Stručni rad

Primljeno: 2. X. 2024.

Prihvaćeno: 3. XI. 2024.

UDK
061.2(497.521.2):030]:004
811.163.42'374:004.031.4
030(497.5):004.031.4

https://doi.org/10.33604/sl.18.35.6

Model digitalizacije arhivskih enciklopedijskih i leksikografskih izdanja za mrežu[1]

Cvijeta Kraus
Leksikografski zavod Miroslav Krleža, Zagreb
cvijeta.kraus@lzmk.hr
Josip Mihaljević
Staroslavenski institut, Zagreb
jmihaljevic@stin.hr
Irina Starčević Stančić
Leksikografski zavod Miroslav Krleža, Zagreb
irina.starcevic.stancic@lzmk.hr

SAŽETAK: Rad prikazuje model koji je osmišljen za digitaliziranje arhivskih tiskanih enciklopedijskih i leksikografskih izdanja Leksikografskoga zavoda Miroslav Krleža radi njihova objavljivanja na mreži. Ta leksikografska izdanja dosad nisu bila dostupna u digitalnome obliku te je njihovu digitalizaciju trebalo započeti skeniranjem tiskanih knjiga. Riječ je o dvanaest arhivskih izdanja, od kojih najstarija uključuju *Pomorsku enciklopediju I. izdanje* (1954–64), *Enciklopediju Leksikografskog zavoda* (1955–64) i *Medicinsku enciklopediju* (1957–65), a najnovije je izdanje *Enciklopedija hrvatske umjetnosti* (1995–96). Skenirana djela razlikuju se po sadržaju, strukturi i prikazu grafičkih dodataka. Zbog toga je bilo potrebno osmisliti korake za uspješnu digitalizaciju i objavu na mreži različitih leksikografskih djela. Prikazani model sastoji se od šest koraka: 1. skeniranje stranica i optičko prepoznavanje znakova, 2. uređivanje teksta i slika, 3. izrada abecedarija, 4. izrada baze podataka, 5. izrada mrežne stranice za prikaz prethodno strukturiranih podataka i 6. objava sadržaja na mreži. Svaki od tih koraka sastoji se od više manjih procesa, koji su određeni dostupnom tehnologijom te ljudskim znanjem i potencijalima. U radu će se svaki korak iscrpno objasniti, analizirati i oprimjeriti primjerima iz leksikografske prakse kako bi se prikazani model mogao primijeniti i na digitalizaciju drugih enciklopedijskih i leksikograf-

https://orcid.org/0000-0002-7927-6020 [Cvijeta Kraus]
https://orcid.org/0000-0002-7482-7663 [Josip Mihaljević]
https://ror.org/00vjz3318 [Leksikografski zavod Miroslav Krleža]
https://ror.org/04sdd2113 [Staroslavenski institut]

[1] Ovaj je rad nastao i u suradnji s projektom Razvoj modela digitalne infrastrukture Staroslavenskoga instituta – DigiSTIN, koji financira Europska unija – NextGenerationEU. Za iznesene stavove i mišljenja odgovorni su samo autori te ti stavovi ne odražavaju nužno službena stajališta Europske unije ili Europske komisije. Ni Europska unija ni Europska komisija ne mogu se smatrati odgovornima za njih.

skih izdanja. Dodatni je prinos ovoga rada prikaz funkcionalne mrežne stranice *Zbirka enciklopedijske baštine* (e-bastina.lzmk.hr), koja je nastala na temelju osmišljenoga modela i koja se sastoji od dvanaest digitaliziranih enciklopedijskih i leksikografskih izdanja te sadržava 57 svezaka dostupnih za pretraživanje natuknica i pregledavanje. Ukupan je broj natuknica u svim izdanjima 95 000, a uz svaku natuknicu u rezultatu pretraživanja prikazuje se izvor, odnosno naziv enciklopedije ili leksikona u kojemu se tražena natuknica nalazi.

Ključne riječi: abecedarij; baze podataka; digitalizacija; digitalizacijski model; mrežna leksikografija; uređivanje teksta i slike

1. Uvod

Digitalizacija je složeni pojam koji ima više značenja. Često se brkaju paronimni nazivi *digitizacija* i *digitalizacija*. Digitizacija se isključivo odnosi na proces prevođenja analognoga signala u digitalni oblik, tj. pretvaranje teksta, slike, videozapisa, zvuka ili trodimenzionalnoga objekta u digitalni oblik.[2] Tipičan je primjer digitizacije skeniranje papira u PDF dokument. Digitalizacija je mnogo širi pojam te osim samoga procesa digitizacije, u kojemu se pretvara tiskani oblik u digitalni, uključuje i ostale dodatne poslove koji se provode s pomoću digitalne tehnologije kako bi se dobiveni digitalni objekt po potrebi učinio jednostavnijim, sigurnijim i/ili dostupnijim za pregled. Digitalizacija u širem smislu podrazumijeva i to da se digitalni objekt dodatno obrađuje, pohranjuje, prenosi i pregledava s pomoću računala te po potrebi prebacuje u druge digitalne oblike (Gorenšek i Kohont 2019, 95–96). Time digitalizacija može podrazumijevati i složenije procese poput izrade mrežnih stranica, aplikacija, digitalnih knjižnica, digitalnih arhiva, baza podataka i drugih računalnih sustava. Proces digitalizacije razlikuje se ovisno o predmetu koji se digitalizira, dostupnoj tehnologiji, znanju i ljudskim potencijalima te očekivanome krajnjem rezultatu.

U ovome radu opisuju se koraci digitalizacije arhivskih tiskanih enciklopedijskih i leksikografskih izdanja Leksikografskoga zavoda Miroslav Krleža objavljenih na mrežnoj stranici e-bastina.lzmk.hr. Ta leksikografska izdanja uključuju *Enciklopediju fizičke kulture* (2 sv., 1975–77), *Enciklopediju hrvatske umjetnosti* (2 sv., 1995–96), *Enciklopediju Leksikografskog zavoda* (I. izdanje, 7 sv., 1955–64). *Enciklopediju likovnih umjetnosti* (4 sv., 1959–66), *Medicinsku enciklopediju* (10 sv., 1957–65), *Otorinolaringologiju* (2 sv., 1965–66), *Poljoprivrednu enciklopediju* (3 sv., 1967–73), *Pomorsku enciklopediju I. izdanje* (8 sv., 1954–64), *Sportski leksikon* (1984), *Šumarsku enciklopediju I. izdanje* (2 sv., 1959–63), *Šumarsku enciklopediju II. izdanje* (3 sv., 1980–87) i *Tehničku enciklopediju* (13 sv., 1963–97). Digitalizacija se provodi kako bi se spomenuta enciklopedijska i leksikografska izdanja učinila dostupnima i pretraživima u digitalnome obliku te (u kasnijoj fazi) povezala i s digitaliziranim izdanjima unutar mrežne stra-

[2] https://enciklopedija.hr/clanak/digitalizacija (pristupljeno 19. VII. 2024.)

nice *Portala znanja*.[3] Izrađena je mrežna stranica za svako spomenuto digitalizirano izdanje te mrežna stranica zbirke *Enciklopedijska baština,* koja omogućuje pretraživanje natuknica u svim izdanjima.[4] Mrežna stranica omogućuje pretraživanje natuknica putem tražilice ili abecedarija i izravnu vezu na skenirane stranice sveska pojedinih izdanja u PDF-u. Određene natuknice sadržavaju i dodatne poveznice na slikovne priloge te poveznice s natuknicama srodnih izdanja na mrežnoj stranici *Enciklopedijska baština*. Povezana su izdanja: *Enciklopedija fizičke kulture* sa *Sportskim leksikonom, Enciklopedija hrvatske umjetnosti* s *Enciklopedijom likovnih umjetnosti* te prvo i drugo izdanje *Šumarske enciklopedije*. Također je omogućeno listanje svakoga sveska u formatu *flipbook*.

Kako bi se uspješno proveo proces digitalizacije i na najbolji način prikazala digitalizirana građa, razvijen je model koji se ne oslanja na postojeće modele vezane za digitaliziranje enciklopedijskih i leksikografskih djela.[5] Model se sastoji od ovih šest koraka:

1. skeniranje i optičko prepoznavanje znakova
2. ispravak teksta i slike
3. izrada abecedarija
4. izrada baze podataka
5. izrada mrežne stranice za prikaz prethodno strukturiranih podataka
6. objava sadržaja na mreži.

Većina koraka izvodi se linearno, ali se neki koraci mogu provesti unaprijed, poput petoga koraka. Međutim, i u njemu se same funkcionalnosti stranice rade nakon što se odredi struktura podataka u bazi koja se izrađuje u četvrtome koraku. Dok prvi korak nije u potpunosti dovršen, ne prelazi se na drugi korak. Treći korak može se raditi usporedno s drugim korakom jer se u procesu provjere cijeloga teksta mogu izlučiti natuknice za izradu abecedarija. Često treći korak može biti dovršen prije drugoga koraka jer je brže izlučiti sve natuknice iz sveska nego provjeriti cijeli tekst. Da bismo prešli na četvrti korak, treći korak mora biti potpuno dovršen jer je potreban popis natuknica po stranicama sveska za inicijalnu bazu podataka. Nakon što je dovršen peti korak, treba odrediti kad i pod kojom domenom ćemo objaviti stranicu digitaliziranoga izdanja te imamo li unesene sve metaoznake koje opisuju sadržaj stranice kako bi stranica bila dobro rangirana na mrežnim tražilicama poput Googla

[3] http://enciklopedija.lzmk.hr/ (19. VII. 2024.)

[4] https://e-bastina.lzmk.hr/ (8. VII. 2024.)

[5] O digitalizaciji leksikografskih djela pisali su Toma Tasovac (2022) te Marijana Horvat i Martina Kramarić (2021), ali se ti radovi uglavnom odnose na rječnike i gramatike.

i Binga. Za navedene korake nije se koristio nijedan od AI programa. Unatoč tome što se AI programi sve više počinju primjenjivati kao pomoć u radu, nisu dovoljno pouzdani za provjeru kvalitete digitalizacije. Umjetna inteligencija snažno je ovisna o velikim količinama podataka kako bi se mogla trenirati i tako donositi točne zaključke. Međutim, tijekom digitalizacije često nedostaju visokokvalitetni, relevantni i označeni podatci. To može dovesti do loših rezultata ili netočnih predviđanja (Aldoseri i dr. 2023, 1–2). Podatci koje je obradila ili stvorila umjetna inteligencija mogu biti nepotpuni, loše strukturirani ili sadržavati pogreške, često zbog nedostatka konteksta informacija, što može leksikografima otežati točan unos i obradu podataka koji se objavljuju.

2. Skeniranje i optičko prepoznavanje znakova

Za skeniranje su korištene neuvezane knjige kako bi se olakšao proces skeniranja. U nekim se slučajevima skeniranje i optičko prepoznavanje znakova (OCR)[6] radi odvojeno. To ovisi o tehnologiji koja je dostupna za rad jer nemaju svi uređaji i programi za skeniranje mogućnost optičkoga prepoznavanja znakova.[7] U analiziranome slučaju, za istodobno skeniranje i optičko prepoznavanje znakova upotrijebljen je program Abbyy Finereader 15, koji se od početaka digitalizacije u Zavodu (2009) koristi zbog svoje jednostavnosti za uporabu, mogućnosti uzastopnoga korištenja više jezika kod optičkoga prepoznavanja znakova[8] i drugih opcija. Abbyy Finereader 15 nudi ove opcije koje su važne za skeniranje i optičko prepoznavanje znakova u leksikografskim djelima:

- spajanje više slika ili PDF dokumenata u jedan zajednički PDF dokument
- raspoređivanje i uređivanje stranica unutar dokumenta
- spremanje radne inačice dokumenta za nastavak rada
- prepoznavanje znakova za različite jezike i pisma te mogućnost prepoznavanja tekstnoga oblikovanja (npr. je li tekst kurziviran ili podebljan).

---

[6] Optičko prepoznavanje znakova (engl. *optical character recognition,* OCR), elektronički ili mehanički proces u kojemu program pronalazi tekstove unutar slika te ih pretvara u računalno čitljiv tekst koji se može pretraživati te uređivati (Panian 2005b, 82–83).

[7] O usporedbi besplatnih i komercijalnih programa za optičko prepoznavanje znakova na hrvatskome jeziku više u radu Mihaljević 2017.

[8] https://www.g2.com/articles/best-ocr-software (pristupljeno 20. VII. 2024)

Program funkcionira tako da se u njega učitaju skenovi stranica koji mogu biti u PDF-u ili slikovnome formatu (.jpg, .png ili .tiff). Stranice se mogu rasporediti i urediti unutar sučelja programa. Grafičko uređivanje omogućuje ispravak orijentacije stranice, brisanje pozadinskih mrlja te izbjeljivanje pozadine. Optičko prepoznavanje znakova može se obaviti u trenutku učitavanja dokumenta ili naknadno unutar programa pritiskom na gumb. U analiziranome slučaju optičko prepoznavanje znakova provedeno je nakon uređivanja slika dokumenta. Neka su izdanja dodatno grafički uređena s pomoću programa Adobe InDesign, koji omogućuje slaganje skeniranih stranica unutar definiranih margina koje su postavljene kao za knjigu. Kako bi optičko prepoznavanje znakova bilo preciznije, važno je da kvaliteta skenirane slike bude visoka. Česti su problemi pri skeniranju na koje treba paziti premala rezolucija slike, crni rubovi, loša orijentacija slike ili bilo koji drugi oblik gubitka sadržaja sa stranice. Treba uvijek provjeriti je li skeniran cijeli sadržaj knjige jer u nekim leksikografskim izdanjima, npr. *Enciklopediji likovnih umjetnosti,* postoje slikovni prilozi koji nisu obrojčeni, tj. umetnuti su između stranica enciklopedije koje su obrojčene. Također neki slikovni prilozi nalaze se na dvije stranice i time skener treba obuhvatiti veće područje za skeniranje te se dobiveni sken po potrebi treba prilagoditi za prikaz u digitalnoj knjizi. Abbyy Finereader podržava 201 prirodni jezik, uključujući i hrvatski, pa je omogućeno prepoznavanje dijakritičkih znakova. Bilo je potrebno osigurati da program prepozna i naglaske. Unatoč tome što je proces optičkoga prepoznavanja znakova u novim programima sve precizniji, te je trenutačna procjena točnosti 99,8 % za tiskane tekstove,[9] i dalje je potrebno provjeriti tekst (posebno kod posebnih znakova koji se ne nalaze u ASCII kodu).[10]

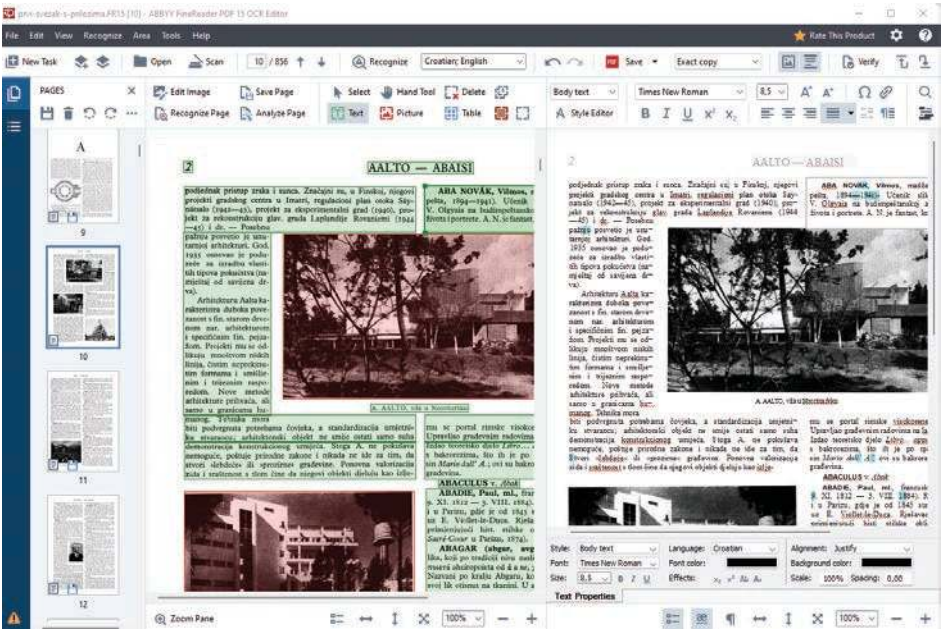
3. Ispravak teksta i slike

Nakon skeniranja treba pregledati i ispraviti rezultate dobivene u prvome koraku. Abbyy Finereader omogućuje istodobni pregled slike i teksta te mogućnost izravnoga uređivanja prepoznatoga teksta i skenirane slike (vidi sliku 1). Sitne pogreške koje nastaju tijekom skeniranja poput mrlja, crnih rubova i iskrivljene orijentacije slike mogu se u programu popraviti izravno s pomoću nekoliko pritisaka mišem. Pri uređivanju slika u nekim slučajevima potrebno je maknuti elemente stranice koji nisu dio izvornoga dokumenta, a nalazili su se na skeniranome papiru. Primjer su toga žigovi te bilješke pisane olovkom ili kemijskom. Složeniji procesi poput spajanja stra-

[9] https://theaccessibilityguy.com/abbyy-finereader-p/ (pristupljeno 20. VII. 2024.)

[10] ASCII *(American Standard Code for Information Interchange)* je način kodiranja koji predstavlja tekst u računalima, komunikacijskoj opremi i drugim napravama. ASCII za kodiranje znakova koristi se samo sa 7 bitova te najviše može prikazati 128 znakova koji se nalaze u engleskoj abecedi (Panian 2005a, 38).

nica kod slikovnih priloga karata radili su se s pomoću GIMP programa za uređivanje slika.[11]



Slika 1. Prikaz pregleda i ispravaka teksta u *Abbyy Finerearderu* / Text review and corrections made using *Abbyy Finereader*

