Душан Тошић
(Математички факултет)

XML-ТЕХНОЛОГИЈЕ И ДИГИТАЛИЗАЦИЈА

Од када је оформљена група W3C (http:/www.w3c.org) за формирање и предлагање стандарда за представљање података на Интернету, урађено је много. Најпре је описан мета-језик XML, а онда је уследила појава читавог низа технологија заснованих на њему (XSLT, DOM, DDT, XHTML, XSL-FO, XML-Shema, SVG, ...). Ако се планира трајније представљање података у електронској форми, онда се не могу заобићи XML-технологије. Све XML-технологије су занимљиве и пружају велике могућности у опису националног културног наслеђа. У раду ће бити описана улога XML-а и представљене неке XML-технологије.

1. Увод

XML је конципиран с идејом да омогући пуну искоришћеност и међуоперативност World Wide Web-а. (видети [4]) Када се родила идеја о његовом креирању, вероватно ни највећи оптимисти нису могли предвидети колики утицај ће овај производ имати на даљи развој информатике и рачунарства. XML (eXtensible Markup Language) је метајезик, који служи за опис других језика, а чија намена може бити различита. Језици који су направљени помоћу XML-а (у складу с правилима у XML-у), називају се XML-апликације. За врло кратак период развијени су језици за приказ мултимедијских садражаја, за опис презентација на Инетрнету, за приказ математичких садржаја итд, а који су засновани на XML-у. Али не само то, развијени су разни софтверски алати за примену ових језика, описана правила за примену и направљена прилагођавања постојећим информатичким производима. И док се у почетку могло говорити само о XML-технологији, након разграњавања примене, може се говорити о читавом низу XML-технологија. Велики произвођачи софтвера (Microsoft, IBM, Sun, ...), врло брзо су увидели значај XML-технологија па су главне своје производе (Dot-net, Системи за управљање базама података, вишеслојне апликације, ...) засновали на XML-у. Ова подршка је веома значајна јер доприноси још већој популарности XML-технологија. Ако се ово има у виду, свако планирање дигитализације података требало би да се заснива на XML-технологијама. Избор ових технологија је гаранција да ће подаци убудуће моћи лако да се обрађују, трансформишу и користе.

2. Стандарди на WWW и W3C

Назив W3C је скраћеница од World Wide Web Consortium. То је организација оформљена 1994. године са задатком да: "World Wide Web води у смеру пуног искоришћења његових могућности развојем општих протокола који промовишу стално ширење ове мреже и обезбеђују њену међуоперативност" ([1]).

Почетком деведесетих година двадесетог века дошло је до наглог развоја Web-а. Основни језик за опис података на Web-у био је HTML. Подаци описани помоћу HTML-а могли су се лако приказати, али када би требало на било који начин оперисати с таквим подацима, јављале су се велике тешкоће. Да би се Web учинио динамичким, интерактивним и приступачним за податке, прављени су разни покушаји допуне и измене HTML-а. То је довело до велике шароликости, самим тим и немогућности примене на различитим платформама, тешком овладавању разним техникама итд. Осим тога, различити произвођачи браузера (пре свих Microsoft и Netscape) су на различите начине допуњавали и мењали језик HTML не поштујући много стандарде. Јавила се потреба за креирањем стандарда у овој области. Да би стандард био прихваћен од широког круга корисника и да би могао да се одржи, мора бити солидно заснован. Дакле, било је потребно направити језик помоћу којег би се лако представљали подаци на Web-у и из којег би лако могли да се преузимају, тј. с којима би се лако могло оперисати. Због тога је оформљен конзорцијум W3C и као један од првих производа овог стручног тела настао је мета-језик XML. Већ у процесу израде XML-а схаватило се да су проблеми с којим се конзроцијум суочио веома озбиљни и да се простиру на разне области рачунарства. Чланови W3C су се ухватили у коштац с овим проблемима и као резултат тога настале су разне технологије, а које су све, на одређен начин, повезане са XML-ом.

W3C није државна организација, нити плаћена међународна организација за прављење стандарда. Стога њени производи (документи) нису званични стандарди, већ препоруке. Без обзира што се препоруке не називају стандардима, то су незванични стандарди на Web-у, а корисницима се препушта да одлуче да ли ће поштовати ове препоруке.

