

Standardi za označevanje dokumentov

Razlogi za standardizacijo označevanja dokumentov, označevalni jeziki:

MARC,
RTF,
TeX, LaTeX,
SGML.

Ponudba e-dokumentov v preteklosti

- ❖ Značilnosti ponudnikov e-dokumentov nekdam:
 - ❖ majhno število, pretežno ponudniki bibliografskih zbirk,
 - ❖ med seboj neodvisni,
 - ❖ samostojen razvoj zbirk e-dokumentov.

Ponudba e-dokumentov v preteklosti

- ❖ Značilnosti e-dokumentov nekdam:
 - ❖ majhno število v primerjavi s številom fizičnih dokumentov,
 - ❖ enostavna oblika in relativno enostavna struktura (npr. bibliografski zapisi),

Ponudba e-dokumentov v preteklosti

- ❖ Značilnosti e-dokumentov nekdam:
 - ❖ za prikaz takih e-dokumentov ni potrebna zmogljiva strojna in programska oprema,
 - ❖ dokler so bili e-dokumenti le ASCII besedila, je zadoščala skoraj vsaka strojna oprema.
 - ❖ Posamezni ponudniki so samostojno razvijali programsko opremo za ponudbo e-dokumentov.
 - ❖ Načini označevanja strukture e-dokumentov prilagojeni potrebam, možnostim posameznih ponudnikov in lastnostim njihovih e-dokumentov.

Posledice razvoja e-dokumentov

- ❖ Razvoj ponudbe e-dokumentov:
 - ❖ hitro povečevanje količine e-dokumentov,
 - ❖ razvoj zapletenejših oblik, predvsem nebesedilnih (multimedijskih) dokumentov,
 - ❖ programska in strojna oprema je morala slediti razvoju e-dokumentov.

Posledice razvoja e-dokumentov

- ❖ Posledica:
 - ❖ zaradi nestandardnih oblik e-dokumentov ti niso bili uporabni na različnih računalniških platformah,
 - ❖ ponudniki so morali izdelovati različne verzije e-dokumentov (in njihovih zbirk) za različne strojne in programske platforme
 - ❖ edini možni izhod - standardizacija.
- ❖ Zaradi ekonomskih razlogov ni bilo pričakovati poenotenja proizvajalcev računalnikov.
- ❖ Edini možni izhod:
standardizacija načinov zapisovanja e-dokumentov.

Posledice razvoja e-dokumentov

- ❖ Nastali so označevalni jeziki.
- ❖ Označevalni jeziki (*mark-up languages*) omogočajo označevanje strukturnih in/ali oblikovnih elementov e-dokumentov.
- ❖ Načeloma tako označeni e-dokumenti funkcionirajo enako na različnih strojnih in programskih platformah.

Kdaj bomo zadovoljni z e-dokumenti?

Če hočemo, da bodo informacije v elektronski obliki vsaj deloma nadomestile papir, morajo biti berljive na enak prostorsko in časovno neodvisen način.

Kdaj bomo zadovoljni z e-dokumenti?

Prostorska neodvisnost e-dokumentov:

- ❖ dokumenti morajo biti enako uporabni v poljubnih okoliščinah.

Časovna neodvisnost e-dokumentov:

- ❖ dokumenti morajo biti uporabni, dokler obstajajo bralci.

Kdaj bomo zadovoljni z e-dokumenti?

- ❖ Izpolnitev omenjenih zahtev je deloma odvisna od razvoja strojne opreme (energetsko avtonomni bralniki velikosti knjige, brezžično povezani s ponudniki e-dokumentov).
- ❖ Bralniki so postali ekonomska kategorija šele z razvojem kontrastnega in energetsko nepotratnega e-črnila.
- ❖ Brez standardizacije e-dokumentov tak razvoj ne bi bil mogoč.

Kdaj bomo zadovoljni z e-dokumenti?

