2. XML jezik za označavanje

Branislava Šandrih
branislava.sandrih@fil.bg.ac.rs

NAPOMENA: Sadržaj ove prezentacije preuzet je od prof. Cvetane Krstev sa http://poincare.matf.bg.ac.rs/~cvetana/kurs-xml/

Šta je XML? (1)

- **eXtensible Markup Language**
  - Proširiv jezik označavanja
- Predstavljanje strukturnih podataka u računaru
  - Novinski članci
  - Crteži
  - Muzičke kompozicije
  - Knjige
  - Korisničke baze podataka
  - Elektronska kola
  - Online porudžbine ...

Šta je XML? (2)

- Osnovna svojstva XML-a:
  - Jednostavan
  - Prenosiv
  - Dobro dokumentovan format za opis podataka
- Ima svojstvo „proširivosti“ novim mogućnostima (koje, takođe koriste osnovni XML):
  - Namespaces (prostor imena) za kombinovanje XML dokumenata iz više izvora
  - XSL za transformisanje XML dokumenata u oblik vidljiv pomoću iteratora
  - DOM definiše hijerarhiju objekata koji čine XML dokument
  - XML Schema okvir za strože deklarisanje strukture i sadržaja XML dokumenata

Šta je XML? (3)

- XML je jezik za označavanje, kao i HTML
- Koja je razlika između XML i HTML?
  - HTML prikazuje podatke na određen način
  - ima predefinisane etikete (body, div, p, br, h1)
  - XML samo čuva podatke
  - nema predefinisane etikete
- XML olakšava:
  - deljenje podataka
  - prenos podataka
  - dostupnost podataka

Šta je XML? (4)

- World Wide Web Consortium preporučuje XML kao format za predstavljanje, čuvanje i razmenu podataka
  - W3C je organizacija koja se bavi standardizacijom Web-a
- XML:
  - **nije** programski jezik
  - **nije** protokol za prenos podataka
    - ali HTTP, FTP, SSH jesu
  - **nije** baza podataka
- Osnovna namena:
  - Cilj XML-a je da olakša automatsku obradu dokumenata i podataka
  - Ideja je da se ostvari takvo strukturiranje informacija koje omogućava i ljudima da ih čitaju na web-u i raznovrsnim aplikacijama da ih automatski obrađuju

Primer XML dokumenta (1)

```
<?xml version="1.0" encoding="ISO-8859-2"?>
<članak>
<naslov>POSLE PORAZA Đoković:
Sada mi se samo ide kući, što pre</naslov>
<datum>19. I 2017.</datum>
<tekst>
<mesto>Melburn</mesto>
Nakon što je od Denisa Istomina izgubio u 2. kolu
Australijan opena sa 7:6(8), 5:7, 2:6, 7:6(5), 6:4,
Novak Đoković je istakao da je njegov uzbekistanski
rival odigrao neverovatno, ali da ga nijednog
trenutka nije potcenio, ioako je u pitanju 117. igrač
sveta.
</tekst>
</članak>
```

POSLE PORAZA Đoković: Sada mi se samo ide kući, što pre

Darko Nikolić | 19. 01. 2017 - 11:20h| Komentara: 256

Nakon što je od Denisa Istomina izgubio u 2. kolu Australijan opena sa sa 7:6(8), 5:7, 2:6, 7:6(5), 6:4, Novak Đoković je istakao da je njegov uzbekistanski rival odigrao neverovatno, ali da ga nijednog trenutka nije potcenio, ioako je u pitanju 117. igrač sveta.

Foto: Reuters
Novak Đoković

Primer XML dokumenta (2)

```
<?xml version="1.0" encoding="ISO-8859-2"?>
<članak naslov="POSLE PORAZA Đoković:
Sada mi se samo ide kući, što pre" datum="19. I 2017." mesto="Melburn">
<tekst>
Nakon što je od <ime vrsta=“osoba”>Denisa Istomina</ime> izgubio u 2. kolu
Australijan opena sa 7:6(8), 5:7, 2:6, 7:6(5), 6:4,
<ime vrsta=“osoba”>Novak Đoković</ime> je istakao da je njegov
uzbekistanski rival odigrao neverovatno, ali da ga nijednog trenutka
nije potcenio, ioako je u pitanju <broj vrsta=“redni”>117.</broj> igrač sveta.
 </tekst>
</članak>
```

Terminologija (1)

- **Element**
  - je oblika
    <PočetnaEtiketa NiskaAtributa> ...sadržaj... </ZavršnaEtiketa>
    Sadržaj, to jest sve ono što se nalazi između početne i završne etikete, može da bude strukturiran i može sadržati druge elemente. U HTML-u su sve etikete unapred date (jer se on zasniva na SGML-u). XML dozvoljava da korisnik po volji dodaje nove etikete

