ANALI
Zavoda za znanstveni i umjetnički rad u Osijeku
Sv. 29, str. 197-210, Zagreb – Osijek 2013.

Stručni članak

UDK [002:008]:004

Mario Essert*, Vlado Cingel**, Nikola Glumac***,
Mario Lončarić**** i Božidar Štimac*****

TEIMARK PROGRAM - OBRADBA TRANSKRIBIRANE STARE KNJIŽNE GRAĐE

Sažetak

U hrvatskim se institucijama čuva poveći korpus digitalizirane baštine koji je u svrhu predstavljanja često dostupan preko interneta. Digitalizirani dokumenti čuvaju tako istinu o bogatoj hrvatskoj kulturnoj i pisanoj baštini. Nakon izgradnje programa DocMark, čija je svrha označivanje digitaliziranih slikovnih dokumenata u svrhu analize tih oznaka na pojedinačnim dokumentima i/ili njihove usporedbe, kao rezultat trogodišnjeg rada načinjen je TEIMark – program za označivanje teksta: utipkanoga, transliteriranoga ili strojno prepoznatoga. Dok je označivanje u DocMarku izvedeno nad slikom dokumenta (u kojem su zanimljiva i dohvatljiva materijalne osobine/vlastitosti, ali ne i sadržaj teksta), u TEIMarku se označivanje odvija nad stvarnim tekstom, a ne njegovom slikom, čime su omogućena lingvistička i druga istraživanja usmjerena na sadržaj dokumenata.

Program je dobio ime po oznakama TEI (*Text Encoding Initiative*), no za razliku od njihova uobičajenoga unosa (pomoću komercijalnih editora oXygen, XMLSpy, XmlBlueprint i sl.) s XML-elementima i pripadajućim atributima (što stvara poteškoće u čitanju i analizi označenog teksta), ovdje je riječ o jednostavnijem, potpuno novom vizualnom pristupu koji isključuje potrebe poznavanja i čitanja XML-a (e*Xtensible Markup Language*) ili XSLT programa za transformaciju (ali ih niti ne odbacuje u naknadnoj analizi i obradbi označenog teksta). Program ima sve napredne generičke osobine pa se osim TEI označivanja može koristiti za tvorbu Wiki stranica, ReST ili Markdown aplikacija i slično. Označivanje dokumenata

* Dr. sc. Mario Essert, red. prof., Fakultet strojarstva i brodogradnje, Zagreb, messert@fsb.hr
** Vlado Cingel, dipl. inž., Čačinci, vladocingel@gmail.com
*** Nikola Glumac, mag. ing. mech., Zagreb, niglumac@gmail.com
**** Mario Lončarić, student, Zagreb, mario.loncaric@gmail.com
***** Božidar Štimac, inž., Zagreb, bozoou@gmail.com

može se provoditi lokalno (s tekstom u HTML formatu), ali i preko interneta, pri čemu je, slično kao i kod DocMarka, omogućeno vizualno označivanje u više nezavisnih slojeva. To omogućuje rad više osoba, npr. stručnjaka iz različitih područja, na istom dokumentu. Za rad je potreban samo WEB preglednik. Rezultati označivanja mogu se izvoziti u XML-u i u drugim formatima te naknadno obrađivati poznatim ili novostvorenim programima za analizu (npr. prebrojavanje oznaka, proučavanje pojmovnih klasa, gramatička istraživanja i slično). TEIMark osim ručnog ima ugrađeno i automatsko označivanje, i to na temelju unaprijed zadanih riječi (npr. iz računalne baze), njihovih dijelova pa čak i fraza (raspršenih riječi). Vizualne oznake moguće je definirati po hijerarhijskoj strukturi u dubinu i po pojmovnim domenama u širinu, te prikazivati skupno, pojedinačno ili po slojevima u označenom dokumentu.

TEIMark program ugrađen je u novu (petu) inačicu elektroničkog izdanja Biblije (© KS, Zagreb) i predstavljen u knjižnici HAZU-a na označivanju i analizi odabranih digitaliziranih dokumenata Instituta za jezik i jezikoslovlje i on-line enciklopedije Leksikografskoga zavoda Miroslav Krleža.

Ključne riječi: digitalizirana baština, TEIMark program, stvarni tekst, sadržaj, digitalizirani dokumenti

1. Uvod

Svaki od istraživača stare knjige, bilo da se radi o knjižničaru, bibliografu, paleografu[1], kodikologu[2], lingvistu, povjesničaru knjige[3] ili čitanja, povjesničaru umjetnosti, antikvaru ili slično, promatra knjigu iz svoga kuta. Proučavanje stare knjige nekad će se temeljiti samo na materijalnim značajkama, a nekad na njezinu sadržaju ili jeziku. Namjena stare knjige, sadržaj koji nam donosi, jako je utjecala na njezin oblik, i to ne samo na format, uvez ili odabir podloge za pisanje, već i na organizaciju cijele knjige ili teksta, odnos margina i tekstnih polja, veličinu slovnih znakova, veličinu i urešenost inicijala i ostalih uresa, i sl. Uz to, sadržaj kodeksa ili knjige tiskane u doba ručnog tiska utjecao je na način i učestalost njihova korištenja, a zaključke o tomu često izvodimo iz analize stanja primjerka ili pak iz tragova koje su čitatelji ili vlasnici ostavljali na njima. Golem je, dakle, broj elemenata knjige u središtu pozornosti mnogobrojnih istraživača: kodikologa će na kodeksu zanimati vrsta pergamene, način liniranja ili raspored stupaca, lingvista će k istom kodeksu privući riječi, sintaksa, analiza uporabe kratica ili

1 Avrin, Leila. Scribes, script and book : the book arts from Antiquity to Renaissance. Chicago: ALA; London : The British Library, 1991. Str. 210

2 Lemaire, Jacques. Introduction a la codicologie. Louvain-la-neuve: Université Catholique de Louvain, 1989. Str.

3 Stipišić, Jakov. Pomoćne povijesne znanosti u teoriji i praksi. Zagreb : Školska knjiga, 1972. Str. 3

neke druge osobitosti jezika, restaurator će promotriti vrstu uveza, a povjesničar knjige sa zanimanjem će proučiti naknadne zapise čitatelja, ekslibrise ili tiskarske znakove[4].