- ❖ Standardi za označevanje e-dokumentov morajo torej omogočati:
 - ❖ prenosljivost med različnimi tipi računalnikov,
 - ❖ prenosljivost med različnimi programi za uporabo dokumentov,
 - ❖ enostavno in ekonomično prenosljivost po računalniških omrežjih,
 - ❖ dolgoživost dokumentov (problem e-arhiviranja).

Kdaj bomo zadovoljni z e-dokumenti?

- ❖ Uveljavi se lahko samo tak standard, ki omogoča uporabniku nove investicije v strojno in programsko opremo, ohrani pa investicije v informacije.

Zgodovina označevalnih jezikov

- ❖ Označevalni jeziki so se razvili v založništvu in tiskarstvu iz oznak, ki so na krtačnih odtisih opisovale potrebne popravke.
- ❖ Izrazno moč so dobili z uvajanjem elektronskega stavljenja besedil (navodila fotostavčnemu stroju).

Zgodovina označevalnih jezikov

- ❖ Označevalni jeziki so umetni jeziki, sestavljeni iz:
 - ❖ oznak za delitev dokumenta na strukturne dele,
 - ❖ oznak, ki opisujejo videz teh strukturnih delov, ter
 - ❖ slovnice, ki določa pravilno rabo oznak.
- ❖ Primer:
 - ❖ najosnovnejše slovnično pravilo določa, kaj je oznaka označevalnega jezika in kaj označeno besedilo.

MARC in izpeljanke

MARC (Machine Readable Catalog):

- ❖ Predvsem format, ki definira strukturo kataložnega zapisa, vendar tudi eden prvih označevalnih jezikov, ki omogoča označevanje strukturnih elementov.
- ❖ Namenjen je izmenjavi informacij med računalniškimi knjižničnimi katalogi.
- ❖ Načeloma je uspešno rešil problem prenosljivosti dokumentov med različnimi računalniškimi sistemi.

MARC in izpeljanke

- ❖ Razvile so se izpeljanke (USMARC, UKMARC, UNIMARC...), kar zmanjšuje popolno prenosljivost zapisov.



MARC: izmišljen primer

```
001 2 0055
010 $a92-2-106396-8
100 $a19890208d1988 f0ENgy103a
101 0 $aeng
200 1 $aSGML comes of age$fTony Smith
210 $aGeneve$cInternational Standards
    Organization$d1988
215 $axii, 180 p
225 2 $Documents of the Text Encoding Initiative$x0257-
3415$v6
461 1 $100120054$12001 $v6
700 1 $aSmith$bTony
```

oznake strukture

- ❖ MARC določa samo strukturo zapisa.

struktura : videz

- ❖ Če označevalni jezik določa le strukturo zapisa, je predstavitev zapisa, njegov videz na zaslonu ali papirju popolnoma prepuščena programom, s katerimi zapis uporabljamo.
- ❖ Struktura zapisa je v takem primeru ločena od definicij za tipe pisave, razmike med vrsticami ipd, s čimer se (v svetu tiskanih publikacij) ukvarjajo tiskarski označevalni jeziki.

RTF: razlogi za uvajanje

- ❖ Zelo pogosta oblika informacij, ki jih ne moremo učinkovito opisati z jeziki, kakršen je MARC, je videz in struktura polnih dokumentov, ki nastajajo v urejevalnikih besedil.
- ❖ Struktura in videz dokumenta sta neločljivo povezana - zelo majhna možnost prenosljivosti med različnimi računalniki in urejevalniki besedil.

RTF : razlogi za uvajanje

V splošnem velja:

- ❖ bolj ko je zmogljiv urejevalnik glede oblikovanja besedila, bolj zaprt sistem je.
- ❖ Tudi pri dokumentih v formatih urejevalnikov besedil obstaja močna potreba po prenosljivosti.
- ❖ Proizvajalci urejevalnikov besedil so se dogovorili za skupen transportni označevalni jezik, ki ga razume večina najpomembnejših urejevalnikov.
- ❖ Ta jezik je RTF (Rich Text Format).

