Digitalizacija arhivskoga gradiva s uključenim procesom optičkog prepoznavanja znakova na primjeru Muzičkih novina Hrvatskog državnog konzervatorija

Kitin, Ilija

SVEUČILIŠTE U ZAGREBU
FILOZOFSKI FAKULTET
ODSJEK ZA INFORMACIJSKE I KOMUNIKACIJSKE ZNANOSTI
SMJER BIBLIOTEKARSTVO
Ak. god. 2022./ 2023

Ilija Kitin

Digitalizacija arhivskoga gradiva s uključenim procesom optičkog prepoznavanja znakova na primjeru Muzičkih novina Hrvatskog državnog konzervatorija

Diplomski rad

Mentor: dr. sc. Željko Trbušić

Zagreb, prosinac 2023.

Izjava o akademskoj čestitosti

Izjavljujem da je ovaj rad rezultat mog vlastitog rada koji se temelji na istraživanjima te objavljenoj i citiranoj literaturi. Izjavljujem da nijedan dio rada nije napisan na nedozvoljen način, odnosno da je prepisan iz necitiranog rada, te da nijedan dio rada ne krši bilo čija autorska prava. Također izjavljujem da nijedan dio rada nije korišten za bilo koji drugi rad u bilo kojoj drugoj visokoškolskoj, znanstvenoj ili obrazovnoj ustanovi.

Za mamu i tatu.

Zahvaljujem se zaposlenicama Knjižnice Muzičke akademije za pomoć pri izradi rada.

Sadržaj

1. Uvod

Diplomski rad obrađuje temu procesa digitalizacije u knjižnici Muzičke akademije u Zagrebu s uključenim procesom optičkog prepoznavanje znakova na primjeru Muzičkih novina Hrvatskog državnog konzervatorija. Tema je odabrana jer se obavljanje studentske prakse izvodilo u knjižnici Muzičke akademije, gdje su zaposlenice predložile digitalizaciju starih novina koje knjižnica posjeduje. U tu svrhu provedeno je istraživanje koje se sastojalo od testiranja dostupnih uređaja s ciljem ubrzanja procesa digitalizacije i ostvarenja više točnosti optičkog prepoznavanja znakova (engl. Optical Character Recognition, OCR). Za potrebe provjere OCR-a, izrađene su dvije verzije temeljnog teksta. Temeljni tekst (engl. *ground truth*) kompletan je i točan zapis svakoga znaka i riječi sa slike. Može se uspoređivati s rezultatima OCR programa kako bi se procijenila njegova točnost i odredilo koliku važnost ima bilo kakvo odstupanje od temeljnog teksta u tom trenutku. Dvije verzije temeljnog teksta korištene su u testiranju dostupnih rezolucija uređaja u odnosu na višu točnost OCR-a. U istraživanje su uključeni optimizacijski parametri, a to su vrijeme u radnim satima i kapacitet pohrane.

Diplomski rad podijeljen je u dva dijela, teorijski i praktični, i četiri poglavlja. U prvom poglavlju opisuje se postupak digitalizacije, koji uključuje odabir, digitalizaciju, obradu i kontrolu, zaštitu, pohranu i prijenos, pregled i korištenje i održavanje. Također se spominju neki od problema koji se pojavljuju. U drugom poglavlju, bit će riječi o postupku OCR-a i izazovima te o temeljnom tekstu. Nakon toga, u sljedećem poglavlju, definirat će se osnovni koncepti vezani uz građu, a to su publikacija, serijska publikacija i novine. U radu se koristi pojam građa naspram pojma gradivo jer je u istraživanju korištena knjižnična građa, a ne arhivsko gradivo. U dijelu s istraživanjem, opisat će se postupak odabira i obrade, skeniranja i OCR-a, s predstavljenim vremenom trajanja postupaka i procijenjenim troškovima projekta koji uključuje vrlo veliki broj stranica kao što je Europeanin projekt zvan Europeana Newspapers. Podatke će se analizirati kako bi se vidjelo koji su od korištenih uređaja, rezolucija itd. korisniji. Na kraju rada nalazi se zaključak sa sintezom rada i prijedlozima za buduće istraživanje.

Digitalizacija (engl. *digitization*) prevođenje je je postupak transfera analognog signala u digitalni oblik. Mogu se digitalizirati tekstovi, slike, zvuk, video ili 3D objekti. Što se tiče teksta, kako bi on postao pretraživ i kako bi ga se moglo indeksirati te kako bi ga se moglo obrađivati, koriste se OCR programi. Optičko prepoznavanje znakova je postupak kojim se

slike, tj. skenovi stranica teksta pretvaraju u računalno čitljive, čime tekst postaje pretraživ. Digitalizacija je alat kojim se ostvaruje naprednije dijeljenje informacija, povećanje dostupnosti, korištenje na daljinu, očuvanje i ušteda na prostoru. OCR softver radi na način da se prepoznaje kontrast teksta i pozadine. Osim korištenjem OCR-a, tekstovi mogu postati pretraživi prepisivanjem. Neki OCR programi su ABBYY FineReader, Recognita i Tesseract.

2. Digitalizacija

U ovome dijelu rada radi se o postupku digitalizacije. Opisuju se postupak i ciljevi digitalizacije, od odabira građe i mjesta, obrade i kontrole kvalitete, zaštite, pohrane, prijenosa, pregleda i korištenja do njezina održavanja. Uz to, prezentiraju se problemi koji se mogu javiti tijekom procesa.

