Gestiunea datelor personale bazată pe microformate

Marius-Gabriel Butuc **Sabin-Corneliu Buraga**
Facultatea de Informatică, Universitatea „A.I.Cuza” din Iaşi
Str. Berthelot, nr. 16, Iaşi 700483
{mbutuc, busaco}@infoiasi.ro

REZUMAT
În stadiul actual al Web-ului resimțim o adevărată înflorire a laturii sociale prin posibilitatea comunicării dintre tot mai multe entități (aparent) separate între ele. În scopul înlesnirii unei comunicări facile şi a înțelegerii între entități, avem nevoie de standarde care să descrie ce spunem, ce ascultăm şi ce stocăm în datele noastre. Astfel, dezvoltarea marcajelor semantice nu vizează doar îmbogățirea înțelesurilor pentru oameni, dar şi la înţelegerea de către maşină a semanticilor definite uman. Scopul acestui proiect este acela de a descrie o aplicație Web – Microwler – care să fie capabilă să ajute utilizatorii să îşi îmbogățească experienţa socială pe Web. Microwler oferă utilizatorului posibilitatea să caute şi să obțină datele personale de contact ale unei persoane ce le-a făcut disponibile, indiferent de restul conținutului paginilor. Utilizatorul are şi posibilitatea de a vizualiza relațiile pe care acea persoană le are în rețeaua sa de prieteni XHTML (XFN).

Cuvinte cheie
Microformate, regăsire de informații, gestiune date personale, interacțiune Web.

Clasificare ACM
H5.4. Hypertext/Hypermedia: Navigation, User Issues.

PREAMBUL
Odată cu creşterea constantă a popularității Web-ului, barierele între autor şi cititor se estompează din ce în ce mai mult. Nivelul responsabilității şi al încrederii creşte permițând cititorului să devină chiar autor şi să contribuie la cantitatea comună de cunoştințe. Tindem să devenim din ce în ce mai sociabili *on-line*, să ne publicăm datele de contact într-un mod care să poată fi uşor accesat de către oricine, să creăm grupuri, rețele de prieteni on-line, să împărtăşim informații despre evenimentele la care participăm etc. Astfel, au apărut diverse modele de structurare informaţională a resurselor publice şi, de asemenea, în concordanţă cu aşteptările utilizatorilor: există numeroase situri focalizate pe *blog*-uri, grupuri sociale, semne de carte colaborative, cunoştinţe colaborative, prezentări şi recenzii de produse sau companii, portaluri de ştiri – toate aliniate Web-ului social (a se consulta [2]).

Web-ul Semantic se regăseşte într-o etapă în care din ce în ce mai multe informații devin disponibile on-line şi în care avem la dispoziție şabloane de proiectare ce oferă semantică acestor informații. În acest context, grupul de inițiativă focalizat asupra microformatelor [7] propune specificarea unor metode de lucru cu aceste şabloane, adică instituirea unor standarde de publicare a informațiilor incluzând semantici repetitive – apar premisele marcării semantice a datelor, conducând Web-ul către un nou stadiu în care informaţia este accesibilă în egală măsură oamenilor şi maşinilor.

Cu o istorie ce a început odată cu articolul lui Vanevar Bush [3] din 1945, regăsirea de informații – *information retrieval* (IR) – a devenit un domeniu vast. Un proces de recuperare de informații începe în momentul în care un utilizator introduce o interogare în sistem. Un motor de căutare este un exemplu de sistem de IR, scurtând atât timpul necesar găsirii informației, cât şi cantitatea de date ce trebuie parcursă. Cele mai vizibile aplicații de tip IR sunt motoarele de căutare Web – ca Google, Yahoo Search sau Live Search – care caută informații în World Wide Web (WWW).

Microwler reprezintă un motor de căutare Web specializat pe microformate ce oferă utilizatorului posibilitatea de a recupera informații de contact din pagini Web îmbogățite semantic prin hCard-uri şi îi facilitează vizualizarea relațiilor între entități marcate cu XFN.