У креирању W3C учествовале су најпознатије софтверске куће (Microsoft, Adobe, Netscape, Sun, ...) и на изради препорука учествовали су (и сада учествују) веома познати научници и стваоци у области рачунарства. Члан корзорцијума може постати било која фирма (или појединац) вољна да се укључи у израду препорука. То значи да је реч о једној отвореној организацији у чијем раду је заступљена потпуна транспарентност. Све што је W3C до сада урадио може се "скинути" с Web-а и све је бесплатно. Такође, на сајтовима W3C (видети [1] и [2]) може се видети шта је следеће на чему се ради и какви су планови за даље.

W3C је интернационални конзорцијум смештен на три континента. Седишта конзорцијума налазе се у Северној Америци, Француској и Јапану. С обзиром на наведене податке, може се рећи да је рад W3C добро осмишљен. Препоруке не представљају само незваничне стандарде, већ и корисна документа из којих се може доста научити о Интернету и XML-технологијама.

3. Зашто је XML постигао велику популарност?

Препорука за XML издата је 1998. године (видети [3]) и ту је описана верзија 1.0, која је још уивек актуелна. XML је настао из SGML - општег и комплексног језика за обележавање (маркирање) података. Сам SGML (Standard Generalized Markup Language) настао је 1986. стандардизацијом језика GML (Generalized Markup Language), који је направљен шездесетих година двадесетог века, а његови творци били су: Charls Goldfarb, Ed Mocher и Rau Lorie. XML је заснован на истим принципима као и SGML, али је знатно једноставнији и прилагођен је Web-у. Напоменимо да је и HTML настао из SGML, али уз знатно већа упрошћавања и модификације.

За XML се може рећи да је добро осмишљен (мета)језик. То је и разлог што је покренуо бујицу нових технологија. Број правила за креирање конструкција XML-а није велики и на основу тога се може закључити да је то једноставан језик. Међутим, садржаји укључени у XML (пратећи елементи језика), који су производи технологија повезаних с XML-ом, чине га комплексним. Ту се крије и главна опасност за даље примене и експанзију XML-а. Наиме, искуство у развоју информатике је показало да све што је претеренао комплексно (поготову ако је тешко за учење), бива напуштено, тј. после одређеног периода престаје да буде актуелно. Ако је сам XML довољно једноставан, поставља се питање да ли су неопходне пратеће технологије које га чине комплексним. Без пратећих технологија XML не би био довољно изражајан па опет губи своју главну намену. (На пример, у XML-у су врло ограничене могућности за рад са сликама и графиком. Међутим, ако се користи скалабилна векторска графика (нова технологија заснована на XML-у!), добија се једно од најмоћнијих средстава за подршку графици.)

XML је добро структуриран и у њему лако може да се оперише објектима. Подаци у XML-у се представљају у облику дрвоидне структуре, при чему сваки чвор у дрвету може да се третира као посебан објекат. Издвајањем чвора са свим гранама које полазе из њега, практично се издваја један објекат који се може на исти начин третирати као и објекти у објектно-оријентисаним језицима. На тај начин подаци описани помоћу XML-а могу се релативно једноставно обрађивати у објектно-оријентисаним језицима. Увођењем појма ентитета, омогућава се третман објеката било које врсте (графичких, табела, линкова, ...) у XML-у.

Пошто постоје правила за опис података у XML-у, постоје и средства (SAX, DOM, ...) помоћу којих се може проверити да ли су поштована правила XML-а за опис података. Ако су та правила задовољена у неком документу, онда се каже да је то *добро-формиран* XML-документ. Да би могле да се примењују XML-технологије, документи с којима се оперише морају бити добро-формирани. Међутим, коришћењем додатних технологија (DTD, XML-Scheme, ...), у XML-у постоји могућност провере валидности описаних података. Шта се овим постиже? Кроз валидацију података, врше се додатне провере података у погледу испуњености појединих захтева. Тиме се знатно смањује могућност да подаци с којима се оперише буду неисправни. Самим тим се и елиминишу грешке приликом коришћења тих података.

