Izdelava XML-shem za slovarske projekte na primeru nastajajočih tipološko raznovrstnih slovarjev

Nina Ledinek, Andrej Perdih

Inštitut za slovenski jezik Frana Ramovša ZRC SAZU
Novi trg 4, 1000 Ljubljana
{NLedinek, APerdih}@zrc-sazu.si

Povzetek

Prispevek prikazuje dileme, ki so se pojavile pri oblikovanju XML-shem za tipološko raznovrstne slovarske podatkovne baze Inštituta za slovenski jezik Frana Ramovša ZRC SAZU (*Slovar sinonimov slovenskega jezika, Slovar slovenskega knjižnega jezika 16. stoletja, Novi slovar slovenskega jezika, Slovar novejšega besedja slovenskega jezika, Slovenski etimološki slovar, Slovar pravopisno težavnega besedja slovenskega jezika*). Pojasnjuje praktične in tehnične dejavnike, ki na izdelavo XML-shem vplivajo z vidika leksikografskega dela pri vnosu jezikovnih podatkov, opisuje pa tudi tiste dejavnike, ki vplivajo na učinkovito izrabo podatkov iz dokončanih slovarskih podatkovnih baz.

Designing XML Schemas for dictionary projects: the case of emerging typologically varied dictionaries

The article deals with dilemmas that have arisen in the designing of XML Schemas of typologically varied dictionary databases of the Fran Ramovš Institute of the Slovene language (Dictionary of Slovene Synonyms, Dictionary of the Literary Slovene Language of the Sixteenth Century, The New Dictionary of Slovene Language, Dictionary of Newer Standard Slovene Words, Slovene Etymological Dictionary, Dictionary of Less Used Slovene Words). It presents the conceptual, practical and technical factors that influence the designing of XML Schemas from the point of view of lexicographical work, but also describes the factors that determine the efficiency of the use of data from the completed dictionary databases.

1. Uvod

V zadnjih desetletjih so se na področju leksikografije zgodili tehnološki in konceptualni premiki, ki so odločilno vplivali na metodologijo leksikografskega dela ter na dojemanje in uporabo slovarskih priročnikov in drugih sorodnih jezikovnih virov. Morda najodločilneje je sodobno leksikografijo zaznamovalo dejstvo, da leksikografi in uporabniki slovarskih priročnikov ne dojemajo več kot (izhodiščno) knjižnih jezikovnih virov, ampak kot večnamenske razširljive strukturirane računalniško berljive podatkovne baze, v katerih so podatki ustrezno hierarhizirani, (standardno) označeni in medsebojno povezani. Vzajemno s konceptualnimi spremembami se je, ob sočasnem razvoju informacijske tehnologije oz. elektronskih medijev, spremenilo tudi dojemanje slovarskih priročnikov in sorodnih jezikovnih virov – ti namreč niso več namenjeni zgolj tehnološko vedno bolj spretnim in zahtevnim človeškim uporabnikom, ampak jih izkoriščamo kot jezikovne vire tudi pri številnih nalogah procesiranja naravnih jezikov.

V prispevku prikazujemo zadrege in dileme, na katere smo naleteli ali se z njimi trenutno srečujemo pri oblikovanju novih računalniško berljivih slovarskih podatkovnih baz v formatu XML, pa tudi na tiste, ki so povezane s pretvorbo slovarskih virov Inštituta za slovenski jezik Frana Ramovša ZRC SAZU v format XML iz drugih elektronskih formatov. Izkazalo se je, da zaradi raznolikosti slovarskih priročnikov, ki nastajajo na inštitutu (*Slovar sinonimov slovenskega jezika, Slovar slovenskega knjižnega jezika 16. stoletja, Novi slovar slovenskega jezika, Slovar novejšega besedja slovenskega jezika, Slovenski etimološki slovar, Slovar pravopisno težavnega besedja slovenskega jezika*, terminološki slovarji ...), oblikovanje podatkovne baze in XML-sheme za vsak slovarski projekt tako s tehničnega kot tudi leksikografskega vidika prinaša svojevrstne dileme.

2. Standardni format XML in slovarske podatkovne baze

Kot standardni format za zapis slovarskih in drugih jezikovnih podatkovnih baz se je zaradi svoje univerzalnosti in fleksibilnosti uveljavil XML (eXtensible Markup Language). Gre za označevalni jezik, ki je zelo primeren za večnivojsko hierarhično strukturiranje podatkov, torej tudi jezikovnih virov, saj ti navadno vključujejo veliko število hierarhično urejenih podatkovnih tipov. XML odlikuje razmeroma preprosta sintaksa, ker pa metaoznake niso definirane vnaprej, lahko vsebino strukturiramo z metaoznakami, ki so logične in intuitivno razumljive, s čimer dobimo dober nadzor nad logično strukturo podatkovne baze. Datoteke XML so navadne besedilne datoteke s privzetim unikodnim kodiranjem znakov, zato so primerne za dolgoročno shranjevanje podatkov ter njihovo izmenljivost in prenosljivost med različnimi orodji in operacijskimi sistemi.