După cum notează Toby Segaran [5], algoritmii pentru căutarea textelor sunt printre cei mai importanți algoritmi de inteligență colaborativă şi ideile inovatoare în acest domeniu au un succes imediat. Se consideră că rapida metamorfoză a Google dintr-un proiect academic în cel mai popular motor de căutare a avut loc grație algoritmului PageRank. Microwler a fost construit pentru parcurgerea informațiilor, indexare şi căutare în datele salvate, păstrând clasificarea în funcție de rang pentru o dezvoltare viitoare.

Lucrarea de faţă este structurată după cum urmează: mai întâi, în secţiunea 2, vom defini noţiunea de microformat şi vom descrie posibilele sale utilizări în contextul vizat. Apoi, în cea de-a treia secţiune vom prezenta algoritmii ce stau la baza dezvoltării proiectului, împreună cu modelul folosit pentru date. Secţiune patru se concentrează pe metodele de prezentare ale aplicației: interfaţa cu utilizatorul, ca design, dar mai ales ca interacţiune. Lucrarea se termină cu o sinteză a problematicilor abordate şi prezintă direcţiile ulterioare de dezvoltare.

MICROFORMATE

Definiții
Microformatele, o abordare evoluționară ce adaugă semantică marcajelor bazate pe *HyperText Markup Language* (HTML) au fost adoptate cu succes de tot mai mulți editori Web şi dezvoltatori de servicii în ultimii 2-3 ani. Actuala definiție a acestora, conform sitului oficial al inițiativei dedicate microformatelor [7], este:

Proiectate în primul rând pentru oameni şi în al doilea rând pentru maşini, microformatele reprezintă un set de formate de date simple si deschise, construite pe baza standardelor existente şi larg adoptate.

Secţiunea de discuţii a aceluiaşi sit furnizează o definiţie mult mai accesibilă:

Microformatele sunt convenţii simple pentru încapsularea de semantică în cadrul marcajelor HTML, în scopul facilitării dezvoltării descentralizate.

Încă mai precis de atât, microformatele reprezintă anumite convenţii de marcare a documentelor Web, pe baza elementelor *eXtensible HyperText Markup Language* (XHTML), via nume şi valori de atribute, având o semantică precisă predefinită.

Caracteristici importante

Principiile de bază în proiectarea microformatelor sunt *simplitatea* – proiectate pentru a rezolva o problemă specifică – şi *interconectivitatea slabă* – reprezintă blocuri mici, specializate, prin a căror compunere se pot obţine blocuri din ce în ce mai largi şi se pot exprima semantici din ce în ce mai complexe, însă a căror interconectare este opţională, nediminuând puterea de expresivitate la nivel semantic.

Microformatele îşi ating scopurile fie prin adăugiri la marcajele XHTML (microformatele elementare), fie prin specificarea de valori pentru atributele asociate unor elemente ale limbajului XHTML şi imbricări de astfel de elemente, construindu-se structuri pentru conţinutul ce trebuie marcat semantic.

Specificarea (definirea) microformatelor este încă în desfăşurare: o parte dintre acestea sunt complet specificate, iar altele sunt în stadiu de lucru. Indiferent de acest aspect, microformatele sunt larg răspândite, fie explicit, fie implicit. Pentru mai multe detalii se poate consulta secţiunea de discuţii a sitului oficial [7] care prezintă, pentru fiecare microformat, şi un index al implementărilor actuale.

Microformate de interes

Lista actuală a microformatelor oficiale este compusă din următoarele: *hCalendar*, *hCard*, *rel-license*, *rel-nofollow*, *rel-tag*, *VoteLinks*, *XFN*, *XMDP*, *XOXO*, *adr*, *geo*, *hAtom*, *hResume*, *hReview*, *rel-directory*, *rel-enclosure*, *rel-home*, *rel-payment*, *Robots Exclusion* şi *xFolk*.

