Informatički praktikum 4

Letnji semestar 2016/2017

Obaveze na predmetu

- 1 test, 15p (u maju)
- 2 seminarska rada
  - BibTeX, 10p
  - XML, 20p
- Prisustvo, 5p
  - dozvoljeno je odsustvovati sa najviše 3 časa
- Pismeni ispit, 50p
  - uslov da bi se ispit položio je bar 20p
- Nema praktikuma

Program predmeta

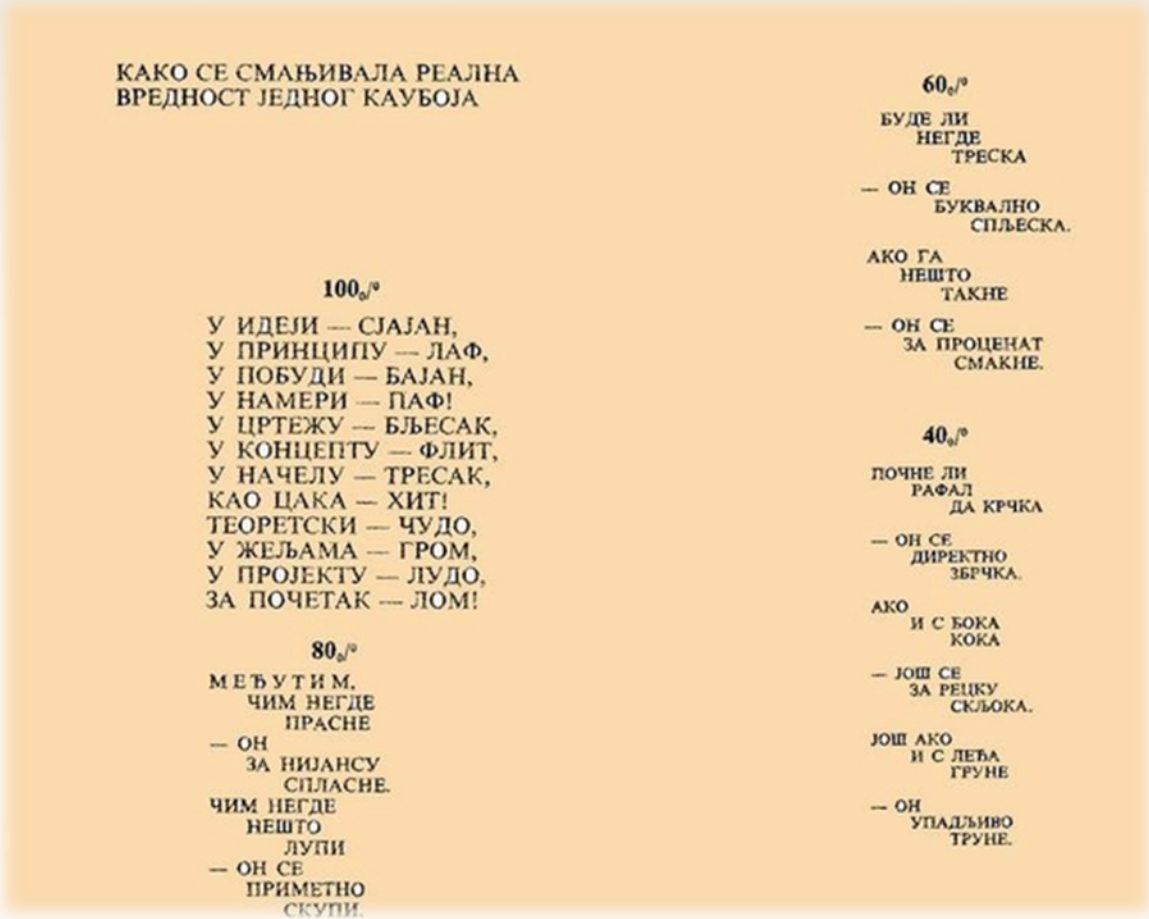
- XML
  - Istorijat
  - Osnovni pojmovi
  - DTD
  - XML Shema
  - Regularni izrazi
  - JSON
  - Primene
- BibTeX

1. Obeležavanje teksta i istorijat XML-a

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/

Kako predstaviti u računaru?