Eden ključnih in hkrati najzahtevnejših korakov pred samo vzpostavitvijo slovarske podatkovne baze je oblikovanje ustreznega slovarskega koncepta, ki bo služil kot vodilo za oblikovanje vseh predvidenih geselskih sestavkov. Predpostavljena struktura geselskih sestavkov mora biti takšna, da podpira logično in intuitivno strukturiran vnos vseh jeziko(slo)vnih podatkov, ki so za opis posamezne leksikalne enote pomembni, hkrati pa dovolj univerzalna in striktna, da redaktorje usmerja k čim bolj sistematični in konsistentni interpretaciji raznovrstnih jezikovnih podatkov na tak način, da so ti uporabni in razumljivi tudi za končne uporabnike slovarskih priročnikov.

Formalno strukturo slovarske podatkovne baze v formatu XML opisuje t. i. shema, ki jo lahko razumemo kot nekakšno projekcijo formalnih značilnosti mikro- in makrostrukture slovarja, kot ju določa slovarski koncept, v standardni računalniški jezik. Shema določa zlasti, kateri

so možni oz. dovoljeni elementi in atributi slovarske podatkovne baze, kakšna so hierarhična razmerja med njimi in po kakšnem vrstnem redu si elementi sledijo, kakšne so omejitve njihove rabe z vidika sopojavljanja oz. izključevanja, kolikokrat se posamezen element na določenem mestu lahko ponovi. Shema ne nazadnje določa tudi, kakšna sme biti formalna vsebina elementov. Predpisuje, da lahko določen element vključuje le druge elemente ali pa morda le slovarsko besedilo ali številke, ki jih v podatkovno bazo vnese redaktor, omejuje dolžino vnesenega besedila, določa, ali je vsebina specifičnega elementa omejena na seznam vnaprej določenih izbir (npr. na določen nabor besednih vrst, kvalifikatorjev), predpisuje rabo obveznih atributov ipd.

Obstajajo sheme različnih formatov – DTD, XML Schema[1] (.xsd) in RELAX NG (.rng). Format XML-shema, ki je nekoliko bolj fleksibilen od formata DTD, uporabljamo pri opisu slovarskih podatkovnih baz Inštituta za slovenski jezik Frana Ramovša ZRC SAZU, saj njegovo uporabo podpira leksikografski program iLEX,[2] s katerim pripravljamo redakcije slovarjev.[3] Ena temeljnih nalog takšnega leksikografskega programa, poleg dejstva, da redaktorjem omogoča urejanje interpretiranih jezikovnih podatkov in njihovo vizualizacijo, je, da nadzoruje skladnost geselskih sestavkov z določili sheme, tako da opozarja na neustreznosti v formalni vsebini in strukturi segmentov slovarske podatkovne baze. Program, podprt z ustrezno shemo, pripomore k zagotavljanju konsistentnosti slovarske podatkovne baze na formalni ravni, seveda pa ne more preprečiti vsebinskih neustreznosti.

Leksikografi običajno ugotavljajo, da je pri načrtovanju XML-sheme za slovarski projekt smiselno upoštevati številne dejavnike. Ključnega pomena je med drugim, da se konceptualizacija slovarske strukture pri redaktorjih čim bolj neposredno odraža v intuitivno razumljivi hierarhično strukturirani podatkovni strukturi, kot je definirana s shemo – da torej posamezen mikrostrukturni element slovarja, npr. zaglavje, vključuje tiste podelemente, ki jih redaktor vidi v tem razdelku, ne pa npr. v okviru posameznih pomenov. Obenem je treba upoštevati tudi praktični vidik obvladljivosti slovarske strukture pri redaktorjih. Število mikrostrukturnih elementov posameznega slovarja lahko namreč doseže trimestno številko, pri čemer se lahko posamezni podatkovni tipi (npr. razdelki, v katerih je izkazan pomen leksikalnih enot), morda strukturno le nekoliko modificirani, pojavijo na različnih mestih v slovarski strukturi. Pri pripravi sheme se je zato treba vprašati, koliko različnih elementov je v shemo smiselno vključiti (tj. ali vsak slovarski podatek označiti kot svoj element ali pa sorodne podatke obravnavati v okviru istega elementa), da bo podatkovna struktura vsebinsko logična, spominsko za redaktorje ne preveč obremenjujoča, hkrati pa tehnično razmeroma lahko obvladljiva (npr. če bi ob analizi dodatnih jezikovnih podatkov redaktor ugotovil, da je treba strukturo geselskega sestavka spremeniti).

Dileme se pojavljajo tudi glede vprašanja, kako hierarhično globoko je posamezen element praktično umeščati. Hierarhično globlje strukturirane enote so po eni strani primernejše, ker omogočajo večjo segmentacijo jezikovnih podatkov, posledično pa tudi iskanje podatkov glede na kompleksnejše iskalne pogoje in podrobnejše postprocesiranje podatkov, po drugi strani pa velja, da pretirano razvejane podatkovne strukture redaktorjem otežujejo navigacijo po slovarskih geslih, zaradi česar se čas redigiranja geselskih sestavkov podaljša, pozornost redaktorja pa je v večji meri usmerjena k vzpostavljanju ustreznih hierarhičnih razmerij, manj pa k vsebinskim vprašanjem. Odločitev o združevanju več elementov v okviru določenega nadelementa je hkrati treba pretehtati še s tehničnega vidika, saj nanj vpliva tudi leksikografski program. Bistvenega pomena je namreč, kakšne možnosti prikaza podatkov omogoča uporabniški vmesnik in kakšne možnosti iskanja po že vzpostavljeni podatkovni bazi program omogoča. Ob pripravi XML-sheme si je poleg tega treba zastavljati še vprašanja o tem, kako predvideno strukturiranje podatkov vpliva na možnosti rabe jezikovnega vira pri uporabnikih, kakšne so možnosti prikaza podatkov v elektronski ali tiskani obliki glede na vzpostavljeno podatkovno strukturo, ali je omogočeno vzpostavljanje sklicev na različne segmente podatkovne baze in kakšno postprocesiranje baze je predvideno.

