



Procesarea documentelor XML in Linux



Sabin-Corneliu Buraga
Facultatea de Informatica
Universitatea “A.I.Cuza” din Iasi, Romania
<http://www.infoiasi.ro/~busaco/>



cuprins

- Ce este XML?
- Caracterizare, aplicatii & instrumente
- Maniere de procesare
- Modelul DOM
- Interfata SAX
- SAX vs. DOM
- Concluzii



ce este XML?

- **XML (Extensible Markup Language)**
- Meta-limbaj de adnotare (marcare)
 - Set de conventii de marcare utilizate pentru codificarea informatiilor
 - Specifica multimea de marcaje (*tag-uri*) obligatorii, identificarea si semantica marcajelor
- Marcaje descriptive
- Tipuri de documente:
 - Specificarea formala a partilor & structurii
 - **DTD (Document Type Definition)**
 - **XML Schema**
- Independenta datelor
- Independent de platforma hard/soft
- Suport pentru uz international



ce este XML?

- Standard W3C (1998, 2000, 2004)

<http://www.w3.org/TR/REC-xml>

- Exemplu:

```
<?xml version="1.0" ?>
```



```
<eveniment data="sep 2005" loc="Arad">
```

```
<tutorial>Procesare XML in Linux</tutorial>
```

```
<autor email="busaco@infoiasi.ro">
```

```
Sabin-Corneliu Buraga
```



```
</autor>
```

```
</eveniment>
```



Element
(Tag)



ce este XML?

• Familia XML

- **XML (Extensible Markup Language)**
meta-limbajul propriu-zis
- **XLL (Extensible Linking Language)**
 - **XLink** – hiper-legaturi intre documente
 - **XPointer** – localizarea relativa a resurselor
- **XSL (Extensible Stylesheet Language)**
transformare/formatare date XML



xml | exemplu

```
<?xml version="1.0" ?>
<antologie pag="...">
  <poem limba="...">
    <titlu>...</titlu>
    <strofa>
      <vers>...</vers>
      <vers>...</vers>
      ...
    </strofa>
  </poem>
  ...
  <!-- mai multe poeme...
        (acesta e un comentariu) -->
</antologie>
```



xml | aplicatii

- **Formatarea continutului**
 - in navigatorul Web: XHTML (Extensible HTML)
 - in medii mobile, fara fir:
WML (Wireless Markup Language)
- **Reprezentarea diferitelor tipuri de continut**
 - expresii matematice: MathML
 - grafica vectoriala: SVG (Scalable Vector Graphics)
 - multimedia sincronizata:
SMIL (Synchronized Multimedia Integration Language)
 - componente ale interfetei-utilizator:
XUL (Extensible User-interface Language)
 - reguli de realizare a afacerilor electronice:
BRML (Business Rules Markup Language)



