UDK 004.056.55:886.3

Literatūrinio paveldo skaitmeninimas, taikant atviruosius standartus

Tomaž ERJAVEC
Žinių technologijų skyrius, Jozefo Stefano institutas, Liubliana, Slovėnija, el. p. tomaz.erjavec@ijs.si
Matija OGRIN
Slovėnų literatūros ir literatūros mokslų institutas, Slovėnijos mokslų ir menų akademijos Mokslinio tyrimo centras, Liubliana, el. p. matija.ogrin@zrc-sazu.si

Straipsnyje pristatoma jungtinio Slovėnijos projekto metu taikyta metodika bei technologija ir pasiekti rezultatai. Šiuo projektu siekta literatūrinio palikimo tekstų peržiūrėtas ir pataisytas laidas publikuoti internete. Medžiagos yra nepaprastai daug, nes dažnai prieinama ne tik dokumento faksimilė, bet ir keletas tarpusavyje susijusių jo transkripcijų variantų, užrašai, glosarijai, žodynai, nuorodos į išorinius išteklius, multimedijos pateiktys ir t. t. Svarbiausi medžiagos rengimo etapai yra du: pirmiausia dokumentai paverčiami į kanoniškąjį ir standartizuotą variantą XML formatu, pritaikant Tekstų kodavimo iniciatyvos metodiką, o paskui šis atminties formatas atverčiamas į HTML kalbą ir publikuojamas. Darbo eiga priklauso nuo to, kokiu būdu vartojami atvirieji standartai ir kaip intensyviai tarpusavyje bendradarbiauja turinio ir technologijų tiekėjai. Taip pat supažindinama su elektroniniais leidiniais, kuriuos šiuo metu galima rasti projekto tinklalapyje, bei aptariama tolesnė veikla, o ypač kalbinių technologijų pritaikymas publikavimo procese.

Reikšminiai žodžiai: atvirieji standartai; skaitmeninimas; literatūrinis paveldas.

1. Įvadas

Įvairių rūšių tekstai – religiniai, literatūriniai, istoriniai ir t. t. – yra kultūrinės atminties priemonės. Jei Europos kultūrai svarbūs tekstai nebūtų išlikę, žmonės nežinotų savo kilmės ir esmės. Jie prarastų svarbiausią savo individualybės dalį. Tačiau tokie tekstai negali atlikti savo funkcijos, jei jie nesuprantami. O dėl istorinių priežasčių šie tekstai dažnai nesuprantami originalia forma, nes yra parašyti senovine kalba ir senoviniais rašmenimis arba paprasčiausiai dauguma skaitytojų nemoka perskaityti senų rankraščių. Tekstų atrankos ir redagavimo tikslas yra panaikinti šiuos sunkumus, dažnai pasitelkiant pagalbinius istorijos mokslus, tokius kaip paleografija.

Tam, kad senieji tekstai taptų suprantami, pirmiausia būtina parengti peržiūrėtą ir pataisytą laidą (vadinamąją *editio maior*), kurioje pateikiami nuo pirminių, jei reikia, pertvarkytų šaltinių kruopščiai nurašyti tekstai ir jų komentarai. Pagal šią laidą vėliau išleidžiamos supaprastintos komercinės laidos (*editio minor*). Tačiau norint išleisti peržiūrėtas ir pataisytas laidas, kurias sudaro faksimilės, transkripcijos, informacija apie teksto variantus ir (jei reikia) vertimai į šiuolaikines kalbas, susiduriama su didelėmis finansinėmis kliūtimis, ypač šalyse, kurių knygų rinkos nėra didelės, pvz., Slovėnijoje. Senesnių slovėniškų tekstų pataisytas laidas būtų geriausia publikuoti skaitmeninė forma ne tik dėl to, kad pasiekiamas geresnis pelno ir sąnaudų santykis, bet ir dėl galimybių, kurias tekstų pertvarkymui, analizei ir pateikimui siūlo skaitmeninė terpė.

Šiam tikslui būtina sukurti specialią metodiką, kuri ir yra pagrindinė šio straipsnio tema. Ji turėtų apimti specifines slovėnų literatūros redagavimo problemas, tačiau ji turi būti paremta atviraisiais tarptautiniais standartais bei metodika, skirta tekstų kodavimui ir keitimuisi jais. Būtent šito ir siekia jungtinis projektas *Elektronske znanstvenokritične izdaje slovenskego slovstva* (http://nl.ijs.si/ezrc/). Šiame straipsnyje supažindinama su pirmojo projekto etapo metu sukurta metodika ir pagrindiniais iki šiol pasiektais rezultatais. Tikimasi, kad pirmosios laidos atspindės svarbiausią šio projekto principą: pritaikyti tradicinius tekstų peržiūrėjimo ir redagavimo standartus, tinkančius seniausiems tekstams slovėnų kalba, elektroninėje terpėje,

t. y. sujungti tradicinius redagavimo ir šiuolaikinius tekstų kodavimo standartus.