V nadaljevanju prispevka na kratko in shematično prikazujemo, kako smo skušali pri pripravi XML-shem za različne slovarske projekte, ki nastajajo na inštitutu za slovenski jezik, omenjene dejavnike (našteti so seveda le nekateri od številnih možnih) prepoznati, jih upoštevati in njihovo součinkovanje čim bolj smiselno uravnotežiti.

Za zapis slovarskih oz. leksikonskih podatkovnih baz so bili po svetu vzpostavljeni različni standardi zapisa, npr. Lexical Markup Framework (LMF),[4] Lexical Interchange Format Standard (LIFT),[5] priporočila iniciative TEI za zapis slovarskih podatkov[6] ipd. Kljub temu da bi bilo podatke večinoma mogoče zapisati v katerem od standardnih formatov, se pri nas za ta korak zaenkrat še nismo odločili. Odločitev je bila posledica presoje, da je tak zapis pri delu slovarskih baz manj smiseln npr. zaradi njihove specifičnosti (*Slovar novejšega besedja slovenskega jezika* je slovar manjšega obsega in je zasnovan predvsem kot dopolnilo *Slovarja slovenskega knjižnega jezika*, ki ni zapisan v skladu z omenjenimi standardi) ali izjemne kompleksnosti slovarske mikrostrukture, zaradi česar bi bila vzpostavitev standardnega zapisa težavna, kljub njegovi fleksibilnosti bi bilo namreč standardu včasih težko slediti (slednje velja zlasti za *Slovar sinonimov slovenskega jezika* in *Slovar slovenskega knjižnega jezika 16. stoletja*). Pri odločitvi smo upoštevali še dejstva, da so številne slovarske baze primarno namenjene (specializiranim) človeškim uporabnikom, v manjši meri pa so uporabne v postopkih procesiranja naravnih jezikov (npr. *Slovenski etimološki slovar*, *Slovar slovenskega knjižnega jezika 16. stoletja*), da so bili koncepti mnogih slovarjev, posledično pa tudi strukturiranje podatkov, oblikovani v času, ko omenjeni

[1] V prispevku je uporabljen poslovenjen zapis XML-shema.
[2] http://www.emp.dk/
[3] Med znanimi leksikografskimi programi so še ABBYY Lingvo Content (http://www.abbyy.com/lingvo_content/), IDM DPS (http://www.idm.fr/products/dictionary_writing_system_dps/27/) in TshwaneLex (http://tshwanedje.com/tshwanelex/), v slovenskem okolju pa je nastala Termania (http://www.termania.net/). IDM DPS uporablja format sheme DTD, iLEX in Termania podpirata XML-shemo, Tshwanelex pa vključuje interni DTD.
[4] http://www.lexicalmarkupframework.org/
[5] http://code.google.com/p/lift-standard/
[6] http://www.tei-c.org/release/doc/tei-p5-doc/en/html/DI.html

standardi še niso bili vzpostavljeni ali splošno razširjeni, poleg tega pa slovarski podatki niso zapisani v formatu XML, zato jih je treba vanj šele pretvoriti, strukturiranje, skladno s standardi, pa je posledično povezano z večjimi finančnimi vložki. Baze so večinoma tudi še v fazi oblikovanja oz. revizije. Deloma je na odločitev vplivala tudi ocena, v kolikšni meri je slovarska podatkovna baza zanimiva z vidika izmenljivosti podatkov in postprocesiranja.

Glede na naštete vidike sta slovarski bazi *Novega slovarja slovenskega jezika* in *Slovarja pravopisno težavnega besedja slovenskega jezika* drugačni, saj sta z vidika izmenljivosti podatkov in predvidenega postprocesiranja bistveno bolj uporabni. Zaenkrat sicer nista oblikovani v katerem od navedenih standardov, sta pa strukturirani na način, ki omogoča relativno enostavno pretvorbo v katerega od njih.

3. Oblikovanje XML-shem za slovarske podatkovne baze

3.1. Slovar sinonimov slovenskega jezika[7]

Slovar sinonimov slovenskega jezika bo prvi specializirani slovarski priročnik za slovenščino, ki bo sistematično izkazoval sinonimna razmerja med leksikalnimi enotami v jeziku. Glede na sklicno naravo slovarja se je pri oblikovanju njegove XML-sheme kot temeljno pojavljalo vprašanje, kako uravnotežiti zahtevo redaktorjev po tem, da bo predpostavljena podatkovna struktura zanje čim bolj logična in intuitivno razumljiva, hkrati pa ustrezna tudi z vidika vzpostavljanja sklicev na različne mikrostrukturne elemente geselskih sestavkov ter predvideni prikaz podatkov v elektronski in tiskani obliki.

