STANDARDIZACIJA ZAPISA JEZIKOVNIH PODATKOV

Tomaž Erjavec
Odsek za inteligentne sisteme
Institut Jožef Stefan
Jamova 39, Ljubljana
tomaz.erjavec@ijs.si

POVZETEK

Standardizirani računališki zapis jezikovnih podatkov poveča njihovo uporabnost, saj spodbudi večnamenskost in izmenljivost podatkov ter poveča njihovo trajnost. V članku argumentiram koristnost standardizacije, nato pa se osredotočim na standard ISO SGML (Standard Generalised Markup Language) ter z njim povezane standarde in pobude. Obravnavam standarde za prenos jezikovnih podatkov po omrežju (HTML/XML), za zapis jezikovnih podatkov v znanstvene namene (TEI) in za zapis terminoloških podatkov (MARTIF, TMX). Te standarde in pobude predstavim na primerih, podam njihovo uporabo na zapisih slovenskega jezika ter pokažem na možne uporabe pri nas.

ABSTRACT

Standardised digital encoding of language data increases its utility as it facilitates multiple uses and interchange of the data and increases its longevity. The paper outlines the benefits of standardisation, and then focuses on the ISO standard SGML (Standard Generalised Markup Language) and standards and initiatives connected with SGML. Described are those for transfering language data over the internet (HTML/XML); for encoding language data for scholarly purposes (TEI); and for encoding of terminological data (MARTIF, TMX). The discussion of these standards and initiatives is accompanied by examples, and possible applications to Slovene language data.

1 UVOD

Z naraščanjem računalniško zajetih besedil postaja vse bolj pomembno, na kakšen način so ta besedila zapisana. Ker so računalniško zapisana besedila uporabna v raznovrstne namene in z različnimi programskimi orodji, je potrebno prevajati med različnimi formati. Pri tem so formati zapisa vedno bolj bogati, obenem pa jih je vedno več. Delno rešitev tega problema nudijo 'industrijski standardi', tj. načini zapisa, ki so sicer v lasti nekega podjetja, se pa uporabljajo tudi s programi drugih proizvajalcev, vsaj tako, da omogočajo uvoz oz. izvoz podatkov v tem formatu. Pri pripravi besedila za tiskalnik je to npr. jezik PostScript in njegov naslednik PDF, oboje proizvod podjetja Adobe. Pri avtorskih besedilih je precej uveljavljen zapis RTF podjetja Microsoft Corporation, v zadnjem času pa se že samoumevno izmenjuje besedila, zapisana kar v eni zadnjih verzij urejevalnika Word.

Posebno zadnja možnost zastrašuje s tesno povezanostjo besedila in orodja, s katerim je to nastalo. V zadnjem desetletju se je na področju jezikovnih tehnologij namreč uveljavilo prepričanje, da so besedila, shranjena na računalnikih, vredna tudi mimo programja, ki nad njimi deluje. Premik k zbiranju in obravnavi jezikovnih virov se je zgodil predvsem na področju računalniškega jezikoslovja oz. jezikovnih tehnologij, pa tudi pri podjetjih, ki imajo opravka z velikimi količinami besedil.

Zato so se začele pojavljati pobude za standardizacijo zapisa jezikovnih podatkov, ki naj bi predpisale javno dostopne in trajne načine zapisa. Tako zapisane jezikovne podatke potem upoprabljajo orodja, ki implementirajo te standarde, bodisi da jih uporabljajo neposredno, ali pa lahko vanje podatke uvažajo oz. izvažajo.

Industrijski standardi imajo namreč pomanjkljivost, da so v lasti podjetja, ki ima nad njimi tudi kontrolo; podjetju je prepuščeno, ali javno objavi specifikacije svojega zapisa, kako in kdaj specifikacije spremeni in ali se drži svojih lastnih specifikacij v orodjih, ki jih ponuja. S tem so podatki vezani na orodje, s katerim so nastali, obenem pa hitro zastarajo.