xml | aplicatii

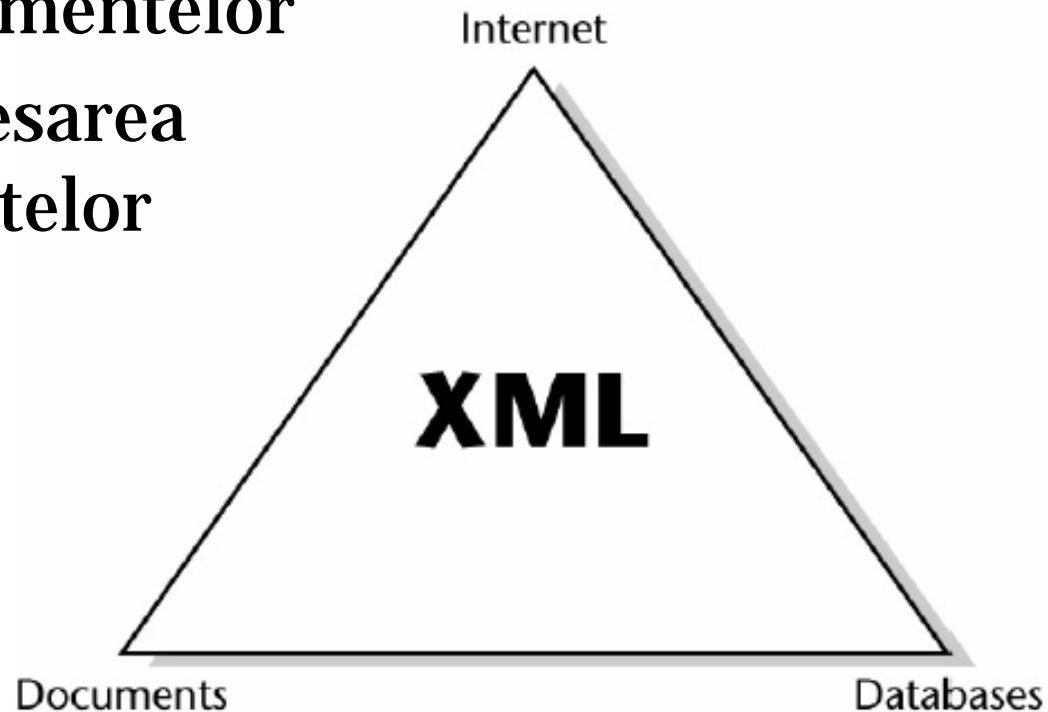
- **Descrierea resurselor Web**
RSS (Rich/RDF Site Summary)
RDF (Resource Description Framework)
OWL (Web Ontology Language)
- **Descrierea serviciilor Web**
WSDL (Web Services Description Language)
SOAP (Simple Object Access Protocol)
- **Realizarea de interogari asupra datelor XML**
XQuery
XQueryX

Detalii: <http://xml.coverpages.org/>



xml | privire de ansamblu

- XML ca principiu unificator al tehnologiilor:
 - Procesarea documentelor
 - Stocarea & procesarea traditionala a datelor
 - Internet-ul





xml | instrumente

- Tipuri:
 - Analizoare (parsere) XML
 - Vizualizatoare & editoare structurale
 - Formataatoare
 - Sisteme de gestiune a bazelor de date orientate-text (baze de date native XML)
 - Sisteme hipertext
 - ...si multe altele



procesare XML

- Tipuri de procesari XML
 - Procesare manuala (e.g., expresii regulate)
 - Procesare obiectuala (DOM & non-DOM)
 - Procesare condusa de evenimente (SAX & non-SAX)
 - Procesare particulara (via interfete specializate – e.g. XLink, RSS, SOAP,...)



procesare XML

- **Procesoare (analizoare) XML**
 - **Fara validare** – verifica doar daca documentul este bine-formatat (**Expat, libxml,...**)
 - **Cu validare** – verifica daca documentul este valid, folosind un DTD sau o schema (**Apache Xerces, Qt,...**)



dom | intro/1

- **DOM (Document Object Model)**
- Scop: procesarea obiectuală a documentelor XML/HTML
- API (interfata de programare a aplicațiilor) abstract pentru XML/HTML
- Independența de platformă & limbaj
- Definiște o structură logică arborescentă a documentelor XML
- Document \equiv set de obiecte (arbore)



dom | intro/2

- Standard al Consorțiului Web
- Niveluri de specificare:
 - **DOM 1** (1998)
<http://www.w3.org/TR/REC-DOM-Level-1/>
 - **DOM Core** pentru XML
 - **DOM HTML**
 - **DOM 2** (2001)
<http://www.w3.org/TR/REC-DOM-Level-2/>
 - **DOM 3** (partial standardizat)
<http://www.w3.org/TR/DOM-Level-3-Core/>



