Con l'HTML 5 la semantica si fa largo nel Web

Saverio Rubini[1]

1. Vantaggi per utenti e per sviluppatori

Il rilascio ufficiale della versione 5 dell'HMTL[2], il linguaggio con il quale vengono create le pagine dei siti Web, è stabilito per il 2014[3] con "*last call*" avvenuta a maggio 2011. Nella "*last call*" (in italiano: "ultima chiamata") tutti vengono invitati a provare la versione delle specifiche e a proporre eventuali modifiche e miglioramenti, sapendo che non verranno introdotte nuove funzionalità, ma solo ottimizzazioni.

Figura 1. Il logo ufficiale del linguaggio HTML 5

Perché si sia deciso di avviare lo sviluppo e il rilascio della versione 5 del linguaggio di marcatura è presto detto: per *migliorare la vita degli utenti e quella di chi realizza materialmente le pagine Web*. Mentre si può ritenere che il primo obiettivo fosse piuttosto scontato, non è detto che dovesse esserlo ugualmente anche il secondo, quello avente come gruppo di beneficiari gli sviluppatori.

A favore degli utenti che navigano in Rete si è lavorato per favorire una migliore *accessibilità* e *usabilità* in senso lato dei contenuti delle pagine Web, aggiungendo diverse nuove caratteristiche funzionali rispetto alla versione precedente. Di esse, una delle più importanti è la migliore gestione del "Web semantico".

2. Il Web semantico

I contenuti delle pagine Web pubblicate in Rete sono composti dall'insieme di una nutrita serie di dati multimediali: scritte, immagini, suoni, filmati. Il *Web semantico* è il Web immaginato da Tim

[1] Ingegnere elettronico, autore di libri e di articoli di informatica, docente in corsi di formazione professionale, funzionario dell'Agenzia delle Entrate (http://www.srubini.it)

[2] HMTL: *HyperText Markup Language*, linguaggio di marcatura per ipertesti

[3] Comunicato stampa del W3C (http://www.w3.org) del 14 febbraio 2011: "*W3C Confirms May 2011 for HTML5 Last Call, Targets 2014 for HTML5 Standard*" (http://www.w3.org/2011/02/htmlwg-pr.html)

Berners-Lee[4], nel quale i dati portano con sé anche il proprio "significato" all'interno del contesto cui appartengono.

Figura 2. Pagina Web del W3C con la biografia di Tim Berners-Lee

Questo permetterebbe alle applicazioni informatiche di riuscire a registrare automaticamente in modo corretto i singoli elementi della pagina.

Attualmente i motori di ricerca non possono fare altro che estrarre i dati rilevanti dai contenuti, utilizzando algoritmi di tipo euristico che scorrono i singoli caratteri del testo. I programmi, però, non hanno la possibilità di "capire" se una parola è il nome e il cognome di una certa persona e se il numero che appare nella stessa pagina è il suo numero di telefono, quello di un'organizzazione alla quale appartiene o una sequenza di cifre che ha un "significato" completamente diverso.

Sarebbe importante, dunque, che i dati più rilevanti fossero accompagnati da indicazioni esplicite, inserite nel codice, che ne definiscono il "significato". Ebbene, quando in una pagina web si parla di una certa persona, nel Web semantico qualsiasi applicazione informatica deve essere in grado di *individuare automaticamente* in un testo il suo nome, il suo indirizzo, il suo numero di telefono, la città in cui vive. I motori di ricerca, quindi, riuscirebbero a *registrare un insieme di dati*

[4] Tim Berners-Lee (http://www.w3.org/People/Berners-Lee/) è l'inventore delle tecnologie software che hanno permesso la creazione dei siti Internet e della navigazione in Rete con i browser: l'HTML e il WWW. Successivamente è stato messo a capo del W3C, il consorzio delle 3 W, il cui scopo è "portare il Web al suo massimo potenziale, mediante lo sviluppo di tecnologie (specifiche, linee guida, software e strumenti)": http://www.w3c.it/w3cin7punti.html

strutturati, legati proprio a quella persona e *indicizzati in modo corretto* all'interno di archivi informatizzati. Di conseguenza, gli utenti riuscirebbero a trovare in modo più efficiente quello che gli interessa cercandolo tra i dati pubblicati in Internet.

Si supponga, poi, di riuscire a raggiungere lo stesso obiettivo con dati relativi a singoli prodotti, servizi, organizzazioni pubbliche e private, notizie di eventi. Con la stessa logica, si possono identificare correttamente i dati significativi di qualsiasi elemento descrivibile con un gruppo di proprietà esposte in una struttura standard e accessibile pubblicamente (automobili, ricette, corsi di formazione e così via di seguito).

Ampliando questa capacità a tutti gli altri tipi di oggetti multimediali, come immagini, suoni e filmati video, chiunque riuscirebbe a ottenere rapidamente e con minori ambiguità possibili qualsiasi tipo di informazione quando la cerca in Internet.