RTF

Del predavanja o označevalnih jezikih, zapisan v RTF.

```
{\rtf1\ansi\deff20\deflang1033{\fonttbl{\f4\froman\charset0\prq2 Times New Roman;}{\f5\swiss\charset0\prq2 Arial;}{\f20\fnil\charset0\prq2 SLOHelvetica;}}{\colortbl;\red0\green0\blue0;\red0\green0\blue255;
```

... sledi še nekaj deset vrstic splošnega opisa tipov pisav, zamikov...

```
\pard\plain \qr\sb40\sa40\tx357 \f20\fs20\lang2057{\fs18 Bib: Računalniško komuniciranje, Informatika 2}\par \pard\plain \s18\sb40\sa40\tx357 b\f20\fs30\lang2057{\i\fs32 Predavanje: Standardi za označevanje dokumentov}\par }\pard\plain \s1\fi-360\li360\sb240\sa40\tx357\b\f20\fs28\lang2057
```

1. **tab Razlogi za standardizacijo označevanja informacij**

RTF označuje izključno videz dokumenta!

Postscript in PDF

Postscript (Adobe):

- ❖ označevalni jezik za krmiljenje laserskih tiskalnikov,
- ❖ označuje izključno videz dokumenta,
- ❖ razume tudi zapis slik.

Postscript in PDF

PDF (Portable Document Format (Adobe)):

- ❖ označevalni jezik, ki omogoča prikaz originalnega videza dokumenta na WWW in na papirju,
- ❖ poenostavljena in dopolnjena različica Postscripta,
- ❖ omogoča prikazovanje člankov iz strokovnih revij v natančno taki obliki, kot jo imajo v reviji,
- ❖ pretežno označuje videz dokumenta, le delno tudi strukturo (hipertekstni kazalci).

TeX, LaTeX

- ❖ Označevalni jezik za elektronsko stavljenje besedil, predvsem takih, ki vsebujejo veliko matematičnih izrazov.
- ❖ Na začetku namenjen krmiljenju fotostavčnih strojev Linotype.
- ❖ Avtor Donald E. Knuth s Stanfordske univerze, skupaj s študenti (10 let dela).
- ❖ V uporabi od l. 1970.

TeX, LaTeX

- ❖ Označevalni jezik, ki definira tako strukturo, kot tudi videz dokumenta, na način, ki omogoča prenosljivost.
- ❖ Zaradi izredne učinkovitosti so bili narejeni prevajalniki za vse pomembnejše modele računalnikov.
- ❖ TeX se drži vseh grafičnih standardov v tiskarski industriji in še vedno presega najsodobnejše urejevalnike besedil.

TeX, LaTeX

```
\documentstyle[12pt,slovene]{article}
\addtolength{\topmargin}{-1cm}
\addtolength{\textheight}{2.5cm}
\input yuset
\def\boxm{\framebox[13.7cm]{\rule[-75mm]{0cm}{7.5cm}}}
\def\boxv{\framebox[13.7cm]{\rule[-160mm]{0cm}{16cm}}}
\def\noparsk{\medskip \noindent}
\begin{document}
\pagebreak
\markright{\hfill {\rm Bib: Ra{\ch}unalni{\sh}ko komuniciranje,
Informatika 2}}
\title {\bf Predavanje: Standardi za ozna{\ch}evanje dokumentov}
\section {\bf 1. razlogi za standardizacijo ozna{\ch}evanja
informacij}
\setcounter{page}{1}
... sledi besedilo, označeno v LaTeXu ...
\end{document}
```

strukturne oznake

TeX, LaTeX

- ❖ TeX ni v splošni rabi predvsem, ker
 - ❖ je prezapleten,
 - ❖ le redki urejevalniki so “WYSIWYG”,
 - ❖ je namenjen pretežno označevanju specialnih besedil in ni zelo primeren za označevanje multimedijskih dokumentov.
- ❖ TeX je še vedno široko uporabljan v nekaterih naravoslovnih strokah, kjer dokumenti temeljijo na matematičnem izrazju.

SGML

SGML (Standard Generalised Markup Language):

- ❖ Standardni splošni označevalni jezik.
- ❖ Mednarodni standard, ki ga je sprejela ISO leta 1986 in je bil od takrat že nekajkrat dopolnjen.
- ❖ Gre pravzaprav za družino standardov, ki urejajo označevanje vseh znanih zvrsti elektronskih informacij.