Microformatele de interes pentru Microwler sunt cele care încapsulează date personale şi relaţiile dintre contacte:

- *hCard*-urile conţin descrieri de entităţi: persoane, companii etc.;
- *XFN* descrie tipurile de relaţii cu persoanele din lista de contacte salvate în format *hCard*.

Exemplificări

Un exemplu modular de marcaj XHTML ce poate descrie autorul unui blog, recurge la *hCard* pentru a modela datele personale după cum urmează:

```
<div class="vcard">
 <img class="photo"
      src="img/me.jpg"
      alt="Marius B"/>
 <a class="url fn n"
 href="http://purl.oclc.org/net/mariusb">
   <span class="given-name">Marius</span>
   <span class="family-name">Butuc</span>
 </a>
 <div class="adr">
  <span class="street-address">
    no. 9, Toma Cozma street
  </span>
  <span class="locality">Ia&#x219;i</span>,
  <abbr class="region"
        title="Ia&#x219;i">IS</abbr>,
  <span class="postal-code">700554</span>
 </div>
 <div class="telecom">
  <span class="tel">+40742754488</span>
  <a class="email"
     ref="mailto:mbutuc@infoiasi.ro">
        mbutuc@infoiasi.ro
  </a>
 </div>
</div>
```

Relaţiile sociale dintre autor şi persoanele din lista de contacte se vor marca în format XFN astfel:

```
<a
href="http://fenrir.info.uaic.ro/~evalica/"
rel="friend met colleague">Ecaterina</a>
<a href="http://ppuuffaann.blogspot.com/"
rel="friend met co-worker">Andi</a>
```

Pe baza microformatelor, putem extrage în mod automat atât datele de contact, cât şi relaţiile sociale.

ALGORITMI ŞI ARHITECTURĂ

Primul pas în dezvoltarea unui motor de căutare este dezvoltarea componentei pentru adunarea informaţiei de care avem nevoie. Dat fiind caracterul ubicuu al datelor de contact în Web-ul actual – cu precădere în blogosferă – singura soluţie viabilă constă în *crawling*: pornim cu un set redus de documente şi mergem din legătură în legătură către altele. Deşi standardul *hCard* prevede numeroase proprietăţi moştenite din standardul *vCard*, acestea se supun totuşi principiului lui Pareto. Din acest motiv alegem doar un subset dintre aceste proprietăţi, cele care se întâlnesc cel mai des în viaţa de zi cu zi. Cât despre vocabularul XFN, acesta va fi considerat în întregime.

Funcţia ce realizează *crawling*-ul în documente poate fi definită recursiv, astfel încât fiecare legătură să apeleze din nou funcţia, dar aplicând algoritmul BFS putem modifica ulterior codul mult mai uşor dacă vrem să căutăm continuu sau să salvăm o listă cu pagini neindexate pentru un *crawling* ulterior.

După ce am adunat informaţiile, le stoca şi le vom indexa. Aceasta implică una sau mai multe tabele care să conţină informaţia şi sursa acesteia.Deşi ne-am limitat la un subset din proprietăţile *hCard*, putem întâlni alte excepţii. Una dintre cele mai des întâlnite probleme apare în momentul în care avem *hCard*-uri diferite care au în comun unele câmpuri. E drept că mai multe persoane pot lucra la aceeaşi organizaţie, sau mai multe organizaţii pot avea aceeaşi poziţionare geografică, dat fiind gradul actual de abstractizare. Totuşi, problemele apar în clipa în care apar duplicate în câmpurile ce se presupun a fi unice, cum ar fi email, URL sau numele formatat. În acest caz vom

considera fiecare câmp ca având o podere, egală cu a celorlalte câmpuri şi adunăm numărul câmpurilor prezente, comparând sumele între cele două *hCard*-uri. Dacă rata de asemănare este peste 80%, vom suprascrie informaţiile vechi cu cele noi. Dacă asemănarea este mai mică, vom combina datele dintre cele două contacte.