2. Uždaviniai

Svarbūs kultūros paveldo dokumentai, tokie kaip senieji literatūriniai tekstai, ypač tinka publikuoti daugialypėje terpėje. Jie tinkami vartoti mokymo tikslams, gali padėti moksliniame darbe, pavyzdžiui, literatūros studijose, lingvistikoje, istorijoje ir t. t. Tačiau pirmiausia tekstą (ir su juo susijusią garso ir (arba) vaizdo medžiagą) reikia išanalizuoti, užkoduoti ir pateikti tokiu būdu, kuris geriausiai perteiktų saugomą informaciją. Pagal tai įvardijami projekto tikslai:

– Surinkti originalias tekstų formas (faksimiles) ir sudėti kartu su smulkiai išnagrinėtomis transkripcijomis, informacija apie tekstų variantus ir galimais vertimais, papildant kitomis raiškos priemonėmis (ištraukomis, vaizdo ekranizacijomis), ir susieti šias teksto pateiktis bei nurodyti išorines sąsajas į susijusius išteklius (nuorodas, istorinius faktus).

– Tekstų kodavimas tokiu formatu, kuriuo įmanoma parodyti redaguotos laidos teksto sudėtingumą, yra veiksnys, labiausiai nepasiduodantis technologiniams pokyčiams, tad jį galima perkelti iš vienos kompiuterinės formos į kitą bei panaudoti lengvai užrašomais ir suprantamais būdais.

– Sukurti prieigą prie tekstų, atskleidžiančią ir palyginančią įvairias tekstų pateiktis ir galimą pritaikyti specifiniams poreikiams bei vartotojų grupėms. Mokomasis aspektas yra ypač svarbus, nes vertindami istorines senųjų tekstų ypatybes ir susipažindami su rašytinės kalbos raida moksleiviai ir studentai gaus galimybę palyginti originalų kūrinį arba jo transkripciją su vertimu į šiuolaikinę kalbą. Poveikis mokymo procesui bus dar stipresnis, jei naudojami garso įrašai ir kitos raiškos priemonės.

3. Metodika

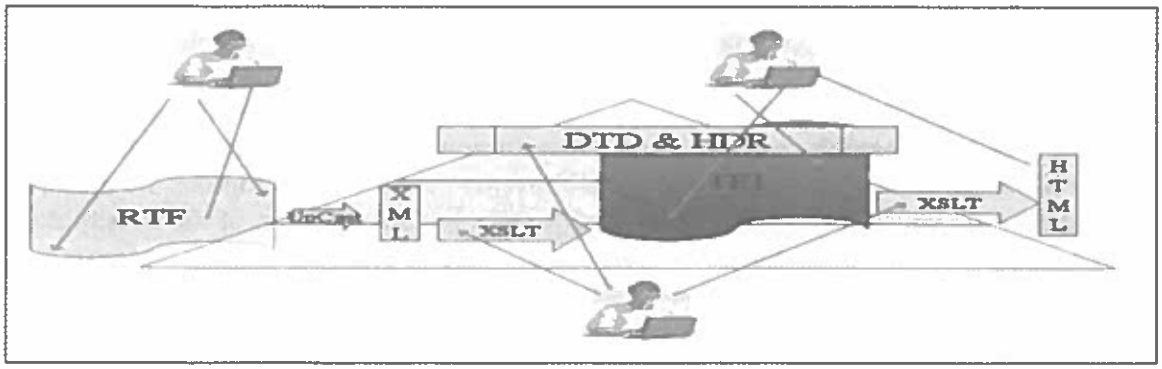
Atsižvelgiant į žmogiškuosius išteklius, šio projekto struktūra yra išties retai pasitaikanti: vienas projekto partneris yra didelis tekstų atrankos ir redagavimo srities žinovas, nors ši procedūra paprastai atliekama klasikiniu būdu, kai kompiuteriu naudojamasi tik tekstams apdoroti, tuo tarpu kitas partneris yra kalbų inžinerijos specialistas, turintis ypač daug patirties, sudarant anotuotų tekstų rinkinius. Kadangi vienas partneris daugiausia užsiėmė turinio teikimu ir rezultatų patvirtinimu, o kitas diegė formalią struktūrą ir į ją arba iš jos vertė turinį, prireikė sukurti vientisą formą, kuri paveiktų kiekvienos skaitmeninės laidos rengimą.

