XML

eXtensible Markup Language

XML se koristi za prenos i smeštanje podataka, za razliku od HTML-a koji je osmišljen za prikaz podataka.Veoma je važno poznavati ga, ali je takođe veoma jednostavan za razumevanje, učenje i korišćenje.

```
<?xml version="1.0"?>
<note>
        <to> Studenti koji slusaju predmet RSZES</to>
        <from> Predrag Teodorovic </from>
        <heading> Prezentacija </heading>
        <body> Molim vas da prisustvujete predavanju jer je izuzetno znacajno da naucite XML </body>
</note>
```

Posmatrajući tag-ove u primeru gore (to, from, heading, ...) oni nisu predefinisani i korisnik mora sam definisati svoje tag-ove. Osim toga, sam XML je osmišljen tako da bude jednostavan, jasan i jednostavan za korišćenje (self-descriptive).

Važno je naglasiti da XML nije osmišljen da zameni HTML jer je cilj XML-a, kao što je već nagovešteno, da prenosi i čuva podatke, sa fokusom na samim podacima, dok je HTML namenjen prikazu podataka, sa fokusom na to kako prikazani podaci izgledaju. U većini web-aplikacija XML se koristi upravo za prenos podataka, dok se za formatiranje i prikaz podataka koristi HTML.
XML je mehanizam prenosa podataka nezavisan od hardvera i softvera. Preporučen je od strane W3C (world wide web consortium) od 10.februara 1998. godine. Danas se XML nalazi svuda: postao je jednako značajan za web aplikacije koliko je i HTML bio značajan za nastajanje web-aplikacija.

Malo je nezahvalno reći, ali činjenica je da XML zapravo ne radi ništa: XML datoteke se ne parsiraju, ne kompajliraju u cilju dobijanja izvršnog programa ili nešto slično. U primeru koji je dat gore, vidi se jednostavnost i jednoznačnost XML-a: sve informacije su zapravo očigledne bez potrebe za nekim dodatnim objašnjenjem. Ali pored svega toga, XML zapis ne radi ništa: on samo predstavlja informacije zapakovane u tag-ove. Da bi se ovakav jedan zapis koristio, neko mora da napiše software koji šalje, prima ili prikazuje podatke iz ovog XML zapisa.

Ono što je takođe značajno dodatno naglasiti jeste da tag-ovi u primeru gore nisu definisani nikakvim standardom: oni su zapravo osmišljeni od strane autora XML dokumenta. Razlog ovome leži u činjenici da XML jezik nema predefinisane tag-ove, za razliku od HTML-a gde se smeju koristiti samo tag-ovi definisani HTML standardom (npr <p>,<h1>,...). XML dozvoljava autoru XML dokumenta da sam definiše svoje tag-ove i svoju strukturu dokumenta.

Korišćenjem XML-a podaci mogu biti skladišteni u više od jedne datoteke. Obzirom da se podaci smeštaju u isključivo tekstualnom formatu, činjenica je da se ti podaci mogu interpretirati nezavisno od softverske i hardverske platforme. Na taj način XML znatno umanjuje kompleksnost prenosa informacija između ne-kompatibilnih sistema na Internetu.

Značajan broj modernih Internet jezika je zapravo baziran na XML-u. U njih spadaju:

- XHTML
- WSDL (koristi se za opisivanje raspoloživih web servisa)
- WAP i WML kao markup-jezici za hendheld uređaje
- RSS (news feeds)

- RDF i OWL za opis resursa
- SMIL za opis multimedijalnog sadržaja na Internetu.

Jednom razumnom strategijom razvoja, u budućnosti ćemo imati softver i baze podataka koji bez problema razmenjuju podatke u XML formatu, bez potrebe konverzije i reformatiranja podataka.