U baštinskim ustanovama u kojima se građa čuva nerealno je očekivati da će pri opisu građe sudjelovati više stručnjaka različitih profila jer ih tolik broj nije zaposlen u pojedinim zbirkama. Osobit se problem pojavljuje pri opisu hrvatske srednjovjekovne građe koja je pisana na više jezika i pisama, te kvalitetan opis takve zbirke iziskuje suradnju stručnjaka za latinsku, ali i glagoljsku i ćirilsku (staroslavensku) paleografiju, kodikologa koji bi uzeli u obzir specifične uvjete postanka hrvatske srednjovjekovne građe i sl. Različiti korisnici traže različite podatke o pojedinim knjigama i baštinska ustanova koja određeni primjerak čuva morala bi pronaći način da te podatke korisnicima i pruži. Rješenje je dakako pronađeno u digitalizaciji knjižne građe što rješava problem dostupnosti, ali ne i problem zajedničkoga istraživanja.

Mnogobrojnost istraživača i njihovih istraživačkih pitanja pri proučavanju stare i rijetke građe nameće potrebu za opisom puno širim od onoga koji nam omogućuje bibliografski ili kataložni opis. Ne niječući važnost potrebe kvalitetnih, mrežno dostupnih strojnočitljivih kataloga stare i rijetke građe i umrežavanja takvih podataka, u ovom su radu prikazana tri računalna programa koji omogućuju detaljniji opis i kolaboracijski istraživački rad na digitaliziranim inačicama starih knjiga, te njihovo digitalizirano sidrenje u vremensko-prostorne koordinate.

Mogućnost da jednom primjerku stare građe, mrežno dostupnom u svojoj digitalnoj inačici, više stručnjaka dodaje opise koji se tiču njihova područja stvara pretpostavku za točniji opis svakog dokumenta, po njegovim materijalnim i sadržajnim značajkama.

2. Problematika i pokušaji njezinog rješavanja

Iako ISBD(A) (*International Standard Bibliographic Description for Older Monographic Publications - Antiquarian*)[5] propisuje navođenje u opisu stare i građe elemenata specifičnih za tu građu (npr. format, a ne samo visinu hrpta, napomene o bibliografskom navodu, podatke o bivšim vlasnicima i sl.), ti se podaci još uvijek ne nalaze u svim zapisima za tu vrstu građe, a nažalost, građa nije pretraživa prema mnogim elementima opisa koji su za nju važni (npr. prema for-

4 Tomić, Marijana; Glumac, Nikola; Essert, Mario. Označivanje i analiza digitaliziranog dokumenta, AKM 14, Poreč, 2010. Str. 2

5 ISBD(A): Međunarodni standardni bibliografski opis starih omeđenih publikacija (antikvarnih) / preporučila Projektna grupa za Međunarodni standardni bibliografski opis starih omeđenih publikacija (antikvarnih); odobrili stalni odbori Sekcije za katalogizaciju i Sekcije za rijetke knjige i rukopise Međunarodne federacije bibliotekarskih društava i ustanova. Zagreb : Hrvatsko bibliotekarsko društvo, Nacionalna i sveučilišna biblioteka, 1995.

matu, uporabi rubrika, vrstama iluminacija, vrsti materijala za pisanje, vodenim znakovima i sl.). Kataložni ili bibliografski zapis, unatoč dodavanju elemenata podataka specifičnih za ovu vrstu građe, nikako ne mogu odgovoriti na sve upite koje će, pretražujući zbirke ove specijalne vrste građe, postaviti korisnici. Stoga je nužno, uz neospornu potrebu za katalogizacijom svih zbirka, pronaći rješenje koje će omogućiti da bar najvažniji dio tih vrijednih primjeraka bude pretraživ prema više elemenata, zanimljivih većem broju specifičnih korisnika.

Jači pokušaj rješavanja te problematike počeo je označivanjem digitaliziranih dokumenata uz pomoć TEI (*Text Encoding Initiative*) [6] smjernica. Svrha je bila označivanje strojnočitljivih tekstova postavljanjem oznaka prihvaćenim od šire (TEI) zajednice. I kod nas se pojavljuje više projekata koji se temelje na postavljanju TEI oznaka na digitalizirane inačice tekstova, na primjer *Edicija: digitalna knjižnica hrvatske tiskane baštine*[7]. U sklopu toga i sličnih projekata stari hrvatski tekstovi su digitalizirani, te su uz pomoć programa za OCR (*Optic Character Recognition*) pretvoreni u strojnočitljive tekstove koji se potom ručno označuju TEI oznakama za osobna imena, toponime i ostale podatke važne za razumijevanje stare hrvatske pisane baštine. Time je omogućeno pretraživanje korpusa tekstova prema označenim pojmovima koji se u njima nalaze, ali i prema uobičajenim bibliografskim elementima.