3. Tecnologia per la semantica

Per andare verso il Web semantico, in HTML 5 nel codice delle pagina Web è previsto l'inserimento di appositi marcatori che fanno riferimento a specifiche ontologie pubblicate in Internet. Il tutto fa capo ai tre termini:

- *microdati*[5]
- *microformati*[6]
- *RDFa*[7]

Le tre voci si riferiscono a tre tecnologie diverse come implementazione tecnica, ma con la stessa finalità: indicare qual è il "significato" delle parti componenti una pagina Web ai sistemi di trattamento automatico dei contenuti pubblicati in Internet.

Parlando di "sistemi di trattamento automatico", in pratica ci si riferisce ai motori di ricerca. Nessuno vieta, però, a qualsiasi altra applicazione di "approfittare" di queste indicazioni per gestire i dati disponibili nella Rete a proprio piacimento. Per esempio, diventerebbe più semplice costituire banche dati di elementi omogenei tra loro.

Per fare un esempio di "indicazione del significato" di specifiche parti di un testo, di seguito viene riportato una porzione del codice di una pagina web gestita con i *microdati*. All'interno è riportata la struttura dei dati di un ipotetico ristorante ("*Organization*") i quali comprendono un'altra struttura, quella dell'indirizzo ("*Address*") riferito al ristorante di cui sopra.

[5] Specifiche microdati in HTML 5 (9 dicembre 2011): http://dev.w3.org/html5/md/

[6] Sito sui microformati: http://microformats.org/

[7] Pagina Web del W3C sull'RDFa: http://www.w3.org/TR/rdf-primer/

```
<div itemscope itemtype="http://data-vocabulary.org/Organization">
    <span itemprop="name">Ristorante La pasta</span>
    Eccone l'indirizzo:
    <span itemprop="address" itemscope
        itemtype="http://data-vocabulary.org/Address">
      <span itemprop="street-address">Via del Corso 100</span>,
      <span itemprop="locality">Roma</span>,
      <span itemprop="region">(Lazio)</span>.
    </span>
    Per prenotare: <span itemprop="tel">06 - 1234 5678</span>.
    <a href="http://www.ristorantelapasta.it"
itemprop="url">http://www.ristorantelapasta.it</a>.
</div>
```

Entrambe le strutture sono definite nel sito http://www.data-vocabulary.org nel quale, tra gli altri, appaiono anche insiemi di proprietà che caratterizzano:

- eventi;
- persone;
- prodotti;
- recensioni;
- offerte.

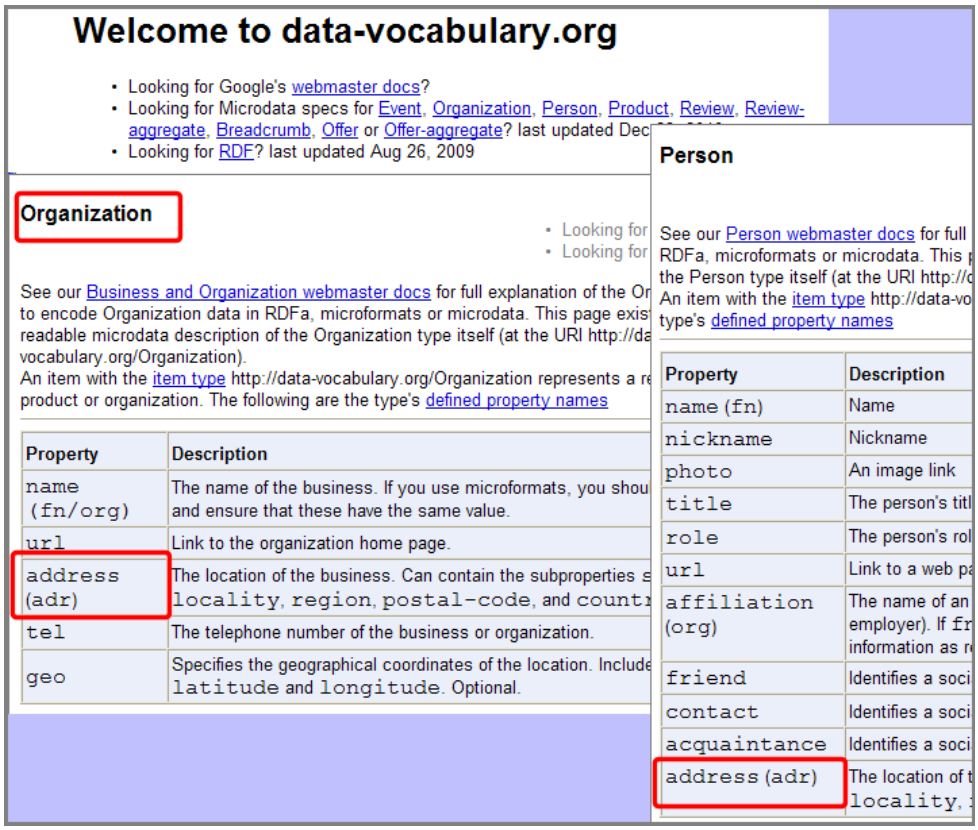

Figura 3. Stralci di pagine web del sito "*data-vocabulary.org*"

Un numero ancora maggiore di strutture dati è disponibile in http://www.schema.org, di cui vengono visualizzate alcune pagine nell'immagine successiva.