Slovarski koncept predvideva geselske sestavke dveh tipov. V dominantnih geselskih sestavkih je poleg osnovnih slovničnih lastnosti v zaglavju za vsako geselsko besedo navedeno, v katerih pomenih in besednih zvezah, katerih skladenjsko jedrni del je, se pojavlja kot nezaznamovan, tj. najbolj običajen leksem za izražanje konkretnega pomena, ob vsakem pomenu in besedni zvezi pa so navedeni še drugi identificirani sinonimni leksemi, ki jih je v različnih okoliščinah mogoče uporabiti za izražanje istega pomena. Slovar prinaša tudi podatke o delnih sinonimih geselske besede v posameznih pomenih, poleg tega pa še poseben sklicni razdelek geselskega sestavka, v katerem je navedeno, v katerih pomenih iztočnica ni dominantni leksem za izražanje določenega pomena, ampak le eden od nedominantnih sinonimov k dominantnemu leksemu, obdelanem v drugem dominantnem geselskem sestavku.

Kazalčni geselski sestavki prinašajo poleg osnovnih slovničnih informacij o iztočnici le sklice na dominantne geselske sestavke, v katerih je iztočnični leksem naveden kot eden od zaznamovanih sinonimov k nezaznamovanemu ali najbolj običajnemu leksemu za izražanje konkretnega pomena. Iztočnica kazalčnega geselskega sestavka je lahko v okviru dominantnega sestavka navedena kot enobesedni sinonim ali pa kot ena od besed v večbesednem sinonimu.

Da bi količino ročnega dela zmanjšali in redakcijski proces pohitrili, smo se odločili, da bodo sklici s kazalčnih geselskih sestavkov na dominantne ob koncu redakcije slovarja na podlagi podatkov v dominantnih geselskih sestavkih samodejno generirani. Z vidika oblikovanja XML-sheme je bila odločitev nekoliko zahtevnejša zaradi zadrege, da med samim redakcijskim procesom ni mogoče z gotovostjo napovedati, ali je konkretna iztočnica izhodišče kazalčnega ali dominantnega geselskega sestavka, saj se strukturni tip konkretnega sestavka lahko določi šele ob redakciji veliko različnih iztočnic. Odločitev o tipu geselskega sestavka, katerega izhodišče je posamezna iztočnica, se namreč vzpostavi na podlagi analize sinonimnih razmerij, ki jih vzpostavljajo vsi leksemi, ki izražajo pomene, ki jih je mogoče izraziti tudi s konkretno iztočnico. Navedeno dejstvo ima neprijetno posledico, da se mikrostruktura oz. strukturni tip posameznega geselskega sestavka lahko spremeni praktično v kateri koli fazi priprave slovarskega besedila.

Pripravljavci XML-sheme so bili zaradi te lastnosti slovarja prisiljeni v iskanje smiselnega ravnovesja med dvema nasprotujočima si težnjama. Po eni strani so bili soočeni z zahtevo, da podatkovne strukture geselskih sestavkov obeh tipov oblikujejo čim bolj enotno in z malo shemskimi elementi, da bi bilo popravkov, če bi redaktor v procesu izdelave slovarja ugotovil, da mora strukturni tip specifičnega geselskega sestavka ali del njegove podatkovne strukture spremeniti, čim manj. Po drugi strani je bilo treba upoštevati vidik sklicevanja. Če bi prevladala odločitev za zelo malo različnih elementov XML-sheme, bi se ti lahko sklicevali na več različnih mest v strukturi dominantnih geselskih sestavkov, kar bi bilo z vidika prikaza možnih tarč sklicev v okviru vmesnika leksikografskega programa in izbire ustreznega tarčnega elementa v pogovornem oknu manj primerno, saj je v tem primeru možnosti za napake več, poleg tega pa bi bilo v veliki meri onemogočeno samodejno vzpostavljanje sklicev. Alternativna možnost bi bila vzpostavitev velikega števila podatkovnih tipov, ki bi bili z vidika sklicevanja zamejeni bolj jasno, vendar bi bilo morebitno spreminjanje strukture določenega geselskega sestavka za redaktorje težavnejše. Pri oblikovanju končne različice XML-sheme smo se zato odločili za kompromisno možnost. Različnih elementov XML-sheme, v katere redaktorji dejansko vpisujejo podatke o sinonimnih razmerjih, je razmeroma malo, vendar pa so ti umeščeni v jasno strukturirane različne nadelemente,[8] ki so redaktorjem logični tudi glede na njihovo konceptualizacijo slovarske strukture, znotraj njih pa so omenjeni elementi na vseh nivojih slovarske strukture hierarhizirani na enak način, zato XML-shema za redaktorje spominsko in tehnično ni preveč obremenjujoča.

Elementom, v katere so vpisani (eno- in večbesedni) zaznamovani sinonimi v dominantnih geselskih sestavkih, bodo ob koncu redakcije slovarja samodejno pripisani atributi z unikatnimi oznakami ID. Skupaj z vsebino nekaterih drugih elementov dominantnih geselskih sestavkov, zlasti neonaglašenih iztočnic, jih bomo nato izvozili v ustrezne, zlasti sklicevalne elemente kazalčnih in dominantnih geselskih sestavkov, na podlagi pripisanih atributov ID pa bodo vzvratno samodejno vzpostavljeni

[7] Natančnejše podatke o oblikovanju XML-sheme za ta slovar prinaša prispevek Ledinek et al. (2012).