    <mesto>Melburn</mesto>
- **Atribut**
  - par ime-vrednost koji se pridružuje početnoj etiketi elementa. Imena se odvajaju od vrednosti znakom jednakosti i opcionim belinama. Vrednosti su okružene dvostrukim ili jednostrukim znacima navoda. Vrednosti su niske karaktera i ne mogu biti strukturirane

    <ime **vrsta="osoba"**>Novak Đoković</ime>

Terminologija (2)

- **Prazan element**
  - je element kod koga je sadržaj između početne i završne etikete prazan. Za takav element se može koristiti posebna početna etiketa
    <PočetnaEtiketa NiskaAtributa />
    Na primer, u XHTML-u elementi za novi red i horizontalnu liniju su
 i <hr/> umesto
 i <hr>. Ne treba misliti da prazni elementi ne nose informaciju, ta informacija je sadržana u atributima
- **Koreni element**
  - je prvi element u XML dokumenu koji sadrži sve ostale elemente. U prethodnom primeru, koreni element je element <članak>.
- **XML imena**
  - koriste se za imenovanje elemenata, atributa i drugih konstrukata. Ona predstavljaju nisku alfanumeričkih karaktera, pod čime se podrazumevaju slova alfabeta, cifre od 0 do 9, kao i slova i cifre iz drugih alfabeta. Dozvoljeno je i korišćenje tri interpunkcijska karaktera: podvlaka, crtica i tačka. Ime može početi samo slovom i podvlakom

Jednostavan sadržaj elementa

- Sadržaj svakog elementa su:
  - njegova deca, to jest drugi elementi ili
  - karakterski podaci, to jest tekst koji ne sadrži druge elemente

```
<?xml version="1.0" encoding="ISO-8859-2"?>
<osoba rođen="23. VI 1912." umro="7. VI 1954.">
  <ime>
    <lično_ime>Alan</lično_ime>
    <prezime>Tjuring</prezime>
  </ime>
  <profesija>informatičar</profesija>
  <profesija>matematičar</profesija>
  <profesija>kriptograf</profesija>
</osoba>
```

Mešoviti sadržaj elemenata

- U primenama vezanim za obradu prirodnojezičkih dokumenata elementi obično imaju mešoviti sadržaj. Sadržaj jednog elementa su i karakterski podaci i drugi elementi

```
<?xml version="1.0" encoding="ISO-8859-2"?>
<biografija> <ime><lično_ime>Alan</lično_ime>
<prezime>Tjuring</prezime></ime>,
rođen <datum><godina>1912.</godina><datum>, jedan je od prvih ljudi
koji zaista zaslužuju naziv <istaknuto>informatičara</istaknuto>. Svi
njegovi doprinos ovom polju ne mogu se nabrojati; napoznatiji
su <istaknuto>Tjuringov test</istaknuto> i <istaknuto>Tjuringova
mašina.</istaknuto>
<ime><prezime>Tjuring</prezime></ime> je bio i vrlo
dobar <profesija>matematičar</profesija> i <profesija>kriptograf</pro
fesija>. Prevashodno uz njegovu pomoć, saveznici su u toku Drugog
svetskog rata uspeli da dekodiraju nemačku mašinu "Enigma". Izvršio
je samoubistvo <datum><dan>7.</dan> <mesec>maja</mesec>
<godina>1954.</godina></datum> godine, pošto je optužen zbog
homoseksualnosti i primoran na ponižvajuće "lečenje".</biografija>
```

Reference entiteta (1)

- U karakterskim podacima znak “<” se ne može koristiti jer se on uvek interpretira kao početak etikete. Isto tako se ne može koristiti ni “&” jer se on interpretira kao početak reference entiteta
- U XML-u postoji pet predefinisanih entiteta:
  - **<** znak za manje ili otvorena šiljasta zagrada (<)
  - **&** znak za ampersand (&)
  - **>** znak za veće ili zatvorena šiljasta zagrada (>)
  - **"** dvostruki znaci navoda (")
  - **'** apostrof ili jednostruki znaci navoda (')

Reference entiteta (2)

```
<?xml version="1.0" encoding="ISO-8859-2"?>
<izdavač>O'Reilly & Associates</izdavač>
<image source='oreilly_koala.gif' alt='Powered by O’Reilly Books'/>
```

```
<?xml version="1.0" encoding="ISO-8859-2"?>
<izdavač>O&apos;Reilly &amp; Associates</izdavač>
<image source='oreilly_koala.gif' alt='Powered by O&apos;Reilly Books'/>
```

CDATA
Odeljci karakterskih podataka (1)