КАКО СЕ СМАЊИВАЛА РЕАЛНА ВРЕДНОСТ ЈЕДНОГ КАУБОЈА

100%

У ИДЕЈИ — СЈАЈАН,
У ПРИНЦИПУ — ЛАФ,
У ПОБУДИ — БАЈАН,
У НАМЕРИ — ПАФ!
У ЦРТЕЖУ — БЉЕСАК,
У КОНЦЕПТУ — ФЛИТ,
У НАЧЕЛУ — ТРЕСАК,
КАО ЦАКА — ХИТ!
ТЕОРЕТСКИ — ЧУДО,
У ЖЕЉАМА — ГРОМ,
У ПРОЈЕКТУ — ЛУДО,
ЗА ПОЧЕТАК — ЛОМ!

80%

МЕЂУТИМ,
ЧИМ НЕГДЕ
ПРАСНЕ
— ОН
ЗА НИЈАНСУ
СПЛАСНЕ.
ЧИМ НЕГДЕ
НЕШТО
ЛУПИ
— ОН СЕ
ПРИМЕТНО
СКУПИ.

60%

БУДЕ ЛИ
НЕГДЕ
ТРЕСКА
— ОН СЕ
БУКВАЛНО
СПЉЕСКА.

АКО ГА
НЕШТО
ТАКНЕ
— ОН СЕ
ЗА ПРОЦЕНАТ
СМАКНЕ.

40%

ПОЧНЕ ЛИ
РАФАЛ
ДА КРЧКА
— ОН СЕ
ДИРЕКТНО
ЗБРЧКА.

АКО
И С БОКА
КОКА
— ЈОШ СЕ
ЗА РЕЦКУ
СКЉОКА.

ЈОШ АКО
И С ЛЕЂА
ГРУНЕ
— ОН
УПАДЉИВО
ТРУНЕ.

Samo tekst (gubitak raznih informacija)

KAKO SE SMANJIVALA REALNA VREDNOST JEDNOG KAUBOJA 100% U IDEJI - SJAJAN, U PRINCIPU - LAF, U POBUDI - BAJAN, U NAMERI - PAF! U CRTEŽU - BLJESAK, U KONCEPTU - FLIT, U NAČELU - TRESAK, KAO CAKA - HIT! TEORETSKI - ČUDO, U ŽELJAMA - GROM, U PROJEKTU - LUDO, ZA POČETAK - LOM! 80% MEĐUTIM, ČIM NEGDE PRASNE - ON ZA NIJANSU SPLASNE. ČIM NEGDE NEŠTO LUPI - ON SE PRIMETNO SKUPI. 60% BUDE LI NEGDE TRESKA - ON SE BUKVALNO SPLJESKA. AKO GA NEŠTO TAKNE - ON SE ZA PROCENAT SMAKNE. 40% POČNE LI RAFAL DA KRČKA - ON SE DIREKTNO ZBRČKA. AKO I S BOKA KOKA - JOŠ SE ZA RECKU SKLJOKA. JOŠ AKO I S LEĐA GRUNE - ON UPADLJIVO TRUNE. 20% DESI SE, STRELA VRISNE

Logička i grafička struktura teksta [Text Encoding Initiative]

```
<TeiHeader>
<titleStmt>
  <title>Utorak</title><author>Duško Radović</author>
<language id='SCO' wsd='ISO-8859-2' alpha='Cyrillic'>
.....
<div2 type='ciklus'>
  <head lang='EN'>Western</head>
<div3 type='pesma'>
  <head>KAKO SE SMANJIVALA REALNA &#RS;&#RE;VREDNOST JEDNOG KAUBOJA</head>
<head rend='bold 14pt'>100%</head>
<lg1 type='strofa' rend='roman 14pt'>
<l>U IDEJI — SJAJAN,</l>
<l>U PRINCIPU — LAF,</l>
<l>U POBUDI — BAJAN,</l>
<l>U NAMERI — PAF!</l>
<l>U CRTEŽU — BLJESAK,</l>
<l>U KONCEPTU — FLIT,</l>
<l>U NAČELU — TRESAK,</l>
<l>KAO CAKA — HIT!</l>
<l>TEORETSKI — ČUDO,</l>
<l>U ŽELJAMA — GROM,</l>
<l>U PROJEKTU — LUDO,</l>
<l>ZA POČETAK — LOM!</l>
</lg1>
<head rend='bold 12pt'>80%</head>
<lg1 type='strofa' rend='roman 12pt'>
<lg2>
<l>MEĐUTIM,<l>
<l rend='1ind'>ČIM NEGDE</l>
<l rend='2ind'>PRASNE</l></lg2>
<lg2>
<l>— ON</l>
<l rend='1ind'>ZA NIJANSU<l>
```

