

XML

Uvod

XML (*Extensible Markup Language*) je markup jezik, sličan HTML-u. Služi za definisanje formata podataka (za razliku od HTML-a koji služi za definisanje prikaza podataka). XML standard propisan je od strane W3C. Ovaj (meta) jezik definiše sintaksu za grupisanje podataka koristeći se takozvanim **etiketama** (engl. *tags*). Mnogi markup jezici se baziraju upravo na XML-u i brojni su softveri i biblioteke koji pružaju podršku za čitanje i pisanje XML-a. Validnost XML dokumenta je lako proveriti, jer XML ima stroga sintaksička pravila.

XML format podataka nezavisan je od platforme i XML dokument može čitati svaka alatka koja rukuje tekstualnim datotekama, s'obzirom da su tekstualni ne samo podaci već i etikete. Dakle, XML generiše podatke koji su prenosivi i nezavisni od hardverske ili softverske platforme. Njegova osnovna zamisao je upravo da olakša razmenu podataka kroz informacione sisteme, koji su potencijalno nekompatibilni.

Struktura i sintaksa

Etikete XML-a nisu predefinisane, tj. Moraju se definisati. Etiketom se daje značenje podatku. Njegova proširivost se ogleda u tome da je dozvoljeno definisanje novih etiketa. Takođe je samodeskriptivan, što znači da svojim etiketama definiše sadržaj obuhvaćen njima. Jedan XML dokument sastoji se iz teksta i etiketa.

U XML dokumentima podatke smeštamo u obliku stringova između etiketa koje ih opisuju i definišu. Ove jedinice se nazivaju **elementima**. Kao što je već napomenuto, njegova proširivost pruža veliku slobodu, ali se moraju poštovati precizna pravila sintakse. Struktura dokumenta je takozvana **stablo** struktura, gde dokument opisuju etikete pomoću kojih se može videti koji elementi su pridruženi drugim. Nepisano pravilo je i da etikete opisuju i semantiku podataka i shodno tome se imenuju. Bitno je zapamtiti da etikete ne sadrže informaciju o tome kako će element biti prikazan.

Hijerarhijska struktura dokumenta, iliti stablo struktura podrazumeva da jedan dokument ima jedan i samo jedan koreni element (engl. *root*). Svi ostali elementi nalaze se unutar korenog elementa. Dozvoljeno je višestruko ugnježdavanje elemenata. Pri pisanju XML dokumenta jako je bitno voditi računa o indentaciji dokumenta, kako bi se razaznalo šta pripada kom elementu i kako su elementi međusobno ugnježdjeni.

Primer jednog XML dokumenta bio bi:

```
<student>

  <ime>Pera</ime>

  <prezime>Perić</prezime>

  <index>RA46</index>

</student>
```

Prva etiketa Student predstavlja otvarajući tag, dok krajnja predstavlja zatvarajući. Svaki tag mora biti otvoren i zatvoren. Zatvarajući tag izgleda isto kao otvarajući sa razlikom u predznaku /. Svi tagovi nalaze se unutar uglastih zagrada <>. Isto važi i za ime, prezime index. Sadržaj unutar njih predstavlja tekst. Imena tagova zavise od velikih i malih slova (engl. *Case sensitive*). Pri dodeljivanju imena se moraju poštovati određena pravila.

XML atributi

Elementima se uglavnom pridružuju i **atributi**. Oni obezbeđuju dodatne informacije o elementima. Vrednost atributa se uvek nalazi između znaka navoda, bez obzira da li su jednostruki ili dvostruki. Moguće je i kombinovanje u slučajevima kada je potrebno prikazati i znake navoda. Razlika u formatima elementa i atributa je sledeća:

```
element:

<ime>Pera</ime>

atribut:

<nesto ime="Pera">
```

Ne postoji jasno definisano pravilo kada se služiti elementom, a kada atributom, ali preporučljivo je da se elementi koriste kada se opisuje podatak koji je već sam po sebi celovita informacija, a ne jedan njen deo. Oba primera koja slede daju istu informaciju, ali su drugačije predstavljeni. Prvi je u formi atributa, drugi elementa:

```
<student tip="novi">

<ime>Pera</ime>

<prezime>Perić</prezime>

</student>

<student>novi</student>

<ime>Pera</ime>

<prezime>Perić</prezime>
```

Naravno, atributi imaju i svojih mana koje treba imati u vidu prilikom korišćenja, kao što su:

- Teška manipulacija iz programskog koda
- Teška proširivost
- Ne mogu da sadrže višestruke vrednosti
- Ne opisuju strukturu

XML deklaracija

Svaki XML dokument mora sadržati deklaraciju na samom početku dokumenta. Ova deklaracija (instrukcija obrade) služi da bi se dokument identifikovao kao XML dokument. Osnovni oblik deklaracije je:

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

Takođe je moguć i sledeći oblik:

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

? označava instrukciju obrade – poruka za program koji procesira XML dokument.