- Odeljci karakterskih podataka mogu da sadrže proizvoljne podatke, uključujući i otvorene i zatvorene šiljaste zagrade
  - ovi podaci se **ne parsiraju**.
  - smeštaju se unutar oznaka **<![CDATA[i]]>**
  - samo se završna oznaka odeljka karakterskih podataka ne sme naći među podacima
  - koriste se za uključivanje HTML ili XML koda u tekst

CDATA
Odeljci karakterskih podataka (2)

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<podaci>
    <![CDATA[
    <?xml version="1.0" encoding="ISO-8859-2"?>
    <osoba rođen="23. VI 1912." umro="7. VI 1954.">
      <ime>
        <lično_ime>Alan</lično_ime>
        <prezime>Tjuring</prezime>
      </ime>
      <profesija>informatičar</profesija>
      <profesija>matematičar</profesija>
      <profesija>kriptograf</profesija>
    </osoba>
    ]]>
</podaci>
```

Komentari

- XML dokumenti mogu sadržati komentare, baš kao i programi
- Komentari se pišu unutar oznaka **<!--** i **-->**

```
<?xml version="1.0" encoding="ISO-8859-2"?>
<osoba rođen="23. VI 1912." umro="7. VI 1954.">
  <ime>
    <lično_ime>Alan</lično_ime>
    <prezime>Tjuring</prezime>
  </ime>
  <profesija>informatičar</profesija>
  <profesija>matematičar</profesija>
  <profesija>kriptograf</profesija>
  <!--Na ovom mestu treba ubaciti sliku kad bude gotova -->
</osoba>
```

Instrukcije za obradu (1)

- XML instrukcije za obradu su neka vrsta komentara koja može da bude od značaja nekim aplikacijama koje dokument čitaju
- Instrukcije za obradu se pišu unutar oznaka **<?** i **?>**
- Ono što neposredno sledi oznaku <? je **meta**, a to je najčešće ime aplikacije kojoj su instrukcije namenjene
- Sve ostalo iza mete su instrukcije namenjene toj aplikaciji u odgovarajućem formatu

Instrukcije za obradu (2)

```
<?robots index="yes" follow="no" ?>
```

- Meta ove instrukcije za obradu je **robots** i ona govori pretraživačima i drugim robotima da li da indeksiraju određenu stranicu
- Ova instrukcija ima dva atributa index i follow čije su dozvoljene vrednosti yes i no
  - ako je vrednost za index „yes“, xml dokument će biti indexiran; u suprotnom neće
  - ako je vrednost za follow „yes“, linkovi iz xml dokumenta će biti praćeni; u suprotnom neće

XML deklaracija

- XML dokumenti bi trebalo da imaju deklaraciju, ali ona nije obavezna
- XML deklaracija izgleda kao instrukcija za obradu kod koje je meta xml i koja ima tri atributa:
  - **version** je verzija XML i, za sada tu treba da stoji 1.0
  - **encoding** govori koja šema kodiranja se koristi (ako nije navedena, pretpostavlja se da je to Unicode)
  - **standalone** govori da li je dokument samostalan (tj. ne zavisi od nekog spoljašnjeg DTD-a)

```
<?xml version="1.0" encoding="ISO-8859-2" standalone="yes"?>
```

Dobro formirani XML dokumenti

- Svaki XML dokument mora da bude dobro formiran, a to znači da mora da zadovolji određene uslove. Neki od tih uslova su:
  - svaka početna etiketa, osim etiketa praznih elemenata, mora da ima završnu etiketu
  - sadržaji elemenata ne mogu da se preklapaju
  - dokument može da ima samo jedan koreni element
  - vrednosti atributa moraju biti unutar navodnika
  - jedan element ne može imati dva atributa s istim imenom
  - komentari i instrukcije za obradu ne mogu se navoditi unutar etikete
  - karakteri < i & ne smeju se pojavljivati unutar vrednosti atributa ni unutar karakterskog sadržaja elemenata.

XML parser

- **XML parser** je program koji otkriva greške u formiranju XML dokumenta i izveštava o njima
- Zadatak XML parsera nije da te greške ispravlja niti da pokuša da protumači šta je autor, u stvari, mislio
- XML parseri imaju, između ostalog, zadatak da zamene reference entiteta karakterima na koje se one odnose. Tako se dobijaju **parsirani karakterski podaci (PCDATA)**
- Razlika između:
  - PCDATA [Parsed Character Data] tekstualni sadržaj koji XML parser **treba** da parsira
  - CDATA [(Unparsed) Character Data] tekstualni sadržaj koji XML parser **ne treba** da parsira
- Sve je PCDATA, osim onoga što je CDATA