[8] Ti se v podatkovni hierarhiji pojavljajo razmeroma visoko, strukturiranje v okviru teh elementov, ki obsega več hierarhičnih podnivojev, pa je praviloma enotno, zato je popravkov ob želji po spreminjanju podatkovne strukture sorazmerno malo.

tudi sklici na (nad)elemente slovarske strukture dominantnih geselskih sestavkov, v katerih so se kot nedominantni sinonimi pojavili. Eno- in večbesedni sinonimi, ki so obravnavani v dominantnih geselskih sestavkih v okviru dopolnilnih razdelkov kot delne sopomenke, se sklicujejo na različne segmente dominantnih geselskih sestavkov (npr. na različne pomene). Ker je z vidika koncepta *Slovarja sinonimov slovenskega jezika* večina iztočnic enopomenskih, hkrati pa večina besednih zvez v slovarju ne nastopa več kot enkrat, bo mogoče tudi večino sklicev z elementov v razdelku delne sopomenke vzpostaviti avtomatsko. Ocenjujemo torej, da bo od nekaj sto tisoč sklicev, ki jih bo slovar prinašal, ročno treba povezati le nekaj tisoč sklicev.

Slovar sinonimov slovenskega jezika je začel nastajati po vzpostavitvi koncepta (Ahlin et al., 2003) brez specializiranih leksikografskih orodij. V programu iLEX, podprtem z opisano XML-shemo, smo slovarsko gradivo pričeli redigirati ob koncu leta 2011, trenutno pa se ukvarjamo z začetno fazo pretvorbe že obstoječega slovarskega gradiva v obsegu približno 30.000 geselskih sestavkov v format XML. Pri pretvorbi se težave pojavljajo zaradi občasne nekonsistentnosti v rabi slogov in ločil, več ročnega dela pa bo v nadaljnjih fazah dela potrebnega zaradi preverjanja ustreznosti segmentacije in lematizacije posameznih segmentov večbesednih enot, saj tudi deli večbesednih enot glede na vzpostavljeni koncept slovarja lahko nastopajo kot iztočnice kazalčnih geselskih sestavkov.

3.2. Slovenski etimološki slovar

Slovenski etimološki slovar je eden od slovarskih projektov, ki so z vidika pretvorbe v standardni format XML in oblikovanja ustrezne XML-sheme zanj porajali najmanj dilem. Podatkovna baza slovarja je bila namreč že izhodiščno oblikovana kot strukturirana podatkovna baza v računalniškem programu Eva[9] ter zelo natančno in premišljeno označena, vendar je bilo strukturiranje in označevanje podatkov vzpostavljeno zlasti z mislijo na možnost ustreznega izpisa slovarskih podatkov in njihovega indeksiranja, zato se v nekaterih segmentih pomembno razlikuje od logike običajnega strukturiranja podatkov v formatu XML.[10] Zaradi omenjenih lastnosti baze se vsemu ročnemu delu pri pretvorbi ni bilo mogoče izogniti.

Že vzpostavljene rešitve v podatkovni bazi so v precejšnji meri pogojevale oblikovanje strukture XML-sheme, ki pa je, glede na običajne slovarske sheme, specifična zlasti zaradi same narave etimoloških jezikovnih podatkov. Zaradi »proznega« zapisa geselskega sestavka etimološkega slovarja, v katerem se rekonstruirane oblike, pomeni besed, tujejezične besede in njihovi pomeni, oznake za jezike, iz katerih izhajajo, in »navadno« pojasnjevalno besedilo med njimi pojavljajo v razmeroma nepredvidljivem zaporedju, se je zdelo pri strukturiranju XML-sheme smiselno uporabiti elemente s t. i. mešano vsebino (mixed content). Z vidika nadaljnje izrabe slovarske baze, zlasti za postprocesiranje, je to lahko problematično, po drugi strani pa je jasno, da bo nadaljnja izraba baze manjša kot pri drugih slovarskih podatkovnih zbirkah. Trenutno predvidevamo izmenljivost podatkov med omenjeno bazo in bazama za *Novi slovar slovenskega jezika* in *Slovar novejšega besedja slovenskega jezika*, mogoča pa bo tudi uporaba podatkov v drugih etimoloških in jezikovnozgodovinskih projektih.

XML-shema, ki vključuje veliko elementov z mešano vsebino, je posledično manj restriktivna, zato je tudi funkcionalnost leksikografskega programa, podprtega s tovrstno shemo, v smislu zagotavljanja konsistentnosti geselskih sestavkov in njihove skladnosti z XML-shemo in opozarjanja na nepravilnosti v hierarhični strukturi baze ter formalni vsebini elementov nekoliko okrnjena.

3.3. Slovar slovenskega knjižnega jezika 16. stoletja

Izdelava *Slovarja slovenskega knjižnega jezika 16. stoletja*[11] kot prvega zgodovinskega slovarja slovenskega jezika je eden ambicioznejših leksikografskih projektov v slovenskem okolju. Pripravljalna dela zanj potekajo že dlje časa, sicer pa je leta 2001 izšel poskusni snopič slovarja (Merše et al., 2001), leta 2011 pa *Besedje slovenskega knjižnega jezika 16. stoletja* (Ahačič et al., 2011), v katerem je skupaj z nekaterimi podatki zbrano besedje, ki bo v nastajajočem slovarju slovarsko obdelano. Slovar bo namenjen nekoliko ožjemu krogu naslovnikov, zlasti tistim, ki se znanstveno ukvarjajo z raziskovanjem zgodovine slovenskega jezika.