Rengimo metu taikoma metodika pavaizduota 1 pav. Pagrindinis jos tikslas – sukurti kanonišką, standartizuotą kiekvieno dokumento laidą, saugomą XML formatu pagal metodikoje [1] nurodytus kriterijus ir specifikacijas, skirtas moksliniam tekstų kodavimui (žr. 4 skyrių). Medžiaga parengiama pirmiausia paverčiant originalų skaitmeninį dokumentą į TEI/XML formatą, o po to šis formatas atverčiamas į HTML.

3.1. Medžiagos parengimas

Pirmiausia teksto redaktorius atlieka išsamią teksto atrankos bei redagavimo analizę ir parengia teksto transkripciją. Kadangi paprastai egzistuoja tik vienas seniausių slovėnų tekstų variantas, laidose siekiama ne sugretinti variantus, o parodyti rankraščio originalą, kuriame išryškėja originali teksto istorinė gramatika, leksika ir rašyba. Šiuo etapu būtina nustatyti, kurias teksto ypatybes reikėtų pabrėžti, t. y. kokią „informaciją reikėtų užkoduoti tekste tokiu būdu, kad programa galėtų ją rasti“ [2].

Išanalizavus ir parengus tekstą, o jo transkripcijas, pataisymus ir pastabas surašius teksto redaktoriuje, dažniausiai „Word“, medžiaga pakeičiama į kanoninį



1 pav. Medžiagos parengimo operacijų eiga: horizontalioji ašis parodo laiką/pastangas, reikalingas ištekliui sukurti, vertikalioji ašis – ištekliaus naudingąją informaciją

formatą (vartojant XSLT lentelių stiliaus redaktorių, žr. 4 skyrių). Pakeitimas vyksta cikliškai, nes duomenys yra automatiškai koduojami TEI (*Text Encoding Initiative*), taikant paskirtąją keitimo priemonę; rezultatas pateikiamas HTML kalba, naudojant stiliaus lentelę, ir galiausiai įvertinamas. Paprastai pasitaiko trijų rūšių klaidos: 1) klaidos originaliame faile; 2) klaidos, atsiradusios konversijos metu; 3) klaidos, atsiradusias kodavimo metu. Jei pasitaiko pirmosios rūšies klaidų – taisomas originalus failas, antrosios rūšies – keičiami veiksmai, o trečiosios rūšies – XML schemos duomenų reikšmė, kuri nurodo, koks elementų žodynas taikomas. Ištaisius pastebėtas klaidas, originalus failas dar kartą konvertuojamas, o ciklas pakartojamas. Toks greitas prototipinis metodas skatina bendradarbiauti ir keistis patirtimi.

Kai tam tikrai laidai skirta medžiaga yra sutvarkoma tiek, kad turinys ir anotacijos dar gali būti taisomi „Word" redaktoriumi, skaitmeninis originalas yra atidedamas ir lieka tik kanoninė TEI versija. Toliau peržiūrima ir tiesiogiai XML redaktoriuje taisoma tik kanoninė versija. Nuo šiol asmuo, redaguojantis medžiagą, turi būti susipažinęs su XML ir TEI kodavimo schemos sąvokomis. Taip pat rašomi ir laidos metaduomenys, t. y. medžiagos aprašas, jo šaltiniai ir pritaikyta kodavimo technika. Visa ši informacija įrašoma į TEI antraštę.