Za razliko od industrijskih standardov so mednarodni javni in večinoma brezplačno dostopni, spreminja pa se jih samo po točno določenem postopku; podobno je z neprofitnimi pobudami, ki jih podpirajo npr. Evropska skupnost in ameriška vlada. Dodati pa je potrebno, da imajo standarni zapisi tudi svoje težave; podpirati moramo nov format zapisa, izbira in implementacija standarda za naše potrebe pa je pogosto tudi zapletena in draga. Poleg tega je zaradi hitro se razvijajoče tehnologije posebno pri novejših pobudah težko vedeti, katere se bodo tudi dejansko obdržale v praksi, katere pa bo neopazno pokopal čas.

V članku predstavim nekaj pobud za standardiziran zapis jezikovnih podatkov. V 2. poglavju je predstavljen standard ISO SGML (Standard Generalized Markup Lan-

guage). Ta ima zadostno podporo tako industrije kot tudi drugih pobud, da se bo razvijal; večje aplikacije tega standarda pa so tudi že prodrle na slovenski prostor. Zaradi pomembnosti omrežja WWW (svetovnega spleta) se zato tu na kratko dotaknem tudi zapisov HTML (Hypertext Markup Language) in XML (eXtensible Markup Language). Prvi je trenutni standard zapisa spletnih strani, drugi pa naj bi to postal v prihodnosti. V 3. poglavju obravnavam zapis jezikovnih podatkov v znanstvene namene po priporočilih TEI (Text Encoding Initiative). Te upošteva večina projektov, ki zbirajo jezikovne vire, npr. korpuse. V 4. poglavju orišem predlog standarda ISO MARTIF (Machine Readable Terminology Interchange Format) in pobude TMX (Translation Memory eXchange). Za oba se bo še pokazalo, v kolikšni meri bosta prodrla, omenjam pa ju zaradi slovenskega priključevanja Evropski zvezi in s tem povezanem narščanjem potrebe po kakovostnih prevodih iz in v slovenski jezik. Na koncu omenim še nekaj drugih pobud za standardizacijo bolj jezikovno odvisnih zapisov, npr. EAGLES (Expert Advisory Group on Language Engineering Standards) in MULTEXT-East (Multilingual Texts and Corpora for Eastern and Central European Languages).

2 STANDARDNI POSPLOŠENI JEZIK ZA OZNAČEVANJE: SGML

SGML (Standard Generalised Markup Language) [6,1] je standard ISO 8879:1986 in določa jezik za predstavitev dokumentov, nad katerimi bodo delovali programi za obdelavo besedil. SGML je prvenstveno jezik za *označevanje* dokumentov, pri čemer lahko oznake opisujejo kakršnokoli informacijo, ki je dodana osnovnemu besedilu, npr. podatek, da je nek niz v besedilu naslov, ime ali beseda, da je neka beseda glagol, da ima nek termin povezavo s svojo razlago, da nek stavek spremlja slika ali njegov prevod ali da nek monolog govori Hamlet v prvem dejanju neke tragedije.
V primerjavi z drugimi jeziki za označevanje dokumentov odlikujejo SGML tri značilnosti:

1) Poudarek na *opisnem* namesto postopkovnem označevanju

Za razliko od mnogih drugih formatov zapisa besedil (npr. RTF) so oznake SGML namenjene opisu lastnosti besedila, ki ga zajemajo, ne pa postopku, ki te lastnosti realizira na konkretnem mediju: oznaka npr. pove, da del besedila, ki ga zajema, predstavlja odstavek, ne pa, da je potrebno izpustiti prazno vrstico in za določeno mero zamakniti začetek naslednje vrstice. Opisno označeni podatki imajo to prednost, da vsebujejo informacije v bolj prečiščeni obliki in jih je zato lažje izmenjavati med ljudmi oz. aplikacijami.

2) Koncept tipa dokumenta

SGML bi lahko poimenovali tudi jezik za metaoznačevanje dokumentov, saj standard ne predpisuje konkretnih oznak in njihovih medsebojnih odnosov. Namesto tega vpelje SGML pojem tipa dokumenta in z njim formalno *definicijo tipa dokumenta:* DTD (Document Type Definition). Šele DTD konkretno določa, kako mora biti nek dokument strukturiran in kako izgledajo njegove oznake. Nek DTD tako predstavlja gramatiko ('jezik') za določeno zvrst dokumentov, npr. za mrežne dokumente, knjige, korpuse, terminološke slovarje itd. Takšen pristop omogoča široko uporabo standarda, saj tako lahko pokriva dokumente z izrazito različno strukturo.