У XML акценат је на опису података. Уопше, XML и пратеће технологије карактерише дескриптивност. Преко прецизног описа и валидације података, смањује се могућност примене процедуралних алата. Тиме се постиже олакшање процеса обраде података и смањење грешака приликом обраде.

XML није везан ни за једну конкретну платформу. Подаци описани у XML-у су независни од платформе на којој се користе. У данашње време, када постоји велики број различитих типова рачунара (платформи) на Интернету, ова особина је јако пожељна. За много кориснике велики проблем је немогућност коришћења података на некој платформи зато што су ти подаци прилагођени другој платформи. Са подацима описаним у XML-у, овакви проблеми се скоро потпуно елиминишу.

Дакле, може се закључити да XML краси низ лепих особина. То је и разлог што је за кратко време (са пратећим технологијама) постигао велику популарност. Због комплексности пратећих технологија, постоји опасаност (према оцени овог аутора!) да велики број корисника не може лако да усвоји и прихвати XML. Међутим, аутори целог концепта смтрају да веома комплесни проблеми (који су захваћени XML-ом) не могу да се решавају без коришћења комплексних средстава. За опис и коришћење података описаних помоћу XML не морају се познавати све пратеће технологије, што је олакшавајућа околност.

4. XML-технологије

Развиј XML-а иницирао развој низа нових технологија. Неке од ових технологија су у почетној фази развоја, док су друге скоро потпуно засноване. Сматраћемо да је XML-технологија заснована ако је за њу издата одоговарајућа препорука. Овде ћемо укратко описати неке XML-технологије уз навођење њихових скраћеница.

Common Markup for Micropayment је технологија везана за е-послове и поједностављења плаћања на Web-у. **CSS (Cascading Style Sheets)** је алат који се користи за представљање података, изражених преко HTML-а и XML-а, на Web-у. Строго говорећи, ово није технологија XML-а, већ средство преузето из HTML-а.

DOM (Document Object Model) омогућава обраду докумената описаних помоћу XML-а. **DTD (Document Type Definition)** намеће одређена ограничења на елементе XML-а (који тагови су дозвољени, из ког скупа се узимају вредности, ...). **P3P (Platform for Personal Privacy Preferences)** је технологија која омогућава да се заобиђу сајтови са нежељеним садржјем. **PICS (Platform for Internet Content Selection)** омогућава да лабеле (метаподаци) буду прикључене Интернет садржајима. Користе се за идентификацију материјала погодних за децу. **RDF (Resource Description Framework)** се односи на оперисање са метаподацима. Сматра се да РДФ-технологија треба да допринесе бољем разумевању података са Web-а и олакша њихову обраду. **SVG (Scalable Vector Graphics)** је релативно нова технологија и односи се на векторску графику и анимацију. **SMIL (Synchronized Multimedia Integrated Language)** је, као што и назив казује, везана за обраду мултимедијских докумената. **XForms** је технологија у развоју. Треба да обезбеди једноставан начин за рад с формама на Web-у. **XHTML (Extensible Hypertext Markup Language)** јесте

језик који представља XML-апликацију и служи за опис података на Web-у. **XIS or Infoset (XML Information Set)** служи за налажење података из добро формираних XML-докумената. **XML Query** је технологија која треба да омогући формирања упита за базе података. Ови упити треба да буду слични упитима у SQL-у. **XML Schema** треба да омогући превазилажење недостатака DTD-а и омогући прецизнији опис података, односно типова података. **XML Signatures** треба да послужи за опис синтаксе докумената која би омогућила да се лако реферише било који **URI (Uniform Resource Indicator)** на Web-у. **XSL-FO (Extensible Stylesheet Language – Formatting Objects)** има сличну улогу као и CSS, тј треба да омпгући приказивање XML-документа на погодан начин. За разлику од CSS-а, ово је потпуно XML-технологија. **XSLT (Extensible Stylesheet Language Transformations)** је језик који омогућава трансформацију XML-документа у неку другу форму (на пример, у XHTML-документ).

Детаљнији опис XML-технологија може се наћи у [5].

5. Репрезентација и презентација података помоћу XML-а.