3.2. Medžiagos pateikimas

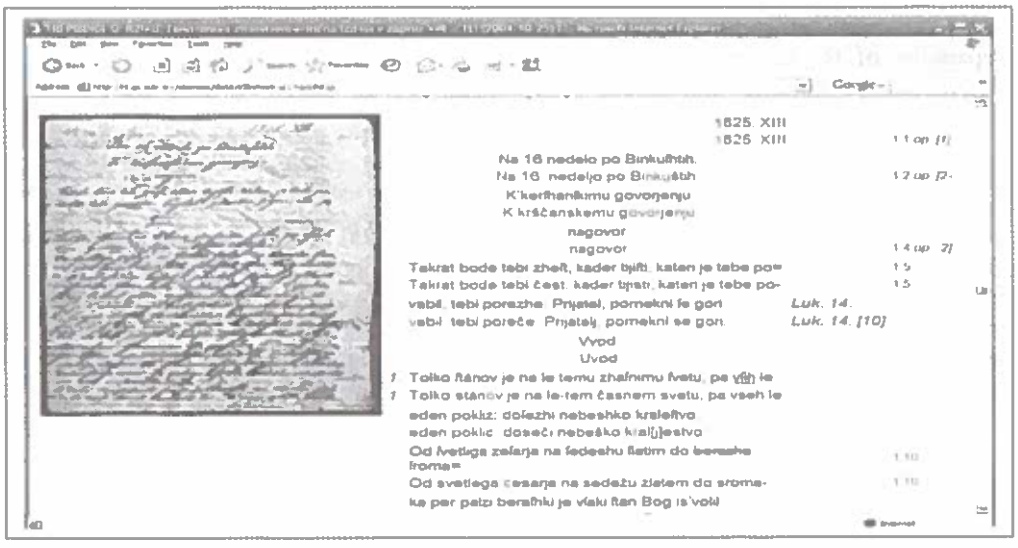
Kai medžiagos kanoninis atminties formatas parengtas, lieka išspręsti klausimą, kaip geriausiai ją panaudoti. Šiuo etapu svarbiausia ne parengti spausdintines versijas, o publikuoti medžiagą žiniatinklyje ir, jei yra tokia galimybė, įrašyti į CD-ROM.

Šiuo metu galima pamatyti po vieną kiekvienos knygos HTML atvaizdą, kuris pagaminamas atjungties būsena, taikant XSLT stiliaus lentelę. 2 pav. matyti, kad pats HTML formatas yra sudėtingas ir atvaizduoja originalą kaip galima tiksliau (pagal struktūrą, išdėstymą, išskyrimą, pataisas ir pakeitimus), pateikia lygiagrečius įvairių transkripcijų atvaizdus bei skaitmeninius priedus, pvz., tikslią nuorodą į Biblijos ištrauką, nurodomą tekste. Stengtasi, kad būtų pateikiama pagal tiesiogiai susijusius prieigos standartus [3].

Dabartinis „kiekvienos knygos vieno HTML atvaizdo" scenarijus naudingas tuo, kad jį galima pamatyti per bet kokį HTML serverį ir publikuoti tiek žiniatinklyje, tiek išleisti CD-ROM. Tačiau jo neįmanoma pritaikyti skirtingoms reikmėms ir vartotojų grupėms arba pasinaudoti visomis kitomis XML formato anotacijų siūlomomis galimybėmis. Šie patobulinimai bus atliekami ateityje, tačiau jau atsiranda programinė įranga, kuri yra specialiai pagaminta taip, kad suteiktų galimybę atvaizduoti sudėtingus TEI koduotus redaguotų laidų atvaizdus [4]. Todėl šiuo etapu protingiau susitelkti ties turiniu, o ne ties pateikimo būdu.

4. Technologijos apibūdinimas

Pagrindinė skaitmeninių laidų kūrimui naudojama technologija yra XML. Ji apima XML elementų žodyno apibrėžimą (XML DTD specifikacija), pirminės ir antrosios konversijos filtrų sukūrimą (XSLT), naudojant XML redaktorių, ir specialių simbolių istorinių tekstų transkripcijose vartojimą. Šiame skyriuje smulkiai apibūdinami šie technologiniai aspektai, taikomi rengiant ir pateikiant medžiagą.



2 pav. Pateikties HTML formatu pavyzdys

4.1. TEI dokumento rūšies apibrėžimas

XML schema apibrėžia elementų žodyną, tinkamą tam tikrai dokumentų rūšiai, bei tarpusavio santykius tarp šių elementų. Tai svarbiausias XML taikymo aspektas, ypač gamybos etapu, nes formatas nustato pavartotų elementų semantiką ir suteikia galimybę oficialiai patvirtinti pagamintus ir pažymėtus dokumentus. Nors įmanoma sukurti ypatingą schemą, kuri patenkintų konkrečią laidą ar projekto poreikius, šis procesas negali būti paviršutiniškai atliekamas, ypač rengiant tokią sudėtingą medžiagą kaip redaguotos laidos. Todėl daug lengviau bei rezultatyviau pritaikyti standartinę schemą konkrečiai dokumento rūšiai, jei tokia atsiranda.