Digitalizacija je u osnovi prevođenje analognog u digitalni oblik. Može se raditi o tekstu, slici, zvuku, video ili trodimenzionalnom objektu, a koji se može obrađivati, pohraniti i prenositi. Digitalizacija slike uključuje fotografije, dijapozitive i grafike, ali i stranice dokumenata i tiskovina te se provodi skeniranjem skenerima i fotografiranjem. Vrijedna i povijesna djela digitaliziraju se fotografskim aparatima, uz posebno osvjetljenje i komore (Digitalizacija, Hrvatska enciklopedija, 2021). Nastale slike teksta mogu se čitati na zaslonima, ali se ne mogu pretraživati i obrađivati. Za pretvaranje takvih digitaliziranih objekata u računalno čitljive koristi se specijalizirani softver za optičko prepoznavanje znakova ili OCR. Može se postići točnost i do 99%, što se smanjuje kod starije građe i posebice rukopisa pa je u tim slučajevima bolje prepisivanje. Nakon toga, pogreške u OCR-u ručno se ispravljaju (Digitalizacija, Hrvatska enciklopedija, 2021).

Digitalizacija se provodi u različitim profesionalnim okruženjima, npr. medicini, inženjerstvu i znanosti, a koriste je i amateri, npr. kod digitalizacije albuma fotografija. Digitalizirana građa često se daje na korištenje javnosti putem mrežnih baza podataka u knjižnicama, arhivima i muzejima (Digitalizacija, Hrvatska enciklopedija, 2021), kao što je slučaj kod građe kojom se ovaj rad bavi.

Digitalizacija u knjižnicama služi za dijeljenje resursa, povećanu dostupnost građe, korištenje na daljinu, očuvanje i uštedu na prostoru. Svrha digitalizacije jest pobrinuti se za zaštitu informacija i pristupa tim informacija sadašnjim i budućim generacijama. U te svrhe mogu se koristiti repozitoriji (Shigwan, 2015), što je upotrijebljeno za građu kojom se ovaj rad bavi. Postoje tri osnovne funkcije digitalnih repozitorija. To su (Shigwan, 2015):

1. nabava digitalne građe ili digitalizacija,
2. pohrana ili upravljanje digitalnim sadržajem,
3. dohvaćanje, tj. prikaz digitalnog sadržaja.

Jedan od izazova s kojim se knjižnice susreću u digitalizaciji je zastarjelost tehnologije. U digitalizaciji potrebni su i softverski i hardverski alati za pohranu građe. Nekad je potrebna i

migracija dokumenata u novi oblik, a namjera je građi osigurati dugoročan pristup. Problem predstavlja propadanje medija na kojem je građa spremljena, ali i konstantne promjene u uređajima za pohranu (Shigwan, 2015).

Digitalizacija se može provoditi radi zaštite, veće dostupnosti, boljih mogućnosti korištenja, razvijanja ponude ili upotpunjavanja fonda (Ministarstvo kulture Republike Hrvatske, 2006 navedeno u Stančić, 2009: 10). Digitalizacija radi zaštite izvornika provodi se zbog toga što se ponudom elektroničke verzije, fizička manje koristi i time manje oštećuje. Elektronička verzija dokumenta služi i kao sigurnosna kopija u slučaju oštećenja. Digitalizacija radi povećanja dostupnosti provodi se zbog mogućnosti pregleda jednog dokumenta na više uređaja, tj. više korisnika može pregledavati isti dokument. Time institucije mogu vidjeti povećanje u broju korisnika, čime se povećava njihova vidljivost i utjecaj u društvu. Moguća je i promocija cijele države digitalizacijom kulturne baštine. Digitalizacija radi stvaranja nove ponude i usluga uključuje razmjenjivanje metapodataka među institucijama, što ubrzava obradu, pretraživanje teksta, analize i organiziranje građe iz fizički udaljenih prostora u virtualne zbirke. Digitalizacija radi upotpunjavanja fonda provodi se npr. kad se radi o nepotpunom, oštećenom ili uništenom dijelu fonda.

U engleskom jeziku postoji razlika između pojmova *digitization* i *digitalization*. Pojam *digitization* javlja se pedesetih godina i označava prijenos analognih podataka (posebice kasnije slika, videa i teksta) u digitalni oblik. S druge strane, pojam *digitalization* definira se kao usvajanje ili povećanje upotrebe digitalne ili računalne tehnologije u nekoj organizaciji, industriji, zemlji itd. (Oxford English Dictionary u Schumacher, Sihn i Erol). Pojam *digitization* je stoga materijalni proces konverzije iz analognog u digitalno, a *digitalization* jest način na koji se sfere društvenog života restrukturiraju oko digitalne komunikacije (Brenner i Kreiss u Schumacher, Sihn i Erol). Digitalizacija se odnosi na automatizaciju procesa poslovanja i operativnih procesa. Radi se na primjer o razlici između pisanja ručno i pisanja na računalu, upotrebe registratora i korištenja servera, skeniranju i korištenju elektronskih potpisa. S druge strane, primjeri digitalizacije bili bi skeniranje i korištenje elektronskih dokumenata umjesto dokumenata u registratorima. Svrha digitalizacije je obrada informacija, njezin cilj je automatizacija poslovnih procesa, što se postiže izradom digitalnih radnih procesa. Njezin izazov je smanjenje troškova, tj. financije. Svrha digitizacije, pak, jest kodiranje. Njezin cilj je pretvorba iz analognog formata u digitalni, što se postiže pretvorbom dokumentacije u papirnatom obliku u digitalni, a njezin izazov je smanjenje obujma, tj.

materijali (Kordić, 2021). Hrvatsko govorno područje i stručna terminologija u načelu ne razlikuje ove pojmove već se koristi pojam digitalizacija za oba koncepta..

2. 1. Zapis

"Zapis je ono što je stvoreno i sačuvano kao dokaz funkcija, aktivnosti i transakcija neke tvrtke ili pojedinca. Da bi se smatrao dokazom, zapis mora imati sadržaj, strukturu i kontekst te mora biti dio sustava za arhiviranje" (National Archives of Australia, 2004 navedeno u Stančić, 2009: 142).

Ovo je suvremena definicija zapisa, nastala zbog potrebe uključivanja elektroničke građe, kako digitalizirane tako i one izvorno nastale u digitalnom obliku, u kojoj se ističe nezavisnost zapisa o njegovom formatu. Definiran na ovaj način, zapis postoji nezavisno o tehnologiji koja se upotrebljava u danom trenutku i postaje virtualan (Stančić, 2009). U osamdesetim godinama prošlog stoljeća, definicija zapisa glasila je ovako:

"Zapis je dokument (uključujući bilo koji rukom pisani ili tiskani materijal) ili objekt (uključujući zvučni zapis, magnetsku traku ili disk, mikrofilm, fotografiju, film, mapu, plan, model, sliku ili neko drugo slikarsko ili grafičko djelo) koji jest ili je bio namjerno sačuvan zbog bilo koje informacije ili sadržaja koje posjeduje, ili značenja koje može biti iz njega izvedeno, ili zbog njegove veze s nekim događajem, osobom, okolnošću ili stvari" (Commonwealth Consolidated Acts).

Elektronička verzija dokumenta nije samo dokument ni samo objekt, već je ona neovisna od medija i razlikuju se njezina fizička i logička struktura. To znači da na računalu ne mora biti zapisan u logičkom slijedu, nego može biti fragmentiran i to na više jedinica, kao što je slučaj u RAID sustavu (*Redundant Array of Inexpensive Drives*). On se sastoji od polja diskova u kojem je zapis podijeljen na blokove, od kojih je svaki zapisan na različiti disk. Ovim postupkom smanjeno je vrijeme zapisivanja i čitanja jer se manje podataka čita odjednom. Primjenjuje se kad je potreban brz pristup građi (Stančić, 2009).

2. 2. Proces digitalizacije

2.2.1. Odabir građe i postavljanje digitalizacijske okoline

Proces digitalizacije započinje odlukom koju građu digitalizirati ili, bolje rečeno, digitalizirati najprije jer se može digitalizirati sve, ovisno o ulaganju. Izbor ovisi o ciljevima institucije.

Ovaj se postupak provodi analizom građe i primjenom kriterija te se odredi i redoslijed digitalizacije. Građa koja bi trebala biti digitalizirana najprije jest ona koja se najčešće koristi i čije digitaliziranje će biti najjeftinije ili, pak, popularna djela jer je time lakše doći do potpore. Važno je voditi računa o pravnim pitanjima. Kako se proces tiče tehnologije, konzervacije i pravnih pitanja, preporuča se osnivanje povjerenstva za odabir koje okuplja stručnjake iz različitih profesija (Vogt-O'Connor, 2000 navedeno u Stančić, 2009: 17). Važno je razviti politiku digitalizacije koja će služiti kao referenca i vodič u implementiranju projekta digitalizacije i koja sadržava ciljeve projekta. Razvijanje ciljeva je važno za nove inicijative. Na primjer, učiniti materijale bolje dostupnima na internetu nije dovoljno specifično. Potrebno je specificirati, posebno što se tiče kategorija korisnika koji će koristiti građu, vrste materijala koja ih može zanimati, na koji će ih način koristiti i koliko korisnika će ih koristiti. Važna je korist koja se donosi korisnicima i instituciji. Jedan od načina na koje se može dobiti ove informacije je kontaktiranje sadašnjih i potencijalnih korisnika putem ankete ili kontaktiranje drugih ustanova koje su digitalizirale sličnu građu te učenje iz njihovih uspjeha i neuspjeha (Pandey i Misra, 2014). Savjetuje se oformiti vijeće za planiranje koje će razviti plan i proračun projekta digitalizacije. Proračun bi trebao uključiti plaće, koje će vjerojatno iznositi otprilike 50% iznosa projekta), osposobljavanje osoblja, opremu, usluge, ugovore i legalne troškove te indirektne troškove, primjerice ured i radni prostor, troškove održavanja, licenci i komunikacije, a otprilike 10% cjelokupnog proračuna trebalo bi se rezervirati za nepredvidive troškove (Oluakachukwu Nneji, 2018).