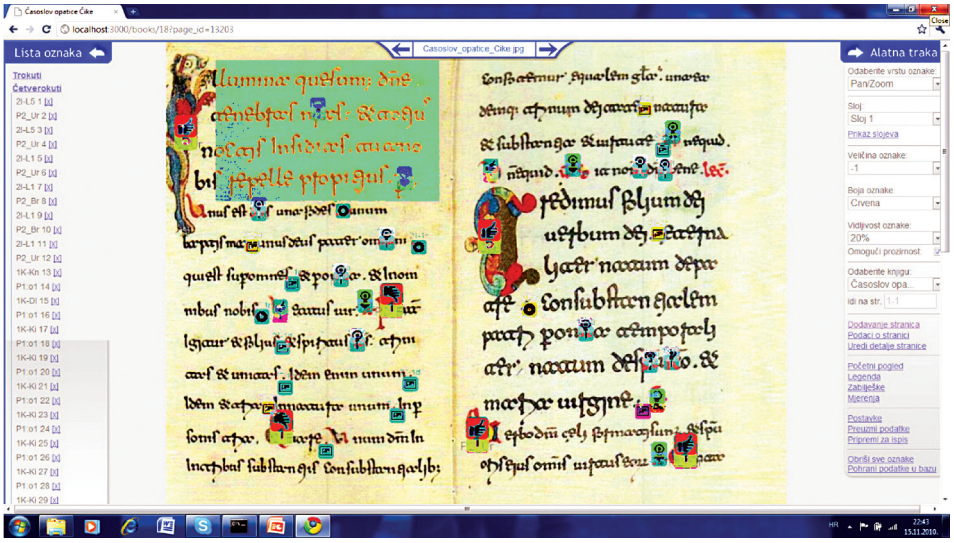
No, unatoč vrlo dobrim rješenjima koje TEI nudi, postoji problem glagoljske baštine i drugih rukopisa koji se ne mogu pretvoriti u strojnočitljive tekstove. Transliteracija ili transkripcija tih tekstova bio bi dugotrajan i težak posao, što zbog različitih vrsta pisama i njihovih razvojnih faza, i pretpostavljalo bi katkad višegodišnji rad stručnjaka na transliteraciji samo jednoga teksta. Jedino rješenje koje se nametalo bilo je označivanje digitalizirane slike, postavljanje oznaka izravno na digitalnu inačicu teksta. Time je dobivena mogućnost obilježavanja ciljanih elemenata na samom dokumentu bez gubljenja mnogih materijalnih obilježja primjerka koja se gube u transliteriranom tekstu, postavljanje većega broja istih oznaka na primjerke pisane različitim jezicima i/ili pismima i njihovu naknadnu analizu (pojedinačnih elemenata, njihove usporedbe i sl.), kao i pretraživanje baze dokumenata prema označenim elementima.

3. DocMark

Označivanje digitaliziranog dokumenta ostvareno je računalnim sustavom *DocMark* koji koristi WebGL, JavaScript i PHP/MySQL tehnologiju, pa se po-

6 TEI: http://www.tei-c.org/index.xml 1995.

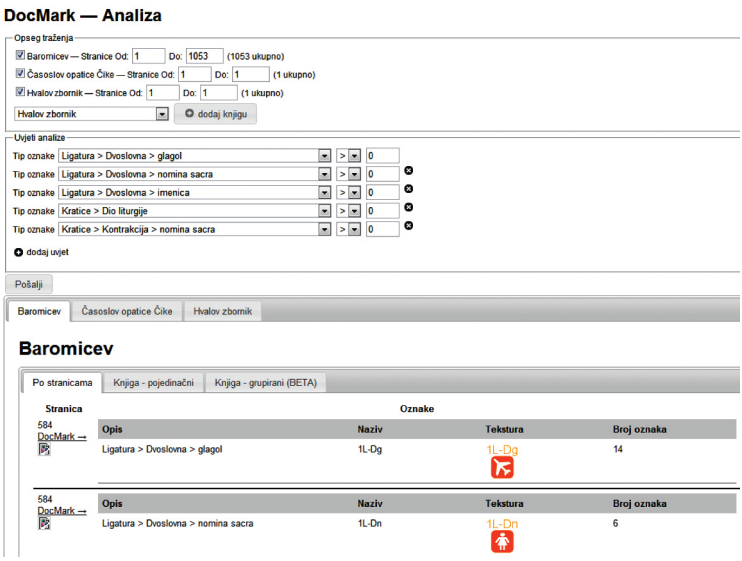
7 Edicija: digitalna knjižnica hrvatske tiskane baštine. Projekt 'Digitalna knjižnica hrvatske baštine tiskane do 1800.' , Odsjek za informacijske znanosti Filozofskog fakulteta u Osijeku, od siječnja 2007. godine, financirano od Ministarstva znanosti, obrazovanja i športa RH [Citirano 2010-12-15] Dostupno na: http://web.ffos.hr/EDICIJA/index.html.