dom | implementari

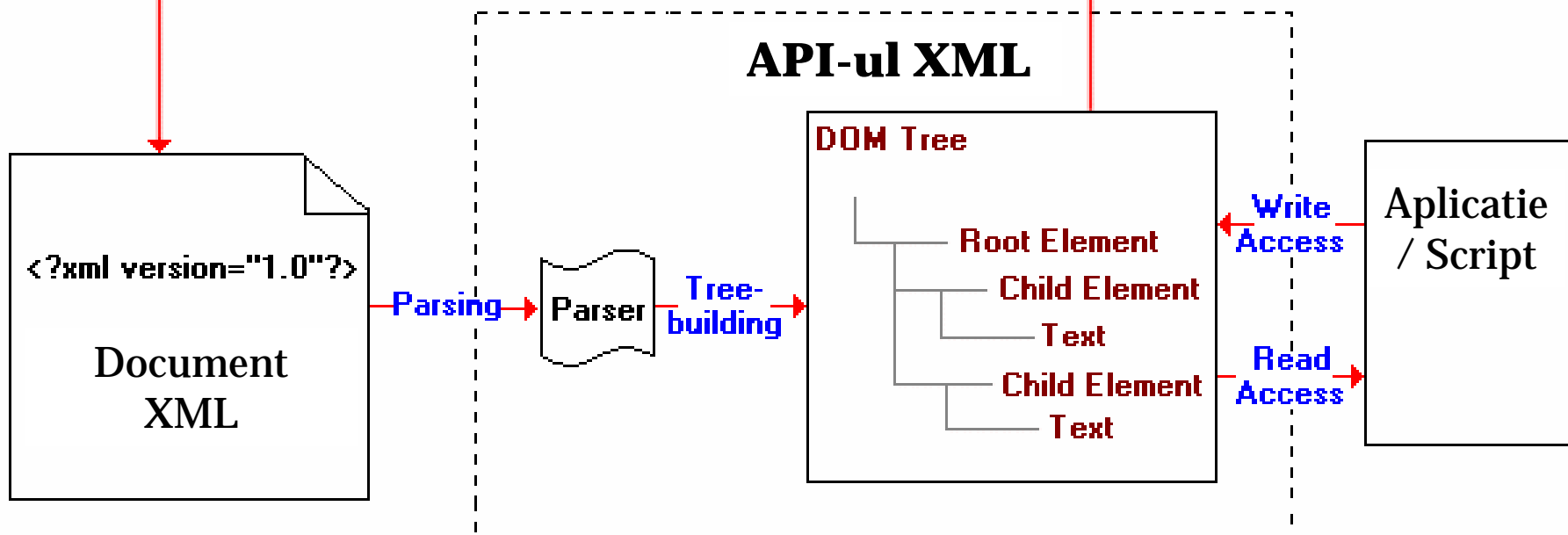
- **domxml** – extensie pentru PHP
- **JDOM** – interfata de programare special construita pentru Java: <http://www.jdom.org>
- **libxml** – API oferit de GNOME: <http://www.xmlsoft.org>
- **QDOM** – parte a Trillian Qt (C++)
- **Xerces DOM API** – platforma XML pentru C++ si Java: <http://xml.apache.org/>
- **XML::DOM** – modul Perl pentru DOM1, bazat pe Expat (XML::Parser)



dom | implementari



Metoda `save()` salveaza arborele DOM ca fisier XML





dom | browser

- Vizualizarea/procesarea documentelor XHTML si XML se realizeaza via DOM in limbajul de scripting acceptat de navigator
 - **ECMAScript** – standard: www.ecma.ch
 - **JavaScript** (Gecko + Expat in Mozilla/Firefox)
- De obicei se folosesc analizoare XML fara validare (**Expat**)
- Exemplu: Inspectarea obiectelor DOM direct in Mozilla/Firefox via **DOM Inspector**



dom | demo

Exemple demonstrative de procesari XML in C++ & PHP



sax | intro

- Scop: manipularea documentelor XML fara ca in prealabil sa fie construit arborele de noduri-obiect
⇒ documentul nu trebuie stocat complet in memorie inainte de a fi prelucrat
- Oferă o procesare XML secventială (liniară), orientată-eveniment



