Validarea documentelor XML cu DTD

Mihai Gabroveanu <mihaiug@central.ucv.ro>

Copyright © 2006 Mihai Gabroveanu

Abstract

Scopul DTD-ului (**D**ocument **T**ype **D**efinition) este de a definii constructia corecta a blocurilor intr-un document XML. El defineste structura documentului impreuna cu lista elementelor permise.

Table of Contents

1. Introducere

Un document XML este ***valid*** daca:

- este un document "bine format" (well-formed)
- refera o gramatica (DTD sau XML Schema), si
- respecta acea gramatica

2. Definirea Tipului de Document

Termenul de documente electronice acopera o multime de tipuri de documente. Un **tip de document** acopera o clasa similara de documente, ca manuale tehnice, carti, cursuri, carti de telefoane, etc.

DTD-ul (**D**ocument **T**ype **D**efinition - definitia tipului de document) defineste forma si sintaxa constructiilor unui document XML dintr-o anumita clasa. Cu alte cuvinte DTD-ul este un set de reguli care definesc modul de structurare a documentelor XML. Trebuie sa facem deosebire intre definitia tipului documentului si declaratia tipului de document care are rolul de a “spune” analizorului ce DTD trebuie s foloseasc pentru verificare si validare.

Definirea DTD-ului devine un foare importanta atunci cand schimbam, procesam sau afisam documente XML in diferite moduri, de exemplu daca intr-un context B2B se doreste interschimbarea de mesaje XML ce respecta un format standard. Creand un DTD si facandu-l disponibil pe WEB, facem ca un limbaj nou sau extins sa fie inteligibil si utilizabil de catre orcine care are un browser care suporta XML.

Standardele DTD-ului sunt definite de World Wide Web Consortium (W3C).

Declararea tipului de document specifica de ce tip este documentul XML. Aceasta declaratie "spune" analizorului unde gaseste DTD-ul. Un document XML poate sa aibe sau nu asociat un DTD. In cazul in care acesta exista, utilizatorul trebuie s se conformeze acestuia. Un DTD poate sa fie *intern* (definit in interiorul fisierului XML) sau *extern* definit intr-un fisier separat.

XML to PDF by RenderX XEP XSL-FO Formatter, visit us at http://www.renderx.com/

2.1. Definirea unui DTD intern

DTD-ul intern se declar imediat dup declaratia XML sau, dac aceast declaratie nu exist , el va fi primul element exceptând comentariile, spatiile de nume sau instructiunile de procesare. Sintaxa declararii unui DTD intern este urmatoarea:

```
<!DOCTYPE element_radacina [
          declaratii de elemente, atribute, entit ti, instructiuni de procesare, notatii
          ]>
```

Example 1. Carte.xml

In exemplul urmator este prezentat un fisier XML ce are definit DTD-ul in cadrul sau:

```
<?xml version="1.0"?>
<!DOCTYPE  CARTE [❶
   <!ELEMENT CARTE (TITLU, AUTOR, EDITURA, AN_APARITIE)>❷
   <!ELEMENT TITLU (#PCDATA)>❸
   <!ELEMENT AUTOR (#PCDATA)>❹
   <!ELEMENT EDITURA (#PCDATA)>❺
   <!ELEMENT AN_APARITIE (#PCDATA)>❻
   ]>
<CARTE>
    <TITLU>XML Bible</TITLU>
    <AUTOR>Elliotte Rusty Harold</AUTOR>
    <EDITURA> IDG Books Worldwide</EDITURA>
    <AN_APARITIE>2002</AN_APARITIE>
</CARTE>
```

unde:

❶ defineste ca acest document este de tip CARTE
❷ defineste elementul CARTE ca fiind format din patru elemente: TITLU, AUTOR, EDITURA, AN_APARITIE
❸ defineste elementul TITLU ca fiind unul de tip #PCDATA
❹ defineste elementul AUTOR ca fiind unul de tip #PCDATA
❺ defineste elementul EDITURA ca fiind unul de tip #PCDATA
❻ defineste elementul AN_APARITIE ca fiind unul de tip #PCDATA

2.2. Definirea unui DTD extern

Un DTD extern este definit intr-un alt fisier text care trebuie s se afle la o adres specificat . Acest fisier extern poate fi referit printr-un i*dentificator public* si/sau unul *sistem*. Un DTD referit identificator system este utilizat exclusiv de o singura aplicatie, in timp ce un DTD identificat printr-un identificator public poate fi partajat de una sau mai multe aplicatii.