Din multitudinea de pagini Web ce conţin date personale de contact ce pot prezenta interes din punct de vedere social, nu toate se bucură de un cod XHTML valid. Pentru a preveni şi rezolva această problemă vom folosi în proiect *Beautiful Soup* [8] – un *parser* HTML/XML ce analizează şi marcaj invalid XHTML (pentru că a fost construit având în vedere marcajele SGML).

O altă problemă ce poate să apară este descoperirea unei versiuni noi al aceluiaşi *hCard* la un URL dat. Pentru a trata această excepţie, nu salvăm în baza de date doar data la care am adăugat URL-ul procesat, ci şi data ultimei modificări. Astfel avem un control mult mai bun asupra versiunilor informaţiilor personale pe care le adunăm.

Ca pas final, informaţia trebuie returnată utilizatorului şi afişată într-o manieră cât mai prietenoasă, clasificând lista de contacte în funcţie de tipul relaţiilor dintre cele două entităţi. Secţiune următoare dezvoltă metodele de prezentare ale aplicaţiei: interfaţa cu utilizatorul, ca design, dar mai ales ca interacţiune.

INTERFAŢA CU UTILIZATORUL

În continuare, vom discuta interfaţa cu utilizatorul a aplicaţiei şi mecanismele de interacţiune dintre utilizator şi aplicaţie. Încă din etapa de proiectare, Microwler a fost gândit avâns în minte utilizatorul final, fie el utilizator uman sau maşină. URL-urile au fost proiectate spre a fi mai curate, simple, mult mai uşor de citit sau chiar de intuit. Spre exemplu:

```
/hcard.php?hcard_id=4
```

reprezintă poate primul mod în care ne-am gândi să accesăm hCardul cu ID-ul 4. Acelaşi efect îl obţinem prin:

```
/hcard/4/
```

Fiind însuşi un motor de căutare, Microwler a fost dezvoltat urmărind tehnicile *Search Engine Optimization* (SEO) [4]. Optimizarea merge mai departe de URL-urile prietenoase amintite anterior. Încă de la nivel de template, în timpul implementării s-a urmărit respectarea standardelor şi ca utilizatorii ce folosesc browsere sau platforme diferite să nu trăiască experienţe diferite.

Întreaga interfaţă (a se consulta Figura 2) a fost împărţită în diviziuni logice:

- Zona superioară – antetul – găzduieşte logo-ul, o scurtă descriere şi sistemul navigaţional principal reprezentat de un meniu orizontal;
- Sub antet regăsim trei coloane:
  - în stânga avem fie câmpurile pe care se efectuează interogarea, fie rădăcina arborelui de căutare – entitatea (persoana/organizaţia) după care utilizatorul a efectuat căutarea. Avem la dispoziţie informaţiile de contact culese de pe situl original. Pentru un plus de interactivitate, utilizatorul are posibilitatea de a salva aceste date in format *vCard*, pentru a le folosi cu orice aplicaţie de gestiune a datelor personale ce suportă acest standard;
  - coloana mijlocie conţine informaţiile de contact ale tuturor entităţilor care au o legătură directă cu persoana/organizaţia menţionată anterior. Relaţiile sociale sunt reprezentate grafic folosind setul de simboluri din Figura 1, create de Wolfgang Bartelme şi Chris Messina [6];
  - în coloana din stângă avem cele mai noi zece date de contact adăugate sau modificate din baza de date;
- Dedesubt, utilizatorul are la îndemână un meniu orizontal ce include cele mai importante acţiuni.

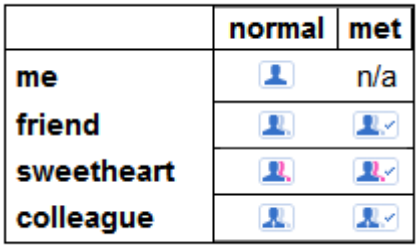

Figura 1. Reprezentarea grafică a relaţiilor sociale dintre persoane

Pentru a facilita accesul la cele mai noi zece date de contact adăugate sau modificate din baza de date cu un dispozitiv de tip *mouse*, s-a recurs la poziţionarea acestei unităţi logice în zona medie pe verticală şi extremă stângă pe plan orizontal. Am luat aceasta decizie pentru a scurta timpul de acces, pentru că se ştie că majoritatea utilizatorilor sunt dreptaci.