Prije početka projekta digitalizacije, mora se osigurati postojanje ili nabava prikladne tehnologije, tj. sve opreme, hardvera ili softvera, koja je potrebna. Također se mora odlučiti i o načinu rada, tj. hoće li se uspostaviti veze s već postojećim digitalnim knjižnicama i hoće li se građa digitalizirati unutar ili izvan ustanove (Oluakachukwu Nneji, 2018). Prednost digitalizacije izvan ustanove je brzo vrijeme dostave, to što može biti više različitih uređaja za skeniranje na raspolaganju, a sama ustanova ne mora brinuti o troškovima zastarijevanja opreme. Nedostaci digitalizacije izvan ustanove su potreba za transportom materijala i osiguranje kvalitete te ugovora (Parekh, 2001). Postoji i potreba za određivanjem vremenskog ograničenja za projekt digitalizacije. Nadalje, važna je verifikacija sadržaja. Nakon odabira građe za digitalizaciju, važno je utvrditi postoje li prethodno izrađeni digitalni primjerci. Ponovna digitalizacija mogla bi biti potrebna ako je digitalna građa bila izrađena koristeći stariju tehnologiju (Oluakachukwu Nneji, 2018).

2.2.2. Digitalizacija

U procesu digitalizacije mogu se upotrijebiti različiti uređaji za različitu građu. Za tekst i slike to su skeneri i digitalni fotoaparati, za zvučne i video zapise hardver za računala ili drugi uređaji, a za trodimenzionalne objekte uz skenere i fotoaparate, mogu se koristiti i 3D skeneri. U ovom istraživanju korišteni su skeneri pa će biti više riječi o njima.

Karakteristike skenera su brzina, rezolucija, dinamički raspon, polje skeniranja, vezni uređaji, softver i opseg skeniranja. Brzina je vrlo važna u digitalizaciji za velike količine građe i na nju utječu različiti faktori, npr. priprema građe, promjena orijentacije stranica, rezolucija i brzina prijenosa podataka s uređaja za skeniranje na računalo (Stančić, 2009). Rezolucija je "prostorna frekvencija uzimanja uzoraka iz okoline" (Stančić, 2009: 43). To je količina piksela koja se može očitati skenerom. Utječe na kvalitetu pa viša rezolucija znači bolje kvalitetu skena. Rezolucija je broj plošne ili linijske gustoće točaka. Može se mjeriti točkama po inču i označavati kao dpi (*dots per inch*), pikselima po inču ili ppi (*pixel per inch*) ili linijama po inču, tj. lpi (*lines per inch*). Ove se oznake koriste za različite upotrebe. Ppi koristi se kod slika, dpi kod printera, a lpi u nijansiranju (*half-toning*) u tiskarstvu. Odabir rezolucije ovisi o kapacitetu pohrane, tome kolika je kvaliteta skena potrebna i OCR softveru. Polje skeniranja označava dimenzije građe koja se može skenirati. Softver za skeniranje sadrži različite opcije kao što su ravnoteža boje ili svjetlina, a može i na primjer omogućiti automatsko uklanjanje točke na skenu nastale zbog probušenih stranica. Softver može popraviti orijentaciju i ukošenost pa utječe na brzinu (Stančić, 2009). U procesu digitalizacije preporuča se planiranje konvencije za imenovanje datoteka, kojom se može spojiti datoteku s originalom. Preporučljivo je imati pričuvne kopije zbog požara, elementarnih nepogoda itd. (Stančić, 2009).

2.2.3. Obrada i kontrola kvalitete

Na početku procesa odlučuje se o karakteristikama slike koja će se izraditi - koja će biti njezina rezolucija i hoće li ona biti u sivoj skali ili u boji. Siva skala jest raspon nijansi između bijele i crne boje (Strugačevac, 1999). Viša rezolucija podrazumijeva i više potrebnog kapaciteta za pohranu, što je također važno u procesu digitalizacije. Kod donošenja odluka u ovom koraku, vodi se računa o predviđenom korištenju digitalizirane građe. Za obrađivanje teksta mogu se koristiti programi za optičko prepoznavanje znakova (Stančić, 2009), kao što je slučaj u ovom istraživanju.

2.2.4. Zaštita elektroničke građe

Cilj zaštite jest zaštita od neovlaštenog pristupa, kopiranja, distribuiranja i dokazivanje autentičnosti (Stančić, 2009). Možemo spomenuti zaštitu i osiguranje identiteta, tj. "dodjeljivanje prava pristupa određenim datotekama" (Stančić, 2009: 95). Za pravo pristupa može se koristiti lozinka (Stančić, 2009).

2.2.5. Pohrana i prijenos

Pohrana i prijenos važni su iz razloga što bi građa u knjižnicama, arhivima i muzejima te komercijalnim institucijama nakon digitalizacije trebala biti dostupna za korištenje, a način pohrane građe utječe na brzinu i način pristupa. Važna je i pohrana građe koja je izvorno nastala u elektroničkom obliku (*born digitally*).

Postoje kriteriji koji se mogu primijeniti kako bi se odabrao sustav za dugoročnu pohranu. To su (Bell i Waugh, 2000 navedeno u Stančić, 2009: 114):

- dugovječnost medija,
- trajnost medija,
- visoki kapacitet,
- mala cijena,
- široka prihvaćenost te
- sustav mora biti izravan (engl. *on-line*) ili poluizravan.

2.2.6. Pregled i korištenje

Pregled i korištenje digitalizirane građe ovisi o načinu pregleda, npr. na računalu ili će se građa ispisivati, i o tome hoće li se građu pretraživati i koristiti lokalno ili na internetu. Moguće je npr. i da se na internetu mogu pretraživati samo metapodaci i čitati tekst, a druga građa dostupna je samo unutar institucije (Stančić, 2009).