Овде нећемо описивати XML и начин записа докумената у XML-у. Међутим, описаћемо како се врши трансформација XML-документа да би на адекватан начин био приказан на Web-у. Тиме ћемо демонстрирати коришћење неких XML-технологија. Нека имамо следећи документ описан у XML (то је позната песма Алексе Шантића "Не веруј ..."):

```
<?xml version="1.0" encoding="UTF-8" ?>
- <Pesma>
    <naslov>Ne veruj ...</naslov>
    <autor>Aleksa Šantić</autor>
    <datum>1905</datum>
  - <strofa>
      <red>Ne veruj u moje stihove i rime</red>
      <red>Kad ti kažu, draga, da te silno volim,</red>
      <red>U trenutku svakom da se za te molim</red>
      <red>I da ti u stabla urezujem ime ...</red>
    </strofa>
  - <strofa>
      <red>Ne veruj! No kasno, kad se mesec javi</red>
      <red>I prelije srmom vrh modrijeh krša,</red>
      <red>Tamo gde u grmu proleće leprša</red>
      <red>I gde slatko spava naš jorgovan plavi.</red>
    </strofa>
  - <strofa>
      <red>Dođi, čekaću te! U časima tijim,</red>
      <red>kad na grudi moje priljubiš se čvršće,</red>
      <red>Osjetiš li, draga, da mi t'jelo dršće,</red>
      <red>I da silno gorim ognjevima svijem,</red>
    </strofa>
  - <strofa>
      <red>Tada veruj meni, i ne pitaj više!</red>
```

```
        <red>Jer istinska ljubav za riječi ne zna;</red>
        <red>Ona samo plamti, silna, neoprezna,</red>
        <red>Niti mari, draga, da stihove piše!</red>
    </strofa>
</Pesma>
```

Приликом приказа помоћу браузера, овај документ се добија у непрегледној форми јер се приказују и тагови. Како се може приказати документ у прегледнијој форми? Један од начина је да се трансформише у XHTML-документ. Трансформација се може урадити коришћењем XSLT-језика. Опис трансформације се наводи у посебној датотеци са наставком XSL. Нека је то датотека **Pesma.xsl** са следећим садржајем:

```
<?xml version="1.0" ?>
- <xsl:stylesheet xmlns:xsl="http://njnjnj.nj3.org/1999/XSL/Transform" version="1.0">
  - <xsl:template match="Pesma">
    - <html>
      - <head>
        - <title>
            <xsl:value-of select="naslov" />
          </title>
        </head>
      - <body>
          <xsl:apply-templates select="naslov" />
          <xsl:apply-templates select="strofa" />
          <xsl:apply-templates select="autor" />
          <xsl:apply-templates select="datum" />
        </body>
      </html>
    </xsl:template>
  - <xsl:template match="naslov">
    - <div align="center">
      - <h1>
          <xsl:value-of select="." />
        </h1>
      </div>
    </xsl:template>
  - <xsl:template match="autor">
    - <div align="left">
      - <h2>
          <xsl:value-of select="." />
        </h2>
      </div>
    </xsl:template>
  - <xsl:template match="datum">
    - <div align="left">
      - <h2>
          <xsl:value-of select="." />
        </h2>
      </div>
    </xsl:template>
  - <xsl:template match="strofa">
    - <p>
```

```
            <xsl:apply-templates select="red" />
        </p>
    </xsl:template>
  - <xsl:template match="red">
        <xsl:value-of select="." />
        <br />
    </xsl:template>
</xsl:stylesheet>
```

Да би се добио документ у прегледнијем облику, потребан је процесор (програм) који обавља трансформацију XML-документа у XHTML-документ. За те сврхе може послужити **процесор Saxon** (видети [6]). Процес превођења се реализује командом (из DOS-а):

>saxon Pesma.xml Pesma.xsl > Pesma.html

и сада је формиран XHTML-документ Pesma.html који се приказује на следећи начин:

Не веруј ...

Не веруј у моје стихове и риме
Кад ти кажу, драга, да те силно волим,
У тренутку сваком да се за те молим
И да ти у стабла урезујем име ...

Не веруј! Но касно, кад се месец јави
И прелије срмом врх модријех крша,
Тамо где у грму пролеће лепрша
И где слатко спава наш јоргован плави.

