Knowledge creation: integrazione di HTML e Semantic Web

W3C ITALIA
Oreste Signore

CREABLE KM

Data Mining

Sommario
Il patrimonio informativo esistente sul web (database, fogli elettronici, pagine HTML) può essere integrato e reso disponibile per derivare nuova conoscenza. In questo lavoro, dopo un breve richiamo ai principi di base del Semantic web, verranno accennate le potenzialità di tecnologie come GRDDL per estrarre conoscenza da pagine web contenenti dati strutturati.

1 Introduzione

Il patrimonio informativo esistente sul web è enorme, ma non è facile utilizzarlo sfruttandone tutte le potenzialità. Le informazioni sono archiviate in maniera eterogenea, e non è facile utilizzare questa conoscenza in applicazioni specifiche. Il Semantic Web si propone come l' ambiente in cui superare queste difficoltà, consentendo il reperimento, l' integrazione e l' utilizzo delle informazioni.

In questo lavoro[1] dopo un breve richiamo dei concetti essenziali del Semantic Web, si mostra come molta conoscenza sia estraibile dalle pagine HTML esistenti.

2 Il Semantic Web: alcuni richiami

2.1 I metadati e RDF

Il Semantic Web si basa sull' ipotesi che le macchine possano accedere ad un *insieme strutturato di informazioni* e a un *insieme di regole di inferenza* da utilizzare per il ragionamento automatico. La sfida del Semantic Web è fornire un linguaggio per esprimere *dati* e *regole* per ragionare sui dati, che consenta l' *esportazione* sul web delle regole da qualunque sistema di rappresentazione della conoscenza.

Nel navigare sul web i link portano a quella che formalmente viene detta *risorsa* (resource) identificata univocamente da un URI. Le informazioni sulla risorsa vengono generalmente dette "*metadati*", definiti come *"informazioni, comprensibili*

[1] La presentazione è disponibile all' URI: http://www.w3c.it/talks/2007/km12/

dalla macchina, relative a una risorsa web o a qualche altra cosa". L'uso efficace dei metadati richiede che vengano stabilite delle convenzioni per la *semantica*, la *sintassi* e la *struttura*. Le *singole comunità* interessate alla descrizione delle loro risorse specifiche definiscono la *semantica* dei metadati pertinenti alle loro esigenze. Resource Description Framework (RDF) è lo strumento base per la codifica, lo scambio e il riutilizzo di metadati strutturati, e consente l'interoperabilità tra applicazioni che si scambiano sul web informazioni *machine-understandable*. Grazie ai metadati, e ad un formalismo per esprimerli, il Semantic Web ha già una struttura semantica *esplicita* distribuita e a grande scala.

2.2 Le ontologie

Il termine ontologia deriva dalla filosofia, dove viene inteso come una spiegazione sistematica dell' essere. Negli anni recenti il termine si è ampiamente diffuso nella comunità del Knowledge Engineering. Un' ontologia include non solo i termini che sono esplicitamente definiti in essa, ma anche la conoscenza che ne può essere derivata mediante un processo di inferenza. Secondo Studer et al.:

> *An ontology is a formal, explicit specification of a shared conceptualisation. A '**conceptualisation**' refers to an abstract model of some phenomenon in the world by having identified the relevant concepts of that phenomenon. '**Explicit**' means that the type of concepts used, and the constraints on their use are explicitly defined. For example, in medical domains, the concepts are diseases and symptoms, the relations between them are causal and a constraint is that a disease cannot cause itself. '**Formal**' refers to the fact that the ontology should be machine readable, which excludes natural language. '**Shared**' reflects the notion that an ontology captures consensual knowledge, that is, it is not private to some individual, but accepted by a group.*

Esistono varie altre definizioni di ontologia, ma questo non deve far ritenere che possa sorgere confusione sul significato che la comunità scientifica attribuisce a questo termine. Le varie definizioni enfatizzano di volta involta alcuni aspetti, ma in realtà forniscono una serie di punti di vista complementari. Va posto invece l'accento su come le ontologie mirino a catturare la conoscenza *consensuale*, e possano essere condivise e riutilizzate tra applicazioni e gruppi di persone diversi.