3) Neodvisnost od konkretnega zapisa besedil

Eden od osnovnih ciljev SGML je, da so v njem zapisani podatki prenosljivi z ene strojne in programske opreme na drugo brez izgube informacije. SGML zato vsebuje splošen način za nadomeščanje nizov med obdelavo dokumenta. Z *entitetami* SGML je mogoče preseči neskladnosti in pomanjkljivosti v naborih znakov različnih specifičnih računalniških sistemov, saj lahko za neprenosljive znake definiramo opisna imena, tj entitete.

V tujini vedno več podjetij, ki imajo opravka z velikimi količinami besedil, prehaja na zapis SGML; obstaja tudi že kar nekaj podjetij, ki se ukvarjajo izključno s SGML, bodisi z izdelovanjem programske opreme ali pa, pogosteje, z omogočanjem končnim uporabnikom, da preidejo na ta standard. V Sloveniji poleg raziskovalnih projektov (npr. MULTEXT-East, [3]) verjetno prve začenjajo z uporabo SGML založbe; tako sta npr. pri DZS v izdelavi angleško-slovenski slovar in korpus slovenskega jezika, ki sta oba zapisana v skladu s SGML.
SGML služi kot osnova množici izvedenih standardov in mednarodnih priporočil, tako tudi večini ostalih, omenjenih v tem članku.

2.1 Jezika omrežja: HTML in XML

Verjetno je vsaj posredno najbolj znana definicija tipa dokumenta SGML tista za HTML (Hypertext Markup Language), ki jo upoštevajo vse pravilno narejene spletne strani mreže WWW (World Wide Web). Čeprav je bila prva inačica HTML DTD narejena kot aplikacija standarda SGML in so take tudi vse ostale, ki jih je objavil konzorcij W3C [14], so izdelovalniki brkljalnikov (prvi Netscape) kaj hitro prekršili določila v DTD. Zaradi tega veliko število strani WWW ni več v skladu s SGML ozrioma tipom dokumenta HTML. Kot primer podajam spodaj začetek dokumenta SGML, zapisanega po HTML DTD, verzija 3.2:

```
<!DOCTYPE HTML PUBLIC
  "-//W3C//DTD HTML 3.2//EN">
<HTML>
  <HEAD>
    <TITLE>Multext-East HTML Corpus
Sample</TITLE>
    <META HTTP-EQUIV="Content-Type"
CONTENT="text/html; charset=ISO-8859-2">
  </HEAD>
  <BODY>
```

```
    <H1>MTE HTML Corpus Sampler: Nineteen
Eighty-Four, Slovene</H1><HR>
    <H2>Prvi del</H2><HR>
      <H3>I</H3>
       <P>Bil je jasen, mrzel aprilski dan in
ure so bile trinajst. <B>Winston Smith</B> je
imel brado zakopano v prsi, da bi ušel
strupenemu vetru, ko je stopil skozi steklena
vrata bloka Zmaga, vendar ne dovolj hitro, da ne
bi vrtinec peščenega prahu vstopil skupaj z
njim.
```

Ker se uporaba omrežja WWW vse bolj širi, nenazadnje kot sredstvo informiranja znotraj organizacij (*intranet*), se kažejo tudi pomanjkljivosti jezika HTML. Pokazala se je potreba po splošnejšem jeziku opisovanja (mrežnih) dokumentov, kot je to SGML. Vendar je standard SGML izjemno zapleten in sedaj star že več kot deset let. Zato je konzorcij W3 izdelal predlog XML [15]. Medtem ko je HTML samo točno določen tip dokumentov SGML, je XML poln, čeprav poenostavljen SGML, primeren za hitro obravnavo preko mreže.

Zaradi zapletenosti standarda SGML in zaradi vse večjega pomena mrežne izmenjave podatkov bo verjetno v prihodnosti XML postal osnova za množico izvedenih standardov in pobud za zapis različnih zvrsti jezikovnih podatkov.

