# STRUCTURAREA DATELOR DE PE WEB FOLOSIND XML

**As. Ing. Drd. Gabriel Dacian CUREA**
**As. Ing. Drd. Daniel MORARIU**

**UNIVERSITATEA "LUCIAN BLAGA" SIBIU,**
**FACULTATEA DE INGINERIE "HERMANN OBERTH",**
**CATEDRA CALCULATOARE ŞI AUTOMATIZĂRI**

## Abstract

Standardele deschise şi concentrarea spre comunicare şi colaborare între oameni şi aplicații au creat mediul necesar trecerii web-ului de la o orientare pe document la o orientare pe date. Structurarea informației cu ajutorul marcajelor generale puse la dispoziție de XML au condus la dezvoltarea serviciilor web, care au devenit platforma pentru integrarea aplicațiilor, şi a web-ului semantic, care încearcă organizarea datelor web, atribuindu-le un înțeles, făcând astfel posibilă procesarea acestora.

## Introducere

Nevoia schimbului de informație şi a accesului la informație a dus la apariția Internetului ca rezultat al unui proiect de cercetare al Departamentului Apărării Statelor Unite, realizându-se astfel suportul fizic de transmitere al informației între sistemele de calcul. Organizarea şi structurarea informației folosind sistemele de calcul au dus la apariția tehnologiilor

hypermedia.Cercetătorii de la Organizația Europeană pentru Cercetări Nucleare (Couseil Européan pour la Recherche Nucléaire - CERN) au fost cei care au combinat tehnologiile hypermedia şi Internetul, având ca rezulatat World Wide Web (WWW).

Conceptul de marcaj generalizat, apărut din nevoia structurării informației, implică utilizarea etichetelor pentru identificarea porțiunilor de informație. Marcajul generalizat cere ca informația să fie cuprinsă între etichete de început şi etichete de sfârşit, acest lucru permițând imbricarea informației şi structurarea sa într-o manieră ierarhică.

Următorul exemplu utilizează marcaje pentru a indica faptul că o carte are un titlu, câțiva autori şi o editură:

```
<carte>
    <titlu>Structuri de date şi algoritmi în Java</titlu>
    <autori>
        <autor>Mitchell WAITE</autor>
        <autor>Robert LAFORE</autor>
    </autori>
    <editură>Teora</editură>
</carte>
```

Utilizarea marcajului pentru reprezentarea informației, face ca aceasta să poată fi uşor de procesat software, deoarece etichetele de început şi de sfârşit delimitează clar locul unde bucățile de informație încep şi unde se termină. Reprezentarea informației prin marcaje face ca aceasta să fie extensibilă.

SGML (Standard Generalized Markup Language – Limbajul de Marcare Generalizat Standard) este prima încercare de definire a marcajului generalizat; acesta însă a produs o specificație foarte complexă. SGML este un metalimbaj deoarece el defineşte modul în care orice limbaj de marcare dat poate fi specificat formal.

**HyperText Markup Language (HTML)**

HTML (Limbaj de marcare hypertext) este o aplicație SGML, fiind totodată limbajul de marcare care domină Web-ul. HTML defineşte structura şi modul de aşezare al informației într-o pagină web (document HTML), prin intermediul marcajelor şi al atributelor acestora. Specificația HTML este deținută de World Wide Web Consortium (W3C). [1]

Hypertext-ul este un tip special de bază de date, în care obiectele (texte, imagini, sunete etc.) pot fi legate unele de altele. Hypertext-ul permite deplasarea de la un obiect la altul, chiar dacă aceste obiecte sunt total diferite (deplasare de la un text la o imagine etc.).

Structura unui document HTML este următoarea:

```
<html>
    <head>
        <title> Document HTML</title>
    </head>
    <body>
        Pagina web.
    </body>
</html>
```

Marcajele folosite de HTML sunt grupate în câteva categorii diferite: etichete de structurare text, etichete de formatare, etichete de legare şi includere şi etichete de introducere date.

Principalele dezavantaje ale HTML-ului sunt: setul de marcaje nu poate fi extins cu marcaje ce ar putea fi folositoare programatorilor; lipsa de semantică a informației.