Obeležavanje teksta (1)

- Uloge:
  - Elementi logičke strukture
  - Funkcije nad elementima
  - Osobine elemenata
  - Različite vrste sadržaja
  - Različite vrste alfabeta
  - Sadržaj iz drugog toka
  - Hipertekstualne veze ...
- Razni formati za obeležavanje teksta:
  - TeX, Word, RTF, HTML, XML, JSON ...

Obeležavanje teksta (2)

- Ukoliko dve strane koriste drugačiji način predstavljanja podataka, mogu da:
  - Usaglase format (nepraktično)
  - Koriste programe za prevođenje različitih formata
    - HTML to LaTeX, XML to JSON, JSON to XML, Word to LateX itd...
    - nije baš uvek dobro jer može doći do gubitka informacija
- Važnost standardizacije mašinski čitljivih dokumenata:
  - prenosivost
  - višestruko korišćenje
  - dugovečnost

Šta je SGML? (1)

- **Standard Generalized Markup Language**
  - Standardni generalizovani jezik za obeležavanje
- Kasnih šezdesetih godina, komunikacija između računara odvijala se razmenom puno različitih formata za obeležavanje, što je bilo dosta problematično
- Zaposleni IBM-a smišljaju pravilo kojim bi se logički opisali podaci, tako da to bude nezavisno od operativnog sistema ili hardverskih komponenti
- Obrada jezika za označavanje podataka svodi se na obradu teksta, što omogućuje sistemsku i hardversku nezavisnost
- SGML **nije jezik**, već sistem za definisanje jezika za označavanje

Šta je SGML? (2)

- Svaki jezik definisan SGML-om naziva se SGML aplikacija
- SGML aplikacije se karakterišu:
  - skupom karaktera i delimitera koji se mogu javiti u jeziku
  - definicijom sintakse elemenata za označavanje
  - opisom semantike sadržaja, u vidu sintaksičkih restrikcija koje nisu mogle biti izražene defincijom sintakse
  - same dokumente koji sadrže podatke i oznake
    - svaki dokument sadrži referencu na svoju definiciju sintakse

Najvažnija svojstva SGML-a

- Apstraktna sintaksa
- Konkretna sintaksa
- Deklaracije oznaka
- Proizvoljnost sadržaja podataka
- Upotreba jedinstvenih idenfitikatora
- Uključivanje i isključivanje delova podataka
- Opciona svojstva jezika

Međutim...

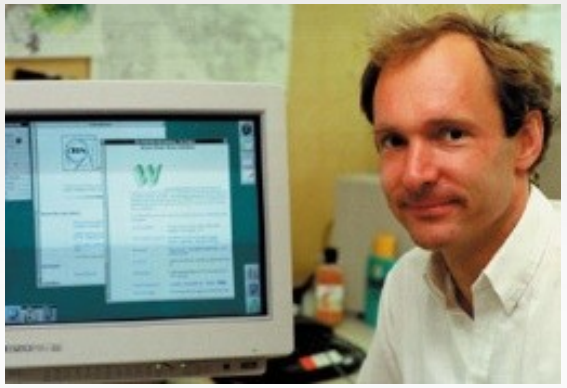
- Iako bi formati bili ukalupljeni, i dalje je svaka organizacija mogla imati svoj format za predstavljanje podataka
- Samo ukalupljivanje (opisivanje sopstvenog formata prema pravilima SGML-a) nije bilo jednostavno:
  - specifikacija na 155 stranica
  - 203 pravila kontekstno-slobodne gramatike
  - složena izrada parsera za takav jezik
- Bilo je i uspešnih primena:
  - Američko ministarstvo odbrane, Američko udruženje izdavača, Međunarodni zavod za standardizaciju, Kancelarija za zvanične publikacije EU ...