2.2.7. Održavanje

S vremenom se pojavljuju problemi jer sustavi, mediji i zapisi zastarijevaju. Ako se fizičku građu čuva u prikladnim uvjetima, ona će i nakon dugo vremena biti čitljiva. Fizičkoj građi može se prepoznati struktura bez obzira na njezinu starost ili ako se radi o tekstualnoj građi, može se prepoznati jezik. Nasuprot tome, elektronička građa može postati nečitljiva čak i u roku od desetak godina. To se događa zbog stalnog razvijanja tehnologije. Kod elektroničke građe može se dogoditi da program novije verzije ne prikazuje ispravno stariji zapis zbog

njegovog formata. U tom slučaju, može se dogoditi da program prikaže građu u neprepoznatljivom obliku ili se javlja greška. Problem se može pojaviti i ako je građa izrađena na starijem ili različitom operativnom sustavu ili ako se nalazi na mediju starijeg formata. Problemi vezani uz kodiranje uglavnom se odnose na komprimiranje. Na primjer, kod slika se još ne zna hoće li dio informacija koji se gubi prilikom kompresije biti potreban za rad aplikacija u budućnosti. Što se tiče međusobne povezanosti sadržaja, gradivo u digitalnom arhivu može biti jedan zapis, a naslov i drugi metapodaci mogu činiti drugi zapis. Važno je zadržati iste veze pri prebacivanju u novi sustav. Mora se osigurati proaktivnost u procesu održavanja i očuvanja digitalne građe (Stančić, 2009).

Mogu se pojaviti problemi (Besser, 2000):

- pregleda,
- kodiranja,
- međusobne povezanosti sadržaja,
- ovlasti za arhiviranje,
- konverzije.

Stalno mijenjanje softvera i hardvera predstavlja pritisak na ustanovu jer se zaštita digitalne građe zasniva na zaštiti digitalnih informacija, migraciji i osiguravanju dugoročnoga pristupa. Jedan od najvećih problema koji prijeti dugovječnosti digitalne građe je propadanje medija za pohranu i brza promjena uređaja za pohranu. Kod analognih informacija, naglasak je na zaštiti fizičkih predmeta, dok se kod digitalne građe čuva njihov informacijski sadržaj. Stalne promjene u hardveru i softveru uzrokuju tehnološko zastarijevanje gubljenjem načina pristupa informacijama u digitalnom obliku, kao što su stalne nadogradnje operativnog sustava, aplikacija jezika za programiranje i medija za pohranu (Pandey i Misra, 2014). Kako ne bi došlo do tehnološkog zastarijevanja, digitalni arhivi trebali bi biti transkribirani svakih deset do dvadeset godina (Alegbeleye, 2009, u Pandey i Misra, 2014).

3. Optičko prepoznavanje znakova

Ovaj dio rada bavi se postupkom optičkog prepoznavanja znakova, njegovom važnošću i problemima koji se mogu pojaviti tijekom provođenja postupka.

Prvi OCR uređaj bio je mrežasti skener (engl. *retina scanner*), koji je izumio Charles R. Carey. Slika se mogla prenijeti putem mozaika fotoćelija (Trbušić, 2022). Do sredine 1950-ih, OCR uređaji postali su komercijalno dostupni. Prvi pravi OCR uređaj koristio je Reader's Digest 1954. godine. U ovom slučaju, strojopisni izvještaji o prodaji prebacivali su se u probušene kartice za unos u računalo. Sustavi za optičko prepoznavanje znakova koristili su se primjerice u bankovnim sustavima za prepoznavanje malog broja tiskanih znakova, obično brojki i nekoliko posebnih znakova. Korišteni su kod prepoznavanja brojeva računa, identifikacije korisnika, iznosa itd. Korišteni su za automatizaciju, na primjer u automatskom prepoznavanju adresa za sortiranje pošte. Njihova uspješnost u velikoj je mjeri ovisila o količini ručno ispisanih adresa. Ipak, bilo je moguće sortiranje trideset tisuća pisama na sat (Eikvil, 1993).

OCR dostupan industrijski može se podijeliti u četiri generacije, ovisno o snazi, efektivnosti i prilagodljivosti. U prvoj generaciji mogli su se prepoznati samo odabrani stilovi teksta i oblici znakova i koristili su se 1960-ih godina (Hamad i Kaya, 2016). Znakovi su bili posebni dizajnirani kako bi bili strojno čitljivi i nisu izgledali vrlo prirodno. S vremenom se počelo raditi s više oblika slova i pojavili su se uređaji koji su podržavali do deset oblika slova (Eikvil, 1993). U drugoj generaciji mogli su se prepoznavati isprintani znakovi i oni ručno napisani, no mogli su prepoznavati jedino brojčane znakove. Koristili su se od sredine 1960-ih do sredine 1970-ih godina (Hamad i Kaya, 2016). Prvi sustav ove vrste bio je IBM 1287 iz 1965. godine. Toshiba je u tom razdoblju razvio prvi automatski sustav za sortiranje pisama, a Hitachi je razvio prvi OCR sustav visokih performansi, a niskih troškova. Ova faza karakterizira se radom na standardizaciji oblika slova. Razvijen je američki standard OCR znakova zvan OCR-A i europski, nazvan OCR-B, koji se razlikovao po tome što je sadržavao prirodnije oblike slova od američkoga. Pokušalo se sjediniti oblike slova u jedan standard, no pojavili su se uređaji kojima je bilo moguće čitati obje vrste (Eikvil, 1993). U trećoj generaciji radilo se s ručno napisanim znakovima slabe kvalitete printa nego ranije i korišteni su od sredine 70-ih do sredine 80-ih (Hamad i Kaya, 2016). U ovoj generaciji, važan je The Kurzweil Reading Machine (KRM) iz 1976. godine, a kojega je izumio Raymond Kurzweil. Bio je to stroj koji je služio slijepima i slabovidnima na način da bi se tiskani tekst prepoznao