„Metodikoje" [1] pateikiama specifikacija, apibrėžianti mokslinį tekstų kodavimo procesą. TEI yra atviroji ilgą laiką didelės vartotojų grupės taikoma kodavimo sistema, kuri tapo standartu „de facto". Ji apima didelę tekstų ir anotacijų rūšių įvairovę: siūlomi žymenų rinkiniai, skirti prozos, poezijos, dramos kūriniams ir žodynams, redaguotiems seniesiems tekstams, pirminių šaltinių transkripcijoms, lingvistinėms analizėms ir t. t. Taip pat pateikiamas antraštės apibrėžimas, suteikiantis galimybę įterpti detalius metaduomenis bei pritaikyti išplėtimo ir modifikacijos mechanizmus.

Apibendrinant TEI, ši sistema nereiškia vienos schemos, kuri apimtų visas dokumentų ir anotacijų rūšis. Dabartinė versija (TEI P4) siūlo daugybę modulių, kuriuos galima sujungti ir toliau tobulinti iki tol, kol sudaroma konkreti schema, įvardijama kaip XML dokumento rūšies apibrėžimas (*Document Type Definition – DTD*). Šiam projektui pasirinkti tokie moduliai:

– *TEI.prose*, skirtas prozos kūriniams koduoti. Jis susideda iš elementų, sudarančių TEI antraštę, pateikiančią detalius metaduomenis apie dokumentą (tokius kaip failas, šaltinis, koduoti ir peržiūrėjimo aprašai), taip pat iš standartinių elementų, skirtų dokumento sandarai (skyrius, paragrafas, lentelė, pastaba, ...), bei specifinių ženklų (kirtis, išryškinimas, ...);

– *TEI.transcr*, papildomas modulis, skirtas pirminių šaltinių, ypač rankraščių, kuriuose esama taisytinų arba keistinų elementų, skirtingomis rašysenomis parašyto teksto įrašų ir t. t., perrašymui;

– *TEI.linking*, papildomas modulis, leidžiantis sukurti sąsajas pačiame dokumente ir su kitais dokumentais bei susidedantis iš elementų ir atributų, kurie susieja skirtingas medžiagos transkripcijas ir pačią medžiagą su išoriniais ištekliais;

– *TEI.figures*, papildomas modulis, skirtas koduoti atvaizdams ir kitai grafiniai medžiagai, vartojamas faksimilei koduoti, t. y. ryšiams su grafiniais failais, kuriuos sudaro faksimilė, pateikta įvairiais dydžiais ir raiškomis;

– *TEI.extensions*, specialiai vartotojui skirtas modulis, kuris įveda vartotojo plėtinius į standartinį TEI. Jis nustato papildomus elementus ir išvardija atributų reikšmes, pvz., informacijos pateikimą pastabose.

Tačiau net ir nustačius TEI parametrus konkrečiam projektui (t. y. pasirinkus reikalingus modulius ir plėtinius, kurie leidžia sukurti XML DTD), vis dar lieka pakankamai laisvės rinktis, kurie elementai bus konkrečiai vartojami. Į tokį TEI DTD, bent jau šiuo metu naudojamą versiją P4, taip pat įeina nemažai medžiagai nereikalingų elementų ir atributų. Toks per daug informacijos apimantis DTD yra naudingas patvirtinant medžiagą ir ja keičiantis, tačiau neigiamai veikia autorizavimo procesą, nes apsunkina DTD atpažįstančio ir meniu valdomo XML redaktoriaus naudojimą tvarkant medžiagą. Dėl šios priežasties, nustačius TEI parametrus, buvo sukurtas griežtas (minimalus) DTD pritaikytas tik prie TEI kodo. Toks DTD buvo naudojamas kūrimo metu, o galutinėje medžiagos versijoje vėl sugrįžta prie „oficialaus" su TEI suderinto DTD.