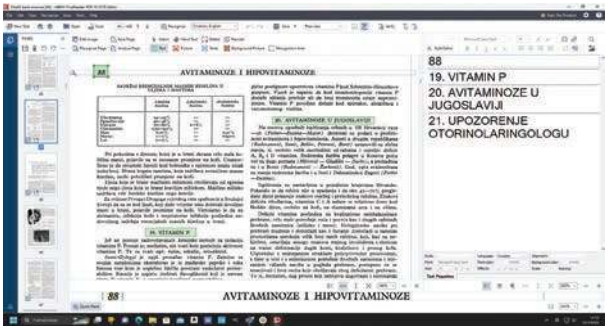
Tekst koji se dobije optičkim prepoznavanjem znakova može se izravno uređivati u sučelju programa. Program će bojom u tekstu označiti određene znakove za koje nije siguran jesu li točno prepoznati, često je riječ o dijakritičkim ili posebnim znakovima, kako bi pomogao uredniku pri pregledu teksta. Česte pogreške pri prepoznavanju teksta događaju se sa znakovima koji nisu dio ASCII kodiranja, što uključuje dijakritičke znakove, znakove s naglascima te druge posebne simbole, npr. slovo ć ako nije dobar sken može se prepoznati kao *ć* ili ĉ. Nakon što su pregledani bojom označeni dijelovi u sučelju programa, cijeli se tekst izvezao iz programa kao dokument u Wordu te je ponovno išao urednicima na pregled. U dokumentu u Wordu često je trebalo ispravljati oblikovanje teksta (podebljanje i kurziviranje teksta). Tekst se oblikuje tako da bude što sličniji tekstu u knjizi. U Wordu je dodatno provedeno odvajanje natuknica s pomoću praznoga paragrafa kako bi se omogućio idući korak.

[11] https://www.gimp.org/ (pristupljeno 9. VII. 2024.)

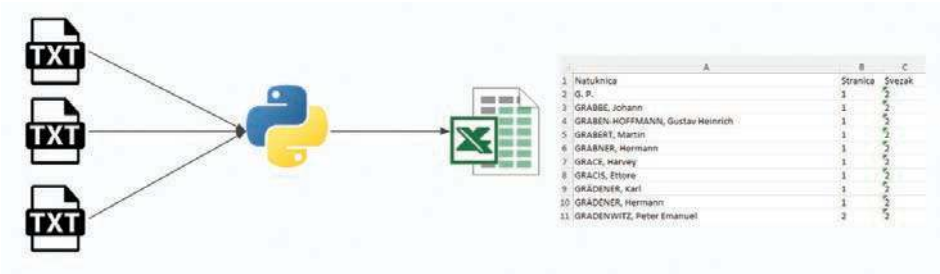
4. Izrada abecedarija

Korak izrade abecedarija može se napraviti nakon drugoga koraka ili raditi istodobno s njime. Prijedlog je da se radi istodobno kako bi se brže mogla stvoriti osnova za izradu baze podataka (četvrti korak). Razlog zašto se ovaj korak izdvaja kao zaseban je taj da traje kraće od drugoga koraka te se može napraviti prije nego što se završi inicijalna provjera teksta. Čak ako provjeru teksta u drugome koraku nismo dovršili, na temelju abecedarija može se izraditi prva inačica digitalnoga enciklopedijskog ili leksikografskog izdanja. Budući da se za svaku natuknicu zapisuje na kojim se stranicama nalazi, mogu se lako napraviti poveznice na odgovarajuće stranice PDF dokumenta. Izrađeni abecedarij mora slijediti strukturu tiskane knjige, što je u većini slučajeva abecedno, ali treba paziti na to da dobivena tablica s natuknicama ima isti redoslijed natuknica kao u knjizi, npr. u *Sportskome leksikonu* u knjizi nakon natuknice *»četiri mušketira«* nalazi se natuknica »420«, pa pri automatskome abecednom sortiranju natuknica treba paziti na to da se ta natuknica ne prebaci na početak abecedarija. Spomenutu natuknicu također na stranici treba staviti pod slovo č.

Pri izradi abecedarija ponovno se koristio Abbyy Finereader program, ali od prepoznatoga teksta jedino su ostavljene oznake za natuknice te brojevi stranica (vidi sliku 2). Tekstovi stranica izvezeni su u .txt datoteke. Ovisno o knjizi, numeracija stranice može biti na vrhu ili na dnu te će prvi ili posljednji redak .txt datoteka sadržavati broj stranice. Natuknice stranica bile su napisane iznad ili ispod broja stranice. Svaka natuknica stavljena je u poseban red, odvojen tipkom *enter*. Nakon što se svaka stranica pretvorila u .txt datoteku, s pomoću skripte napisane u programskome jeziku Python sve te datoteke automatski su spojene u Excel tablicu, u kojoj je za svaku natuknicu napisana stranica na kojoj se nalazi u knjizi (vidi sliku 3). Ta tablica služi kao osnova u idućemu koraku pri izradi baze podataka. Kod izdanja koje se sastoje od više svezaka (npr. *Enciklopedija likovnih umjetnosti* ima četiri sveska), abecedarij svakoga sveska radio se zasebno te su naknadno abecedariji spojeni unutar Excel datoteke.



Slika 2. Prikaz označavanja natuknica i brojeva stranica za abecedarij unutar izdanja *Otorinolaringologija* / Entry and page number marking for the purposes of creating the alphabetic index of the publication Otorhinolaryngology



Slika 3. Proces automatskoga pretvaranja .txt datoteka s natuknicama po stranicama u jednu tablicu koja sadržava sve podatke abecedarija / The process of the automatic conversion of .txt files with entries per page into a single table containing all the data in the alphabetic index

Za knjigu *Otorinolaringologija* posljednji korak izrade abecedarija drukčije je izgledao jer je riječ o izdanju koje nije klasično leksikografsko izdanje te nema podjelu po natuknicama nego hijerarhijski organiziran sadržaj podijeljen na poglavlja i potpoglavlja. Stoga su se stvarala poglavlja i potpoglavlja po stranicama te se ti podatci nisu spojili u Excel tablicu nego u Word dokument, koji se poslije koristio za izradu baze podataka u formatu JSON.