Version je atribut označava XML verziju.

Opcioni atribut **encoding** označava znakovni kod u kom je XML dokument napisan (UTF – 16 Unicode, UTF – 8 kompresovana verzija Unicode-a).

Za verifikaciju oformljenosti XML dokumenta zadužen je **XML parser**. On čita dokument i vrši konverziju u hijerarhijsku strukturu. On uglavnom čini sastvani deo aplikacije i prenosi tako parsirani dokument do same aplikacije.

Entiteti

Entiteti u XML-u su uslovno rečeno varijable korišćene za definisanje teksta. Oni se obrađuju prilikom parsiranja od strane XML parsera. Imena entiteta su prilično intuitivna, a pravilo je da uvek počinju kodom **&** kojem sledi tekst koji opisuje značenje. Neki od predefinisanih XML entiteta su:

- **<** – manje od (engl. *Less than*)
- **>** – veće od (engl. *Greater than*)
- **&** – ampersand
- **"** – dupli znaci navoda

Imenski prostori (engl. *Namespaces*)

U XML imenski prostori koriste se kako bi se izbegli konflikti pri zadavanju imena elementima. Upotrebljavaju se tako što se dodaju prefiksi vezani za specifični element, zatim : (dvotačka), pa ime elementa i etikete.

```
<Tabela>

<tr>

  <td>Ime</td>

  <td>Prezime</td>

</tr>

</Tabela>
```

```
<Tabela>

  <Ime>Pera</Ime>

  <Prezime>Perić</Prezime>

  <Godište>1982</Godište>

</Tabela>
```

Oba elementa sadrže element „Tabela“, iako je kontekst drugačiji. Međutim, ukoliko bi se ova dva fragmenta našli unutar istog dokumenta, nastao bi konflikt. Rešenje je jednostavno – upotrebom imenskih prostora:

```
<h:Tabela>

  <h:tr>

    <h:td>Ime</h:td>

    <h:td>Prezime</h:td>

  </h:tr>

</h:Tabela>

<f:Tabela>

  <f:Ime>Pera</f:Ime>

  <f:Prezime>Perić</f:Prezime>

  <f:Godište>1982</f:Godište>

</f:Tabela>
```

Prilikom korišćenja prefiksa u XML – u, imenski prostor za prefiks se mora definisati. To se radi uz pomoć **xmlns** atributa na otvarajućem tagu elementa. Sintaksa je sledeća: xmlns: *prefix* = „URI“. Kada se definiše imenski prostor elementa, svi deca – elementi sa istim prefiksom su povezani sa tim imenskim prostorom. Imenski prostori se takođe mogu definisati i u korenskom elementu XML dokumenta, koristeći navedenu sintaksu i razdvajajući ih razmakom.

```
<root>
```

```
<h:Tabela xmlns:h="URI">
```

```
<h:tr>
```

```
<h:td>Ime</h:td>
```

```
<h:td>Prezime</h:td>
```

```
</h:tr>
```

```
</h:Tabela>
```

```
<f:Tabela xmlns:f="URI">
```

```
<f:Ime>Pera</f:Ime>
```

```
<f:Prezime>Perić</f:Prezime>
```

```
<f:Godište>1982</f:Godište>
```

```
</f:Tabela>
```

```
</root>
```

Primer jednog složenijeg XML dokumenta, sa upotrebom deklaracije, elemenata sa atributima i imenskim prostorom:

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

```
<bookstore>
```

```
<author xmlns:daunt="http://www.dauntbooks.co.uk"
```

```
name="Eric Arthur Blair" pseudonym="George Orwell">
```

```
<daunt:book daunt:title="Animal Farm" daunt:year="1945">
```

```
<price>500.00</price>
```

```
</daunt:book>
```

```
<book title="1984" year="1949"/>
```

```
</author>
```

```
<author name="Samuel Langhorne Clemens" pseudonym="Mark Twain">
```

```
<book title="The Adventures of Tom Sawyer" year="1876"/>
```

```
<book title="A Dog's Tale" year="1904"/>
```

```
</author>
```

```
<author name="Stephen Edwin King">
```

```
<book title="Carrie" year="1974"/>
```

```
<book title="The Shining" year="1977"/>
```

```
<book title="Misery" year="1987"/>
```

```
</author>
```

```
</bookstore>
```