3 pav. pateikiamas pagal šiame projekte naudotą DTD koduotos medžiagos pavyzdys. Pavyzdyje matyti, kad elementai gausiai koduoti, kiekvienam skyriui, puslapiui ir eilutei priskirtas atskiras identifikavimo numeris bei nuoroda į jį atitinkantį kitos transkripcijos identifikavimo numerį.

```
<div id="sl1d" corresp="sl1k" n="1" type="dipl">
  <head>Diplomatični prepis</head>
  <page id="sl1d-f.1" corresp="sl1f.1" n="1">
    <line id="sl1d.1" corresp="sl1k.1" n="1" rend="right">1825. XIII</line>
    <line id="sl1d.2" corresp="sl1k.2" n="2" rend="center">Na 16 nedelo po Binkuſhtih.</line>
    <line id="sl1d.3" corresp="sl1k.3" n="3" rend="center">K'kerſhanſkimu govorjenju</line>
    <line id="sl1d.4" corresp="sl1k.4" n="4" rend="center">nagovor.</line>
    <line id="sl1d.5" corresp="sl1k.5" n="5">Takrat bode tebi zheſt, kader tijifti, kateri je tebe
po&#301F;</line>
    <line id="sl1d.6" corresp="sl1k.6" n="6">vabil, tebi porezhe: Prijatel, pomekni ſe gori.
      <note place="right">Luk. 14.</note></line>
    <line id="sl1d.7" corresp="sl1k.7" n="7" rend="center">Vvod.</line>
    <line id="sl1d.8" corresp="sl1k.8" n="8"><note place="left">1.</note> Tolko ſtánov je na le temu
zhaſnimu ſvetu, pa <emph>vſih</emph> le</line>
    <line id="sl1d.9" corresp="sl1k.9" n="9">eden pokliz: doſezhi nebeshko kraleſtvo.</line>
    <line id="sl1d.10" corresp="sl1k.10" n="10">Od ſvetliga zeſarja na ſedeshu ſlatim do
      <del hand="AMS">berazha</del> froma&#301F;</line>
    <line id="sl1d.11" corresp="sl1k.11" n="11">ka per palzi berafhki je vſaki ſtan Bog is'volil</line>
...
```

3 pav. Faksimilinė transkripcija TEI/XML formatu – plg. 2 pav.

4.2. XSLT filtrai

Kaip jau minėta, yra du etapai, kurių metu konversijos scenarijai taikomi duomenims, t. y. pirminė konversija, kai skaitmeninis šaltinis verčiamas į TEI XML, ir antrinė konversija, kai TEI XML verčiamas į HTML.

Pirminės konversijos metu duomenys pirmiausia paverčiami į „įprastą" formatą, pvz., iš „Word" į XML, naudojant vieną iš daugelio keitiklių (pvz., Open Office arba UpCast), RTF formatą keičiančių XML formatu. Po to kiekvienai laidai parašomos tam tikros formos, pagal kurias pateikti skirtas šaltinis XML kanale paverčiamas tiksline TEI koduote. Šie filtrai daugiausia parašyti XSLT kalba, XML transformavimo kalba, kuri taip pat yra W3C rekomendacija, ir dėl to susidaro standartizuotos specifikacijos, kurios įgyvendinamos naudojantis įvairiomis priemonėmis, pvz., „IE Explorer". Tačiau nors XSLT idealiai tinka XML struktūros konversijoms koduoti, ji mažiau tinka tais atvejais, kai tam tikri eilučių pavyzdžiai turi pradėti kurti XML struktūras. Tokiems atvejams filtrai yra užrašyti „Perl" programavimo kalba.

Antrinė konversija, paverčiant medžiagą atskirai iš HTML, taip pat atliekama taikant XSLT, nors šiuo atveju naudojama ir nemaža dalis kitų kodų. Be to, grafiškai vaizduojant įvairius TEI šaltinio elementus (pavyzdžiui, TEI <del> elementus keičiant į HTML <s> elementus, t. y. perbraukiant), antrinės konversijos metu taip pat sudaromas turinio sąrašas ir, svarbiausia, pagaminamas greta pateiktų faksimilės ir teksto atvaizdas, sugretinus skirtingų transkripcijų atvaizdus.

Mokslininkams, norintiems turėti daugiau informacijos apie (skaitmeninę) laidą, būtina žinoti, kaip HTML vaizduojama TEI antraštė. XSLT šablonas išplečia antraštės žymenas į jų lokalizuotus eilučių aprašus (pvz., <respStmt> į „atsakomybės duomenis"), o paskui susieja kiekvieną žymeną su jos apibrėžimu, pateiktu TEI gairėse. TEI antraštę sudaro ir dokumento pagrindinėje dalyje vartojamų žymenų sąrašas, taigi visi medžiagoje vartojami elementai turi tiesiogiai prieinamą dokumentaciją.

Kadangi medžiagą galima peržiūrėti žymiai lankstesniu būdu, šiuo metu vartojamas modelis yra parankus tuo, kad jį galima taikyti dirbant atjungties būsena (visos TEI gairės taip pat atsispindi kiekvienoje knygoje), todėl nereikia specialios programinės įrangos.

4.3. Ženklų kodavimas

Tam tikra problema kyla dėl sudėtingų (istorinių ir fonetinių) ženklų, kuriuos reikia vartoti perteikiant medžiagą. Didžiąją dalį tokių ženklų galima perteikti taikant Unicodą, bet ne visada. Tada vartojama laisva Unicodo kodų sritis (privačioji sritis) ir specialus viešai prieinamas ZRCola ženklų rinkinys ir šriftas [5].