5. Izrada baze podataka

Da bi se započelo s provedbom ovoga koraka, nužno je da prethodni korak bude završen. Baza podataka skup je podataka koji su organizirani tako da program može brzo odabrati željeni podatak. Podatci u tradicionalnim bazama podatka organizirani su u polja u tablici (Panian 2005, 144). Abecedariji izrađeni u prethodnome koraku kao Excel tablice služili su kao osnova za daljnju izradu baza jer su podatci abecedarija već organizirani u polja. Kako bi više suradnika moglo zajedno i neometano ispunjavati bazu, Excel tablice prebačene su na mrežu te je pristup tablicama omogućen samo urednicima. U našemu slučaju prethodno stvorene Excel tablice pretvorene su u Google tablice jer Google tablice omogućuju izravno povezivanje podataka s mrežnim stranicama u CSV formatu. Važno je kod baza odrediti koja su sve polja potrebna na temelju sadržaja knjige. Kod određivanja tih polja, osim osnovnih polja za natuknicu, broj sveska, stranicu natuknice i teksta koji smo dobili u drugome koraku, važno je gledati i na funkcionalnosti koje bi se u budućnosti mogle ugraditi poput međusobnoga povezivanja natuknica unutar istoga ili vanjskoga izvora te poveznice s grafičkim prilozima ili drugim dodatcima. Ako se natuknice različitih izdanja povezuju (npr. povezivanje *Enciklopedije likovnih umjetnosti* s *Enciklopedijom hrvatske umjetnosti*), važno je da svaka natuknica u svakome izdanju ima određenu jedinstvenu identifikacijsku oznaku ili ID. Time je u tablicama za svako izdanje napravljeno polje ID te su se automatski po redoslijedu natuknicama pridružili brojevi kao ID oznake. Te ID oznake dalje mogu služiti za povezivanje natuknica

između baza enciklopedijskih i leksikografskih izdanja te se mogu koristiti kako bi se napravile izravne veze na natuknicu s pomoću hiperveze (u prikazanome slučaju to se radilo s pomoću oznake #).

Pregledani tekst koji se dobio u drugome koraku trebalo je pretvoriti u HTML oznake kako bi se ispravno prikazao na mreži. Unatoč tomu što postoje mnoga programska rješenja za pretvaranje teksta u HTML kod (npr. https://wordtohtml.net/ i https://wordhtml.com/), pri izboru pretvarača treba paziti na to da on točno pretvori cijeli tekst s njegovim oblikovanjima u HTML. Problem mogu prouzročiti oblikovanja određenih dijelova teksta (podebljanje, kurziv itd.), znakovi koji nisu dio ASCII koda te razmaci. U prikazanome slučaju koristili smo se modulom unutar Python skripte koji se zove Mammoth,[12] koji je automatski ubacio tekst iz dokumenta u Wordu u odgovarajuća polja u tablici.

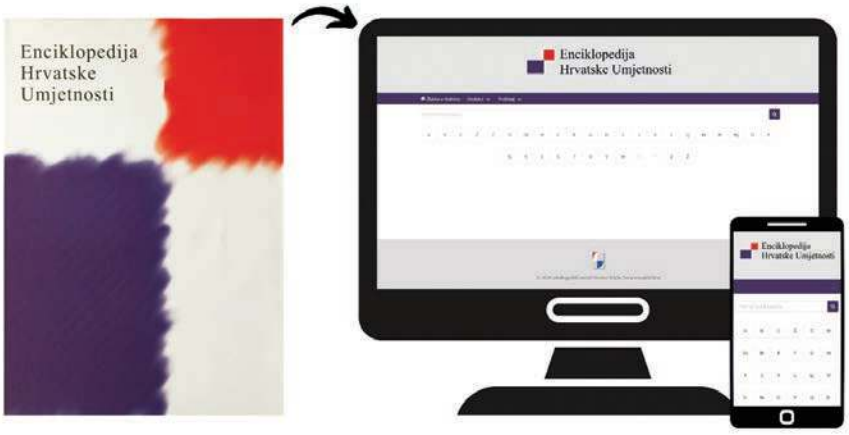
U slučaju izdanja *Otorinolaringologija,* u kojoj nema natuknica nego je sadržaj hijerarhijski organiziran unutar poglavlja, baza je složena u JSON formatu jer omogućuje ugnježđivanje podataka, tj. može se duboko u više slojeva zapisati i učitati podatke. Za svako poglavlje napisane su stranice te je izrađena izravna poveznica na njih u PDF dokumentu. Ti podatci izravno se učitavaju na mrežnu stranicu te prikazuju s pomoću padajućih izbornika (vidi sliku 4).

Slika 4. Prikaz prijenosa podataka zapisanih iz *JSON* datoteke u padajući izbornik na mrežnoj stranici / The transfer of data formatted as a *JSON* file into a drop-down menu on the website

[12] https://pypi.org/project/mammoth/ (pristupljeno 25. VII. 2024.)

6. Izrada mrežne stranice za prikaz prethodno strukturiranih podataka

Osnovni dizajn za mrežne stranice počeo se izrađivati istodobno s drugim korakom. Kad je završen četvrti korak, dodatno se prilagodio dizajn stranice s obzirom na strukturu podataka te dodatne funkcionalnosti stranice. Grafički dizajn stranice temeljio se na dizajnu korica izdanja. Koristile su se boje korica te logo koji se nalazi na koricama kako bi se uspostavio isti vizualni identitet izdanja i mrežne stranice (vidi sliku 5). Budući da su sva spomenuta izdanja koja su prebačena na mrežu povezana na stranicama zbirke *Enciklopedijska baština,* odlučeno je da će dijeliti isti predložak za dizajn kako bi se znalo da pripadaju istoj cjelini. Dizajn je rađen tako da bude što responzivniji, pa se pregled stranice može prilagođavati na uređajima s različitim veličinama ekrana.

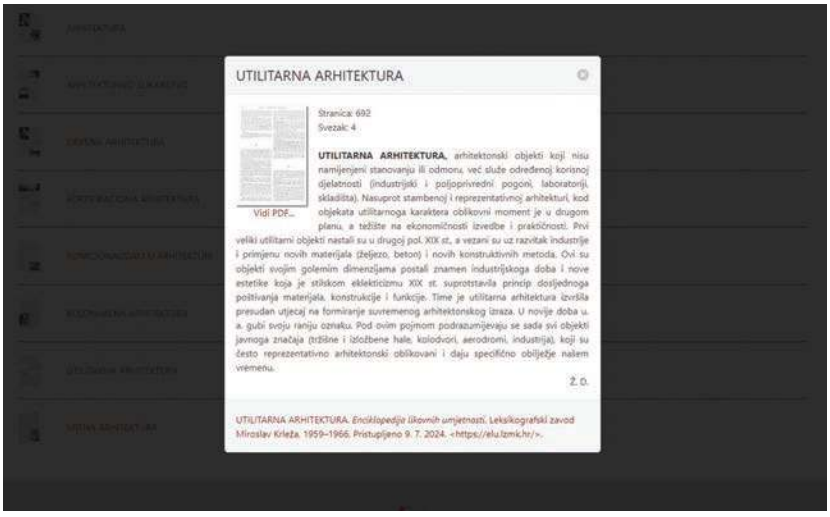


Slika 5. Usporedba korica *Enciklopedije hrvatske umjetnosti* s mrežnom stranicom / Comparison of the cover of the *Encyclopedia of Croatian Art* with the website

Podatci iz baze podatka dinamično se učitavaju nakon što se otvori mrežna stranica. Sve natuknice raspoređene su abecedno te se stvorio indeks za njihovo pretraživanje. Natuknice je moguće pretraživati abecedno pritiskom na tipku slova ili s pomoću tražilice. Tražilica s pomoću indeksa nudi korisniku moguće rezultate na temelju znakova koje je utipkao (vidi sliku 6).