Predpostavljena hierarhična podatkovna struktura XML-sheme slovarja je v veliki meri prilagojena slovarski strukturi, kot je bila zasnovana ob oblikovanju poskusnega snopiča. Čeprav je bila naknadno sprejeta odločitev, da se izkazovanje podatkov glede na poskusni snopič v precejšnji meri poenostavi, pa XML-shemo slovarja kljub vsemu zaznamuje izredna kompleksnost.[12] Ta je v največji meri posledica potrebe po natančnem in sistematičnem prikazu rabe leksikalnih enot, ki je bila zaradi nekodificiranosti jezika – zapisani oz. knjižni jezik se je v 16. stoletju šele vzpostavljal in uveljavljal – raznolika in neustaljena. Prav navedeno dejstvo narekuje vzpostavitev logike strukturiranja podatkov, ki se precej razlikuje od strukturiranja elementov v drugih slovarskih podatkovnih bazah. Če namreč želimo celostno in sistematično predstaviti vso variabilnost, ki jo v besedilih izpričane leksikalne enote izkazujejo na pomenski in slovnični ravni, pri čemer je seveda treba vedenje leksikalne enote zaradi nekodificiranosti knjižnega jezika opisati z več vidikov, kot je to običajno v razlagalnih slovarjih (npr. tudi z vidika (ne)regularnosti stranskih oblik in naglasa, zapisa skupaj – narazen, zapisa velike začetnice, stabilnosti besednovrstne kategorije, zapisa posameznih glasov v besedi, pomenske neustaljenosti, opozarjanja na

[9] Avtor programa je dr. Primož Jakopin.

[10] Omeniti velja, da je slovarska podatkovna baza nastajala v času, ko XML še ni bil vzpostavljen kot standardni format za zapis slovarskih podatkovnih baz, zato je opisano neskladje v strukturiranju podatkov seveda pričakovano in upravičeno.

[11] Kot nekakšno izhodiščno različico zgodovinskega slovarja za slovenščino bi bilo mogoče dojemati tudi podatkovno zbirko *Jezikovni viri starejše slovenščine IMP* (http://nl.ijs.si/imp/).

[12] Morda velja zgolj za ilustracijo opozoriti, da ima XML-shema za slovar besedja 16. stoletja kljub siceršnji težnji po sorodnem strukturiranju podatkov kot pri nastajajočem enojezičnem razlagalnem slovarju (*Novi slovar slovenskega jezika*) in težnji po poenostavitvah, kjer je to mogoče, približno 2,5-krat več shemskih elementov.

nepričakovan zapis ali napačen zapis), je treba v hierarhičnem smislu podatkovne strukture oblikovati precej bolj plosko, kot je običajno sicer. V nasprotnem primeru bi bilo kombinatoričnih možnosti toliko, da bi bila shema v tehničnem smislu za redaktorje praktično neobvladljiva – onemogočena bi bila orientacija znotraj geselskih sestavkov, hkrati bi se povečalo število napak v bazi in podaljšal čas izdelave slovarja. Opisano dejstvo seveda pomembno vpliva na možnosti vzpostavitve iskalnih možnosti, ki so redaktorjem in uporabnikom slovarja na voljo, na možnosti postprocesiranja podatkovne baze ipd. – vse te možnosti so nekoliko okrnjene.

Poseben izziv je predstavljalo oblikovanje shemske strukture za oblikoslovno zaglavje, ki prinaša podatke o besednih oblikah. Pri obravnavah sodobnega jezika to ne bi predstavljalo posebnih zadreg, obravnava oblik iz 16. stoletja, ko knjižni jezik še ni bil standardiziran in tako enoten, kot je danes, pa se je izkazala za posebej zapleteno tako z jezikoslovnega kot tudi s tehničnega vidika. Zaradi izredno raznovrstnih, številnih in za ta slovarski projekt specifičnih podatkovnih tipov je bil glavni poudarek pri izdelavi XML-sheme zlasti na čim večji preglednosti in praktičnosti sheme za sestavljavce slovarja, obenem pa je bilo treba predvideti možnost kasnejše dopolnitve sheme, ki ne bi pretirano dodatno zapletla vnosa podatkov, če bi se izkazalo, da niso bili predvideni vsi potrebni podatkovni tipi. Danes obstoječih 552 mest najnižjega hierarhičnega nivoja za vnos besednih oblik je namreč obvladljivih samo v logično urejeni strukturi. V to število niso vključeni elementi za dodatne podatke o variantnosti in posebnosti besednih oblik in njihovega zapisa.

Spodnja slika prikazuje shematični zapis oblikoslovnega zaglavja pridevnika *bel* v formatu XML, ki ustreza XML-shemi. Struktura je zaradi preglednosti skrajšana, hkrati pa ne vključuje nekaterih specifičnih elementov, ki opozarjajo na dodatne lastnosti jezikovnih elementov.

```
<oblikoslovno_zaglavje>
 <pridevniško>
  <nedoločna_obl>
   <m>
    <ednina>
     <imenovalnik>b | é/e<i>j</i>/e/ee/ee | l</imenovalnik>
     <rodilnik>b | é/e/e<i>j</i> | liga</rodilnik>
<!-- ... -->
    </ednina>
    <dvojina>
     <imenovalnik>bela</imenovalnik>
<!-- ... -->
    </dvojina>
    <množina>
     <imenovalnik>b | e/é/e<i>j</i> | li</imenovalnik>
<!-- ... -->
    </množina>
   </m>
  </nedoločna_obl>
  <določna_obl>
<!-- ... -->
  </določna_obl>
 </pridevniško>
</oblikoslovno_zaglavje>
```