sax | intro

- Efort independent (de W3C) de standardizare a procesarii XML condusa de evenimente
 - **SAX 1.0**
 - **SAX 2.0** (spatii de nume + extensii)
- Initiator: **David Megginson**
- SAX larg acceptat ca standard industrial
- <http://www.megginson.com/SAX/>
- <http://www.saxproject.org>



sax | procesare

- Modelul procesarii:
 - Pentru fiecare tip de constructie XML (inceput de *tag*, sfirsit de *tag*, continut, instructiune de procesare, comentariu,...) se va “aprinde” un eveniment care va fi tratat de o functie/metoda (*handler*)
 - Functiile de tratare se specifica de catre programator, pentru fiecare tip de constructie in parte
 - Programul consuma si trateaza evenimente produse de procesorul SAX



sax | implementari

- **libxml** – API oferit de GNOME (C)
- **org.xml.sax** – API pentru Java
- **QSAX** – parte a Trillian Qt (C++)
- **Xerces SAX API** – platforma XML pentru C++ si Java: <http://xml.apache.org/>
- **XML::Parser** – modul Perl (bazat pe Expat)
- **xml_*()** – functii PHP



sax | demo

Exemple demonstrative de procesari XML in C++, Perl si PHP



sax vs. dom

- Cind trebuie folosit SAX?
 - Procesarea unor documente de mari dimensiuni
 - Necesitatea abandonarii procesarii
(procesorul SAX poate fi oprit oricind)
 - Extragerea unor informatii de mici dimensiuni
 - Crearea unei structuri noi de document XML
 - Utilizarea in contextul unor resurse computationale reduse
(memorie scazuta, largime de banda ingusta,...)



sax vs. dom

- Cind trebuie folosit DOM?
 - Accesul direct la datele dintr-un document XML
 - Cautari complexe
 - Necesitatea efectuării de transformari XSL
 - Filtrarea complexa a datelor via XPath
 - Necesitatea modificării și salvării documentelor XML
 - In contextul procesării XML direct in cadrul navigatorului



sax vs. dom

- DOM necesita incarcarea completa a documentului XML in vederea procesarii ca arbore
- SAX necesita pentru procesare existenta unor fragmente reduse din document, efectuindu-se o prelucrare liniara (sir de evenimente)
- SAX poate fi utilizat pentru generarea de arbori DOM; invers, arborii DOM pot fi traversati pentru a se emite evenimente SAX
- In cazul unor structuri XML sofisticate, modul de procesare SAX poate fi inadecvat
- Unele implementari SAX ofera suport pentru validari si transformari
- Uzual, se folosesc ambele API-uri



concluzii

- Orice distributie Linux ofera o multitudine de modalitati de procesare a documentelor XML, via biblioteci (API-uri) pentru diverse limbaje (C, C++, Perl, PHP, Python,...)
- De asemenea, pot fi folosite biblioteci sau platforme *open source*, unele chiar independente de sistem (cazul Java ori PHP)



referinte

- S. Buraga, *Tehnologii Web*, Matrix Rom, Bucuresti, 2001:
<http://www.infoiasi.ro/~busaco/books.html>
- S. Buraga (coord.), *Aplicatii Web la cheie. Studii de caz implementate in PHP*, Polirom, Iasi, 2003:
<http://www.infoiasi.ro/~phpapps>
- S. Buraga et al., *Programare Web in bash si Perl*, Polirom, Iasi, 2002: <http://www.infoiasi.ro/~cgi>
- * * *, *Apache XML*: <http://xml.apache.org/>
- * * *, *SAX*: <http://www.saxproject.org/>
- * * *, *Consortiul Web*: <http://www.w3.org/>



rezumat

- Ce este XML?
- Caracterizare, aplicatii & instrumente
- Maniere de procesare
- Modelul DOM
- Interfata SAX
- SAX vs. DOM
- Concluzii



Mulumiri pentru atentie! Intrebari...?