Slika 6. Primjer funkcioniranja tražilice na mrežnoj stranici *Enciklopedije likovnih umjetnosti* / Example of the search engine functionality of the *Encyclopedia of Visual Arts* website

Sadržaji natuknica otvaraju se izravno na stranici s pomoću infoprozora (vidi sliku 7). U slučaju određenih natuknica s mnogo teksta omogućeno je klizanje sadržaja (engl. *scrolling*) unutar infoprozora. Na dnu infoprozora za svaku natuknicu automatski se stvara tekst za citiranje natuknice, s datumom pristupa natuknici, koji se pritiskom može kopirati u međumemoriju[13] računala. Infoprozori također mogu sadržavati poveznice na natuknice koje se nalaze u srodnome enciklopedijskom ili leksikografskom izdanju te slikovne priloge koji su spremljeni kao zaseban PDF dokument.



Slika 7. Prikaz infoprozora za natuknicu »utilitarna arhitektura« unutar *Enciklopedije likovnih umjetnosti* / Pop-up window for the entry on »utilitarian architecture« in the *Encyclopedia of Visual Arts*

[13] Međumemorija ili engl. *clipboard* privremena je memorija u kojoj se podatci privremeno čuvaju prije kopiranja na drugo mjesto ili do zapisivanja novih podataka u memoriju (Panian 2005a, 91).

Kod svake natuknice postoji i veza sa stranicama u knjizi s pomoću poveznice za PDF dokumente. Unatoč tomu što mrežni preglednici mogu izravno otvoriti PDF dokumente, za njihov izravan pregled na mrežnoj stranici koristio se dodatak PDF.js[14] jer se kod mobilnih uređaja poveznice na PDF dokumente često moraju preuzeti prije nego što se mogu otvoriti i čitati. PDF.js omogućuje da se PDF dokument izravno otvori na mrežnoj stranici bez potrebe za preuzimanjem dokumenta. Korisnicima je omogućeno i da na stranici listaju knjigu s pomoću *flipbook* dodatka DearFlip Lite Version.[15] Taj dodatak učitava PDF te omogućuje korisnicima da ga listaju u digitalnome izdanju kao knjigu (vidi sliku 8). PDF za *flipbook* inačicu sadržava cijelu knjigu s koricama, impresumom, grafičkim prilozima i ostalim dodatcima.

Slika 8. Prikaz listanja PDF-a *Enciklopedije likovnih umjetnosti* u *flipbook* formatu / Turning the pages of the *Encyclopedia of Visual Arts* in the PDF flipbook format

Početna stranica zbirke *Enciklopedijska baština* sadržava tražilicu koja pretražuje 95 000 natuknica iz 11 digitaliziranih izdanja u zbirci. Iznimka je *Otorinolaringologija*, 12. izdanje, koja zbog specifičnosti sadržaja (prikazuje se s pomoću kazala sadržaja, a obuhvaća 1200 poglavlja) nije uključena u rezultate pretraživanja. U rezultatima pretraživanja uz natuknicu prikazan je naziv i korice enciklopedijskoga ili leksikografskoga izdanja (vidi sliku 9). Odabirom natuknice u rezultatu pretraživanja otvara se infoprozor odabranoga izdanja s pripremljenim prikazom odabrane natuknice.

[14] https://mozilla.github.io/pdf.js/ (26. VII. 2024.)

[15] https://github.com/dearhive/dearflip-jquery-flipbook (26. VII. 2024.)

Slika 9. Početni rezultati pretraživanja za natuknicu »arhitektura« u tražilici zbirke *Enciklopedijska baština* / Top search results for the ***Architecture*** entry in the search engine of *The Encyclopedic Heritage Collection*

Osim mogućnosti pretraživanja cijeloga sadržaja zbirke moguće je odabrati zasebno izdanje koje se želi pretražiti i pregledati.

7. Objava sadržaja na mreži

Nakon što se mrežni sadržaji izrade i testiraju njihove funkcionalnosti, potrebno je dodati odgovarajuće metaoznake u HTML formatu (naslov, opis stranice, ključne riječi, autori) na sadržaje kako bi sadržaj bio dobro opisan te se dobro rangirao na mrežnim tražilicama. Programi koji se bave dohvaćanjem metapodataka te njihovim uvrštavanjem u mrežne tražilice zovu se mrežni pauci (eng. *web crawler*). Način na koji mrežne tražilice nakon što kroz paukove indeksiraju mrežnu stranicu određuju njihovu relevantnost u pretraživanju ovisi o mnogim čimbenicima koji uključuju povezanost sadržaja stranice s upitom iz pretraživanja (temelji se često na istim riječima), broj posjeta i korištenja stranicom (stranice s više posjeta bolje su rangirane u pretraživanju), kvalitetu sadržaja (lošije su rangirani sadržaji koji imaju prijavljene lažne ili nepotpune informacije, pokušaji prijevare ili sadržaji koji nisu prilagođeni za prikaz na mobilnim uređajima), lokaciju na kojoj je nastala mrežna stranica i upit pretraživanja (korisniku se češće nude mrežna mjesta koja su nastala u njegovoj nego u stranoj zemlji) (Google 2024).