5. Rezultatai

Svarbiausias šių pastangų rezultatas – slovėnų literatūros kūrinių redaguotų laidų internetinė biblioteka (http://nl.ijs.si/e-zrc/). Šiuo metu ją sudaro pirmosios trys redaguotos elektroninės laidos, o konkrečiai – Anton Martin Slomšek (1800–1862) *Tri pridige o jeziku* („Trys pamokslai apie kalbą") kartu su faksimile, tikslia ir redaguota transkripcijomis bei pastabomis [6], Sigismund Zois (1747–1819) korespondencijos dalis kartu su faksimilėmis, tikslia transkripcija ir vertimu į slovėnų kalbą (korespondencija yra vokiškai) bei pastabomis ir hipersaitiniu laiškuose minimų asmenų vardų žinynu ir Alojz Gradnik (1882–1967) poezijos rinkinys, kurį sudaro 15 skirtingų rinkinio variantų transkripcijos (įvairūs spaudiniai ir autoriaus pataisymai). Šiuo metu rengiama ir nemažai kitų laidų, iš kurių svarbiausia ir sudėtingiausia yra Freisingo rankraščiai (972–1039), t. y. trys religiniai tekstai, kurie yra seniausi tekstai slovėnų kalba ir seniausi slaviški tekstai, užrašyti lotynų abėcėle. Dėl jų svarbos redaguotas laidas sudarys daugybė elementų: faksimilės, tiksli, redaguota ir fonetinė transkripcijos, vertimai į lotynų, senąją bažnytinę slavų ir penkias šiuolaikines Europos kalbas; žodynas, kurį sudarys redaguotos transkripcijos, o prie kiekvieno žodyno žodžio pateikiama fonetinė forma, gramatinė informacija, vertimai, žodžiai su citatomis iš redaguotos transkripcijos ir t. t. Papildomai spausdintoje laidoje pateikiamos įžangos, pastabos ir bibliografija.

Išskirtinis bibliotekos bruožas yra laisva prieiga prie medžiagos (išskyrus A. Gradnik eilėraščius, nes tai draudžia autoriaus teisių apribojimai) HTML formatu ne tik naršant, bet ir parsisiųsdinant ją kaip TEI/XML šaltinį kartu su grafiniais faksimilės failais. Laisva prieiga prie originalių tekstų galima dėl to, kad jie paprastai yra daugiau nei 100 metų senumo, ir transkripcijų ir žymenų autoriai bei redaktoriai sutiko suteikti laisvą prieigą.

6. Privalumai

Didžiausias elektroninės bibliotekos privalumas yra tas, kad ji lengvai prieinama, todėl nereikia pirkti ar skolintis knygos, o kiekvienas, prisijungęs prie interneto, bet kuriuo metu gali peržiūrėti medžiagą. Be to, multimedija pateikti tokie tekstai aiškiau atskleidžia istorinę patirtį ir yra patrauklesni nei spausdintinė jų forma. Reikšmingas elektroninių laidų privalumas yra tas, kad lygiagrečiai pateikiamos įvairios tekstų transkripcijos, o originalus tekstas sugretinamas su labiau suprantamomis jo formomis. Šiuo metu priemonės toliau tobulinamos, o konkrečiai Freisingo rankraščių kiekvieno fragmento garso įrašais. Tad moksleiviai ir studentai galės aiškiai, o tuo pačiu ir daugiaprasmiškai bei dinamiškai suvokti istorinį kalbos ir tautos kultūros vystymąsi. Autorių nuomone, būtent tokie

yra labiausiai akivaizdūs sudėtingo redaguoto senovinio teksto elektroninės laidos, skirtos skirtingo lygio mokymo tikslams bei tolesniems tyrimams, privalumai.

7. Išvados

Senųjų tekstų ir istorinių dokumentų komunikacinę vertę galima geriausiai atskleisti, susiejant į vieną visumą jų vaizdinius, visatekščius ir garsinius pateikimo būdus. Šį uždavinį, reikšmingą tokioms sritims kaip švietimas, muziejininkystė, archyvai, humanitariniai mokslai ir t. t., galima pasiekti suskaitmeninus ir užkodavus medžiagą, griežtai taikant atviruosius kodavimo standartus ir publikuojant gautą medžiagą internete. Metodikos ir technologijos, kuriomis siekta šį uždavinį įgyvendinti, yra: 1) pirminė konversija į TEI/XML/Unicodą bendrojo ir cikliško tobulinimo proceso, daugiausia įgyvendinto taikant XSLT transformavimo priemones, metu; 2) antrinė konversija į vartotojui patogų HTML, pateiktą viešai prieinamame URL. Tokia metodika buvo išbandyta trijose užbaigtose elektroninėse laidose. Šiuo metu rengiamos dar kelios.