3 POBUDA ZA ZAPIS BESEDIL: TEI

Za standardiziran zapis besedil, ki naj bi se uporabljala pretežno v znanstvene namene, je naredil velik korak TEI (Text Encoding Initiative), katerega priporočila TEI P3 [13,9] za široko paleto zvrsti besedil določajo konkretne oznake SGML in strukturo teh oznak; poln opis TEI P3 obsega preko 1200 strani. TEI P3 sestavlja nabor definicij tipov dokumentov in entitet, dokumentacija pa podaja pomen posameznih oznak in opisuje DTD-je ter izpelje načine za njihovo kombiniranje ter nadgradnjo.

TEI P3 loči več vrst oznak; nekatere so obvezne v vseh s TEI skladnih dokumentih (npr. glava dokumenta), od drugih lahko za naš konkreten projekt izberemo po eno (npr. proza ali slovar), tretje, ki predstavljajo interpretacijo besedila (npr. skladenjska analiza), pa lahko dodajamo na to osnovo.

Kot primer podajam spodaj element ENTRY, ki pri izbranem slovarskem modulu TEI zajema, en geselski članek slovarja:

```
<entry type="hom" n="2" key="a">
  <form lang="en">
    <orth type="headword">a</orth>
    <pron>eI</pron>, <pron>@</pron>
  </form>
  <gramgrp><pos>article</pos></gramgrp>
  <trans>
    <usg lang="sl">nedolo&ccaron;ni
&ccaron;len</usg>
    <tr lang="sl">neki, eden, en</tr>&comma;
  </trans>
  <xr><ref>few</ref>&comma;</xr>
  <xr><ref targtype="homograph">
little<label>1</label></ref>&comma;
  </xr>
  <xr><ref targtype="homograph">
lot<label>1</label></ref>&comma;
  </xr>
  <xr><ref>many</ref></xr>
</entry>
```

Kot vidimo, posebej še če smo seznanjeni s TEI, je v geselskem članku za drugo sopomenko besede *a* zapisano geslo skupaj z dvema izgovorjavama, pripada pa besedni vrsti *article*. Ima eno prevodno enoto, ki vsebuje uporabo tega prevoda in prevedeno geslo. Poleg tega obsega povezave na geselske članke *few, little[1], lot[1]* in *many*. Seveda pa ni namen besedil, zapisanih v TEI, da jih moramo brati v 'izvirniku', pač pa ustrezen program pretvori standardiziran zapis v želeno obliko, npr. knjižno ali spletno. Vendar je posebej za izmenjavo pomembno, da je besedila TEI možno brati v izvirniku, poleg tega pa so robustno zapisana. Za naše črke so, kot se vidi iz zgornjega primer, uporabljene entitete SGML, in sicer nabor znakov `ISO 8879:1986//ENTITIES Added Latin 2//EN`.

Na TEI se dandanes že samoumevno sklicujejo projekti, ki ustvarjajo jezikovne vire, predvsem korpuse. Slovenski jezikovni viri, zapisani po TEI P3, so del korpusov mednarodnih projektov s področja jezikovnih tehnologij, MULTEXT-East in TELRI. Šestjezični korpus MULTEXT-East [3] vsebuje okoli 300.000 besed v slovenskem jeziku: domače leposlovje ('Galjot', D.Jančar), časopisne članke (Dnevnik) ter prevedeno leposlovje ('1984', G.Orwell, v prevodu A.Puhar). Slednji je še posebej zanimiv, saj je roman v šestih prevodih po stavkih poravnan z izvirnikom v angleščini, roman v vseh sedmih jezikih pa je tudi oblikoslovno označenih. Korpus je označen po CES (Corpus Encoding Specification), ki je TEI prirejen za opis korpusov, namenjenih izmenjavi in zbranih za namene jezikovnih tehnologij. Korpus projekta TELRI, ki je zapisan v TEI, obsega Platonovo 'Republiko' v dvajsetih jezikih, med njimi tudi slovenskem (besedilo so digitalizirali na Inštitutu za slovenski jezik, ZRC SAZU). TELRI je izdal CD-ROM [4], ki vsebuje tudi oba korpusa. Nadaljnji primer uporabe TEI pri nas pa je nastajajoči korpus FIDA [11].Zaradi izredno bogatega obsega besedilnih zvrsti, ki jih pokriva TEI P3, in zaradi nadaljnjega delovanja iniciative TEI (tako je npr. v teku izdelava inačice P3 v XML), se s TEI povezuje tudi večina standardov, izvedenih iz SGML.