SLIKA 1.
Označena stranica Časoslova opatice Čike, latinica, Zadar, 11. st.

ziva preko interneta. Korisniku je omogućen rad u jednom ili više slojeva nad slikom, stranicom dokumenta. U svakom sloju, nakon izbora oznake, korisnik klikom miša postavlja vizualnu oznaku na željeno mjesto slike dokumenta, ne mijenjajući pritom samu sliku, tj. digitalni dokument ostaje sačuvan u svom izvornom obliku. Razlikuje se više tipova oznaka, a riječ je o grafičkim znakovima u obliku kvadrata, trokuta ili kruga, različitih boja i prozirnosti. Svakom tipu oznake korisnik prije označivanja pridružuje neko svojstvo/kategoriju (analogno izboru neke TEI oznake). Svakoj posebnoj oznaci unutar istog tipa može se dodati i poseban opis. Za mjesta u kojima je bitna veličina (npr. istaknuta glagoljička slova, inicijali), margine, razmaci između stupaca i druge bjeline na slici, sustav nudi alat za precizno mjerenje (do stotinke milimetra). Među oznakama posebno mjesto zauzimaju linije koje korisnik može povlačiti od jedne pozicije do druge, te rastezljiva polja kojim označuje zanimljiva područja dokumenta (obično s visokom prozirnošću, kako bi se dokument ispod njih mogao vidjeti). Sve oznake, sva mjerenja i svi opisi se automatski spremaju u bazu, a na poziv automatski iz nje dohvaćaju, što omogućuje naknadno uređivanje dokumenta. U slučaju transliteracije teksta, oznake se s lakoćom mogu pretvoriti u bilo koju TEI oznaku (tag).

Sve oznake grupirane su po tzv. slojevima, odijeljenim prema vrstama oznake, što je određeno unaprijed, pri definiranju sheme oznaka. Slojevi su omogućili da se, nakon što su se oznake postavile, na slici dokumenta prikažu oznake svih



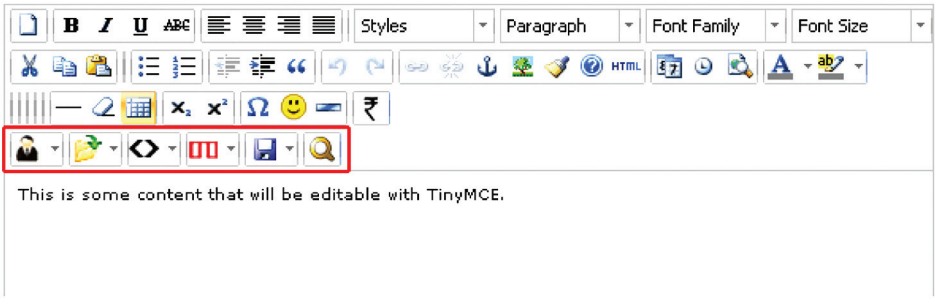
SLIKA 2. Analiza grafetičkih sredstava u dokumentima pisanim na trima pismima

slojeva, samo pojedinih slojeva prema odabiru ili da se pak sakriju sve oznake. (Slika 1.)

Nakon postavljenja oznaka slijedi analiza. *DocMark* omogućuje jednostavno i brzo prebrojavanje oznaka po vrstama, slojevima i stranicama/slikama dokumenta, računa udaljenosti željenih oznaka različitih tipova, obrađuje zadana područja dokumenata (npr. koliko oznaka nekog tipa ima u određenim poljima/područjima), te brzo pozicioniranje i prikaz dokumenta prema odabranoj (pojedinačnoj) oznaci, uz mogućnost njegova automatskoga zumiranja na središte stranice. Analizirati se istodobno mogu svi označeni dokumenti, ili samo pojedini, prema jednom ili više kriterija, a izbor je omogućen odabiranjem dokumenta i traženog kriterija. Na taj je način omogućena analiza prema iznimno veliku broju kriterija, i to na svim ili pojedinim dokumentima, ovisno o broju postavljenih oznaka i o upitu. Uz to, odabiranjem dokumenata koji se analiziraju, a rezultati za koje su prikazani odvojeno u usporednim stupcima, moguće je pratiti određenu značajku u njezinoj pojavnosti u jednom dokumentu ili ju uspoređivati u različitim dokumentima. (Slika 2.)

Jedno takvo opsežno znanstveno istraživanje, koje je rezultiralo doktorskim radom, provela je Marijana Tomić iz Zadra[8]. Pod mentorstvom prof. dr. sc. Matea Žagara i prof. dr. sc. Zorana Velagića, na odabranom korpusu (Mavrov brevijar,

8 Tomić, Marijana. *Organizacija i aproprijacija tekstova hrvatskoglagoljskih brevijara na razmeđu rukopisne i tiskane tradicije*, doktorski rad, Zagreb 2013