1. XML stablo (XML tree)

Svi XML dokumenti imaju istu strukturu koja uključuje polaznu tačku (the “root”) i grana se prema listovima (“leaves”). U malo modifikovanom primeru koji smo dali ranije:

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<note>
        <to> Studenti koji slusaju predmet RSZES</to>
        <from> Predrag Teodorovic </from>
        <heading> Prezentacija </heading>
        <body> Molim vas da prisustvujete predavanju jer je izuzetno znacajno da naucite XML </body>
</note>
```

Prva linije predstavlja XML deklaraciju. U njoj je definisana verzija XML-a (1.0) kao i kodni raspored karaktera koji se koriste u okviru dokumenta (ISO-8859-1 odgovara Latin-1/West European kodnom rasporedu). Sledeći red opisuje “root” XML dokumenta i treba da označi da ovaj dokument zapravo predstavlja podsetnik (`<note>`). Naredna četiri reda opisuju 4 “child” elementa (to, from, heading i body). Konačno, poslednji red definiše kraj “root” elementa: `</note>`. Trebalo bi da bude očigledno da ovaj XML primer zapravo predstavlja podsetnik asistenta studentima da treba da prisustvuju prezentaciji XML jezika uz dodatno objašnjenje zašto je to značajno.

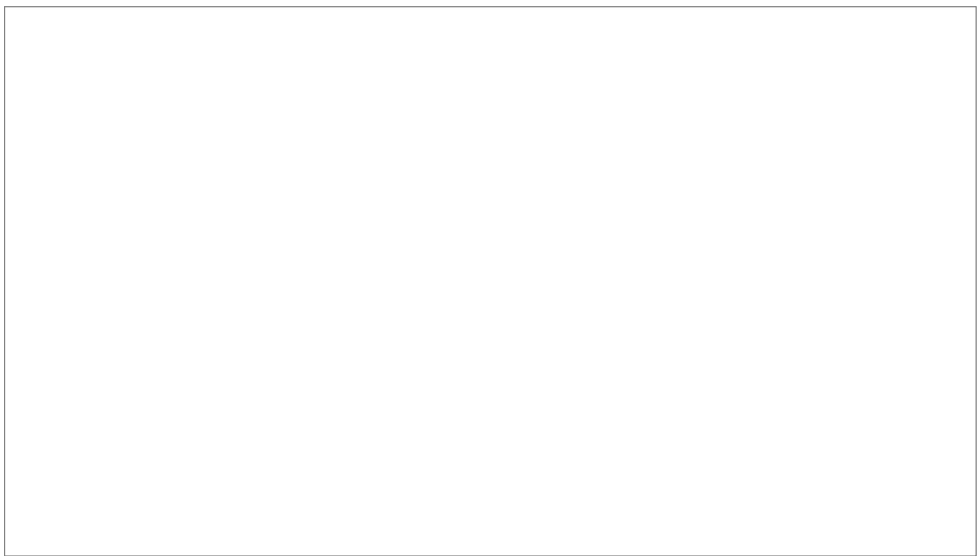
Svaki XML dokument mora da sadrži “root” element. Ovaj element je “parent” svim ostalim elementima. XML dokument se grana od “root”-a do najnižih nivoa u hijerarhiji. Svi elementi mogu imati “child” elemente.

```
<root>
  <child>
    <subchild>.....</subchild>
  </child>
</root>
```

Termini “parent”, “child” i “siblings” se koriste sa ciljem da se opišu relacije između XML elemenata. “sibling” su zapravo dva “child” elementa na istom “nivou” XML hijerarhije.

Svi elementi mogu imati tekstualni sadržaj ili atribute (ovo je isto kao i u HTML-u).

Primer:



Slika gore predstavlja jednu knjigu iz knjižare koja bi sa ostalima u XML-u mogla biti zapisana:

```
<bookstore>
  <book category="COOKING">
    <title lang="en">Everyday Italian</title>
    <author>Giada De Laurentiis</author>
    <year>2005</year>
    <price>30.00</price>
  </book>
  <book category="CHILDREN">
    <title lang="en">Harry Potter</title>
    <author>J K. Rowling</author>
    <year>2005</year>
    <price>29.99</price>
  </book>
  <book category="WEB">
    <title lang="en">Learning XML</title>
    <author>Erik T. Ray</author>
    <year>2003</year>
    <price>39.95</price>
  </book>
</bookstore>
```

“root” element u zapisu gore je <bookstore>. Svi <book> elementi u dokumentu su zapravo sadržani unutar <bookstore> elementa. <book> element sadrži 4 “child” elementa: <title>, <author>, <year> i <price>.

2. XML sintaksa

- svi XML elementi moraju imati zatvarajući tag.

U HTML-u ovo nije slučaj i sintaksa tipa

```
<p>This is a paragraph.
<br>
```

je potpuno regularna. U XML-u nije dozvoljeno isključiti zatvaranje tag-a. Svi elementi **moraju** imati zatvoren odgovarajući tag.

```
<p>This is a paragraph.</p>
```

```
<br />
```

Napomena: XML deklaracija nema zatvoren tag, ali ovo nije greška jer deklaracija zapravo nije deo XML dokumenta u strogom smislu, te stoga predstavlja izuzetak po tom pitanju.

- XML tagovi su case-sensitive

```
<Message>This is incorrect</message>
<message>This is correct</message>
```

- XML elementi moraju biti ugnježđeni kako treba.

U HTML-u se često može videti nešto slično kao:

```
<b><i>This text is bold and italic</b></i>
```

U XML-u svi elementi moraju biti adekvatno ugnježđeni jedan unutar drugog:

```
<b><i>This text is bold and italic</i></b>
```

- XML dokument **mora** imati “root” element

- XML atributi moraju biti pod znacima navodnika.

XML elementi mogu imati atribute u stilu *ime_atributa/vrednost_atributa*. Međutim, svi atributi moraju uvek biti pod znacima navodnika. U primeru dole, prvi slučaj nije korektan, dok je drugi korektan sa aspekta ovoga:

```
<note date=12/11/2007>
  <to>Tove</to>
  <from>Jani</from>
</note>
```

```
<note date="12/11/2007">
  <to>Tove</to>
  <from>Jani</from>
</note>
```

- specijalni karakteri. Neki karakteri u XML-u imaju specijalno značenje.

Npr. ukoliko se karakter “<” stavi unutar XML elementa, rezultat toga će biti greška jer XML parser to interpretira kao početak novog elementa.

```
<message>if salary < 1000 then</message>
```

Da bi se ovo izbeglo, neophodno je kritičan karakter zameniti njegovim XML ekvivalentom:

```
<message>if salary &lt; 1000 then</message>
```

Postoji 5 predefinisanih specijalnih karaktera u XML-u:

<	less than
>	greater than

&amp;		ampersand
&apos;		apostrophe
&quot;		quotation mark

Napomena: samo karakteri “<” i “&” su strogo zabranjeni od strane XML standarda. “>” karakter je legalan, ali je dobra praksa ne koristiti ga kao takvog.

- komentari u XML-u

sintaksa za komentare u XML-u je slična onoj koja se koristi u HTML-u:

<!-- This is a comment -->

- white-space karakteri u XML-u

U HTML-u više takozvanih white-space karaktera su uvek zamenjeni jednim karakterom:

HTML:	Hello	Tove
Output:	Hello Tove	

U XML-u takvi karakteri se ne sažimaju i svi će biti adekvatno sadržani u XML dokumentu.

- u XML-u novi red je predstavlje jednim karakterom LF

U Windows operativnom sistemu novi red se čuva kao grupa karaktera CR+LF. Na MacOSX i Unix operativnim sistemima u tu svrhu se koristi LF, baš kao i u slučaju XML-a.

3. XML elementi

XML element je sve ono što se nalazi (uključujući) start tag do (uključujući) end tag-a. Elementi mogu da sadrže druge elemente, tekst, atribute ili miks svega navedenog.