4 TERMINOLOGIJA IN PREVODI: MARTIF IN TMX

Zaradi aktualnosti prevajanja v Sloveniji in s tem povezane potrebe po terminoloških virih (slovarjih) orišem tem poglavju dve zvrsti dokumentov SGML, ki naj bi standardizirala računalniški zapis terminoloških baz (MARTIF) in pomnilnikov prevodov (TMX) [12].

MARTIF (Machine Readable Terminology Interchange Format) je izšel iz pobude TEI in je trenutno v fazi poslednjih osnutkov dveh mednarodnih standardov ISO/FDIS 12620 in ISO/FDIS 1220. Prvi definira podatkovne elemente, drugi pa njihove medsebojne

relacije, oboje za zapis terminoloških baz, tj. dokumentov, v katerih so definirana pomenska polja, znotraj njih pa identificirani termini in koncepti. Primer javno dostopne večjezične terminološke baze je EURODICAUTOM [5] prevajalske službe komisije EU.
MARTIF je za Slovenijo zanimiv toliko, kolikor bi bila smiselna pretvorba večjega števila terminoloških slovarjev v ta format in iskanje po takšni terminološki bazi večjega števila uporabnikov. Prevajanje terminologije v slovenščino in iz nje je namreč pogosto pereča točka prevajalskega postopka, obenem pa v Sloveniji obstaja precej delnih in različno računalniško zapisanih terminoloških slovarčkov. Ponudba servisa, kot je EURODICAUTOM za slovenski jezik, bi bila prevajalcem v pomoč, saj bi s tem vsaj za zajeta področja lahko hitreje napisali boljše prevode. Ponudba terminoloških baz v standardiziranem formatu pa tudi olajša ponovno uporabo takih virov za druge programe jezikovnih tehnologij.
V zadnjem času postajajo programi s pomnilnikom prevodov aktualno pomagalo prevajalcem tudi za slovenski jezik, predvsem pri prevajalskem oddelku Službe Vlade RS za evropske zadeve [8]. Pomnilnik prevodov hrani prevodne enote, tj. segmente (ponavadi povedi) nekega originala in njihove prevode. Pomnilnike polavtomatsko ustvarjamo na podlagi že prevedih besedil in njihovih izvorov.
Za zapis pomnilnikov prevodov trenutno prevladujejo 'industrijski standardi' zapisa (izdelki Trados in, širše, Word ter Microsoft), postopek standardizacije ISO pa se šele začenja. Leta 1997 je bila v okviru združenja LISA (Localisation Industry Standards Association) ustanovljena skupina OSCAR (Open standards for Container/Content Allowing Re-use). Prva naloga skupine OSCAR je definicija formata za izmenjavo baz pomnilnikov prevodov; osnutek le-tega se imenuje TMX (Translation Memory eXchange).
TMX uporablja XML kot osnovni standard in znotraj XML določi tip dokumenta, ki je primeren za izmenjavo baz pomnilnikov prevodov. Del TMX, ki je že izdelan, zajema specifikacijo formata vrhnjih elementov, ki opisujejo dokument kot celoto (glava dokumenta) ter formata vnosov, tj. prevodnih enot. Za zgled podam spodaj primer enostavne prevodne enote, zapisane v TMX:

```
<TU ID="0002">
  <PROP NAME="Domain">Cooking</PROP>
  <TUV LANG="EN">
    <SEG>menu</SEG>
  </TUV>  <TUV LANG="FR-CA">
    <SEG>menu</SEG>
  </TUV>  <TUV LANG="FR-FR">
    <SEG>menu</SEG>
  </TUV>
</TU>
```