Siekta, kad laidos būtų kuo naudingesnės jų prieinamumo prasme, pasireiškiančia lengva prieiga prie medžiagos ir neribotomis galimybės kopijuoti dokumentus pirminiu XML formatu. Šis metodas veikia panašiai kaip gerai žinoma atviroji programinė įranga, kuri leidžia dokumentus leisti kaip komercinius, o šiuo atveju – spausdintas ar CD-ROM laidas. Kadangi galimybė gauti atsiperkančių pajamų yra beveik lygi nuliui ir šį projektą finansuoja vyriausybė, maksimaliai atviras kreipties leidimas, atrodo, yra geriausias būdas siekti vieno svarbiausių projekto tikslų, t. y. padaryti laidas kuo labiau prieinamas.

Kaip jau minėta, toliau bandoma papildomai įtraukti garso įrašus, t. y. planuojama sujungti rankiniu būdu suskirstytus garso failus su įvairių etapų elementais tekste (pvz., žodžiais išreikšti ribas, pažymėtas fonetinėje Freisingo rankraščių transkripcijoje) ir rezultatus paversti HTML formatu. Vėliau, jei multimedijos turinys bus sudėtingesnis, planuojama pritaikyti SMIL standartą [7]. Be garso, pirmą kartą planuojama pabandyti įtraukti į baroko laikotarpio Škofja Loka „Passion" pjesės laidą ir vaizdo įrašą. Tolesniame darbe pageidaujama įtraukti ir ženklinimą, skirtą lingvistinei tekstų struktūrai pateikti [8, 9]. Tokios priemonės leis įvesti tekstus į interneto pagrindu sukurtą atitinkamą mechanizmą (kaip jau įgyvendinta kitiems rinkiniams: http://nl2.ijs.si/), taip pat iš transkripcijų išskirti lygiagretų žodyną. Visomis šiomis analitinėmis priemonėmis siekiama perteikti skaitmeninėse laidose išspausdintų tekstų sudėtingumą, mokslinį potencialą ir istorinę vertę.

Iš anglų kalbos vertė S. Racevičiūtė

Versta iš: Innovation and the knowledge economy: issues, applications, case studies. Amsterdam [etc.], 2005. Pt. 2, p. 999-1006.

[1] Sperberg-McQueen, C.M., Burnard, L. (eds.) (2002). Text Encoding Initiative: Guidelines for Electronic Text Encoding and Interchange, TEI P4, the XML-compatible edition. TEI Consortium. http://www.tei-c.org/P4X/

[2] Hockey, S. (2000). Electronic Texts in the Humanities. Principles and Practice. Oxford University Press.

[3] Wymer, K. (2005). Why Universal Accessibility Should Matter to the Digital Medievalist. *Digital Medievalist* 1(1). http://www.digitalmedievalist.org/

[4] Schreibman, S., Kumar, A., McDonald, J. (2003). The Versioning Machine. *Literary and Linguistic Computing* 18(1), 101-107. http://mith2.umd.edu/products/ver-mach/

[5] Weiss, P. (2004). Vnašalni sistem ZRCola. (The text input system ZRCola) In: *Language Technologies: Proceedings B of the 7th Intl. Conf. Information Society*, IS 2004. Ljubljana: Jožef Stefan Institute, p.124. http://zrcola.zrc-sazu.si/

[6] Erjavec, T., Ogrin, M., Faganel, J. (2005). E-Slomšek: A TEI Encoding of a Critical Edition of 19th Century Slovenian Rhetoric Prose. *Review of the National Center for Digitization*. 6(4), 31-41.

[7] W3C. The Synchronized Multimedia Integration Language (SMIL). http://www.w3.org/AudioVideo/

[8] Erjavec, T. (2002). The IJS-ELAN Slovene-English Parallel Corpus. *International Journal of Corpus Linguistics*. 7(1), 1-20.

[9] Erjavec, T., Džeroski, S. (2004). Machine Learning of Morphosyntactic Structure: Lemmatising Unknown Slovene Words. *Applied Artificial Intelligence* 18(1), 17-40.