Slika 1: Zgled strukture oblikoslovnega zaglavja *Slovarja knjižnega jezika 16. stoletja*

Omeniti velja še eno specifično lastnost XML-sheme slovarja, ki je posledica neustaljenosti rabe obravnavanih enot in dejstva, da slovar izkazuje le v besedilih izpričano rabo. Gre za dejstvo, da so skoraj vsi elementi XML-sheme neobvezni. V tehničnem smislu to zaradi številnih možnih kombinacij elementov postavlja zahteve pred snovalce delovnega in tudi končnega izpisa slovarskih podatkov v elektronski ali knjižni izdaji slovarja. Oblikovanje besedila, njegova segmentacija, različna pomagala za uporabnike in uporabljeni simboli morajo namreč biti taki, da redaktorju in uporabniku omogočajo neovirano orientacijo tudi znotraj najkompleksnejših geselskih sestavkov.

3.4. Slovar novejšega besedja slovenskega jezika

Slovar novejšega besedja slovenskega jezika je priročnik v obsegu približno 6000 geselskih sestavkov s podgesli, ki z besediščem, ki se je v slovenščini dokumentirano pojavilo po izidu *Slovarja slovenskega knjižnega jezika*, omenjeni priročnik dopolnjuje. Gre za slovar z – glede na tip slovarja, tj. enojezični razlagalni slovar – razmeroma preprosto mikrostrukturo. Slovarsko besedilo je bilo izhodiščno oblikovano v slovarski aplikaciji SlovarRed,[13] vendar je bilo zaradi njene neprilagojenosti predpostavljeni slovarski strukturi[14] iz baze v zaključni fazi redakcije ponovno izvoženo in dopolnjeno v urejevalniku besedil. Z vidika pretvorbe v standardni format XML so ravno v sicer konsistentno podatkovno bazo naknadno dodani elementi povzročali največ težav, saj je pri njih zaradi manjših nedoslednosti pri kodiranju stilov in ločil nastalo največ napak, ki jih je bilo treba odpraviti ročno. Končna redakcija slovarskega besedila v formatu XML poteka v programu iLEX.

Zaradi že omenjene relativne preprostosti slovarske mikrostrukture se je pri oblikovanju XML-sheme slovarja pojavljalo razmeroma malo dilem. Glede na nekoliko manjši obseg priročnika smo ocenili, da je podatkovna baza manj zanimiva za postprocesiranje kot drugi priročniki inštituta, zato smo se odločili nekatere slovarske razdelke, zlasti zaglavje, ki vključuje podatke o izgovorjavi, tonemskosti itd., segmentirati nekoliko manj podrobno, kot je to sicer običajno za razlagalne slovarje večjega obsega. Tudi sicer so že vzpostavljene rešitve in uporabljena programska oprema – slovarski vir je bil v format XML pretvorjen v zaključni fazi redakcije – precej vplivale na oblikovanje XML-sheme. Kljub temu da je to sicer manj običajno, so bile npr. številke pomenov in homonimov, ki so običajno razumljene kot del izpisa slovarskega besedila, v podatkovno bazo vključene kot podatkovni tip, saj bi bilo treba v nasprotnem primeru za zagotovitev pravilnosti podatkov v zaključni fazi redakcije v okviru programa iLEX vzpostaviti tudi sistem sklicevanja. Razdelek, v katerem so navedeni podatki o etimologiji, je strukturiran podobno kot razlagalni razdelek *Slovenskega etimološkega slovarja* in etimološki razdelek *Novega slovarja slovenskega jezika*, da bi med obstoječimi slovarskimi bazami omogočili izmenljivost podatkov.

[13] Avtor programa je Tomaž Seliškar, vsebinsko zasnovo zanj je pripravila Borislava Košmrlj-Levačič (2004).

[14] SlovarRed je bil izdelan kot specializirani program za oblikovanje terminoloških slovarjev, ne pa kot univerzalni leksikografski program, ki bi podpiral uporabo XML-formata in vključitev shem za formalni opis podatkovnih baz.

3.5. Drugi slovarji

V letu 2011 je bila oblikovana tudi XML-shema za nastajajoči *Novi slovar slovenskega jezika*, enojezični razlagalni slovar v obsegu približno 70.000 gesel, ki naj bi bil nekoliko manj ambiciozen naslednik *Slovarja slovenskega knjižnega jezika*. Pri njeni pripravi smo se srečali z izzivom, kako vzpostaviti XML-shemo, ki bo omogočala ohranjanje leksikografske tradicije predhodnika v segmentih, ki so se izkazali za dobre in ki so jih uporabniki vajeni, in sicer tudi na ravni izmenljivosti podatkov med obema bazama, hkrati pa omogočala vzpostavitev novih leksikografskih praks, ki so se kot ustrezne potrdile v praksi sodobne, tudi tujejezične leksikografije, pri čemer naj bi bila shema oblikovana čim bolj striktno in preudarno, tj. tako, da v čim večji meri preprečuje nesistematično interpretiranje podatkov, hkrati pa njihovo predstavitev na uporabniku čim bolj prijazen način v elektronski obliki.