Kot vidimo, je zapis precej drugačen od slovarskega zapisa TEI tudi v delih, kjer bi si bila lahko podobna. Oznake v TMX so prilagojene specifiki pomnilnikov prevodov, medtem ko je zapis TEI precej bolj splošen in s tem tudi bolj dolgovezen. Prevedba med raznimi tipi dokumentov pa je olajšana s tem, ker je vsem skupen SGML kot osnovni standard.
Največji težava pri izdelavi TMX predstavlja želja po ohranitvi formatiranja in označitev pomnilnikov prevodov, ki izhajajo iz konkretnih orodij, s katerimi je bil pomnilnik narejen (npr. RTF), in sožitju teh oznak z oznakami, definiranimi s standardom. Tudi večjezikovnost dokumentov prinaša težave. Kot se že vidi iz uporabe XML (raje kot SGML), je tudi tu TMX zazrt v prihodnost: za predstavitev naborov znakov predpisuje TMX uporabo Unicode (ISO 10646), ki je opisno sicer izredno močen, ga pa trenutno podpira še zelo malo orodij.
TMX sicer še ni zrel za uporabo, vendar novi poskusi standardizacije nudijo možnost, da posamezna država sodeluje pri njihovem nastajanju. Potencialna uporaba TMX je precejšnja, saj pomnilniki prevodov lahko služijo ne samo kot pomoč pri prevajanju, pač pa tudi kot dragocen večjezični vir besedil; v njihovo izdelavo je potrebno vložiti precej truda, vsebujejo pa obilico znanja o jezikovnem paru. Na osnovi pomnilnikov prevodov je mogoče izdelovati boljše in bolj ažurne (ne samo terminološke) slovarje, lahko bi pa služili tudi kot učna množica pri avtomatskem prevajanju.

5 ZAKLJUČEK

Področje standardizacije je izredno široko in obsega tudi načine zapisa govornih podatkov, obstajajo pa tudi (še) bolj jezikovno odvisne pobude, kot je npr. zapis oblikoslovnih oznak v leksikonih in korpusih. Tu je bila dejavna evropska skupina EAGLES [2], ki je pripravila priporočila, ki skušajo ponuditi skladen zapis jezikovnih virov, ob tem pa ohraniti prilagodljivost, potrebno za opis različnih jezikov. Njihova priporočila upošteva (slovenski) leksikon projekta MULTEXT-East [8].
Oblikoslovne oznake MULTEXT-East je zanimivo primerjati z istonamenskimi oznakami, ki so jih izdelali na Institutu za slovenski jezik ZRC [10]; primerjavo z MULTEXT-East najdemo prav tam, na str.515–516. Primer slovenske oblikoslovne oznake MULTEXT-East je npr. Vcps-sma, ki jo lahko pripišemo besedni obliki *bil*. V oblikoslovni oznaki pomeni prva črka besedno vrsto, ostale pa, glede na mesto v nizu, podajo vrednost od besedne vrste odvisnega atributa. Niz je formalno enak daljšemu, pa bolj preglednemu zapisu PoS: Verb, Type: copula, VForm: participle, Tense: past, Person: -, Number: singular, Gender: masculine, Voice: active. Ista oznaka v ZRC zapisu je GLBme, ki pomeni GLagol, opisni deležnik na *-l biti*, moški spol, ednina. Za razliko od angleških oznak MULTEXT-East so oznake ZRC berljive v slovenskem jeziku, krajše, pa tudi bolj v skladu s slovensko slovnico.
Čeprav je pri vseh v članku omenjenih standardizacijskih pobudah večjezičnost v ospredju, pa ostaja metajezik zapisov angleški, kot se vidi tudi na primerih zapisov HTML, TEI in TMX. Podobno tudi oznake MULTEXT-

East privzemajo sistem, ki temelji na angleškem jeziku in formalizmih, ki so bili najprej razviti za angleški jezik in jezike evropske unije. So pa zato te oznake v primerjavi z oznakami ZRC bolj formalizirane in primerljive z oznakami drugih evropskih jezikov. Pri prevzemanju mednarodnih pobud je tako potrebno vložiti ne samo finančna sredstva, pač pa se v teh zapisih tudi odreči polni uporabi slovenskega jezika.