SLIKA 3. Osnovno TEIMark korisničko sučelje

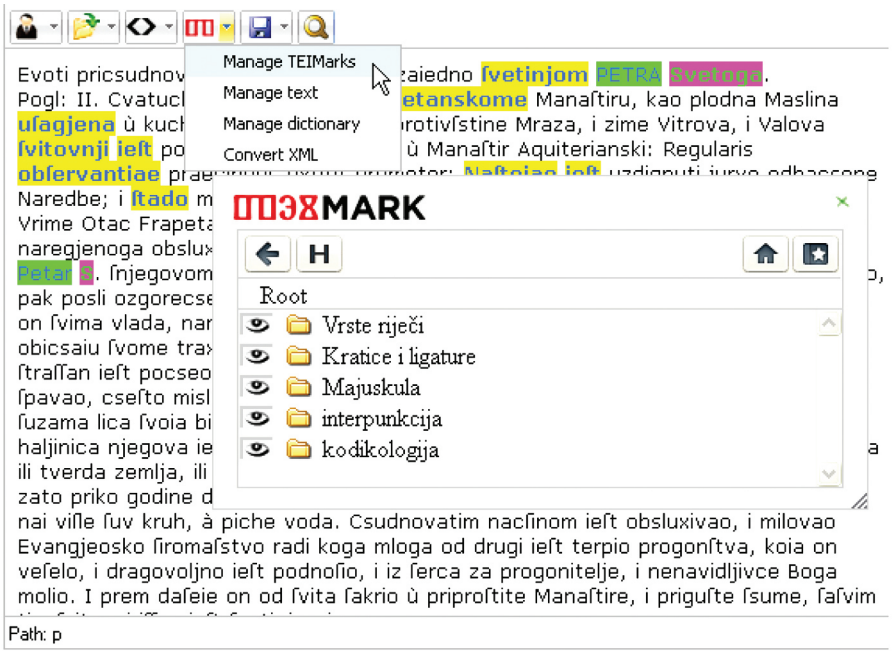
Baromićev brevijar, Drugi novljanski brevijar, Brozićev brevijar) postavila je 15.900 oznaka iz svoje, vlastito definirane sheme, koja se sastoji od 387 vizualnih oznaka razvrstanih u sedam slojeva (kratice i ligature, majuskula, interpunkcija, praznine i bjeline, spacioniranje/kondenziranje, organizacija teksta/aproprijacija, kodikologija). Uz to, uz svaki su brevijar istraživačkoga korpusa sačuvane bilješke i svi rezultati mjerenja koji se odnose na stranični postav, ali i bibliografski opis brevijarâ i bilješke vezane za pojedine osobitosti organizacije teksta uočene tijekom označivanja.

4. TEIMark

Dok je označivanje u DocMarku izvedeno nad slikom dokumenta (u kojem su zanimljiva i dohvatljiva materijalna svojstva, ali ne i sadržaj teksta), u TEIMarku se označivanje odvija nad stvarnim tekstom, a ne njegovom slikom, čime su omogućena lingvistička i druga istraživanja usmjerena na sadržaj dokumenata. Građu stare knjige potrebno je utipkati ili postupkom *OCR*a pretvoriti u strojno čitljivi tekst. Označivanje teksta izvodi se u WORD-u sličnom WEB-editoru pod imenom *TinyMCE* u koji su ugrađeni gumbi za otvaranje različitih dijaloških okvira. (Slika 3.)

Označivanje tekstova s pomoću *TEIMark*a ima prednost pred sličnim *TEI/XML* alatima u sljedećem:

- za rad je potreban samo web preglednik (eng. *browser*), a broj suradnika/istraživača ostvaren je preko intraneta (unutar ustanove) ili interneta (na globalnom prostoru)
- moguće je definirati više slojeva/kategorija i skupina oznaka za svaki sloj, slično slojevima i vizualnim oznakama nad slikovnim dokumentom DocMarka
- oznake se postavljaju klikom miša na tekstu editorske stranice WEB pregled-



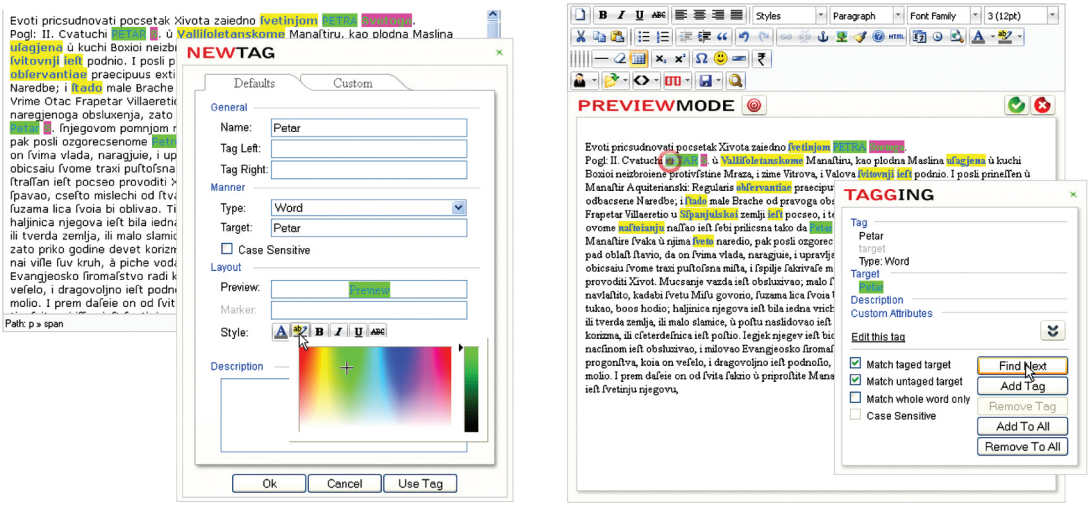
SLIKA 4. Kartica za upravljanje slojevima i oznakama

nika, umjesto mukotrpnim upisom TEI-tagova u nekom lokalnom XML-editoru

- ugrađeno je automatsko označivanje, a u slučaju višestrukog značenja, program nudi izbor mogućih rješenja
- slojevi se mogu pojedinačno ili skupno prikazivati nad istim tekstom
- označeni tekstovi (čak i oni s klasičnim TEI oznakama) mogu se učitati u TEIMark i nastaviti označavati na lakši način, te izvesti (export) u vlastitom ili klasičnom (TEI/XML, JSON) formatu
- analiza označenog teksta može biti klasična (XSLT) ili moderna (preko Python/Ruby programa)