Дођи, чекаћу те! У часима тијим,
кад на груди моје приљубиш се чвршће,
Осјетиш ли, драга, да ми т'јело дршће,
И да силно горим огњевима свијем,

Тада веруј мени, и не питај више!
Јер истинска љубав за ријечи не зна;
Она само пламти, силна, неопрезна,
Нити мари, драга, да стихове пише!

Алекса Шантић 1905

Овако приказан документ је знатно прегледнији.

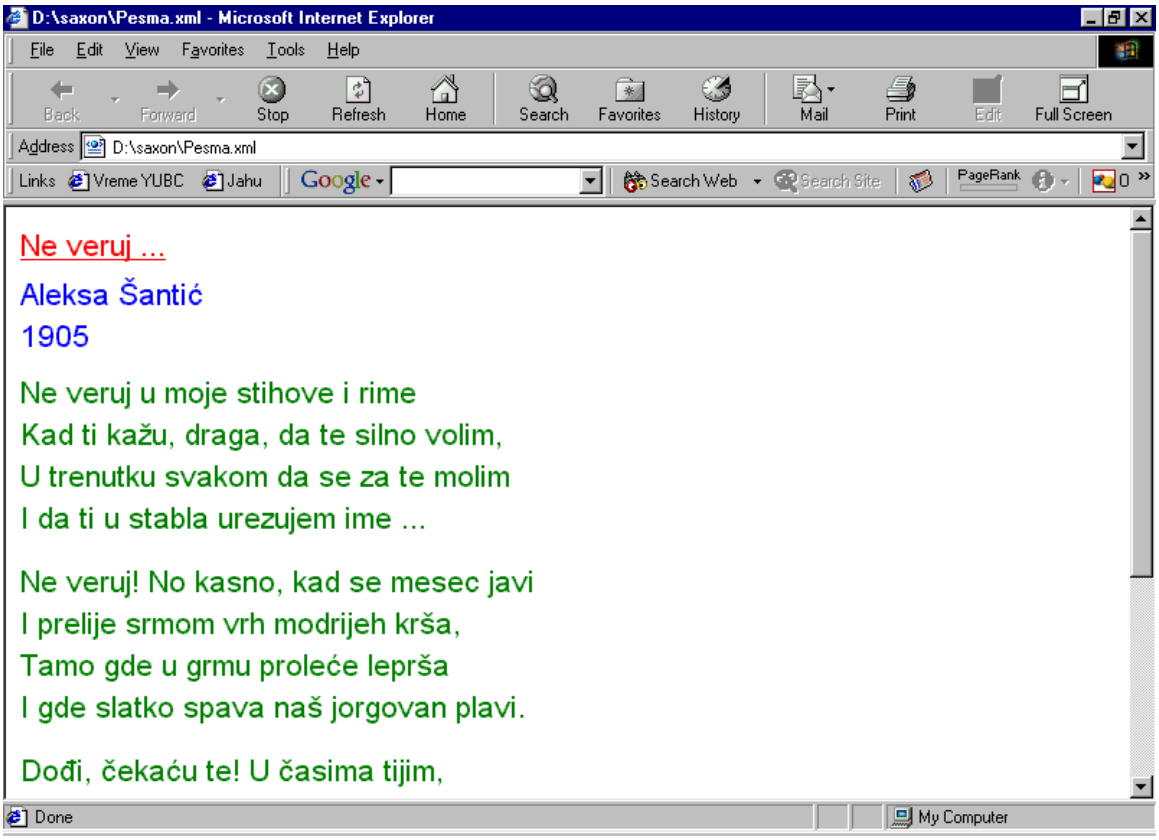
Постоје и други начини за прегледнији приказ XML-документа. На пример, ако искористимо CSS-технологију, потребно је да формирамо датотеку са наставком css у којој се налази опис сваког елемента из XML-документа. Нека је то датотека Pesma.css са следећим садржајем:

```
naslov{
  display: block;
  margin-bottom: 10px;
  font-family: Arial, Helvetica;
  color: red;
  text-decoration: underline
  }
autor{
   display: block;
   margin-bottom: 5px;
   font-family: Arial, Helvetica;
   color: blue
   }
datum{
   display: block;
   margin-bottom: 15px;
   font-family: Arial, Helvetica;
   color: blue
   }
red{
   display: block;
   margin-bottom: 5px;
   color:green;
   }
strofa{
   display: block;
   margin-bottom: 20px;
   ;}
```

Ако имамо овако формирану датотеку, треба извршити измену у оригиналном XML-документу додајући један ред (иза првог реда) облика:

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

Прегледањем документа **Pesma.xml** помоћу браузера, добијамо излаз облика:



6. Коришћење наших слова у XML-документима

XML је заснован на Уникоду. То значи да се у документима могу користити разни знаци, само је потребно знати њихове кодове. Уникодне вредности, за разне скупове знакова (па и за ћирилицу), могу се наћи на Web-у (видети [7]). У претходном примеру, почетак текстовне верзије XML-документа изгледа овако:

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

<Pesma>
<naslov>Ne veruj ...</naslov>
<autor>Aleksa  &#352;anti&#263;</autor>
<datum>1905</datum>

<strofa>
<red>Ne veruj u moje stihove i rime </red>
<red>Kad ti ka&#382;u, draga, da te silno volim, </red>
.....
```

Запажа се да су за слова Ш, ћ, и ж коришћени уникодови. (Ови кодови почињу симболима &#, а завршавају се тачком-запетом (;). Уколико су хесадецималном облику почињу знацима &#x.)

Ако желимо да користимо ћирилицу у документу, приципи обраде остају исти, једино се мењају кодови за слова у документу. Нека почетак нашег документа (у текстовном облику) изгледа овако:

```
<?xml version='1.0'  encoding='UTF-8'?>
<Pesma>
<naslov>&#x41D;&#x435;  &#x432;&#x435;&#x440;&#x443;&#x458; ...</naslov>
<autor>&#x410;&#x43B;&#x435;&#x43A;&#x441;&#x430;
&#x428;&#x430;&#x43D;&#x442;&#x438;&#x45b;</autor>
<datum>1905</datum>
<strofa>
<red>&#x41D;&#x435;  &#x432;&#x435;&#x440;&#x443;&#x458; &#x443;
&#x43C;&#x43E;&#x458;&#x435;
 &#x441;&#x442;&#x438;&#x445;&#x43e;&#x432;&#x435; &#x438;
&#x440;&#x438;&#x43c;&#x435;</red>
<red>Ka&#x434; &#x442;&#x438; &#x43a;a&#x436;&#x443;,,
&#x434;&#x440;a&#x433;a, &#x434;a
&#x442;&#x435; &#x441;&#x438;&#x43b;&#x43d;&#x43e;
&#x432;o&#x43b;&#x438;&#x43c;, </red>
<red>.....</red>
<red>.....</red>
</strofa>
</Pesma>
```

У овом документу су коришћени хексадекадни уникодови за ћирилична слова. Овде се појављује проблем уноса уникодова. Данас постоје програми помоћу којих се откуцан документ латиничним словима може искодирати ћириличним уникодовима (такав програм се лако може направити у било ком процедуралном програмском језику). Сада приказ XML-докумета (помоћу браузера) изгледа овако:

```
<?xml version="1.0" encoding="UTF-8" ?>
- <Pesma>
    <naslov>Ne veruj ...</naslov>
    <autor>Aleksa Šantić</autor>
    <datum>1905</datum>
  - <strofa>
      <red>Ne veruj u moje stihove i rime</red>
      <red>Kad ti kažu, draga, da te silno volim,</red>
      <red>.....</red>
      <red>.....</red>
    </strofa>
  </Pesma>
```

Ако се изврши превођење помоћу процесора Saxon, добија се у облику:

Service started.

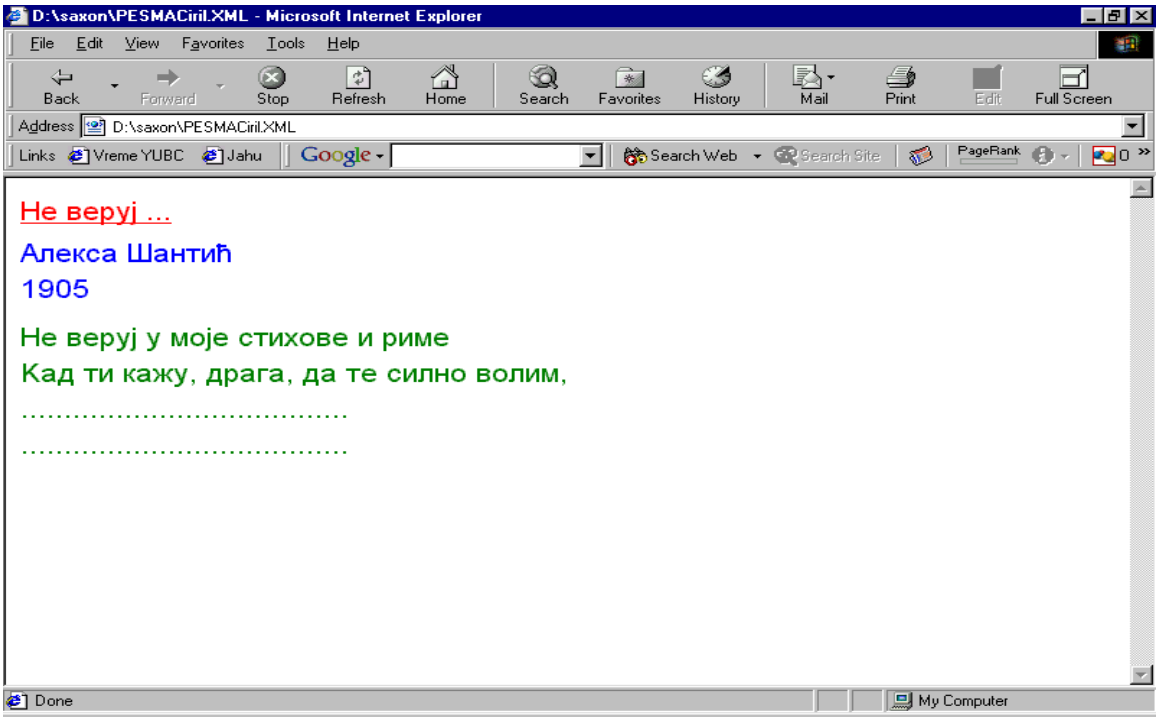
Не верuj ...

Не верuj у моје стихове и риме
Кад ти кажу, драга, да те силно волим,
.....
.....

Алекса Шантић 1905

Service started.

Ако извшимо измену у оригиналном текстовном документу додајући наредбу:
<?xml-stylesheet type="text/css" href="pesma.css" ?>,
онда приказ овог документа, помоћу браузера, изгледа овако:



Из наведених примера види се да принципи обраде нису мењани приликом рада с ћирилицом.

7. Закључак

У претходном одељку укратко је описано како се користе неке XML-технологије (XHTML, XSLT, CSS). И из наведених примера може се уочити да примена XML-технологија захтева познавање посебних језика и техника. Међутим, ове технологије пружају могућност трансформације података на разне начине. То је главна предност XML-технологија.

Имајући у виду претходно речено, можемо закључити да дигитализација података (представљање података у електронском облику), треба да буде заснована на XML-у и пратећим технологијама. То је нека врста гаранције да ће будуће генерација лако приступати тим подацима. Уједно, обрада тако представљених података биће максимално поједностављена.

Литература и линкови

[1] **http://www.w3c.org**
[2] **http://www.w3.org**

[3] http://www.w3.org/TR/1998/REC-xml-19980210+
[4] Hunter D.: *Od početka ...XML*, CET,Beograd 2001.

[5] Ladd E, O'Donnell J, Morgan M. and Watt A.H.: *Using XHTML, XML and Java 2* , Que, 2000.

[6] **http://users.iclway.co.uk/mhkay/saxon/**

[7] **http://www.unicode.org/charts/**

dtosic@matf.bg.ac.yu

Abstract

Since the group W3C (http:/www.w3c.org) is formed, for the purpose of working on making and proposing standards for representation the data for Internet, it is done a lot. First of all, the meta-language XML is described and after that the whole line of new technologies (XSLT, DOM, DDT, XHTML, XSL-FO, XML-Shema, SVG, ...), based on XML, is followed. If the representation of data in electronic form is planed for a durable period, the XML-technologies could not be avoided. All XML technologies are interesting and provide the great possibilities in describing national cultural inheritance. In this paper the role of XML is described and some XML technologies are presented.