SGML

- ❖ Prednost SGML je v njegovi splošnosti, ker sta logična struktura in videz dokumenta povsem ločena.
- ❖ Videz dokumenta prepušča programski opremi za oblikovanje in tiskanje dokumenta.
- ❖ V tej lastnosti je SGML zelo podoben MARC.

SGML

Osnovna filozofija SGML:

- ❖ obstaja končno število različnih tipov dokumentov (razredi dokumentov),
- ❖ razrede dokumentov je mogoče definirati v SGML,
- ❖ vsak dokument sodi v enega od razredov dokumentov,
- ❖ če nekega dokumenta ne moremo uvrstiti v nobenega od razredov, mora SGML omogočiti definiranje novega ustreznega razreda.

SGML

Dokument v SGML, je sestavljen iz treh delov:

- ❖ Deklaracije, ki določa
 - ❖ nabor znakov, ki ga uporablja dokument (znaki latinice, cirilice...) in
 - ❖ znake, ki imajo poseben pomen za SGML in torej niso del vsebine dokumenta.

SGML

- ❖ Definicije: (DTD, Document Type Definition)
 - ❖ definira pomen strukturnih delov dokumenta
 - ❖ opiše oznake (tage), ki označujejo strukturne dele.
- ❖ DTD definira tudi hierarhične odnose med strukturni deli dokumenta.
- ❖ Vsebine dokumenta, označene z oznakami, naštetimi v DTD.

SGML

- ❖ Primer dela DTD za tip dokumenta "memo".
- ❖ Primer vsebuje opis elementov dokumenta in njegovih lastnosti (atributov).

```
1. <!ENTITY % doctype "MEMO" --document type generic id. ->
2. <!--          ELEMENTS      MIN          CONTENT          (EXCEPTIONS) ->
3. <!ELEMENT    %doctype;    --          ((TO &FROM), BODY, CLOSE?) >
4. <!ELEMENT    TO            -O          (#PCDATA)        >
5. <!ELEMENT    FROM          -O          (#PCDATA)        >
6. <!ELEMENT    BODY          -O          (P)*             >
7. <!ELEMENT    P             -O          (#PCDATA | Q)*  >
8. <!ELEMENT    PREF          -O          EMPTY            >
9. <!ELEMENT    Q             --          (#PCDATA)        >
10. <!ELEMENT   CLOSE         -O          (#PCDATA)        >
11. <!--          ELEMENTS      NAME        VALUE            DEFAULT        -->
12. <!ATTLIST   %doctype;     STATUS      (CONFIDEN|PUBLIC) PUBLIC          >
13. <!ATTLIST   P id          ID          #IMPLIED        >
14. <!ATTLIST   PREF         refid       IDREF            #IMPLIED        >
```

SGML

```
<!ENTITY %doctype "MEMO" -->
```

- ❖ Opis najvišjega elementa (doctype).

```
<!ELEMENT MEMO - - ((TO&FROM), BODY, CLOSE?) >
```

- ❖ - - uporaba začetne in končne značke je obvezna,
- ❖ (TO&FROM), BODY, CLOSE elementi , iz katerih je sestavljen tip dokumenta MEMO,
- ❖ TO&FROM elementa se morata pojaviti enkrat v poljubnem zaporedju,
- ❖ CLOSE? Element se pojavi nič- ali večkrat.

SGML

```
<!ATTLIST %doctype STATUS (CONFIDENT|PUBLIC) PUBLIC >
```

- ❖ Lastnost STATUS ima možni vrednosti CONFIDENT in PUBLIC.
- ❖ Privzeta lastnost je PUBLIC.

SGML

- ❖ SGML je spremenil razmišljanje o nujnih lastnostih sodobnih označevalnih jezikov.
- ❖ Vsi označevalni jeziki v rabi na spletu sledijo načelom, ki jih je uveljavil SGML.
- ❖ To velja tudi za HTML in XML.