Deo SGML deklaracije za HTML

```
<!SGML  "ISO 8879:1986 (WWW)"
  --
     SGML Declaration for HyperText Markup Language version HTML 4

     With support for the first 17 planes of ISO 10646 and
     increased limits for tag and literal lengths etc.
  --

  CHARSET
      BASESET  "ISO Registration Number 177//CHARSET
            ISO/IEC 10646-1:1993 UCS-4 with
            implementation level 3//ESC 2/5 2/15 4/6"
      DESCSET 0      9      UNUSED
            9      2      9
            11     2      UNUSED
            13     1      13
            14     18     UNUSED
            32     95     32
            127    1      UNUSED
            128    32     UNUSED
            160    55136  160
            55296  2048   UNUSED  -- SURROGATES --
            57344  1056768 57344

CAPACITY       SGMLREF
           TOTALCAP       150000
           GRPCAP         150000
           ENTCAP         150000

SCOPE    DOCUMENT
SYNTAX
      SHUNCHAR CONTROLS 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
       17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 127
      BASESET  "ISO 646IRV:1991//CHARSET
            International Reference Version
            (IRV)//ESC 2/8 4/2"
      DESCSET  0 128 0

      FUNCTION
            RE          13
            RS          10
            SPACE       32
            TAB SEPCHAR    9

      NAMING   LCNMSTRT ""
            UCNMSTRT ""
            LCNMCHAR ".-_:"
            UCNMCHAR ".-_:"
            NAMECASE GENERAL YES
                  ENTITY  NO
      DELIM    GENERAL  SGMLREF
            HCRO "&#38;#x" -- 38 is the number for ampersand --
            SHORTREF SGMLREF
      NAMES    SGMLREF
      QUANTITY SGMLREF
            ATTCNT   60     -- increased --
            ATTSPLEN 65536  -- These are the largest values --
            LITLEN   65536  -- permitted in the declaration --
            NAMELEN  65536  -- Avoid fixed limits in actual --
            PILEN    65536  -- implementations of HTML UA's --
            TAGLVL   100
            TAGLEN   65536
            GRPGTCNT 150
            GRPCNT   64

FEATURES
  MINIMIZE
    DATATAG  NO
    OMITTAG  YES
    RANK     NO
    SHORTTAG YES
  LINK
    SIMPLE   NO
    IMPLICIT NO
    EXPLICIT NO
  OTHER
    CONCUR   NO
    SUBDOC   NO
    FORMAL   YES
  APPINFO NONE
>
```

Šta je HTML?

- **Hypertext Markup Language**
  - Hipertekstualni jezik obeležavanja
  - http://www.w3schools.com/html/
- Jedna jednostavna SGML aplikacija
- Samo jedna u nizu ideja CERN-ovog zaposlenog, Tim Berners-Lee:
  - World Wide Web (1989)
  - Specifikacije za URL, HTTP, HTML (1991)
  - Osnovao W3C pri MIT-u (Massachusetts Institute of Technology) (1994)



Osnovne ideje World Wide Web-a

- WWW omogućava korisnicima da pronađu, pregledaju i pretražuju multimedijalne dokumente (tekstualne, grafičke, animacije, audio i video zapise) bilo kog sadržaja, na način nezavisan od tipa dokumenta
- To je tehnologija deljenja podataka upotrebom tekstualnih dokumenata koji sadrže hiper-linkove
  - WWW nije isto što i internet!
- Klijent-server model kao osnovna arhitektura
  - Korisnik je „klijent“, a podaci se nalaze na računaru koji ima ulogu „servera“
- Usklađivanje formata

Nastanak XML-a

- 1996. godine odvijaju se aktivnosti za pojednostavljenje SGML-a
- Ciljevi su da se zadrže osnovni koncepti, a uklone:
  - duplirana svojstva
  - svojstva koja se teško implementiraju
  - svojstva koja zbunjuju korisnike
  - svojstva koja se empirijski nisu pokazala korisnim
- Rezultat: