Introducere in XML

Mihai Gabroveanu

Cuprins

Rezumatul Prezentarii

Prezentarea cuprinde:

- Concepte de baza ale limbajului XML
- Sintaxa limbajului XML
- Validarea documentelor XML

Ce este XML?

XML - eXtensible Markup Language

- Un limbaj pentru crearea altor limbaje
- Are o structura bine definita
- Poate fi utilizat pentru a crea limbaje de marcare precum HTML, XHTML

Scopul limbajului XML

XML a fost elaborat pentru:

- separarea **sintaxei** de **semantica** pentru a furniza un cadru comun de structurare a informatiei
- **construirea de limbaje de mark-up** pentru aplicatii din orice domeniu
- **structurarea informatiei** in viitor
- **asigurarea independentei** de platforma si suport pentru **internationalizare**

XML vs. HTML

Un exemplu de HTML

```
<html>
    <body>
        <h2>Mihai Gabroveanu</h2>
            <p>A.I. Cuza<br>
            http://inf.ucv.ro/~mihaiug<br>
            mihaiug@central.ucv.ro<br>
            </p>
    </body>
</html>
```

XML vs. HTML

Acesta va fi afisat astfel

Mihai Gabroveanu
A.I. Cuza
http://inf.ucv.ro/~mihaiug
mihaiug@central.ucv.ro

HTML

- Specifica modul de randare a documentului, si nu ce tip de informatie este continuta in document
- Este dificil pentru o masina sa extraga informatia continuta in pagina. Este relativ simplu pentru om

XML vs. HTML

Sa consideram acum urmatoarea reprezentare

```
<contact>
    <name>Mihai Gabroveanu</name>
    <address>A.I. Cuza</address>
    <web-page>http://inf.ucv.ro/~mihaiug</web-page>
    <email>mihaiug@central.ucv.ro</email>
</contact>
```

In acest caz

- Informatia continuta este scoasa in evidenta si nu modul de randare
- Continutul este separat de prezentare
- Informatia poate fi "*inteleasa*" atat de oameni cat si de masini

XML vs. HTML

Concluzii:

- HTML este utilizat pentru a marca textul astfel incat el sa poata fi afisat
- XML este utilizat pentru marcarea datelor astfel incat ele sa poata fi procesate automat de calculator
- HTML descrie atat structura (ex. `<p>`, `<h2>`, `<em>`) cat si modul de reprezentare (ex. `<b>`, `<font>`, `<i>`)
- XML descrie numai continutul sau "intelesul"
- HTML utilizeaza un set fix, neschimbat, de tag-uri
- In XML putem sa definim propriile tag-uri

De ce XML?

Avantajele utilizarii limbajului XML:

- Este un limbaj bazat standard open
- Este un limbaj extensibil
- Este transformabil
- Disponibilitatea unui numar mare de unelte de dezvoltare

GML, SGML

Generalized Markup Language (GML)

- Charles Goldfarb, Ed Mosher si Ray Lorie, 1969
- Este un limbaj de formatare-editare utilizat de IBM ce descrie partile unui document in mod structurat precum si relatiile dintre acestea.
- utilizat in 90% din documentele IBM

GML, SGML

Standard Generalized Markup Language (SGML)

- Charles Goldfarb, 1974
- Urmas al limbajului GML
- Extrem de puternic, dar complex
- Adoptat ca standard ISO in 1986
- Utilizat pentru publicarea documentelor electronice

HTML

Hyper Text Markup Language (HTML)

- Tim Berners-Lee si Anders Berglund, 1989
- Limbaj de marcare pentru publicarea documentelor pe Internet
- Este o versiune simplificata a limbajului SGML
- Are un numar fix de tag-uri ce sunt utilizate in principal pentru definirea modului in care este afisat continutul
- Este un caz paticular al unui limbaj de marcare
- Nu poate fi utilizat pentru a definii noi limbaje de marcare

XML

Extensible Markup Language (XML)

- Jon Bosak , 1996
- Consortiu World Wide Web (W3C) combina puterea SGML-ului cu simplitatea HTML-ului realizand limbajul XML
- In 1998 este impus ca standard de W3C, ultima versiune de standard din Februarie 2004 fiind XML 1.0.
- Include multe din caracteristicile SGML-ului, printre care: structura si validarea

Structura documentelor XML

Un document XML este format din:

- marcaje (tag-uri)
- date caracter

Un *marcaj (tag)* este un sir de caractere delimitat de caracterele "<" si ">". *Datele caracter* reprezinta continutul marcajelor.