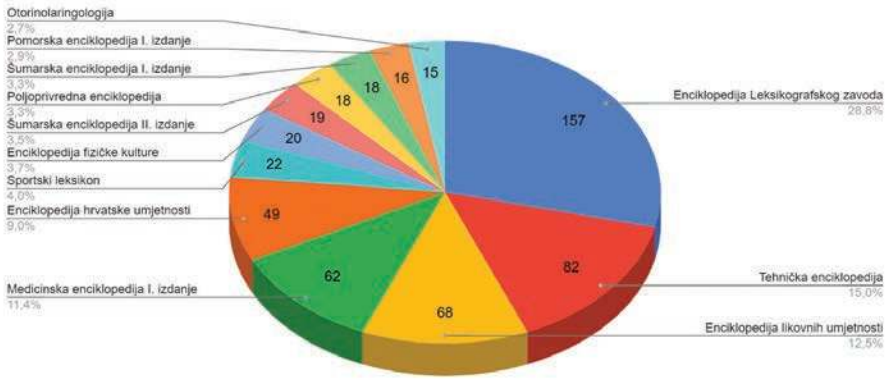
Osim osnovnih metapodataka koji opisuju sadržaj stranice za mrežne pauke, trebalo je dodati i posebne metaoznake za društvene mreže, koje se zovu *Open Graph Meta Tags*. One određuju kako će se prikazivati poveznice kad se dijele na društve-

nim mrežama (Pecánek, 2020). S pomoću njih može se oblikovati poveznice tako da se prikazuju kao ikone s naslovom, kratkim opisom i sličicom (vidi sliku 10).

Slika 10. Prikaz izgleda poveznice koja sadržava *Open Graph* metaoznake kad se objavljuje na društvenoj mreži *Facebook* / A link containing *Open Graph* meta tags when published on the Facebook social network

Posljednje što treba dodati na stranicu prije objave mogućnosti su praćenja posjeta i angažiranosti korisnika na stranicama. Za to se najviše upotrebljava Googleova analitika (W3Tech 2024). Googleova analitika omogućuje anonimno praćenje vremena posjeta stranici, a može se pratiti i ponašanje korisnika na stranici (koliko dugo se zadržavaju na određenim stranicama te kako se kreću sa stranice na stranicu), demografski podatci o korisnicima te podatci o mrežnim preglednicima te tipovi uređaja kojima se koriste pri pregledavaju stranica (Google Analytics 2024). Na svako digitalizirano izdanje postavljen je kod za praćenje, s time da se aktivnosti korisnika

mogu pratiti samo ako oni to odobre pri posjetu stranici. Postoji obavijest na stranicama o praćenju koju korisnik tijekom prvoga posjeta stranicama unutar zbirke *Enciklopedijska baština* može odobriti ili odbiti. Zbirka *Enciklopedijska baština* sa svim digitaliziranim enciklopedijskim i leksikografskim izdanjima objavljena je u prosincu 2023. godine, a dopunjena u ožujku 2024. U pet analiziranih mjeseci (od veljače do srpnja 2024) najviše je pregleda imala *Enciklopedija Leksikografskog zavoda* (157), pa *Tehnička enciklopedija* (82) te *Enciklopedija likovnih umjetnosti* (68), što prikazuje grafikon 1.



Grafikon 1. Podatci o pregledu digitaliziranih enciklopedijskih i leksikografskih izdanja od veljače 2024. do srpnja 2024. / Data on the number of views of digitised encyclopedic and lexicographic editions between February 2024 and July 2024

8. Zaključak

Cilj je ovoga rada prikazati model digitalizacije arhivskih tiskanih enciklopedijskih i leksikografskih izdanja koji su autori rada razvili za objavu enciklopedijskoga i leksikografskoga sadržaja na mreži. Rezultat je ovog modela *Zbirka enciklopedijske baštine* u kojoj se mogu pretraživati i pregledavati arhivska enciklopedijska i leksikografska izdanja Leksikografskoga zavoda Miroslav Krleža. Razvijeni model sastoji se od šest koraka. Unatoč tome što je model nastao u Zavodu te je utemeljen na njegovim izdanjima, može se uz male izmjene primijeniti i na enciklopedijske i leksikografske projekte drugih ustanova. Uporaba tehnologije te način i redoslijed provođenja spomenutih koraka ovisi o tipu sadržaja koji se digitalizira, jer nije isto digitalizirati za mrežu enciklopediju, udžbenik, glosar ili časopis zbog različite strukture i prikaza sadržaja. Također, nemaju sve ustanove jednako dostupna tehnološka sredstva i ljud-

ske potencijale za digitalizaciju. Važno je samo da se unutar koraka digitalizacije strogo odrede procesi izvođenja radnji s odabranom tehnologijom kako bi rezultat bio uspješan i učinkovit bez mnogo ispravaka. Izrađeni model u ovom radu primijeniti će se u daljnjoj digitalizaciji enciklopedijskih i leksikografskih izdanja u Zavodu, pa će proces njihove objave na mreži ići brže, ali također moguće je da će doći do novih spoznaja koje će unaprijediti model. Ako se druge ustanove koriste spomenutim modelom u svojemu radu, moguće je da će ga trebati prilagoditi svojim potrebama ili stvoriti vlastiti model koji poslije i drugi mogu preuzeti te dalje prilagođivati svojim potrebama. Također treba napomenuti da unatoč tomu što je većina koraka napravljena uz pomoć tehnologije, tj. alatima, npr. optičko prepoznavanje teksta, pretvaranje teksta u HTML format te automatsko upisivanje podatka u bazu s pomoću zadanoga algoritma, ona i dalje nije dovoljno razvijena da se bez ljudske provjere njezini rezultati objavljuju kao konačna inačica. U posljednje vrijeme dosta se govori o uporabi umjetne inteligencije u različitim računalnim poslovima kako bi se ubrzao rad. Istraživanja pokazuju da se proces pregleda kvalitete digitalizacije s pomoću umjetne inteligencije još uvijek temelji na generativnim prethodno treniranim pretvaračima (engl. *Generative pre-trained transformer*). Programi kao ChatGPT nisu još dovoljno razvijeni da rade provjere i ispravke nakon digitalizacije zbog nedovoljnoga razumijevanja konteksta različitih sadržaja te mogućnosti proizvodnje neispravnih podataka (Mitra 2023, 7). Moguće je unaprijediti model umjetne inteligencije dodavanjem većega uzorka te količine obrađenih podataka koji se mogu koristiti za trening i unaprjeđenje obavljanja poslova, ali to zahtijeva mnogo vremena i novca (Martinez, 2014). Proces digitalizacije ubuduće će se vjerojatno unaprijediti s razvojem umjetne inteligencije te će se ona više upotrebljavati i time će doći do usavršenih modela digitalizacije koji će se vjerojatno sastojati od manje koraka. Nakon nastanka novih modela trebat će utvrditi nove standarde i načine na koje se digitalizira s obzirom na mogućnosti nove tehnologije. Čak i kad budući modeli budu mogli primijeniti umjetnu inteligenciju, leksikografi će i dalje još dugo raditi velik dio posla provjere i ispravljanja valjanosti podataka, a ključni su i pri izboru tehnologije i određivanju strukture baze podataka te funkcionalnosti koje mrežna stranica mora imati.

