipo libro. L'autore dell'opera libro qui considerata è Tizio, secondo una convenzione universalmente riconosciuta e definita in dettaglio [a un indirizzo in-



ternet specifico]. Il titolo dell'opera è Xyz, secondo una convenzione... ecc.”.

Ecco allora che 'autore', 'titolo' e qualsivoglia attributo non sono più delle semplici definizioni o metadati a uso interno e scelte in modo arbitrario dal web designer, ma degli identificatori universali a cui è associato un significato specifico.

Così come i motori di ricerca sono oggi in grado di separare le pagine web dalle immagini e dai video in base al formato riconosciuto, in futuro sarà possibile dare istruzioni per la ricerca di tutti i libri di un determinato autore o con un determinato titolo in tutti i siti internet disponibili, senza includere tra i risultati altre opere di diversa natura quali film, album, blog.

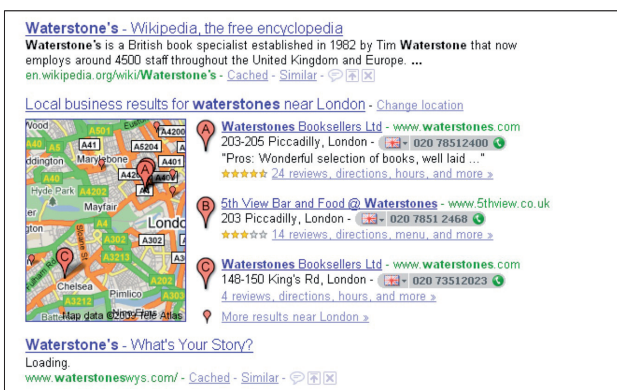


Figura 1. Un esempio di come Google si sta impegnando nel fornire risultati di ricerca sempre più accurati e ad alto valore semantico. Secondo quanto annunciato dalla stessa Google, in futuro i contenuti arricchiti di metadati mediante microformati e Rdf tenderanno ad avere una particolare rilevanza tra i risultati di ricerca.

Applicativi comunicanti con la rete

Le applicazioni delle tecnologie semantiche sono innumerevoli non solo per quel che riguarda i motori di ricerca, ma in generale per tutti i software che si interfacciano con la rete.

I contenuti di una pagina di eventi su internet possono, ad esempio, essere marcati in modo tale che un software per la gestione degli appuntamenti (come Outlook, iCal, o Google calendar) li possa riconoscere e importare su richiesta. Alcuni siti consentono già questa funzionalità, con possibilità di trasferire i dati anche verso dispositivi portatili come palmari e cellulari.

Anche i dettagli di una persona (nome, cognome, professione, recapito telefonico) possono essere marcati in modo tale da poter essere riconosciuti e importati da un software in forma di biglietti da visita elettronici.

La fine di Facebook e LinkedIn?

I siti di social networking come Facebook e LinkedIn devono il loro successo alla possibilità di aggregare contenuti al loro interno consentendo agli utenti di condividere le loro esperienze, foto, competenze.

Si tratta però di reti chiuse, che richiedono una registrazione, anche se gratuita, e l'adesione a specifiche condizioni d'uso. Tale procedura va ripetuta ogni volta che si intenda entrare a far parte di un nuovo network, affinché i propri dati vengano salvati sull'ennesimo server privato.

Gli scenari che abbiamo descritto ci portano invece a considerare la possibilità di pubblicare i propri dati personali in uno spazio online che non è gestito da un'azienda, per esempio il nostro sito personale o un aggregatore aperto come Foaf, in modo che tali dati siano resi disponibili a qualsivoglia applicazione o motore di ricerca sappia interpretarli, previa autorizzazione da parte dell'utente cui appartengono.

In futuro, dunque, saranno disponibili aggregatori di contenuti che funzioneranno in modo simile ai motori di ricerca e passeranno in rassegna tutto il materiale pubblicato online indicizzandolo e proponendolo in risposta a richieste specifiche.

È ipotizzabile dunque che i siti di social network come realtà chiuse tenderanno a ridimensionarsi e che si potranno effettuare ricerche mirate per trovare, ad esempio, persone che condividono determinati tipi di interessi e competenze e che vivono in una determinata area, e via dicendo.

Differenze tra microformati e Rdf: alcuni aspetti controversi

Microformati e Rdf nascono con obiettivi diversi. Senza avventurarsi nei numerosi cavilli tecnici, ci interessa sottolineare un'importante differenza che è anche fonte di un acceso dibattito: mentre gli standard definiti attraverso i microformati possono divenire tali solo dopo essere approvati da un organismo centrale, la definizione delle risorse in Rdf ha una natura decentralizzata e chiunque può creare dei vocabolari condivisi purché vengano pubblicati online, ed eventualmente utilizzare quelli creati da altri.

Si tratta indubbiamente di due tecnologie complementari e alcuni sforzi sono stati fatti per creare strumenti che permettono l'intercomunicazione tra i due formati.

Ci sono anche coloro che vedono con occhio critico l'evoluzione del web nella direzione fin qui descritta, prefigurando uno scenario in cui saranno i comitati tecnici dei consorzi internazionali a stabilire le convenzioni a cui il mondo intero dovrà adeguarsi. Osservazioni tutt'altro che marginali, soprattutto quando provengono da uno dei pionieri dell'ipertesto, Ted Nelson, che già in passato ha rivolto aspre critiche al modello con cui il web e le tecnologie informatiche si sono evolute in ciò che oggi noi conosciamo.

Indubbiamente, a prescindere da quali tecnologie e paradigmi prevarranno sul lungo termine, si preannunciano tempi affascinanti per il connubio tra strumenti informatici e conoscenza.