3 Come creare l' infrastruttura di metadati?

3.1 I metadati sono già nelle pagine Web

Le pagine Web scritte in XHTML contengono dati intrinsecamente strutturati: eventi di calendario, informazioni anagrafiche, didascalie di foto, titoli di canzoni, informazioni sul copyright, etc. Se gli autori delle pagine sono in grado di esprimere questi dati in maniera precisa, e i tool possono leggerli con affidabilità, si possono immaginare scenari di utilizzo del tutto nuovi, in cui i dati strutturati possono essere trasferiti dal web alle applicazioni, in modo che, per esempio, i dati di calendario, o i diritti della licenza, possono essere comunicati direttamente all' utente, per essere registrati nella sua agenda personale o per informarlo dei suoi diritti.

Uno degli aspetti dei recenti sviluppi, che alcuni chiamano "Web 2.0", riguarda applicazioni che si basano sulla combinazione - in "mashup" - di varie tipologie di dati che sono diffusi in tutto il Web. Varie comunità molto attive nell'innovazione sul Web

hanno come obiettivo la condivisione di dati, come informazioni temporali, sociali e di georeferenziazione, e hanno sviluppato numerose pratiche sociali che soddisfano le loro specifiche esigenze. Ad esempio i motori di ricerca hanno riscosso un grande successo utilizzando metodi statistici, mentre le persone che condividono i loro album fotografici sul Web hanno trovato molto utile "taggare" le loro fotografie manualmente, con brevi etichette di testo. Molte informazioni possono essere catturate attraverso i "*microformati*", basati su standard esistenti e largamente adottati, inclusi HTML, CSS e XML. Questa ondata di attività è strettamente connessa con l'essenza stessa del Semantic Web. Le community basate sul Semantic Web hanno cercato i modi per migliorare la qualità e la disponibilità dei dati sul Web, perché sia possibile una più massiccia integrazione dei dati e perché applicazioni differenti possano scalare alla dimensione del Web e permettere mashup più potenti.

I metadati possono essere inclusi nelle pagine HTML in vari modi (microformati, embedded RDF, RDFa). Per estrarli e portarli nelle applicazioni, servirebbero di volta in volta procedure ad hoc, mentre è possibile un approccio unitario. Nel seguito di questo lavoro descriveremo in dettaglio solo l' inclusione di metadati mediante microformati, trascurando, per brevità, quella basata su "*embedded RDF*" (in cui un sottoinsieme di RDF viene inserito nell' (X)HTML) e quella basata su *RDFa* (in cui i dati strutturati sono espressi usando gli attributi XHTML e alcuni nuovi attributi).

3.2 Cosa sono i microformati?

Un microformato è un approccio web-based alla formattazione dei dati che mira a riutilizzare come metadati il contenuto esistente, usando unicamente le classi e attributi di (X)HTML. In tal modo alcune informazioni possono essere elaborate automaticamente dal software. In effetti il contenuto delle pagine è sempre stato (tecnicamente) elaborato automaticamente, ma con ovvie limitazioni, perché i tag tradizionalmente usati per il markup sono privi di semantica. I microformati vogliono proprio ovviare a questo problema, arricchendo di semantica i tag per permettere l' indicizzazione, l' estrazione e il riutilizzo dell' informazione, senza ricorrere a metodi più complessi, come l' elaborazione del linguaggio naturale. Le prossime versioni dei browser probabilmente includeranno il supporto nativo dei microformati.