LITERATURA

Aldoseri, Abdulaziz, Al-Khalifa, Khalifa N., Hamouda, Abdel Magid. 2023. »Re-Thinking Data Strategy and Integration for Artificial Intelligence: Concepts, Opportunities, and Challenges«. *Appl. Sci,* 13(12), 1–33. https://doi.org/ 10.3390/app13127082

Gorenšek, Tilen, Kohont, Andrej. 2019. »Conceptualization of digitalization: opportunities and challenges for organizations in the Euro-Mediterranean area«. *International journal of Euro-Mediterranean studies*, 12(2): 95–96.

Horvat, Marijan, Kramarić, Martina. 2021. »Retro-Digitization of Croatian Pre-Standard Grammars«. *Athens Journal of Philology*, 8(4): 297–310.

Panian, Željko. 2005a. *Informatički enciklopedijski rječnik: @-L*. Zagreb: Europapress holding d.o.o.

Panian, Željko. 2005b. *Informatički enciklopedijski rječnik: M-Z*. Zagreb: Europapress holding d.o.o.

Mihaljević, Josip. 2017. »Može li računalo pročitati tekst na hrvatskome jeziku?«. *Hrvatski jezik: znanstveno-popularni časopis za kulturu hrvatskoga jezika,* 4(4): 19—23.

Mitra Manu. 2023. »ChatGPT: Capabilities, Limitations, and Ethical Considerations from the Perspective of ChatGPT«. *Oriental Journal of Computer Science and Technology,* 16(2): 1—11.

INTERNETSKI IZVORI

Abbyy Finereader PDF – The Accessibility Guy: https://theaccessibilityguy.com/abbyy-finereader-p/ (pristupljeno 20. VII. 2024)

DearFlip jQuery Flipbook Plugin – GitHub: https://github.com/dearhive/dearflip-jquery-flipbook (pristupljeno 26. VII. 2024)

Digitalizacija – Hrvatska enciklopedija: https://www.enciklopedija.hr/natuknica.aspx?id=68025 (pristupljeno 19. VII. 2024)

G2 – 21 Best OCR Software in 2023. https://www.g2.com/articles/best-ocr-software (pristupljeno 20. VII. 2024)

How Google Analytics works – Google Analytics: https://support.google.com/analytics/answer/12159447?hl=en (pristupljeno 9. VII. 2024)

mammoth – PyPI: https://pypi.org/project/mammoth/ (pristupljeno 25. VII. 2024)

Martinez, Jorge – DocuClipper: https://www.docuclipper.com/blog/ocr-vs-ai/ (pristupljeno 10. VII. 2024)

PDF.js – GitHub: https://mozilla.github.io/pdf.js/ (pristupljeno 26. VII. 2024)

Pecánek, Michal – Open Graph Meta Tags: Everything You Need to Know (2020): https://ahrefs.com/blog/open-graph-meta-tags/ (pristupljeno 9. VII. 2024)

Portal znanja LZMK: https://enciklopedija.lzmk.hr/ (pristupljeno 19. VII. 2024)

Ranking Results: How Google Search Works – Google: https://www.google.com/search/howsearchworks/how-search-works/ranking-results/ (pristupljeno 9. VII. 2024)

Word to HTML: https://wordhtml.com/ (pristupljeno 25. VII. 2024)

Word to HTML: https://wordtohtml.net (pristupljeno 25. VII. 2024)

Zbirka enciklopedijske baštine: https://e-bastina.lzmk.hr/ (pristupljeno 8. VII. 2024)

Tasovac, Toma – Capturing, Modeling and Transforming Lexical Data: An Introduction. Version: https://campus.dariah.eu/resource/posts/capturing-modeling-and-transforming-lexical-data-an-introduction (pristupljeno 31. X. 2024)

THE MODEL OF DIGITISING ARCHIVAL LEXICOGRAPHIC PUBLICATIONS FOR THE WEB

Cvijeta Kraus
The Miroslav Krleža Institute of Lexicography, Zagreb
cvijeta.kraus@lzmk.hr
Josip Mihaljević
The Old Church Slavonic Institute, Zagreb
jmihaljevic@stin.hr
Irina Starčević Stančić
The Miroslav Krleža Institute of Lexicography, Zagreb
irina.starcevic.stancic@lzmk.hr

ABSTRACT: The paper presents a model designed to digitise old lexicographic publications from The Miroslav Krleža Institute of Lexicography for online publication (e-bastina.lzmk.hr). It includes twelve archival publications, the oldest of which are the first edition of the *Maritime Encyclopedia* (1954–1964), the *Encyclopedia of the Institute of Lexicography* (1955–1964), and the *Medical Encyclopedia* (1957–1965), while the most recent publication is the *Encyclopedia of Croatian Art* (1995–1996). These lexicographic publications have not been available in digital form until now, and their digitisation began with the scanning of printed books. Each lexicographic work varies in content, structure, and presentation of graphic supplements. Therefore, it was necessary to design a digitisation and online publication model that would work for different lexicographic publications. The presented model consists of six steps: 1) page scanning and optical character recognition, 2) text and image editing, 3) creation of an alphabetic index, 4) creation of a database, 5) development of a website for displaying previously structured data, and 6) online publication. Each of these steps incorporates multiple processes, which are determined by available technology, human knowledge, and resources. The paper explains, analyses, and illustrates each step with examples from the authors' lexicographic practice, to ensure that the presented model can also be applied to the digitisation of other lexicographic publications. Another outcome of this paper is the presentation of the functional website called *The Encyclopedic Heritage Collection* (e-bastina.lzmk.hr), which was developed using this model. This site consists of twelve digitised lexicographic editions and contains 57 volumes available online with searchable entries. The total number of entries in all editions is 95,000. For each search result, the source is displayed, indicating the encyclopedia or lexicon in which the entry is found.

Keywords: entry index; database; digitisation; digitisation model; web lexicography; text and image editing