i sintetizirao u zvuk. Uređaj je prvi koji je doživio široku dostupnost, a bio je i pristupačan pa je vodio k primjeni OCR-a u svakodnevnici i daljnjem razvoju tehnologije (Trbušić, 2022). U četvrtoj generaciji znakovi se prepoznaju iz kompleksnih dokumenata. Prepoznaju se matematički simboli, rukopis, dokumenti loše kvalitete i s puno buke. Kvalitetniji OCR dostupan je za arapski, kineski, japanski i latinicu (Hamad i Kaya, 2016). U današnje doba, pojavila se i peta generacija sustava za optičko prepoznavanje znakova, koja se bazira na radu u oblaku (engl. *cloud computing*). Takvi sustavi dostupni su kod Google Cloud Visiona i Amazon Textracta (Trbušić, 2022).

OCR ili optičko prepoznavanje znakova koristi se kako bi se slike, tj. skenove stranica teksta prebacilo u oblik računalno kodiranog teksta, koji računala mogu procesirati i time tekst postaje pretraživ. Druga metoda kojom se ovom može postići je prepisivanje originalnog teksta, postupak poznat kao *keying*. Ovaj postupak provodi se ručno, dok je OCR automatski, što znači da ručno prepisivanje može biti i deset puta skuplje od skeniranja i provođenja OCR-a (Chapman, 2000). Prepisivanje je dugotrajan i vrlo skup proces, iako može biti i najisplativiji u slučaju rukopisa, starih i požutjelih stranica ili stranica na kojima nema dovoljno kontrasta ili tekstovima s rukopisom na marginama. Tada automatizirane tehnike nisu upotrebljive ili dovoljno učinkovite pa su potrebni dodatni ispravci. Time je automatizirani proces digitalizacije skuplji i vremenski zahtjevniji od digitalizacije prepisivanjem (Stančić, 2009).

OCR se u načelu koristi za slike teksta ispisanog korištenjem elektroničkih (računalni pisač) ili mehaničkih (pisaća mašina) uređaja, a nailazi na probleme kod rukom pisanih tekstova. Njih se treba prepisati kako bi postali pretraživi. Drugi izazovi u optičkom prepoznavanju znakova su kompleksni tekstovi. To su oni koji sadrže više oblika slova, mnogo stupaca i ilustracije uz tekst (Chapman, 2000). Nadalje, kvaliteta skenova teksta utječe na kvalitetu OCR-a. Nekoliko istraživanja pokazalo je da obučena osoba može ispraviti od šest do deset stranica na sat, što je prednost naprema prepisivanju (Chapman, 2000). Nadalje, izazov su nearapski znakovi, stariji, neki suvremeni i smanjeni fontovi te kompleksni rasporedi stranica (Gifford Fenton, 2000).

OCR softver radi na principu prepoznavanja kontrasta između teksta i pozadine. Program prepoznati tekst pretvara u obradiv, pretraživ i indeksibilan. Može se provoditi tijekom skeniranja ili nakon (Stančić, 2009). U ovom istraživanju proveden je nakon što su izrađeni svi uzorci. OCR program radi na način da prvo analizira raspored stranice teksta i podijeli ga

u zone koje generalno odgovaraju odlomcima u izvornom tekstu. Nakon toga, određuje se redoslijed odlomaka. Sljedeći je korak analiza znakova. Većina programa za OCR radi na način da traži grupe znakova, tj. riječi i uspoređuje ih s unosima u rječniku programa. Kad je riječ pronađena, program je pridružuje dokumentu. U slučajevima kad se ne može dobro odrediti, program odabire najviše moguće sličnu riječ i označuje je kao nisko pouzdanu (engl. *low confidence output*). Postoji i treći slučaj, a to je situacija u kojoj se riječ ili znak uopće ne mogu prepoznati pa se kao zamjenski simbol (engl. *placeholder)* koristi zadani znak (Gifford Fenton, 2000). Zamjenski simbol je objekt čiju se vrijednost može specificirati kasnije (Isthiaq i Saif, 2020). Segmentacija je proces lociranja regija ispisanog ili ručno napisanog teksta, tj. proces izdvajanja znakova ili riječi. Proces uključuje i razlikovanje teksta od grafika (Mithe, Indalkar i Diverkar, 2013). Važna je jer netočno segmentirani znakovi neće biti ispravno prepoznati. Za segmentaciju riječi koristi se prazni prostor između riječi jer je on u latiničnom pismu veći od prostora između znakova u riječi (Shinde i Chougule, 2012). Prosječna točnost suvremenih OCR programa jest 99,5% (Stančić, 2009).