Tot din punct de vedere al accesibilităţii, pentru a îmbunătăţi experienţa atât a utilizatorilor cu deficienţe de vedere cât şi a acelora ce folosesc unelte alternative de vizualizare, în implementarea interfeţei aplicaţiei Web se va folosi un design relativ, *em-based* [1]. În acest mod, toate zonele logice ale aplicaţiei au dimensiuni relative la mărimea caracterelor. Dacă un utilizator va dori să mărească dimensiunea caracterelor, toate diviziunile logice ale aplicaţiei se vor redimensiona automat direct proporţional cu noile dimensiuni. De asemenea, suntem datori să luăm în calcul şi scenariul în care un utilizator va folosi un dispozitiv mobil sau alt tip de dispozitiv cu afişaj de dimensiuni reduse ori rezoluţii mici pentru a accesa aplicaţia. Acesta va putea să beneficieze în continuare de experienţa completă, datorită faptului că întreaga interfaţă va fi scalată relativ la dimensiunea implicită, pentru acel dispozitiv, a caracterelor.

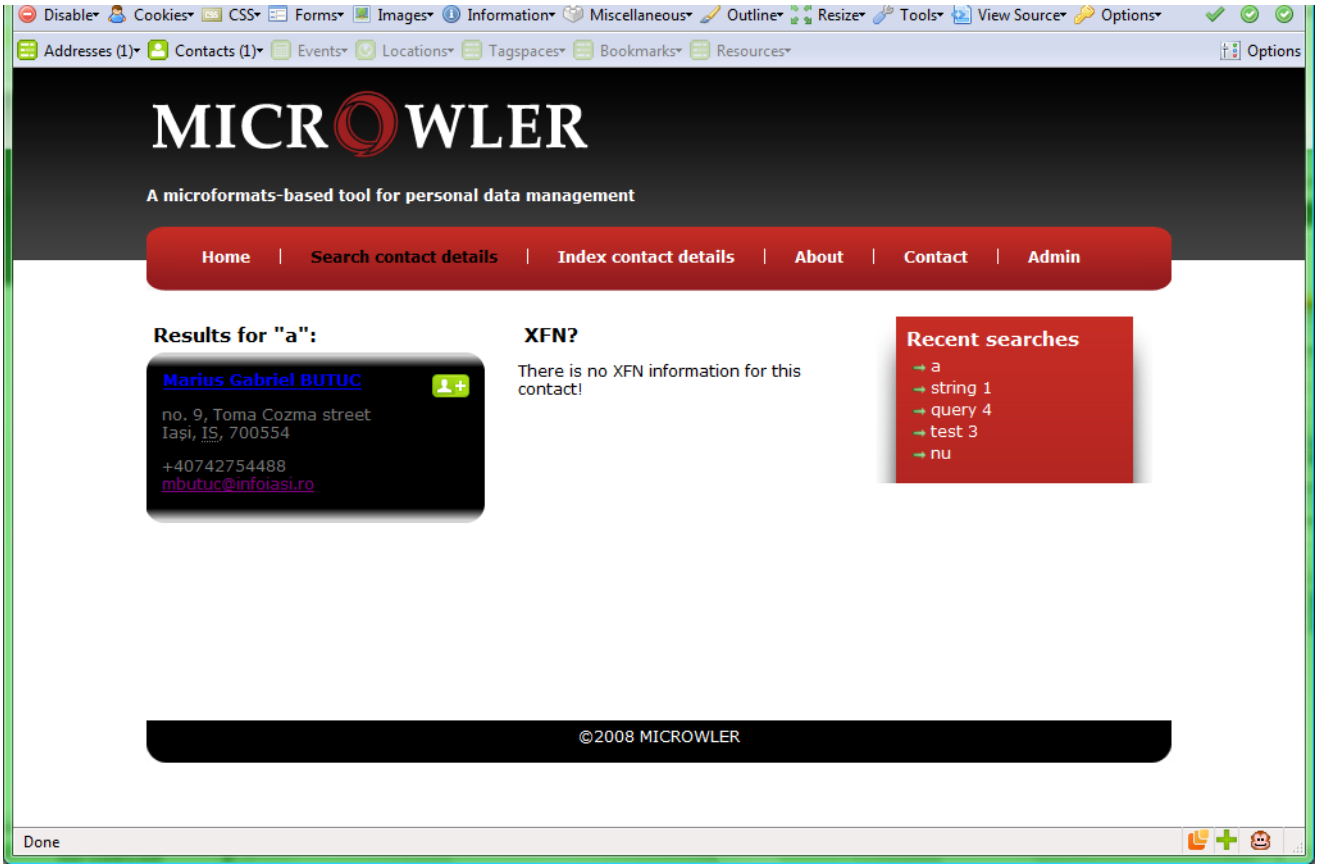

Figura 2. Interfaţa aplicaţiei în momentul căutării unui contact

CONCLUZII ŞI DIRECŢII VIITOARE DE DEZVOLTARE

Acest articol descrie un pas proaspăt în evoluţia unui Web Social, în care utilizatorul devine autor: *microformatele*.

Când ritmul de viaţă tot mai accelerat ne determină să socializăm cu persoane/organizaţii uneori chiar de pe alte continente, datele personale de contact devin un lucru din ce în ce mai de preţ.

O direcţie importantă în care se poate avansa este aceea a recomandărilor: aplicaţia să ofere posibilitatea regăsirii de entităţi deja cunoscute de utilizator din cercul de cunoştinte apropiate ale contactelor sale.