```
<bookstore>
  <book category="CHILDREN">
    <title>Harry Potter</title>
    <author>J K. Rowling</author>
    <year>2005</year>
    <price>29.99</price>
  </book>
  <book category="WEB">
    <title>Learning XML</title>
    <author>Erik T. Ray</author>
    <year>2003</year>
    <price>39.95</price>
  </book>
</bookstore>
```

U primeru iznad, <bookstore> i <book> su elementi sa sadržajem, jer se sastoje od drugih

elemenata. <book> element takođe ima i atribut (category="CHILDREN"). Elementi <title>, <author>, <year>, i <price> imaju tekstualni sadržaj.

Xml elementi dobijaju nazive u skladu sa sledećim pravilima:

- imena smeju sadržati slova, brojeve i ostale karaktere
- ne smeju počinjati brojem ili znakom interpunkcije
- imena ne mogu počinjati slovima xml, XML, Xml i slično
- imena ne smeju sadržati space-ove
- sve reči mogu biti korišćene za imena elemenata-ne postoje rezervisane reči u tom smislu.

Preporuka prilikom davanja imena elementima jeste da ona treba da budu dovoljno deskriptivna i ako je potrebno da koriste _ za spajanje više reči <first_name>,<last_name>. Takođe, poželjno je da budu što je kraća moguća: <book_title> je svakako bolje od <the_title_of_the_book>. Treba izbegavati "-" karaktere jer softver koji će parsirati XML datoteku može sa pravom očekivati da je potrebno uraditi aritmetičko oduzimanje kada se naiđe na ovakav karakter. Osim toga, treba izbegavati i "." jer je takva notacija često korišćena u OOP. Ni ":" nije poželjan da se nađe u okviru imena elementa jer je ovaj karakter rezervisan za korišćenje namespace-ova (biće uskoro reči i o tome). Ukoliko XML dokument predstavlja interfejs za bazu podataka, dobra je praksa da se imena elemenata daju u skladu sa imenima korišćenim u okviru baze podataka. Ono što je nama posebno interesantno, to je da su i ćirilični i latinični karakteri potpuno podržani od strane XML standarda i ukoliko se koriste to bi samo trebalo jasno naglasiti u prvom redu XML dokumenta (XML deklaracija).

XML elementi su lako proširivi sa ciljem da sadrže više informacija. Na primer:

```
<note>
<to>Tove</to>
<from>Jani</from>
<body>Don't forget me this weekend!</body>
</note>
```

Zamislimo da aplikacija za čitanje ovog XML fajla treba da iz nje izvadi podatke određene sa poljima <to>, <from> i <body> sa ciljem da ih ispiše kao u:

MESSAGE

To: Tove
From: Jani

Don't forget me this weekend!

Ukoliko bi autor XML dokumenta dodao informacije tipa:

```
<note>
<date>2008-01-10</date>
<to>Tove</to>
<from>Jani</from>
<heading>Reminder</heading>
<body>Don't forget me this weekend!</body>
</note>
```

Postavlja se pitanje da li bi to bio problem sa stanovišta aplikacije. Odgovor je NE. Aplikacija bi bez ikakvih problema pronalazila elementa od interesa i samim tim proizvodila isti izlaz i nakon dodavanja dodatnih informacija. Upravo ovo je jedna od glavnih prednosti XML-a: moguće ga je proširivati bez ikakvih posledica po softver koji ih koristi.

4. XML atributi

XML elementi mogu sadržati atribute baš kao i HTML. Uloga atributa je da obezbede dodatne informacije o samom elementu. Atributi najčešće označavaju informacije koje nisu sastavni deo podataka koji se prenosi. U primeru dole, tip datoteke je nevažan sa stanovišta samog podatka, ali može biti izuzetno značajan za softver koji manipuliše tim elementom:

```
<file type="gif">computer.gif</file>
```

Vrednosti atributa moraju uvek biti pod navodnicima. Pri tome, u obzir dolaze i jednostruki i dvostruki navodnici. Na primer pol osobe se može iskazati atributima na dva načina potpuno ekvivalentno:

```
<person sex="female">
```

ili

```
<person sex='female'>
```

Ukoliko atribut sam po sebi sadrži dvostruke navodnike, jednostruki navodnici mogu biti korišćeni da bi se ta vrednost prikazala:

```
<gangster name='George "Shotgun" Ziegler'>
```

Alternativno, mogu biti korišćeni specijalni karakteri:

```
<gangster name="George &quot;Shotgun&quot; Ziegler">
```

Prilikom pisanja XML datoteka, treba biti svestan da se iste informacije mogu prikazati i kao atributi i kao elementi:

```
<person sex="female">
  <firstname>Anna</firstname>
  <lastname>Smith</lastname>
</person>
```

```
<person>
  <sex>female</sex>
  <firstname>Anna</firstname>
  <lastname>Smith</lastname>
</person>
```

Ne postoji pravilo povodom toga kada treba koristiti elemente, a kada atribute. Iako su atributi zgodni u HTML-u, preporuka je da se izbegavaju u okviru XML-a.

Naredna tri elementa sadrže iste informacije:

```
<note date="10/01/2008">
  <to>Tove</to>
  <from>Jani</from>
  <heading>Reminder</heading>
  <body>Don't forget me this weekend!</body>
</note>
```

Atribut date se koristi u prvom primeru. U drugom, on je zamenjen sa elementom:

```
<note>
  <date>10/01/2008</date>
  <to>Tove</to>
  <from>Jani</from>
  <heading>Reminder</heading>
  <body>Don't forget me this weekend!</body>
</note>
```

Proširen element se koristi u trećem primeru:

```
<note>
  <date>
    <day>10</day>
    <month>01</month>
    <year>2008</year>
  </date>
  <to>Tove</to>
  <from>Jani</from>
  <heading>Reminder</heading>
  <body>Don't forget me this weekend!</body>
</note>
```

Neki od problema koji nastaju prilikom korišćenja atributa su sledeći:

- atributi ne mogu da sadrže višestruke vrednosti dok elementi mogu
- atributi ne mogu da se sastoje od stabla kao elementi
- atributi nisu jednako proširivi (za buduće potrebe).

Atributi su složeni za čitanje i održavanje, te je stoga preporučljivo koristiti elemente za podatke, dok atribute treba koristiti za prikaz informacija koje nisu značajne za same podatke sadržane u okviru XML fajla.

Ono što bi svakako trebalo izbeći to je nešto tipa:

```
<note day="10" month="01" year="2008"
to="Tove" from="Jani" heading="Reminder"
body="Don't forget me this weekend!">
</note>
```

Često je potrebno dodavati ID reference elementima. Ovi ID brojevi se mogu kasnije koristiti kako bi se identifikovao neki konkretan XML element, slično kao što se to radi i u okviru HTML-a. Na primer:

```
<messages>
  <note id="501">
    <to>Tove</to>
    <from>Jani</from>
    <heading>Reminder</heading>
    <body>Don't forget me this weekend!</body>
```

```
  </note>
  <note id="502">
    <to>Jani</to>
    <from>Tove</from>
    <heading>Re: Reminder</heading>
    <body>I will not</body>
  </note>
</messages>
```

ID brojevi se koriste da identifikuju različite podsetnike. Oni, samim tim, nisu delovi podsetnika. U ovakvim slučajevima metapodaci (podaci o podacima) treba da budu zapisani u formi atributa, dok podaci treba da budu upisani kao XML elementi.

5. XML namespace

Kao što je već naglašeno ranije, imena elemenata su definisana od strane osobe koja pravi XML dokument. Ovo, nažalost, često rezultuje u konfliktima kada se pokušaju spojiti XML dokumenti iz različitih XML aplikacija. Na primer, ovaj XML sadrži podatke sa HTML tabelom:

```
<table>
  <tr>
    <td>Apples</td>
    <td>Bananas</td>
  </tr>
</table>
```

Dok ovaj XML sadrži informacije o stolu (eng. table isto kao i tabela).