Može se upotrijebiti programe za prepoznavanje rukom pisanog teksta (engl. *Handwriting Recognition* ili *HWR*) koji se dijele na dva tipa. *Online* HWR prepoznaje rukom pisane tekstove stvorene na računalima, pametnim telefonima i sličnim elektroničkim uređajima. S druge strane, *offline* HWR prepoznaje takve tekstove sa skeniranih ili fotografiranih dokumenata. Ovaj je tip zahtjevniji zbog varijabilnosti kvalitete rukopisa osobe koje ja pisala tekst (Kumar i Pati, 2023). Postoje i HTR programi (engl. *Handwritten Text Recognition*), koji se koriste za osobne stilove rukopisa. Iznimno su zahtjevni jer se kod rukopisa pojavljuje mnogo više varijacija nego kod elektronički ispisanih oblika slova. Može se spomenuti platforma Transkribus, koja olakšava izradu temeljnog teksta i treniranje HTR modela velikih razmjera. Iako se postigao znatan napredak kod HTR programa, još uvijek nisu naišli na širu primjenu u knjižnicama i arhivima (Ströbel et al., 2022).

OCR se koristi npr. kod CAPTCHA sustava. CAPTCHA (Completely Automated Public Turing Test to tell Computers and Humans Apart) koristi se za zaštitu protiv zlonamjernih programa kao što su botovi, programi koji imaju mogućnost pokretati automatizirane zadatke putem mreže (Xiao i Zhang, 2012 u Singh i Pal, 2014). Mogu se zasnivati na slikama ili tekstu. Radi se o obrnutom Turingovom testu (Singh Saini i Bala, 2013 u Singh i Pal, 2014) u kojem je program sudac. Ako korisnik padne na testu, onda se smatra uređajem. U suprotnom, korisnika se smatra autentičnim korisnikom ili ljudskim bićem. Danas je CAPTCHA koja se zasniva na upotrebi OCR-a ranjivija jer su na internetu dostupni alati

kojima se mogu proći takvi testovi. Način na koji se to može smanjiti je izobličiti tekst, no onda je korisnicima teže riješiti test. Zato je uvedena CAPTCHA koja se zasniva na slikama. CAPTCHA koja se ne bazira na OCR-u koristi prirodnu vještinu ljudi da identificiraju što je na slikama (Raj et al., 2010 u Singh i Pal, 2014). CAPTCHA koja se bazira na tekstu uključuje elemente koji otežavaju prepoznavanje korisnicima kao što su različiti oblici i veličine slova i zamagljena slova (Singh i Pal, 2014), s pozadinama kojima je cilj distrakcija, arbitrarnim podjelama, isticanjima i bukom, a sve kako se tekst ne bi mogao prepoznati OCR-om (Hamad i Kaya, 2016), no mogu se lako zaobići tehnikama OCR-a (Singh i Pal, 2014). ATMA ili Android Travel Mate Application omogućava prepoznavanje i prijevod teksta na cestovnim znakovima, stranicama knjiga, menijima, transparentima itd. za lakše snalaženje u stranoj zemlji čiji jezik ne poznajemo. OCR se koristi i kod automatskog prepoznavanja automobilskih tablica. Koristi ga policija za elektroničko naplaćivanje cestarina (Hamad i Kaya, 2016).

Sustave za optičko prepoznavanje znakova može se podijeliti na sustav otvorenog kôda (engl. *open source*) i komercijalne sustave (engl. *proprietary*) (Trbušić, 2022). Sustav otvorenog kôda je na primjer Tesseract 5, a komercijalni je sustav ABBYY FineReader PDF 15. Kako bi se ispravno donijela odluka koju vrstu sustav odabrati za projekt digitalizacije, treba se sagledati više aspekata. Jedan od njih su troškovi. Početni troškovi, primjerice nabava tehnologije, kod komercijalnih sustava su viši. Mogu postojati opcije testne verzije, kada je sustav dostupan na određeno vrijeme ili ima određen maksimalan broj stranica koje se mogu obraditi. S druge strane, sustavi otvorenog kôda su besplatni i ne zahtijevaju početno ulaganje. Njihova mana je to što su potrebna tehnička znanja za korištenje. Usto, dokumentacija sustava može biti nepotpuna. Stoga, kod sustava otvorenog kôda početni trošak može biti nula, no njegovo dugoročno korištenje može biti skupo. Može se dogoditi da je kod komercijalnih sustav potrebno daljnje financiranje i proizvođač mora održavati sustav, a kod sustava otvorenog kôda institucija ulaže u vještine zaposlenika. Sustavi otvorenog kôda su bolji u ovom pogledu zbog znanja koje zaposlenici steknu instalacijom i rukovanjem sustavom (Trbušić, 2022).

Jedan sustav otvorenog kôda jest Tesseract. To je sustav za optičko prepoznavanje znakova otvorenoga kôda koji je razvijao HP između 1984. i 1994. godine. HP je izdao program kao sustav otvorenog kôda 2005. godine (Smith, 2007). Recognita je također sustav otvorenog kôda. Koristi napredne algoritme i tehnologiju strojnog učenja za analizu znakova. Podržava

više jezika kao što su engleski, španjolski, francuski, njemački, japanski, kineski itd. (Recognita Standard OCR 3.2, updatestar).

Za potrebe OCR-a u ovom postupku digitalizacije koristio se program ABBYY FineReader 12 Professional, koji je primjer komercijalnog sustava za optičko prepoznavanje znakova. Kako bi program izradio tekstualni zapis iz slike, prvo se analizira struktura slike i stranica se dijeli na element kao što su dijelovi teksta, tablice, slike itd. Linije se dijele na riječi i zatim na znakove, koje program uspoređuje s uzorcima (engl.*pattern images*) (Learning Center, ABBYY FineReader PDF).