Gli standard XHTML e HTML consentono di codificare la presenza di metadati e la loro semantica utilizzando gli attributi `class`, `rel` e `rev`. Ad esempio una pagina web potrebbe contenere i dati di riferimento di una persona nel seguente modo:

```
<div>
   <div>Oreste Signore</div>
   <div>Ufficio Italiano W3C - CNR</div>
   <div>+39 (050) 315 2995</div>
   <a href="http://www.w3c.it/">http://www.w3c.it/</a>
 </div>
```

che con il formato di markup hCard diventa:

```
<div class="vcard">
   <div class="fn">Oreste Signore</div>
   <div class="org">Ufficio Italiano W3C - CNR</div>
   <div class="tel">+39 (050) 315 2995</div>
   <a class="url" href="http://www.w3c.it/">http://www.w3c.it/</a>
 </div>
```

In questo esempio sono stati identificati mediante specifici nomi di classe il nome (*fn* - formatted name), l' organizzazione (*org*), il numero di telefono (*tel*) e l' indirizzo web (*url*). Non è possibile ambiguità, perché il tutto è stato inviluppato in una classe specificata con l' attributo `class="vcard"`, che indica che il nome delle altre classi non è casuale, ma che esse sono alcune di quelle definite per una `hCard` (abbreviazione per "HTML vCard", dove vCard indica, come è noto, un formato file standard per lo scambio di dati personali, in particolare le *electronic business cards*). Grazie a questo accorgimento appositi software, come i plug-in dei browser, possono estrarre le informazioni rilevanti e riversarle in altre applicazioni, per esempio un' agenda elettronica. Sono stati definiti svariati microformati, che consentono il markup semantico di particolari tipi di informazioni. Tra questi ricordiamo `hCalendar` (per gli eventi) e `hCard` (per le informazioni di riferimento alle persone) che include i microformati `adr` e `geo`, per gli indirizzi postali e le coordinate geografiche (latitudine e longitudine).

4 GRDDL

Con "**G**leaning **R**esource **D**escriptions from **D**ialects of **L**anguages", o GRDDL (pronunciato come la parola inglese "griddle"), un software può automaticamente estrarre l'informazione da pagine Web strutturate, per renderla parte del Semantic Web. Chi utilizza dati strutturati con microformati in XHTML può aumentare il loro valore dati trasferendoli nel Semantic Web, con costi veramente minimi[2].

GRDDL è il ponte per trasformare i dati espressi in un formato XML (come XHTML) in dati Semantic Web. Con GRDDL gli autori trasformano i dati che desiderano condividere in un formato che può essere usato e trasformato a sua volta in applicazioni più formali. Una volta che i dati sono parte del Semantic Web, possono essere fusi con altri dati (per esempio, una base di dati relazionale) per query, inferenze, e conversione ad altri formati.

GRDDL introduce un markup per dichiarare che un documento XML include dati che possono essere estratti, e per indicare quale algoritmo, normalmente scritto in XSLT, va utilizzato per estrarre i dati RDF dal documento.

5 Casi d' uso[3]

5.1 Fissare un incontro

Supponiamo che un utente utilizzi un *calendaring service* che pubblica la sua agenda sotto forma di feed RSS 1.0[4], e voglia organizzare un incontro con tre colleghi che vivono in altre città, ma seguono spesso le stesse conferenze. Supponiamo anche

---

2 "*A volte una sola riga di codice può fare una grande differenza. Come un foglio di stile rende le pagine Web più leggibili, GRDDL rende le pagine Web, i tag dei microformati, i documenti XML ed altri dati, più comprensibili per le applicazioni Semantic Web, rendendo disponibile un numero maggiore di dati per nuove possibilità ed un riuso creativo.*" (Tim Berners-Lee, Direttore W3C)

3 Gli esempi seguenti, in particolare le figure, sono estratti in gran parte dalla documentazione ufficiale W3C: *GRDDL Use Cases: Scenarios of extracting RDF data from XML documents*, (http://www.w3.org/TR/grddl-scenarios/)

4 *RSS* (acronimo di *RDF Site Summary* ed anche di *Really Simple Syndication*) è uno dei più popolari formati per la distribuzione di contenuti Web; è basato su XML, da cui ha ereditato la semplicità, l'estensibilità e la flessibilità.

che ognuno di questi pubblichi il suo calendario personale, utilizzando meccanismi diversi (microformato hCalendar, eRDF e RDFa).