Un fisier XML cuprinde urmatoarele sectiuni:

- Prolog
- Definitia tipului de document (optionala)
- Elementul radacina

Structura documentelor XML

Exemplu: Fisierul `mail.xml`

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE MAIL SYSTEM "MAIL.DTD">
<MAIL id="10" date="01-03-2019">
    <FROM>mihaiug@central.ucv.ro</FROM>
    <TO>giurca@inf.ucv.ro</TO>
    <SUBJECT>Hello</SUBJECT>
    <MESSAGE>
        Hello Giurca
    </MESSAGE>
</MAIL>
```

Structura documentelor XML

Prologul:

```
<?xml version="1.0" encoding="UTF-8"?>
```

Este o instructiune de procesare. Ea informeaza ca urmeaza descrierea unui fisier XML ce respecta versiunea de specificatie 1.0 iar setul de caractere utilizat este encodat `UTF-8`.

Definitia Tipului de Document:

```
<!DOCTYPE MAIL SYSTEM "MAIL.DTD">
```

Precizeaza ca fisierul `MAIL.DTD` contine declaratia tipului de document (DTD-ul), document ce are ca radacina tag-ul `MAIL`. Acesta este un set de reguli ce defineste structura unui fisier XML.

Structura documentelor XML

Elementul radacina:

```
<MAIL id="10" date="01-03-2019">
    <FROM>mihaiug@central.ucv.ro</FROM>
    <TO>giurca@inf.ucv.ro</TO>
    <SUBJECT>Hello</SUBJECT>
    <MESSAGE>
        Hello Giurca
    </MESSAGE>
</MAIL>
```

Un document XML are un singur element radacina.

Sintaxa unui document XML

Un document XML poate contine urmatoarele tipuri de marcaje:

- Elemente
- Atribute
- Comentarii
- Entitati
- Sectiuni `CDATA`
- Instructiuni de procesare
- Declaratia tipului de document

Sintaxa unui document XML

Elementele:

<nume_tag> ..... </nume_tag>

- sunt blocurile de baza ale unui document XML
- Retin informatii sau definesc structura
- Fiecare element are un tag de inceput si unul de sfarsit
- Elementele pot fi imbricate

Elemente vide:

<nume_tag/>

Sintaxa unui document XML

Exemplu

```
<?xml version="1.0"?>
<BIBLIOTECA>
    <CARTE>
        <TITLU>XML Bible</TITLU>
        <AUTOR>Elliotte Rusty Harold</AUTOR>
        <EDITURA>IDG Books Worldwide</EDITURA>
        <AN_APARITIE>2002</AN_APARITIE>
    </CARTE>
</BIBLIOTECA>
```

Se observa ca

- elementele <TITLU>, <AUTOR>, <EDITURA>, <AN_APARITIE> contin informatii
- <BIBLIOTECA>, <CARTE> sunt folosite doar pentru a defini structura datelor

Sintaxa unui document XML

Atributele:

```
<nume_tag numeAtr1="val1" ... numeAtrN="valN">
          . . .
</nume_tag>
```

- sunt localizate in tag-ul de start al unui element
- au rolul de a descrie elementele
- adesea contin metadate, ex: `id`

Sintaxa unui document XML

Exemplu:

```
<?xml version="1.0"?>
<BIBLIOTECA>
    <CARTE cota="12345">
        <TITLU>XML Bible</TITLU>
        <AUTOR>Elliotte Rusty Harold</AUTOR>
        <EDITURA> IDG Books Worldwide</EDITURA>
        <AN_APARITIE>2002</AN_APARITIE>
    </CARTE>
</BIBLIOTECA>
```

- cota="12345" este un atribut

Sintaxa unui document XML

Comentarii:

```
<!-- comentariu -->
```

- sunt secvente de caractere ce pot aparea oriunde in document ignorate la parsare
- Ele nu fac parte din datele caracter ale documentului

Sintaxa unui document XML

Exemplu:

```
<?xml version="1.0"?>
<!-- Documentul retine cartile dintr-o biblioteca -->
<BIBLIOTECA>
    <CARTE cota="12345">
        <!--  titlul cartii -->
        <TITLU>XML Bible</TITLU>
        <!-- Autorul Cartii -->
        <AUTOR>Elliotte Rusty Harold</AUTOR>
        <!-- Editura in care a aparut cartea -->
        <EDITURA> IDG Books Worldwide</EDITURA>
        <!-- Anul de aparitie a cartii -->
        <AN_APARITIE>2002</AN_APARITIE>
    </CARTE>
</BIBLIOTECA>
```

Sintaxa unui document XML

Entitati:

`&nume_entitate;`

- sunt unitati de text, unde o unitate poate fi orice, de la un singur caracter la un intreg document sau chiar o referinta la un alt document.

Sintaxa unui document XML

Entitate	Referinta la entitate
<	<
>	>
&	&
'	'
"	&quote;

Table: Entitati definite in XML

Sintaxa unui document XML

Exemplu:

```
<TITLE>Tom &amp; Jerry</TITLE>
```

Dupa analizarea textului de catre analizorul XML, va rezulta:

```
Tom & Jerry
```

Sintaxa unui document XML

Instructiuni de prelucrare:

```
<?numeaplicatie instructiune="valoare" ?>
```

- contin informatii despre anumite aplicatii ce urmeaza a fi executate
- incepe cu <? urmeaza numele aplicatiei si se incheie cu ?>
- numele aplicatei trebuie sa fie diferit de `xml` sau `XML`

Sintaxa unui document XML

Exemplu:

```
<?xml-stylesheet type="text/css" href="mySheet.css"?>

sau

<?xml-stylesheet type="text/xsl" href="fisier.xsl"?>
```

Sintaxa unui document XML

Sectiuni CDATA:

```
<![CDATA[
....
]]>
```

- utilizate pentru a include blocuri de text continand caractere care altfel ar fi recunoscute ca marcaje
- sunt folosite in general atunci cand dorim ca datele incluse in interiorul lor sa nu fie interpretate de catre analizor, ci sa fie considerate date caracter

Sintaxa unui document XML

Exemplu:

```
<?xml version="1.0" encoding="UTF-8"?>
<exemplu>
Un exemplu de creare a unui tabel in HTML:
 <![CDATA[
        <table align="center">
                <tr>
                    <td>Coloana 1</td>
                    <td>Coloana 2</td>
                </tr>
        </table>
]]>;
</exemplu>
```

Sintaxa unui document XML

Declaratia tipului de document:

- marcaj special ce poate fi inclus in documentele XML cu rolul de a specifica existenta si locatia **definitiei tipului de document** (**DTD** -**D**ocument **T**ype **D**efinition)
- DTD-ul este un set de reguli care definesc structura unui document XML
- se face intre prolog si elementul radacina
- sintaxa declaratiei tipului de document difera in functie de tipul DTD-ului: intern sau extern.

Sintaxa unui document XML

Sintaxa declararii tipului de document cu DTD intern:

```
<!DOCTYPE element_radacina [
<!-- Setul de reguli-->
]>
```

Sintaxa declararii tipului de document cu DTD extern:

```
<!DOCTYPE root SYSTEM "reguli.dtd">
```

Documente bine formate (Well-Formed Documents)

Un document XML este un document bine format daca satisface urmatoarele conditii sintactice:

- au exact un singur element radacina (root element)
- fiecare element are un tag de inceput si unul de sfarsit
- tag-urile sunt inchise corect, adica primul tag deschis trebuie sa fie ultimul care este inchis
- numele atributelor sunt unice in cadrul unui element

Exemple de limbaje de marcare

Exemple de limbaje de marcare bazate pe XML:

- **MathML** Mathematical Markup Language.
- **CML** Chemical Markup Language.
- **SpeechML** Speech Markup Language.
- **XFRML** Extensible Financial Reporting Markup Language.
- **SMIL** Synchronized Multimedia Interface Language.
- **PDML** Product Data Markup Language.
- **Microsoft Office Open XML**

Exemplu de document MathML

Un exemplu de document in MathML:

```
<html>
    <body>
        <math>
            <mrow>
                <msup>
                    <mi>x</mi>
                    <mn>2</mn>
                </msup>
                <mo>=</mo>
                <mn>2</mn>
            </mrow>
        </math>
    </body>
</html>
```

Documentul MathML anterior reprezinta:

$x^2 = 2$

Editoare XML

Cele mai cunoscute editoare XML:

- **oXygen XML Editor** - *http://www.oxygenxml.com*
- **Altova XML Spy** - *http://www.altova.com*
- **Stylus Studio** - *http://www.stylustudio.com*

Bibliografie

XML

- Extensible Markup Language (XML) 1.0 (Third Edition) W3C Recommendation 4th February 2004, François Yergeau, Tim Bray, Jean Paoli, C. M. Sperberg-McQueen, Eve Maler
- Elliotte Rusty Harold, XML Bible. IDG Books Worldwide, Inc, 919 E. Hillsdale Blvd., Suite 400, Foster City, CA 94404
- http://www.w3schools.com/xml/

Q & A

- Intrebari?
- Comentarii?