TEIMark posjeduje i više funkcija koje su usmjerene povezivanju označenih tekstova s podatcima iz računalnih baza, npr. enciklopedije Leksikografskog zavoda Miroslava Krleže ili Strune – strukovnoga nazivlja Instituta za hrvatski jezik i jezikoslovlje, ali to je izvan okvira obradbe stare knjige, pa će ovdje biti ispušteno. Između velikog broja kartica i mogućnosti koje *TEIMark* nudi, izdvojit ćemo samo sljedeće:

A) Kartica „TEI MARK“ je sučelje za upravljanje slojevima, a nalazi se u izbor niku pod nazivom „Manage TEIMarks“. (Slika 4.)
Unutar ovoga sučelja korisniku je omogućeno stvaranje novih slojeva (kate-



SLIKA 5.
Tvorba oznaka i on-line automatizirano označivanje zadane vezne oznake

gorija), navigacija kroz slojeve, stvaranje novih vizualnih oznaka (tagova), kopiranje, premještanje i brisanje slojeva ili oznaka, te upravljanje vidljivošću slojeva.

Vidljivost stila označenoga teksta upravlja se preko simbola "oka" (lijevo od sloja i taga na Slici 3.): ugašeno oko ne prikazuje stil (boju ili uzorak teksta i pozadine) označenog objekta (dio riječi, riječ ili skupina slova), dok upaljeno prikazuje.

Slojevi su hijerarhijski složeni do bilo koje dubine, a unutar bilo kojeg sloja mogu se spremati nove oznake i drugi slojevi. Uključenje 'oka' odnosi se na taj sloj i sve njegove podslojeve.

B) Korisnik može definirati neograničen broj oznaka, pridružiti im imena, zamisliti vizualni izgled te spremiti u kategoriju/folder oznaka bilo kojega sloja. Funkcijski gledano, oznake mogu biti „vezane" i „slobodne". Vezane oznake imaju ciljno svojstvo/pojam, tj. moguće ih je koristiti samo za označivanje unaprijed definiranih objekata (odredišta ili meta). To svojstvo se najviše koristi kod automatskoga označivanja tekstova. Moguće vrste *vezanih* oznaka su:

1. Word - obuhvaća (eng. *wrap*) znak, riječi ili niz riječi, paragraf ili dijelove teksta, što odgovara klasičnom označavanju povlačenjem miša uz pritisnutu njegovu lijepu tipku (*mark*).
2. Point - umeće oznaku na neko mjesto (npr. među dva slova). To je slučaj označivanja kratice ili umetanja novih riječi, retka ili slično.
3. Phrase – za skupine nesusjednih riječi ili znakova. Tako će za automati-

SLIKA 6. Pretraživanje oznaka

SLIKA 7. Izvoz i zamjena oznaka

zirano označivanje fraze: "Ja hodam", "Ja brzo hodam", "Ja katkad sporo hodam" u sva tri slučaja program moći prepoznati frazu "Ja hodam" i na njoj primijeniti zadanu oznaku.

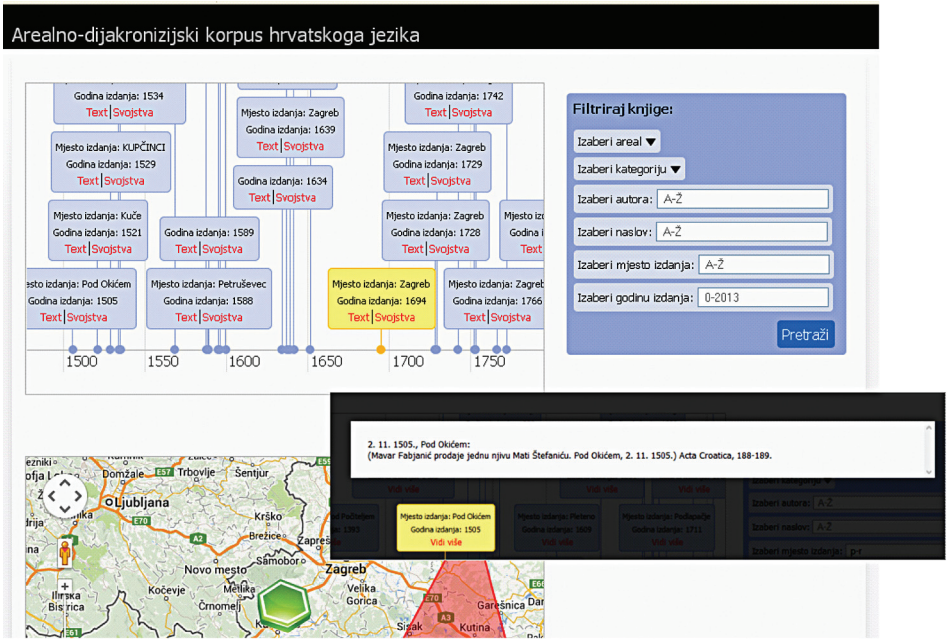
Moguće vrste *slobodnih* oznaka koje nisu povezane pojmom koju riječ ili rečenica ima, sukladne su nevezanim oznakama, pa razlikujemo: Free Word, Free Point i Free Phrase.

C) Snaga TEIMark sustava leži u automatiziranom, kolaboracijskom web sučelju preko kojega skupine istraživača rade na istom dokumentu transkribirane stare građe pri čemu sve oznake i postupak označivanja mogu provesti bez ikakvog poznavanja programskoga alata, čak ni na razini XML ili TEI zapisa. Na Slici 5. vidljiv je dio postupka automatiziranog označivanje riječi 'TEI' (dakako, moguće i bilo koje druge ili čak neke fraze). (Slika 5.)

Dobro je spomenuti kako se sve promjene u radu, prije povratka u osnovni editor, mogu spremiti ili odbaciti (otuda *Preview mod* na Slici 5.).

D) S pomoću opcije za prebrojavanje oznaka, korisnik ima mogućnost uvida u sve oznake koje je koristio za označivanje tekstova koji se nalaze na njegovom korisničkom računu i to s obzirom na oznake i dokumente (vidljive ili nevidljive/sakrivene).

Unutar kartice za prebrajanje oznaka postoje četiri aktivatora uz pomoć kojih se može prilagoditi pretraga oznaka, kao i preko parova atributa i njihovih pripadajućih vrijednosti, što pokazuje Slika 6.



SLIKA 8. Stara knjiga u vremenu i prostoru

E) Jednom označen tekst moguće je izvesti (eng. *export*) u različite oblike koji su pogodni za daljnju obradbu teksta (Slika 7).

Nekoliko funkcijskih dijelova TEIMark programa ugrađeno je sa svrhom označivanja biblijskog teksta u novoj inačici Biblije[9].

4. Arealno-dijakronizijski korpus hrvatskoga jezika

U tijeku je stvaranje mrežnoga softvera koji će se brinuti za same dokumente, bez obzira jesu li digitalizirani i spremljeni u slikovnom (PDF, JPG) ili u tekstnom (TXT) formatu. Označivanje korpusa takvih dokumenata bit će moguća uz pomoć opisanih DocMark i TEIMark programa, a sami dokumenti lako će se pretraživati po vremenskoj ljestvici (*TimeLine*) ili po prostornim arealima (*Gmaps*), kao što se vidi na Slici 8. Voditelj toga (tek prijavljenoga) projekta je dr.sc. Jurica Budja s Instituta za hrvatski jezik i jezikoslovlje. (Slika 8.)

9 Copyright - Kršćanska sadašnjost, Zagreb. (očekivano pojavljivanje proizvoda: početak 2014.)

5. Prijedlozi za buduća istraživanja

1. Objedinjavanje fragmenata koji su bili dijelom istog dokumenta

Fragmenti su iznimno vrijedan dio baštine jer je često riječ o dijelovima odbačenih rukopisa koji su upravo kao „nepotrebni" bili upotrebljavani za uveze drugih knjiga te su nam do danas očuvani samo tako razjedinjeni. U postupanju s fragmentima važno ih je precizno opisati kako bi bilo moguće usporediti fragmente pronađene na različitim mjestima (unutar jedne zbirke ili više njih) i povezati one koji su pripadali istom rukopisu. Dakle, osim datacije i ubikacije koja se uvijek pokušava dokučiti uz pomoć paleografske i drugih analiza, valja pokušati pronaći ostale dijelove teksta. Činjenica da pripadaju istom dokumentu utvrđuje se usporedbom kodikoloških, paleografskih, leksičkih i mnogih drugih značajka pronađenih fragmenata. Tu DocMark može naći značajnu primjenu te povezati baštinske ustanove unutar Domovine, ali i inozemstva.

2. Atribucija različitih autora u istom djelu

U današnje se vrijeme često susrećemo s kodeksima ili starom knjigom u kojima se prepliće rad više pisara ili autora, na primjer, za tzv. Njujorški misal A.R. Corin je analizom utvrdio da je u njegovu pisanju sudjelovalo vjerojatno jedanaest pisara[10]. Da bi se utvrdili dijelovi istoga kodeksa koje su pisali različiti pisari, radi se paleografska analiza kodeksa i utvrđuju se paleografske i druge značajke koje su svojstvene ovom ili onom pisaru. Nakon što se te značajke definiraju, označivanjem digitaliziranog dokumenta dobiva se pri analizi mogućnost lakog uvida u dijelove teksta s istim značajkama. Time je olakšana atribucija tekstova pojedinim pisarima. Označe li se varijante istog razlikovnog obilježja različitim oznakama na digitalnoj inačici, bilo slikovne, bilo tekstovne stare građe (s pomoću DocMark ili TEIMark programa) na označenim stranicama vizualne oznake omogućuju lako razlikovanje dijelova teksta s obilježjima pojedinog pisara, što se statističkom analizom oznaka može lako i kvantitativno dokazati. Utemeljenim izborom oznaka ista se metodologija potencijalno može primijeniti i na atribuciju dokumenata pisarskim školama ili skriptorijima, a analizom TEIMark oznaka s lingvističkog stajališta moguće je uspoređivati autore i otkriti značajke njihova pisanja.

3. Život hrvatske riječi

Zahvaljujući vremensko-prostornoj sistematizaciji stare knjige za transkribirane digitalne dokumente moći će se pratiti postajanje i nestajanje hrvatske riječi kroz arealno-dijakronizijski korpus hrvatskih jezika i narječja.