Sintaxa declararii unui DTD extern este urmatoarea:

```
<!DOCTYPE element_radacina SYSTEM "SYSTEM-URI">
```

sau

```
<!DOCTYPE element_radacina PUBLIC "PUBLIC-URI" "SYSTEM-URI">
```

unde, element_radacina este numele elementului radacina.

Exemplu: Presupunand ca avem un DTD pentru un document de tip CARTE care este disponibil la adresa URL `http://inf.ucv.ro/CARTE.dtd` atunci un declararea tipului unui document XML de acest tip este de forma:

```
<?xml version="1.0">
<!DOCTYPE CARTE SYSTEM "http://inf.ucv.ro/CARTE.dtd" >
<CARTE>
    <TITLU>XML Bible</TITLU>
```


XML to PDF by RenderX XEP XSL-FO Formatter, visit us at http://www.renderx.com/

```
    <AUTOR>Elliotte Rusty Harold</AUTOR>
    <EDITURA> IDG Books Worldwide</EDITURA>
    <AN_APARITIE>2002</AN_APARITIE>
</CARTE>
```

Fisierul `CARTE.dtd` ce contine definirea tipului de fisier

```
<?xml version="1.0"?>
<!ELEMENT CARTE (TITLU, AUTOR, EDITURA, AN_APARITIE)>
<!ELEMENT TITLU (#PCDATA)>
<!ELEMENT AUTOR (#PCDATA)>
<!ELEMENT EDITURA (#PCDATA)>
<!ELEMENT AN_APARITIE (#PCDATA)>
```

3. Sintaxa DTD-urilor

Reaminim ca un fisier XML este construit din urmatoarte tipuri de blocuri simple:

- Elemente
- Atribute
- Entitati
- Date caracter neparsabile CDATA
- Date caracter parsabile PCDATA

Un fisier DTD contine tipuri de definitii:

- ELEMENT
- ATTLIST
- ENTITY
- NOTATION

3.1. Declararea elementelor - ELEMENT

Intr-un fisier DTD, elementele sunt declarate printr-o declaratie de tip element cu rumatoarea sintaxa:

```
<!ELEMENT nume-element (continut-element)>
```

sau

```
<!ELEMENT nume-element categorie>
```

Elementele vide

Elementele vide sunt declarate utilizand cuvantul cheie de categorie EMPTY:

```
<!ELEMENT nume-element EMPTY>
```

Exemplu: Daca conideram elementul XML
, atunci declararea lui este:

```
<!ELEMENT br EMPTY>
```

Elementele numai cu date caracter

Elementele numai cu date caracter se declara astfel:

```
<!ELEMENT nume-element (#PCDATA)>
sau
<!ELEMENT nume-element (#CDATA)>
```


XML to PDF by RenderX XEP XSL-FO Formatter, visit us at http://www.renderx.com/

Exemplu: Daca conideram elementul XML AUTOR din exempul anterior, atunci declararea lui este:

```
<!ELEMENT AUTOR (#PCDATA)>
```

Elementele cu orice continut

Elementele ce contin orice tip de date parsabile sau neparsabile se declara astfel:

```
<!ELEMENT nume-element ANY>
```

Elementele ce au copii

Daca un element are o multime de elemente copii atunci ele trebuie enumerate exat in ordinea in care apara in document. Sintaxa este urmatoarea:

```
<!ELEMENT nume-element (nume-copil-1,nume-copil-2, ...)>
```

Exemplu: Daca conideram elementul XML CARTE din exempul anterior, atunci declararea lui este:

```
<!ELEMENT CARTE (TITLU, AUTOR, EDITURA, AN_APARITIE)>
```

Elementele copil trebuie sa fie decalarate si ele la randul lor:

```
<!ELEMENT TITLU (#PCDATA)>
<!ELEMENT AUTOR (#PCDATA)>
<!ELEMENT EDITURA (#PCDATA)>
<!ELEMENT AN_APARITIE (#PCDATA)>
```

Numarul de aparitii ale elementelor

Putem specifica daca dorim numarul de aparitii ale unui element

Table 1. Specificarea cardinalitatii

Simbol	Numar de aparitii
nimic	elementul apare o singura data
*	0 sau mai multe ori
+	1 sau mai multe ori
?	0 sau 1 ori