**Extensible Markup Language (XML)**

XML (Limbaj de marcare extensibil) este creația celor de la W3C şi a fost creat din nevoia de a simplifica SGML şi nevoia controlării evoluției HTML.

XML este similar cu SGML prin aceea că păstrează noţiunea de marcaj generalizat, totuşi specificaţia acestuia este mult mai simplă. Avantajul major este acela că utilizatorul îşi poate defini propriile marcaje.

Tehnologiile XML au două arii largi de aplicaţie: aplicaţiile centrate pe document şi aplicaţiile centrate pe date.

În aplicaţiile centrate pe document XML-ul este folosit ca un mecanism pentru reprezentarea documentelor semistructurate (manuale tehnice, cataloage de produse etc.). În continuare este prezentat un exemplu de document semistructurat folosind marcaje XML:

```
<H1>Cerinţe pentru participarea la cursul de microelectronica</H1>
<P>Pentru participarea la cursul de microelectronică trebuie:</P>
<LISTA>
<ITEM>să cunoaşteţi limba engleză</ITEM>
<ITEM>să fiţi liberi sâmbăta şi duminica</ITEM>
<ITEM>să fi absolvit cursul de Bazele electronicii</ITEM>
</LISTA>
<P>Dacă îndepliniţi condiţiile de mai sus puteţi continua cu
<LINK HREF=”Program.xml”>”Programul cursului”</LINK>
</P>
```

În cazul aplicaţiilor centrate pe date XML-ul este utilizat pentru a marca informaţie “înalt structurată” (reprezentarea textuală a datelor relaţionale din bazele de date, structurile de date ale limbajelor de programare etc.). În aceste cazuri documentele XML sunt generate de către aplicaţii şi sunt destinate prelucrării de către aplicaţii. În continuare prezentăm un exemplu (un ordin de cumpărare) de document XML centrat pe date:

```
<ord_cump id=”2398” transmis=”2003/10/07”>
  <adresa_livrare>
    <societate>S.C. PARTICULARUL S.R.L.</societate>
    <strada>Eroilor</strada>
```

```
    <nr>7</nr>
    <localitate>Mediaş</localitate>
    <județ>Sibiu</județ>
    <cod_poştal>56890</cod_poştal>
  </adresa_livrare>
  <ordin>
    <element cod="P4190" cantitate="2">
      <descriere>Ştampilă rotundă tip P4190</descriere>
    <element cod="R1133" cantitate="2">
      <descriere>Ştampilă dreptunghiulară tip R1133</descriere>
  </ordin>
</ord_cump>
```

În acest exemplu se poate vedea că marcajul este utilizat pentru descrierea unei unități de informație, în loc de modul în care trebuie prezentată aceasta unei persoane. Exemplu anterior de utilizare centrată pe date a XML ar putea fi reprezentat, într-un limbaj obiectual, ca o structură de date astfel:

```
Class ord_cump
 {
   int id;
   Date transmis;
   Adresa adresa_livrare;
   Element ordin[];
 }
```

La început Web-ul s-a dezvoltat foarte rapid ca un mediu al documentelor utile pentru oameni, în detrimentul datelor şi informației care pot fi procesate automat. XML-ul stă la baza a două tehnologii folosite pentru prelucrarea automată a informației de pe Web: Web-ul semantic (Semantic Web) şi servicii web (Web Services). [2]

Web-ul semantic încearcă să organizeze datele şi informaţiile web într-un mod natural pentru oameni şi în acelaşi timp facil pentru procesarea automată. Pentru dezvoltarea web-ului semantic sunt folosite două elemente: XML-ul şi RDF-ul (Resource Desciption Framework – cadru de descriere de resurse).

XML permite utilizatorilor să adauge o structură arbitrară documentelor , dar nu spune nimic despre semnificaţia structurii. RDF codifică semnificaţii în seturi de triplete, fiecare triplet conţinând un subiect, un verb şi un obiect al unei secvenţe elementare. RDF este un cadru pentru descrierea şi interschimbarea metadatelor (informaţii despre informaţii). [1]

RDF este construit pe următoarele reguli: o resursă poate fi orice structură care poate avea un URI (Universal Resource Identifier – Identificator de resurse universal); o proprietate este o resursă care are un nume şi o declaraţie care este o combinaţie între o resursă, o proprietate şi o valoare.

