Archeologia e Calcolatori
Supplemento 2, 2009, 245-249

BLUEOCEAN: UN FRAMEWORK PER LA REALIZZAZIONE DI CMS SEMANTICI

1. La condivisione dell'informazione nel Web 2.0: possibilità e limitazioni

L'ultimo decennio è stato teatro di una rivoluzione nell'ambito della gestione dei media, avvenuta grazie alle tecnologie digitali. È possibile creare a basso costo librerie di media (documenti testuali, foto, filmati, etc.) e, grazie alla pervasività di Internet, rendere questi contenuti accessibili ad un vasto pubblico. Le tecnologie genericamente note con il termine "Web 2.0" hanno permesso la realizzazione di siti web dinamici in grado di presentare agli utenti vari tipi di contenuti, rendendo accessibili in modo facilitato due attività fondamentali:

1. La ricerca. Questa funzione rende possibile navigare all'interno di database anche di grandi dimensioni, cercando i documenti ed i media desiderati secondo vari criteri. Tipicamente questa funzionalità è supportata da un database relazionale (Wikipedia1) sul quale viene eseguita una query costruita automaticamente da alcuni parametri inseriti dall'utente.

2. La redazione collaborativa. La possibilità di ogni sistema web di operare in modalità multi-utente rende possibile l'inserimento di documenti da un numero elevato di utenti; mentre l'esempio più comune, quello dei wiki (Wikipedia2), è facilmente criticabile per quanto riguarda il controllo della qualità dei contenuti e la loro affidabilità, la possibilità di implementare meccanismi di gestione del workflow (Network World), cioè una sorta di "controllo editoriale", permette di garantire l'autorevolezza dei contenuti di un sito Web 2.0. Questa è una caratteristica molto importante qualora l'ente che gestisce il sito debba garantire la qualità dell'informazione, caratteristica irrinunciabile per qualsiasi sito gestito da un'entità con una seria reputazione scientifica.

Se si analizza la situazione attuale del Web 2.0 si può notare come questo modello abbia portato alla nascita di un gran numero di "isole" non connesse o scarsamente connesse tra loro: siti in grado di offrire grandi quantità di informazione, si pensi per esempio al Getty, ma non in grado di interoperare tra di loro. Un utente che debba eseguire una ricerca su più fonti non può fare altro che ottenere l'accesso ad ognuna, eseguire una ricerca indipendente su ogni sito (spesso con modalità differenti) ed eseguire manualmente l'aggregazione dei risultati. Una situazione paradossale, che non sfrutta minimamente le potenzialità che il web aveva promesso (ipertestualità, collegamenti senza soluzione di continuità, etc.).

Volendo risolvere questo problema rimanendo nel mondo Web 2.0, esistono due soluzioni:

1. L'imposizione di un unico sito. Dal punto di vista tecnologico, si potrebbe ipotizzare di "fondere" tutti i siti Web 2.0 esistenti, almeno per un dato dominio di conoscenza, fornendo quindi una modalità di accesso unificata. Si tratta tuttavia di un'ipotesi puramente teorica, spesso irrealizzabile dal punto di vista politico per problemi di *ownership*: infatti, ogni singolo sito web è anche un "biglietto da visita" dell'istituzione che lo gestisce, che ha quindi tutto l'interesse a mantenerne l'identità (non solo di sostanza, ma anche di forma). Inoltre, approfondendo l'aspetto tecnologico, è frequente verificare che le modalità di archiviazione dei dati, pur in presenza di un minimo denominatore comune in uno certo dominio di conoscenza, presentano variazioni tipiche di ogni entità (ad es. campi in più, vocabolari diversi, etc.).

2. La realizzazione di un aggregatore. Rinunciando alla presenza di un'unica interfaccia di riferimento, e lasciando quindi ad ogni sito la propria autonomia gestionale, è possibile realizzare un sito "terzo" in grado di estrarre le informazioni da ogni fonte e presentarle in forma aggregata, con un'interfaccia di ricerca unificata. Un esempio notevole di aggregatore in campo scientifico è Google Scholar (Google), un motore di ricerca che opera in un database di letteratura accademica. Soluzioni di questo tipo, tuttavia, presentano vari inconvenienti, di cui il principale è la necessità di "collaborazione" da parte dei siti originari, che devono esporre un'interfaccia tecnologica verso l'aggregatore, come ad esempio avviene con i flussi RSS dei blog (Wikipedia3). Questo però implica la definizione tecnica di un formalismo comune per descrivere i dati pubblicati. Viceversa, l'aggregatore sarà limitato a ricerche generiche sui testi, così come avviene per il motore di ricerca di Google, limitando però pesantemente la qualità della ricerca a causa delle inerenti limitazioni dei linguaggi naturali e a causa di eventuali barriere linguistiche. Gli stessi inconvenienti valgono per la tecnica nota come "tagging", che consiste nell'applicare una serie di semplici etichette ad ogni documento; essa è oltretutto povera dal punto di vista espressivo, non potendo rappresentare concetti complessi (il successo del tagging nel contesto di "social network" come Flickr o Facebook non tragga in inganno: per la stessa natura di quei contesti non si rendono necessari i requisiti di completezza e precisione di un'attività di ricerca scientifica). Inoltre, se un sito richiede l'autenticazione degli utenti per la consultazione, si pone un ulteriore problema di accesso, legato alla granularità: tipicamente un documento web tradizionale è completamente accessibile o non accessibile, e non è possibile controllare la modalità di fruizione solo per una parte dei suoi contenuti. La necessità di impostare esplicitamente tutte queste forme di cooperazione limita e spesso vanifica alcune delle migliori proprietà del web, che sono la cooperazione spontanea e la possibilità di aggregazione indipendente da parte di terzi.

Alcuni consorzi hanno definito validi formalismi standard di rappresentazione di dati in alcuni domini specifici: per esempio IPTC (International Press Telecommunication Council) e più recentemente Dublin Core (DC) operano nel campo dei media, definendo vocabolari per descrivere soggetti, autori, titolari di copyright, eccetera. Tuttavia, nessuno di questi standard affronta il problema dell'organizzazione della conoscenza che "circonda" il documento digitale. In altre parole, grazie ad IPTC o DC è possibile creare metadati intelleggibili relativamente al soggetto, alla data di creazione, al possessore dei diritti d'autore e così via, ma non offrono di per sè la possibilità di descrivere completamente ed in modo formalizzato la conoscenza relativa al soggetto (per esempio, la storia di un'opera d'arte riprodotta in una foto).

Da queste premesse si può concludere che il modello del Web 2.0 presenta delle limitazioni insormontabili quando si parla di ampia condivisione di modelli conoscitivi ricchi.

2. blueOcean – Tecnologie semantiche (verso il Web 3.0)

blueOcean è un framework per la realizzazione di sistemi di gestione di contenuti (CMS) basati su un database "semantico". Oltre a contenere componenti pronti per l'uso in grado di creare documenti ipertestuali e redarre un intero sito web con capacità collaborative e gestire i metadati tradizionali (EXIF, IPTC, dati geografici, etc.), funzioni tipiche di un CMS, il database semantico permette di definire a piacere un repository di conoscenze associato ai media catalogati.

Le "tecnologie semantiche" sono state introdotte come parte del cosiddetto Web Semantico, proposto da Tim Berners-Lee (l'inventore del World Wide Web) come la naturale evoluzione del web (Wikipedia4). Evitando di entrare in tecnicismi troppo spinti, l'idea che sta dietro al Web Semantico è innanzitutto la definizione di una notazione formale per la rappresentazione della conoscenza: le cosiddette "triple RDF" (Wikipedia5). Esse sono tutte composte da asserzioni nella forma "soggetto – predicato – oggetto"; i componenti di ogni tripla sono a loro volta descritti da vocabolari formali (ontologie), che associano ad ogni elemento un valore semantico preciso (Wikipedia6). I siti che vogliono collaborare ad un web semantico possono tranquillamente mantenere la propria indipendenza ed identità, anche estetica, semplicemente affiancando alla pubblicazione delle pagine originali un numero opportuno di documenti RDF.

Sarà così possibile realizzare un aggregatore in grado di comprendere il valore semantico dell'informazione manipolata – questo può essere sia un sito esterno (una specie di "motore di ricerca semantico" – PuntoInformatico), sia una possibilità offerta da ogni singolo sito (ad es. offrendo all'utente la possibilità di aggregare risultati provenienti da Internet direttamente dalle proprie pagine di ricerca). Per questo RDF è in grado di trasformare le isole

del Web 2.0 in vere e proprie "federazioni" di fonti dati, pur lasciando ad esse la propria identità e responsabilità di ownership.

L'uso del formalismo RDF comporta un ulteriore vantaggio, relativo alla flessibilità della memorizzazione delle informazioni. Il database relazionale, tipicamente utilizzato nel Web 2.0, prevede una rigida definizione della propria struttura (o schema) sotto forma di un numero definito di tabelle con un numero definito di colonne. Ogni modifica allo schema è tipicamente un'operazione amministrativa che ha delle conseguenze sul resto del sistema (e spesso implica un aggiornamento del software). Viceversa, l'unica struttura di RDF è data dalle triple, che sono libere: aggiungere nuove strutture di informazione vuol dire semplicemente aggiungere nuove triple.

Infine, non è neanche necessario che due partner di una federazione concordino a priori su un numero predefinito di ontologie standard (per quanto auspicabile: standard come il già citato Dublin Core sono basati su triple): RDF mette a disposizione la possibilità di definire l'eguaglianza semantica di concetti appartenenti a differenti ontologie. Questo vuol dire che se due siti definiscono diversamente il concetto di "reperto", è comunque possibile effettuare aggregazioni di dati definendo opportunamente l'equivalenza dei concetti di "reperto" per il primo ed il secondo sito. Un aspetto interessante è dato dalla possibilità di definire soggettivamente l'equivalenza semantica: l'utente può impostarla al momento di eseguire ogni nuova ricerca, aprendo varie possibilità di consultazione dei dati.

3. Altre caratteristiche di blueOcean

blueOcean è realizzato esclusivamente con software open source (FLOSS). L'aspetto open del software presenta diversi vantaggi, dalla assenza di licenze per l'installazione alla facilità di customizzazione ed integrazione con altri sistemi informativi.

Siccome blueOcean è basato su tecnologie standard, le federazioni che è in grado di creare possono interoperare con altri partner basati su RDF, ma non realizzati con blueOcean.

blueOcean sarà disponibile entro l'estate 2009.

Augusto Palombini
CNR – ITABC – Roma
Fabrizio Giudici
Tidalwave S.a.s.

SITI WEB

Wikipedia1. http://it.wikipedia.org/wiki/Modello_relazionale.
Wikipedia2. http://it.wikipedia.org/wiki/Wiki.

WIKIPEDIA3. http://it.wikipedia.org/wiki/Really_simple_syndication.
WIKIPEDIA4. http://it.wikipedia.org/wiki/Web_semantico.
WIKIPEDIA5. http://it.wikipedia.org/wiki/Resource_Description_Framework.
WIKIPEDIA6. http://it.wikipedia.org/wiki/Ontologia_(informatica).
NETWORKWORLD. http://www.nwi.it/nwi_arretrati/ap090001.htm.
GETTY. http://www.getty.edu/research/conducting_research/.
GOOGLE. http://scholar.google.it/.
PUNTOINFORMATICO. http://punto-informatico.it/2235663/PI/Interviste/che-cos-un-motore-ricerca-semantico.aspx.

ABSTRACT

The so-called Web 2.0 offers good methods for sharing knowledge, but it does not provide adequate tools for performing automated, complex searches on the Internet with the quality needed for scientific research. blueOcean is a software product for managing knowledge with the technologies of "semantic web" and offers an effective solution to the problem.