Exemplu: Sa presupunem ca dorim sa declaram un document ce contine cartile dintr-o biblioteca. Atunci DTD-ul documentului respectiv ar putea fi de forma:

```
<?xml version="1.0"?>
<!ELEMENT BIBLIOTECA (CARTE*)>
<!ELEMENT CARTE (TITLU, (AUTOR+ | EDITOR), EDITURA, AN_APARITIE, REZUMAT?)>
<!ELEMENT TITLU (#PCDATA)>
<!ELEMENT AUTOR (#PCDATA)>
<!ELEMENT EDITOR (#PCDATA)>
<!ELEMENT EDITURA (#PCDATA)>
<!ELEMENT AN_APARITIE (#PCDATA)>
<!ELEMENT REZUMAT (#PCDATA)>
```

Un fisier XML valid ar fi urmatorul:

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


XML to PDF by RenderX XEP XSL-FO Formatter, visit us at http://www.renderx.com/

```
        <TITLU>The XML Handbook </TITLU>
        <AUTOR>Charles F. Goldfarb</AUTOR>
        <AUTOR>Paul Prescod</AUTOR>
        <EDITURA> Prentice Hall PTR</EDITURA>
        <AN_APARITIE>2003</AN_APARITIE>
        <REZUMAT>Prezinta o introducere in XML</REZUMAT>
    </CARTE>
    <CARTE>
        <TITLU>XML Exercises </TITLU>
        <EDITOR>Charles F. Goldfarb</EDITOR>
        <EDITURA> Prentice Hall PTR</EDITURA>
        <AN_APARITIE>2003</AN_APARITIE>
    </CARTE>
</BIBLIOTECA>
```

3.2. Declararea atributelor - ATTLIST

Intr-un DTD, atributele se declara utilizand adnotarea ATTLIST. O declarare a unui atribut specifica elementul caruia ii este asociat, numele lui, tipul, si posibilele valori implicite. Sintaxa de declarare este urmatoarea:

```
<!ATTLIST element-name attribute-name attribute-type #DEFAULT  default-value>

<!ATTLIST element-name  attribute-name attribute-type #FIXED   fixed_value>

<!ATTLIST element-name attribute-name attribute-type (Val1|Val2|..) default_val>

<!ATTLIST element-name attribute-name attribute-type  #IMPLIED>

<!ATTLIST element-name attribute-name attribute-type  #REQUIRED>
```

unde

- **Tipul atributului**

  Tipul unui atribut poate fi:

  Table 2.

Tip	Descriere
CDATA	Date caracter
ENTITY	Valoarea este o entitate
ENTITIES	Valoarea este o lista de entitati
ID	Vloarea este un id unic
IDREF	Valoarea este o reverinta la un alt id
IDREFS	Valoarea este o lista de referinte la alte id-uri
NMTOKEN	Valoarea este un nume valid XML
NMTOKENS	Valoarea este o lista de nume valide XML
NOTATION	Numele unei notati
(val1|val2|...)	Lista de valori
xml:	Valoarea este una predefinita in XML

- **Valoarea implicita**

  Valoare implicita poate fi:

XML to PDF by RenderX XEP XSL-FO Formatter, visit us at http://www.renderx.com/

Table 3.

Valoare implicita	Descriere
#DEFAULT value	daca valoarea nu exista, se va prelua valuarea default definita
#FIXED value	daca in document exista o alta valoare atunci se genereaza eroare la validare
#IMPLIED	atributul poate sa lipseasca fiind optional
#REQUIRED	daca elementul nu contine atributul respectiv se genereaza eroare

Exemplu:

```
<!ATTLIST persoana sex CDATA #DEFAULT "masculin">
<!ATTLIST persoana sex CDATA #FIXED "masculin">
<!ATTLIST persoana sex CDATA #REQUIRED>
<!ATTLIST persoana sex CDATA #IMPLIED>
<!ATTLIST persoana sex (masculin|feminin) "masculin">
```

3.3. Declararea entitatilor - ENTITY

O entitate poate fi declarata *local* sau *extern*.

Sintaxa declararii entitatilor in interiorul DTD-ului este:

```
<!ENTITY entity-name entity-value>
```

Exemplu:

```
<!ENTITY website "http://inf.ucv.ro">
<!ENTITY copyright "Copyright (c) 2006 Mihai Gabroveanu.">
```

Utilizare intr-un fisier XML

```
<author>&copyright;&website;</author>
```

Sintaxa declararii entitatilor in interiorul DTD-ului este:

```
<!ENTITY entity-name SYSTEM "entity-URL">
```

Exemplu:

```
<!ENTITY website SYSTEM "http://inf.ucv.ro/entity.dtd">
```

4. Bibliografie

- 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/

XML to PDF by RenderX XEP XSL-FO Formatter, visit us at http://www.renderx.com/