```
<table>
  <name>African Coffee Table</name>
  <width>80</width>
  <length>120</length>
</table>
```

Ukoliko bi neko spojio ove XML fragmente, došlo bi do definitivnog konflikta. Oba sadrže <table> element, ali sami elementi imaju drugo značenje i sadržaj. U tom slučaju XML parser ne bi znao kako da se nosi sa takvim dvosmislenostima.

Ovakvi konflikti se u XML-u rešavaju korišćenjem prefiksa. XML koji sledi sadrži informacije o HTML tabelama i delovima nameštaja istovremeno:

```
<h:table>
  <h:tr>
    <h:td>Apples</h:td>
    <h:td>Bananas</h:td>
  </h:tr>
</h:table>

<f:table>
  <f:name>African Coffee Table</f:name>
  <f:width>80</f:width>
  <f:length>120</f:length>
</f:table>
```

U primeru navedenom gore, neće biti konflikta, jer dva <table> elementa imaju zapravo različita imena.

Prilikom korišćenja prefiksa u XML-u, takozvani *namespace* mora biti definisan.

Namespace se definiše korišćenjem atributa **xmlns** u okviru start tag-a elementa. Sintaksa je sledeća:

xmlns:*prefix*="*URI*".

```
<root>

<h:table xmlns:h="http://www.w3.org/TR/html4/">
  <h:tr>
    <h:td>Apples</h:td>
    <h:td>Bananas</h:td>
  </h:tr>
</h:table>

<f:table xmlns:f="http://www.w3schools.com/furniture">
  <f:name>African Coffee Table</f:name>
  <f:width>80</f:width>
  <f:length>120</f:length>
</f:table>

</root>
```

U gornjem primeru, **xmlns** atribut u okviru <table> tag-a definiše za h: i f: prefikse adekvatan *namespace*.

Kada se definiše namespace za element, "child" elementi sa istim prefiksom se dodeljuju automatski istom namespace-u.

Namespace-ovi mogu biti definisani u okviru elementa u kome se koriste, ili u okviru "root" elementa:

```
<root xmlns:h="http://www.w3.org/TR/html4/"
xmlns:f="http://www.w3schools.com/furniture">

<h:table>
  <h:tr>
    <h:td>Apples</h:td>
    <h:td>Bananas</h:td>
  </h:tr>
</h:table>

<f:table>
  <f:name>African Coffee Table</f:name>
  <f:width>80</f:width>
  <f:length>120</f:length>
</f:table>

</root>
```

Napomena: namespace URI se ne koristi od strane parsera kako bi se proveravale informacije. Uloga ovog link-a je zapravo da se dodeli namespace-u jedinstveno ime. Ipak, često kompanije koriste link za namespace kao pokazatelj na web stranicu na kojoj se nalaze dodatni podaci o tom konkretnom namespace-u (pokušajte otići na http://www.w3.org/TR/html4/).

Definisanje default-nog namespace-a za dati element omogućava da ne moramo unositi prefikse za sve "child" elemente. Sintaksa je sledeća:

```
xmlns="namespaceURI"
```

Tako, XML koji sledi predstavlja HTML tabele:

```
<table xmlns="http://www.w3.org/TR/html4/">
  <tr>
    <td>Apples</td>
    <td>Bananas</td>
  </tr>
</table>
```

dok ovaj sadrži informacije o nameštaju:

```
<table xmlns="http://www.w3schools.com/furniture">
  <name>African Coffee Table</name>
  <width>80</width>
  <length>120</length>
</table>
```