Conceput şi dezvoltat ca o aplicaţie extensibilă, Microwler este pregătit pentru permanenta evoluţie a microformatelor. Cum propunerea unor microformate noi este deschisă larg comunităţii Web, apariţia unor specificaţii care să încapsuleze noi semantici reprezintă o posibilitate de a îmbunătăţi aplicaţiile dedicate prin facilitarea accesului la date care sunt momentan „ascunse” de lipsa unui marcaj corespunzător. În clipa de faţă, pentru a suporta un microformat nou este nevoie doar sa scriem un nou modul pentru Microwler pentru a putea parcurge structura corespunzătoare.

O altă oportunitate de dezvoltare o prezintă integrarea coordonatelor geografice regăsite in *hCard*-uri cu hărţi Google, pentru a spori experienţa utilizatorului.

Astăzi, când instalăm o aplicaţie software, există riscul ca versiunea noastră să fie deja învechită. Planurile pe termen lung includ transformarea Microwler într-un serviciu Web complet funcţional.

REFERINŢE

[1] J. Allsopp. *Microformats: Empowering Your Markup for Web 2.0*. Friends of Ed, 2007.

[2] S. Buraga (coord.), *Tendinţe actuale în proiectarea şi dezvoltarea aplicaţiilor Web* (în limba română), Matrix Rom, 2006.

[3] V. Bush. *As we may think.* O’Reilly Network, 1945.

[4] J. Grappone and G. Couzin. *Search Engine Optimization: An Hour a Day*. Wiley Publishing, Inc., 2006.

[5] T. Segaran. *Programming Collective Intelligence: Building Smart Web 2.0 Applications*. O’Reilly Media, 2007.

[6] C. Messina and W. Bartelme. *Microformats Icons*. *http://factorycity.net/projects/microformats-icons/*

[7] * * *, *Situl oficial dedicat microformatelor*: *http://microformats.org/*

[8] *** . *Beautiful soup*. *http://crummy.com/software/BeautifulSoup*