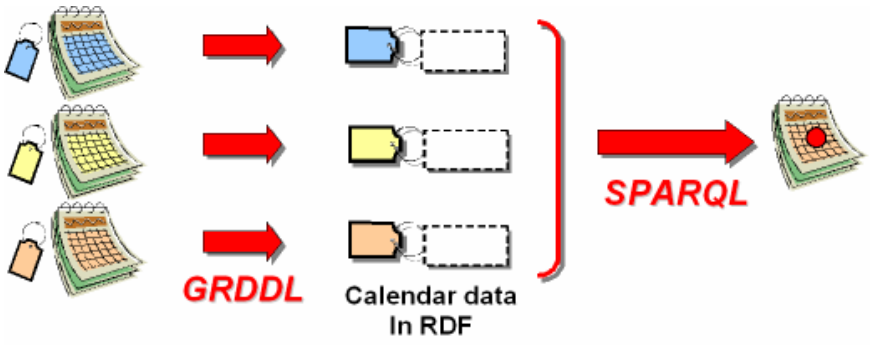

Figura 1. Trasformazione delle agende e interrogazione in SPARQL

Nonostante utilizzino formati differenti, i calendari di tutti e quattro possono essere usati come documenti di input per GRDDL e convertiti in RDF, utilizzando un *GRDDL-aware agent*, cioè un software agent in grado di identificare la trasformazioni GRDDL specificate nel documento sorgente per estrarre automaticamente i dati dalle varie pagine, per poi caricarli in un RDF store e combinarli in un unico modello. Il modello risultante può essere interrogato con il linguaggio SPARQL (*Simple Protocol And RDF Query Language*), un query language per RDF con una sintassi simile a quella ben nota per interrogare i database relazionali, per trovare le date in cui tutti e quattro si trovano nella stessa città (Fig. 1).

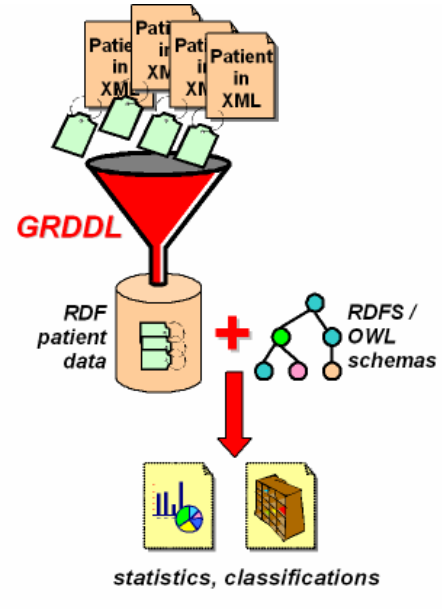

Figura 2. Un esempio di accesso ed elaborazione di dati biomedici

5.2 Accesso a dati clinici

In uno scenario più complesso (Fig. 2), possiamo immaginare il caso di un ricercatore nel settore biomedico, che, in un ambiente *decentralizzato*, ha necessità di accedere ai dati clinici dei pazienti, memorizzati in formato XML, e trova proficuo

utilizzare delle query RDF per le sue attività di ricerca. La conversione dei dati XML in grafi RDF permette di avere accesso alla conoscenza contenuta in ontologie come HL7, ma pone problemi di elaborazione, spazio e sincronizzazione. Un approccio basato su GRDDL può essere di grande aiuto nell' affrontare questi problemi: per ogni tipo di documento sorgente viene definita una trasformazione che estrae i dati clinici e li rappresenta in RDF utilizzando un vocabolario universale, consentendo quindi l' integrazione dei dati e supportando l' interoperabilità.

5.3 GRDDL con i microformati: la tecnica

Nel file XHTML che descrive l' agenda vanno inserite le opportune istruzioni per specificare che il file contiene metadati GRDDL e per indicare quali trasformazioni utilizzare, come nell' esempio seguente:

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.1//EN"
"http://www.w3.org/TR/xhtml11/DTD/xhtml11.dtd">
<html xmlns="http://www.w3.org/1999/xhtml" xml:lang="it" lang="it">
 <head profile="http://www.w3.org/2003/g/data-view">
    <title>Agenda Oreste Signore</title>
    <link rel="transformation" href="http://www.w3.org/2002/12/cal/glean-hcal"/>
  </head>
  <body>
    <ol class="schedule">
      <li>2006
         […]
      </li>
      <li>2007
        <ol>
          […]
          <li class="vevent">
            <strong class="summary">Conferenza KM12</strong> in
            <span class="location">Milano, Italia</span>: da
            <abbr class="dtstart" title="2007-11-27">27 nov</abbr> a
            <abbr class="dtend" title="2007-11-28">28 nov</abbr>
          </li>
        </ol>
      </li>
    </ol>
  </body>
</html>
```

L' URI specificato come valore dell' attributo `profile` significa che il software che riceve il documento può cercare degli elementi `<link>` contenenti un attributo `rel` con valore `transformation`, e può usare uno o tutti quei link per determinare come estrarre i dati in formato RDF (Fig. 3).

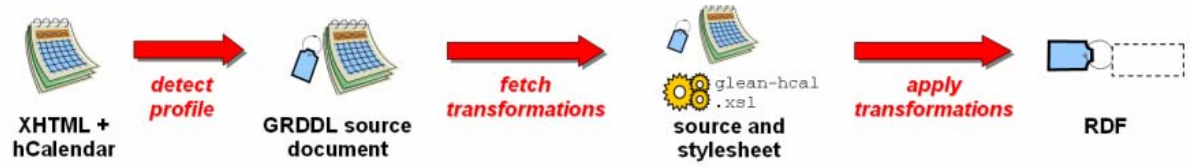

Figura 3. Il meccanismo della trasformazione in RDF

Nel caso in esempio, il risultato della trasformazione è questo file RDF