Două baze de date pot folosi identificatori diferiţi pentru ceva ce reprezintă de fapt acelaşi concept. Un program trebuie să aibă o cale să descopere astfel de înţelesuri comune, indiferent de bazele de date întâlnite. O ontologie este un document sau un fişier care în mod formal defineşte relaţiile între termeni. Cel mai tipic fel de ontologie pentru web are o taxonomie şi un set de reguli logice de inferenţă. Taxonomia defineşte clasele de obiecte şi relaţiile dintre ele.

Web-ul semantic este foarte flexibil, două programe pot ajunge să-şi pună în comun înţelegerile prin schimbarea de ontologii, care furnizează vocabularul necesar pentru discuţie.

Serviciile web sunt construite pentru realizarea procesării distribuite pe Internet. Serviciile web XML prezintă o funcţionalitate folositoare utilizatorilor de web prin intermediul unui protocol web standard (în cele mai multe cazuri SOAP – Simple Object Access Protocol – Protocol simplu de acces la obiecte). SOAP este un protocol simplu folosit pentru schimbul de informaţii într-un mediu distribuit. Are la bază XML-ul şi conţine trei părţi: un pachet care

defineşte un cadru pentru descrierea a ceea ce se află în mesaj, şi cum se procesează; un set de reguli de codare pentru exprimarea instanţelor de tipuri de date definite de aplicaţie; o convenţie pentru reprezentarea răspunsurilor şi apelurilor de procedură la distanţă.

Serviciile web XML pune la dispoziţie o cale de descriere a interfeţelor sale în detalii suficiente pentru a permite utilizatorului să creeze o aplicaţie client care să interacţioneze cu el (în mod reactiv, chiar proactiv). Această descriere este pusă la dispoziţie în mod uzual într-un document WSDL (Web Services Description Language – Limbaj de descriere a serviciilor web).

Serviciile web sunt înregistrate astfel încât utilizatorii potenţiali să le poată găsi uşor. Această înregistrare este făcută cu ajutorul UDDI (Universal Discovery Description and Integration – Descoperire, descriere şi integrare universale).

## CONCLUZII

Structurarea informaţiei de pe Web, folosind marcajele generale puse la dispoziţie de XML, deschide drumul procesării ei pentru realizarea sistemelor distibuite şi pentru extragerea de cunoştinţe.

Întreaga putere a Web-ului semantic va fi pusă în valoare atunci când oamenii vor crea programe care vor colecta conţinutul web-ului din diverse surse, vor procesa informaţia şi vor face schimburi de rezultate cu alte programe.

Aplicaţiile vor putea fi construite folosind servicii web XML multiple din surse variate, care vor lucra împreună indiferent de locul unde se află şi cum sunt implementate.

În continuare autorii îşi propun studierea căutării informaţiei pe web şi extragerea informaţiei utile din paginile web.

## MULȚUMIRI

Acest articol a fost elaborat in cadrul programului de doctorat urmat de autori la Universitatea "Lucian Blaga" din Sibiu (conducător ştiinţific prof. univ. dr. ing. VINȚAN Lucian).

Autorii îşi exprimă întreaga gratitudine companiei SIEMENS AG, CT IC, München, Germania în special domnului vicepreşedinte Dr. H.C. mat. Hartmut Raffler pentru sponsorizarea programului de doctorat precum şi pentru sugestiile profesionale extrem de utile.

## BIBLIOGRAFIE

**[1] Thuraisingham B. – "XML Databases and the Semantic Web", CRC Press, 2002**

**[2] Graham S.,Simeonov S., Boubez T., Davis D., Daniels G., Nakamura Y., Neyama R. – "Building Web Services with Java™ – Making sense of XML, SOAP, WSDL and UDDI", SAMS Publishing, 2002**

**[3] Fensel D., Hendler J., Liberman H., Wahlster W. – "Spinning the Semantic Web – Bringing the World Wide Web to its Full Potential", MIT Press, 2003**

**[4] Cagle K., Dix C., Hunter D., Kovack R., Pinnock J., Rafter J. – "Beginning XML 2nd Edition", Wrox Press Ltd., 2001**