Največjo pobudo bodo pri prevzemanju standardnih zapisov verjetno dali tisti, ki jim je v interesu, da se omogoči izmenjava in dolgoročno arhiviranje jezikovnih podatkov. Ker njihovo združevanje in izmenjava omogočata razvoj jezikovnih tehnologij, te pa spodbujajo uporabo nacionalnih jezikov, zastopa to stališče Evropska unija, lahko pa bi jo tudi slovenska vlada. Kjer pa so standardi šele v nastajanju, lahko udeležba slovenskih predstavnikov v ustreznih delovnih skupinah tudi vpliva na končno obliko standardov, posebej še tam, kjer drugače ne bi upoštevali posebnosti slovenskega jezika.

V članku sem tako obravnavali nekaj standardov in priporočil za zapis pisanih jezikovnih virov: SGML, HTML, XML, TEI, MARTIF in TMX. Pokazali sem na težavnost uvajanja teh standardov pri zapisu (slovenskih) besedil, predvsem pa predstavili prednosti, ki jih prinaša njihova uporaba. Prihodnost bo pa pokazala, v kolikšni meri bodo opisani standardi k nam tudi res prodrli.

6 ZAHVALA

Za koristne pripombe se zahvaljujem dr. Jerneji Gros in Vojku Gorjancu ter dvema anonimnima recenzentoma. Za vse napake v članku je seveda odgovoren avtor.

7 VIRI

[1] Batagelj,.V. Uvod v SGML. http://vlado.mat.uni-lj.si/vlado/sgml/sgmluvod.htm, 1995.

[2] EAGLES: Expert Advisory Group on Language Engineering Standards. http://www.ilc.pi.cnr.it/EAGLES96/home.html,1996.

[3] Erjavec, T., Ide, N. The MULTEXT-East Corpus. V Proceedings of the First International Conference on Language Resources and Evaluation, LREC'98. Granada. 971-974. 1998.

[4] Erjavec, T., Lawson, A., Romary, L (ur.).East meets West: A Compendium of Multilingual Resources. CD-ROM, ISBN: 3-922641-46-6, 1998.

[5] EURODICAUTOM. European Commission Translation Service.http://www2.echo.lu/edic/, 1998.

[6] Goldfarb, C.F. The SGML Handbook. Clarendon Press, Oxford, 1990.

[7] Ide, N., Tufis, D., Erjavec, T. Development and Assessment of Common Lexical Specifications for Six Central and Eastern European Languages V Proceedings of the First International Conference on Language Resources and Evaluation, LREC'98. Granada. 223-240, 1998.

[8] Erbič, D., Milan, Ž. (1998) Vzpostavitveni dokument projekta 2. IDC – Prevodi in terminologija. VDP-CVI|PRO – 024, Vlada RS, junij 1998.

[9] Ide, N., Veronis, J. (ur.). The Text Encoding Initiative: Background and Context. Kluwer Academic Publishers, Dordrecht, 1995.

[10] Jakopin, P., Bizjak, A. O strojno podprtem oblikoslovnem označevanju slovenskega besedila. Slavistična Revija, 45(3-4):513–532, 1997.

[11] Krek, S., Stabej, M., Gorjanc, V., Erjavec, T., Romih, M., Holozan, P. FIDA: korpus slovenskega jezika. http://www.fida.net, 1998.

[12] Melby, A. Data exchange standards from the OSCAR and MARTIF projects. V Rubio, A., Gallardo, N., Castro, R.,Tejada, A., ur. Proceedings of the First International Conference on Language Resources and Evaluation, LREC'98. Granada. 3-8. 1998. (glje tudi http://www.lisa.unige.ch/tmx/).

[13] Sperberg-McQueen, C.M., Burnard, L. (ur.). Guidelines for Electronic Text Encoding and Interchange. Chicago and Oxford. http://www.uic.edu/orgs/tei/, 1994.

[14] W3C: World Wide Web Consortium. http://www.w3.org/

[15] XML: Extensible Markup Language. http://www.w3.org/XML/