```
<?xml version="1.0" encoding="UTF-8" ?>
<rdf:RDF xmlns:_6="http://www.w3.org/2002/12/cal/icaltzd#"
         xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
<rdf:Description rdf:nodeID="jTCXOrie6">
  <_6:location xml:lang="it">Milano, Italia</_6:location>
  <rdf:type rdf:resource="http://www.w3.org/2002/12/cal/icaltzd#Vevent" />
  <_6:url rdf:resource="http://www.w3c.it/calendar/oreste-hcal-grddl.html" />
  <_6:summary xml:lang="it">Conferenza KM12</_6:summary>
  <_6:dtstart rdf:datatype="http://www.w3.org/2001/XMLSchema#date">2007-11-
27</_6:dtstart>
  <_6:dtend rdf:datatype="http://www.w3.org/2001/XMLSchema#date">2007-11-
28</_6:dtend>
  </rdf:Description>
  </rdf:RDF>
```

rappresentato sotto forma di grafo come:

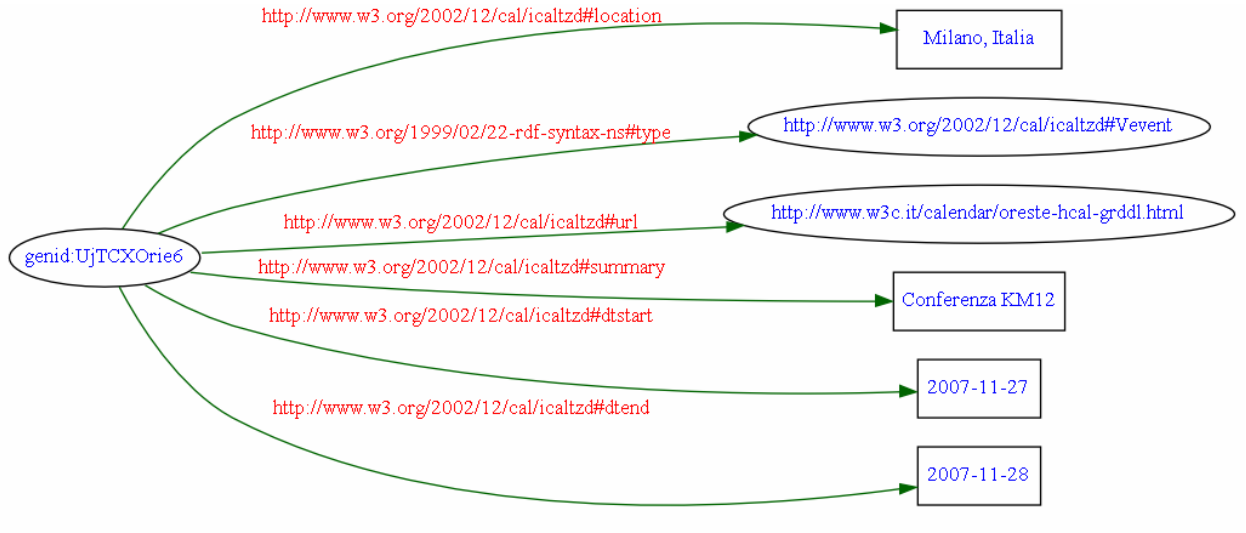

Figura 4. Il grafo RDF risultante

6 Conclusioni

Un problema chiave per lo sviluppo del Semantic Web è la creazione di una infrastruttura di metadati. Molte informazioni strutturate esistono già nelle pagine web, e poterle esportare verso altre applicazioni, per combinarle e utilizzarle, costituisce una formidabile molla di sviluppo per la crescita del Semantic Web.

Tecnologie come GRDDL permettono di estrarre le informazioni strutturate presenti nelle pagine web e codificate utilizzando microformati, embedded RDF o RDFa, e combinarle per alimentare la conoscenza.

Riferimenti bibliografici

Asunción Gómez-Pérez, Mariano Fernández-López, Oscar Corcho, *Ontological Engineering*, Springer-Verlag (2004), ISBN 1-85233-551-3

Connolly, Dan (Editor), *Gleaning Resource Descriptions from Dialects of Languages (GRDDL)* - W3C Recommendation 11 September 2007 - Latest Version: http://www.w3.org/TR/grddl/

GRDDL Use Cases: Scenarios of extracting RDF data from XML documents , F. Gandon, Editor, W3C Working Group Note, 6 April 2007, http://www.w3.org/TR/2007/NOTE-grddl-scenarios-20070406/ . Latest version available at http://www.w3.org/TR/grddl-scenarios/ .

Khare, R., "Microformats: the next (small) thing on the semantic Web?," *Internet Computing, IEEE* , vol.10, no.1, pp. 68-75, Jan.-Feb. 2006

Microformati, http://microformats.org/ , http://en.wikipedia.org/wiki/Microformats

RDFa Primer - Embedding Structured Data in Web Pages - W3C Working Draft 26 October 2007 - http://www.w3.org/TR/2007/WD-xhtml-rdfa-primer-20071026/

Resource Description Framework (RDF), http://www.w3.org/RDF/

Rudi Studer and V. Richard Benjamins and Dieter Fensel **in** *Knowledge Engineering: Principles and Methods*, Data Knowl. Eng. 25(1-2): 161-197 (1998)

Signore, Oreste, *Semantic Web: il futuro è già qui?* - 10th Knowledge Management Forum - Siena, 24-25 Novembre 2005, http://www.w3c.it/papers/km10.pdf

Signore, Oreste: *Strutturare la conoscenza: XML, RDF, Semantic Web* - Clinical Knowledge 2003 (1st edition) - Udine, 20-21 September 2003 http://www.w3c.it/papers/ck2003.pdf, http://www.w3c.it/talks/ck2003/

SPARQL Query Language for RDF , E. Prud'hommeaux, A. Seaborne, Editors, W3C Working Draft (work in progress), 26 March 2007, http://www.w3.org/TR/2007/WD-rdf-sparql-query-20070326/ . Latest version available at http://www.w3.org/TR/rdf-sparql-query/

VCard Ontology, H. Halpin, B. Suda, and N. Walsh, W3C Semantic Web Interest Group Note (in progress). Latest version available at http://www.w3.org/2006/vcard/ns.

XSL Transformations (XSLT) Version 1.0 , J. Clark, Editor, W3C Recommendation, 16 November 1999, http://www.w3.org/TR/1999/REC-xslt-19991116 . Latest version available at http://www.w3.org/TR/xslt .

W3C completa l'unione tra Microformati HTML e Semantic Web, comunicato stampa: http://www.w3c.it/pr/2007/grddl-pressrelease-it.html