6. Zaključak

S obzirom na to da su korisnici stare i rijetke građe stručnjaci raznih profila koji

10 Andrew R. Corin. The New York Missal: a paleographic and phonetic analysis. Columbus: Slavia Publishers, 1991. Str. 49Andrew R. Corin. The New York Missal: a paleographic and phonetic analysis. Columbus: Slavia Publishers, 1991. Str. 49

građi prilaze s vrlo specifičnim upitima koji nadilaze mogućnosti bibliografskog opisa, uz pomoć posebno dizajniranih računalnih programa *DocMark* i *TEIMark* označuju se različiti elementi građe na digitaliziranim inačicama dokumenata. Oznake (stvorene na temelju pretpostavljenih zahtjeva širokog kruga korisnika) u *DocMark*u se postavljaju na slike dokumenta kako bi se izbjegla kompliciranа transliteracija srednjovjekovne hrvatske građe, dok se s *TEIMark*om vizualne oznake postavljaju na samom tekstu. Oba programa omogućuju da na isti dokument oznake postavljaju putem interneta i kompetentni udaljeni korisnici kojima je to dopušteno. Analiza dokumenta omogućuje pretraživanje dokumenata prema različitim kriterijima, i to kvantitativnu analizu u smislu prebrojavanja elemenata u pojedinačnim dokumentima i skupno, te komparativnu analizu u smislu usporedbe uporabe određenih elemenata u dijelovima pojedinih dokumenata kao i ukupno među dokumentima. *DocMark*om se omogućuju i precizna mjerenja straničnog postava (uporabe inicijala i njihova veličina, raspored bjelina i sl.), grafetička sredstva i definiranje posebnih straničnih polja, što je česta potreba u obradbi stare knjige. Uporaba istih oznaka omogućila bi preko *Arealno-dijakronizijskog korpusa hrvatskoga jezika* WEB-programa usporedbu velikog korpusa hrvatske stare baštine koja je u mnogim svojim aspektima još uvijek neistražena i skrivena u trezorima i zbirkama mnogih baštinskih ustanova, skrivajući tajnu iznimno važnih razdoblja hrvatske nacionalne povijesti. Mogućnost dodavanja novih oznaka otvara mjesta za nove elemente koji se nužno pojavljuju pri svakom istraživanju slabo poznate građe. Realizacijom programa u tehnologiji internetske aplikacije pruža se mogućnost da se na isti način obrade i oni primjerci stare građe koji se ne nalaze u Hrvatskoj, ali su nam dostupni u digitalnim inačicama.

LITERATURA

1. Avrin, Leila. 1991. *Scribes, script and book : the book arts from Antiquity to Renaissance*. Chicago: ALA; London: The British Library.
2. Corin, Andrew R. 1991. *The New York Missal: a paleographic and phonetic analysis*. Columbus: Slavia Publishers.
3. ISBD(A): Međunarodni standardni bibliografski opis starih omeđenih publikacija (antikvarnih) / preporučila Projektna grupa za Međunarodni standardni bibliografski opis starih omeđenih publikacija (antikvarnih); odobrili stalni odbori Sekcije za katalogizaciju i Sekcije za rijetke knjige i rukopise Međunarodne federacije bibliotekarskih društava i ustanova. 1995. Zagreb: Hrvatsko bibliotekarsko društvo, Nacionalna i sveučilišna biblioteka.
4. Lemaire, Jacques. 1989. *Introduction a la codicologie.* Louvain-la-neuve: Université Catholique de Louvain, str. 3.
5. Stipišić, Jakov. 1972. *Pomoćne povijesne znanosti u teoriji i praksi*. Zagreb: Školska knjiga.

Web izvori:

Edicija: digitalna knjižnica hrvatske tiskane baštine. [Citirano 2010-12-15] Dostupno na URL: http://web.ffos.hr/EDICIJA/index.html

Summary

Mario Essert, Vlado Cingel, Nikola Glumac, Mario Lončarić i Božidar Štimac
TIEMARK PROGRAM – THE PROCESSING OF THE TRANSCRIBED OLD LIBRARY MATERIALS

In Croatian institutions, there is a great quantity of digitalized legacy, which is, for the purpose of presentation, often accessible via internet. Digitalized documents hold the truth of rich Croatian cultural and written heritage. After the completion of the program DocMark, whose purpose is to mark digitalized picture documents, and to analyze these marks on individual documents and/or to compare them, the TEIMark program was created – which served to mark text: be it typed, translitarated or machine – recognized. While the marking in DocMark was executed over the document picture (in which the key points are material properties/singularities, but not the content of text), in TEIMark the marking is done on a real text, and not on it's image, and in this way, linguistic and other types of research focused on the content of document are made possible.

Program was given it's name because of the marks TEI (Text Encoding Initiative), however, unlike the usual input (through commercial editors such as oXygen, XMLSpy, XmlBlueprint, etc.) with XML elements and belonging attributes (which creates difficulties in reading and analysis of the marked text), here we have a more simple, completely new visual approach which excludes the need of knowing and reading XML (eXtensible Markup Language) or XSLT program for transformation (however, they are not rejected in the supplemental analysis and processing of the marked text). Program has all advanced generic attributes, so it can be used, aside from TEI markings, for creating Wiki sites, ReST or Markdown applications and similar. Document markings can be made locally (with text in HTML format) but also through internet, which has, similar to DocMark, enabled the visual markings in several independent layers. This enables the work of more than one person, ie several experts from various fields on the same document. The work requires only a web browser. The results of markings can be exported in XML and other formats, and additionally be processed with classic or newly made programs for analysis (ie counting of marks, studying of conceptual classes, grammatical research and such). Aside from manual, TEIMark has installed the automatic marking option, based on words given in advance (ie from the computer base), their parts and even phrases (words in dispersion). Visual marks are possible to define according to the hierarchical structure in depth, and according to conceptual domains in width, and to be displayed in groups, individiually, or in a layered fashion, within a marked document.

TEIMark program is built into the new (fifth) version of electronic edition of BIBLE (© KS, Zagreb) and demonstrated in the HAZU library, for the purposes of marking and the analysis of selected digitalized documents of Institute of Croatian Language and Linguistics and online encyclopedia of The Miroslav Krleža Institute of Lexicography.

Keywords: digitalized legacy, TEIMark program, real text, content, digitalized documents