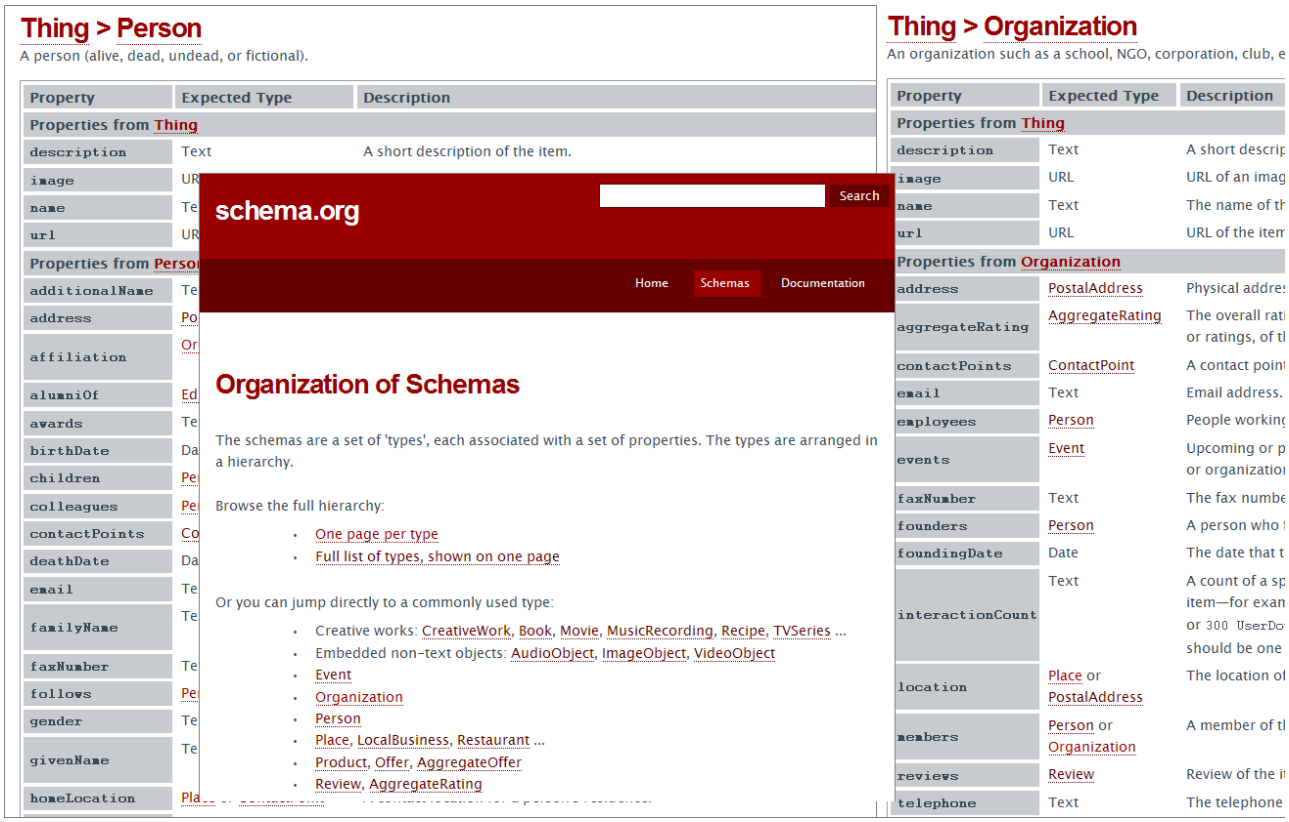

Figura 4. Strutture dati definite in "*schema.org*"

Per avere i vantaggi di cui si è parlato, è necessario che coloro che creano le pagine Web lavorino per inserire i marcatori necessari per raggiungere questo importante risultato. Nel caso dei CMS (*Content Management System*) il lavoro può essere reso un po' più automatico, ma deve comunque essere impostato correttamente dal tecnico che si occupa del codice. Ai componenti delle redazioni e agli sviluppatori viene chiesto, comunque, lo sforzo di implementare le nuove caratteristiche del linguaggio nelle proprie pagine Web, per ottenere i vantaggi promessi, che non sono pochi.