Uporabniška prijaznost in nedvoumnost predstavitve podatkov je tudi eno od temeljih načel *Slovarja pravopisno težavnega besedja slovenskega jezika*, katerega XML-shema je v fazi testiranja, da bi lahko že oblikovano slovarsko gradivo v obsegu približno 15.000 geselskih sestavkov v prihodnjih mesecih pretvorili v standardni format XML. Slovarsko gradivo je bilo izhodiščno oblikovano kot relacijska baza v programu Mravljica,[15] zato večjih težav pri pretvorbi ne pričakujemo.

V elektronski obliki kot relacijske baze s pomočjo aplikacije SlovarRed že dobro desetletje nastajajo tudi inštitutski terminološki slovarji, ki pa zaenkrat še niso pretvorjeni v standardni format XML.

4. Zaključek

Tehnološki napredek na področju računalništva je odločilno vplival tudi na leksikografijo na prehodu v novo tisočletje. Zaradi napredka informacijske tehnologije, ki pogojuje tako metodologijo sodobnega leksikografskega dela kot konceptualizacijo nastajajočih slovarskih priročnikov, obenem pa tudi dojemanje nastajajočih jezikovnih virov pri uporabnikih, je postalo nepogrešljivo, da so slovarski priročniki in drugi sorodni jezikovni viri izhodiščno oblikovani kot strojno berljive razširljive hierarhično strukturirane podatkovne baze, zapisane v standardnem formatu, saj lahko le tako zagotavljamo izmenljivost, povezljivost podatkov in njihovo večkratno izrabo, s čimer pripomoremo k učinkovitejšemu redakcijskemu procesu, ob ustrezni ciklični aktualizaciji jezikovnih podatkov pa tudi k aktualnejšim slovarskim priročnikom. Pri oblikovanju XML-shem za slovarske podatkovne baze je treba upoštevati številne vidike (konceptualizacija slovarske strukture, kompleksni načini iskanja podatkov, praktičnost za vnos podatkov, postprocesiranje podatkovne baze, možnosti vzpostavljanja sklicev, prikaz podatkov v elektronski in tiskani obliki, značilnosti leksikografskega programa …), zato delo poraja številne dileme, kljub temu pa je vzpostavitev strojno berljivih slovarskih podatkovnih baz v standardnem formatu XML razumljiv korak na poti sodobne slovenske leksikografije.

[15] Avtor programa je Uroš Parazajda, podatkovno strukturo zanj pa sta zasnovali dr. Helena Dobrovoljc in dr. Nataša Jakop.

5. Literatura

ABBYY Lingvo Content <http://www.abbyy.com/lingvo_content/>.

Ahačič, K., A. Legan Ravnikar, M. Merše, J. Narat, F. Novak, 2011. *Besedje slovenskega knjižnega jezika 16. stoletja*. Ljubljana: Založba ZRC, ZRC SAZU.

Ahlin, M., B. Lazar, Z. Praznik, J. Snoj, 2003. *Slovar sinonimov slovenskega jezika: splošna določila in opis zgradbe slovarskih sestavkov z vzorčno predstavitvijo*. Ljubljana: ZRC SAZU, Založba ZRC SAZU.

Erjavec, T., 2012. The goo300k corpus of historical Slovene. In *Eight International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul.

Hunter, D., J. Rafter, J. Fawcett, E. van der Vlist, D. Ayers, J. Duckett, A. Watt, L. McKinnon, 2007. *Beginning XML*. Indianapolis: Wiley Publishing.

IDM DPS <http://www.idm.fr/products/dictionary_writing_system_dps/27/>.

iLEX <http://www.emp.dk/>.

Jezikovni viri starejše slovenščine IMP <http://nl.ijs.si/imp/>.

Košmrlj-Levačič, B., T. Seliškar, 2004. Uporabniški računalniški program SlovarRed 2.0. In M. Humar (ur.), *Terminologija v času globalizacije*. Ljubljana: Založba ZRC, ZRC SAZU.

Ledinek, N., A. Perdih, 2012. Uporaba XML-formata v leksikografiji na primeru oblikovanja XML-sheme za Slovar sinonimov slovenskega jezika. *Jezikoslovni zapiski*, 18/1:157-176.

Lexical Interchange Format Standard (LIFT) <http://code.google.com/p/lift-standard/>.

Lexical Markup Framework (LMF), ISO-24613:2008 <http://www.lexicalmarkupframework.org/>.

Merše, M., F. Novak, F. Premk, 2001. *Slovar jezika slovenskih protestantskih piscev 16. stoletja : poskusni snopič*. Ljubljana: Založba ZRC, ZRC SAZU.

Smrž, P., 2001. Slovníková data ve formátu XML. In A. Jarošová (ur.), *Slovenčina a čeština v počítačovom spracovaní. Zborník referátov zo seminára. Bratislava 26. – 27. októbra 2001*. Bratislava: Veda.

Snoj, M., 2003. *Slovenski etimološki slovar*. Ljubljana: Modrijan.

Standard XML <http://www.w3.org/standards/xml/>.

TEI Consortium, 2007. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. <www.tei-c.org/release/doc/tei-p5-doc/en/html/>.

Termania <http://www.termania.net>.

Thompson, H. S., D. Beech, M. Maloney, N. Mendelsohn, 2004. *XML Schema Part 1: Structures: W3C Recommendation 28 October 2004*. <http://www.w3.org/TR/2004/REC-xmlschema-1-20041028/>.

TshwaneLex <http://tshwanedje.com/tshwanelex/>.

w3schools.com <http://www.w3schools.com/>.