3.1. Izazovi OCR-a

Kako bi se postigla dobra kvaliteta i visoka točnost prepoznavanja znakova, potrebne su slike visoke kvalitete ili rezolucije, u kojima se tekst značajno razlikuje od pozadine. Izrada slika ili skenova već je vrlo značajan faktor u točnosti i uspjehu optičkog prepoznavanja znakova jer se radi o razlikama u kvaliteti slike. Obično OCR izrađen od skenova postiže visoku točnost i dobre performanse, dok slike koje se izrade korištenjem kamera nisu toliko dobre za OCR zbog faktora povezanih uz elemente koji okružuju tekst ili uz kameru. Na uspjeh OCR-a mogu utjecati neki od sljedećih faktora.

1. Kompleksnost scene

Ovaj faktor podrazumijeva prepoznavanje teksta koji se nalazi na slikama koje uključuju druge predmete i pozadinu, a ne na slikama s isključivo tekstom. Drugi elementi koji se mogu nalaziti na slici su slike, građevine ili simboli. Otežavaju prepoznavanje znakova jer je teško razdijeliti tekst od drugih netekstualnih elemenata (Hamad i Kaya, 2016). Tekst ispisan na pozadini koja ima uniformnu boju vodi k višoj točnosti od slika koje sadržavaju netekstualne elemente jer one mogu biti obilježene vrlo niskim kontrastom između znakova i pozadine (Nwokoma et al., 2021).

2. Neravnomjerno osvjetljenje

Slike nastale u prirodnim okruženjima često imaju neravnomjerno osvjetljenje i sjene, što uzrokuje manju točnost segmentacije i prepoznavanja (Hamad i Kaya, 2016). Na fotoaparat više utječu okolni uvjeti kao što je osvjetljenje (Trémeau et al., 2011 i Mhaske i Sadavarte, 2016 u Nwokoma et al., 2021). Slike stvorene digitalnim fotoaparatom razlikuju se od

skeniranih. Često imaju defekte kao što je distorzija na rubovima i prigušeno svjetlo, što otežava prepoznavanje u OCR programima (Mithe, Indalkar, i Divekar, 2013).

3. Rotacija

Rotacija ima veliki utjecaj na OCR. Kako bi se odstranila, rotacija se može ispraviti u programu za uređivanje slika, u programu OCR-a ili tehnikama kao što su Projection Profile, RAST algoritmom, Hough transformacijom, Fourier transformacijom itd. Tradicionalna metoda Projection Profile bazira se na horizontalnoj upotrebi profila. Niz horizontalnih profila računa se na rasponu kutova. Profil s najvećom varijacijom odnosi se na najbolje poravnanje s linijama teksta (Jain i Borah, 2014). RAST algoritam (Recognition by Adaptive Subdivison of Transformation Space) brza je i fleksibilna metoda koja koristi geometrijske modele. Koristio se za procjenu ravnih linija teksta na skeniranim dokumentima (Breuel, 2002 u Ulges, Lampert i Breuel, 2005). Hough transformacija (engl. *Hough transform* ili *HT*) tehnika je koja se koristi za detektiranje linija, krugova i drugih elemenata slika (Touj, Amara i Amiri, 2005). Između ostalog, koristi se u OCR-u za procjenjivanje iskrivljenog kuta (Touj, 1999 u Touj, Amara i Amiri, 2005). Fourier transformacija je naširoko korištena tehnika procesiranja slika, koja se često koristi kod poboljšanja informacija o opisu slika i kod vizualnih efekata (Manjunath Aradhya, Hemantha Kumar i Noushath, 2008).

4. Zamagljenost i degradacija

Važan faktor je fokusiranje jer se za bolji uspjeh OCR-a traži što veća oštrina znakova. Uglavnom se radi o situacijama u kojima objekt nije u fokusu ili kad se objekt pomiče.

5. Nakošenost ili distorzija perspektive

U skeneru je slika dokumenta uvijek paralelna ravnini senzora, što se ne događa uvijek s fotoaparatima. Linije teksta udaljenije od senzora čine se manjima, no pametni telefoni mogu imati funkciju prepoznavanja nakošenosti.

6. Oblik slova

Italic stil i neki fontovi mogu uzrokovati preklapanje znakova pa je npr. teško provesti segmentaciju.

7. Savijanje

Savijanje se češće događa kod ručnih uređaja, no može se dogoditi i kod skenera gdje se sken teksta savije jer se nalazi pokraj uveza jako velike knjige (Hamad i Kaya, 2016).

Povijesni dokumenti razlikuju se od suvremenih dokumenata na nekoliko načina, koji otežavaju njihovu digitalizaciju. Na primjer, mogu biti oštećeni, imati različite fontove, biti izrađeni od različitih materijala, a za njih može biti dostupno tek malo podataka o jeziku ili oni mogu biti nedostupni (Nunamaker et al., 2016). Predlaže se upotreba alata i standarda za treniranje modela za prepoznavanje (Boenig et al., 2019).

3.2. Temeljni tekst (g*round truth*)

Predstavlja se koncept izrade temeljnog teksta i njegova važnost za istraživanje.

Temeljni tekst (engl. *ground truth*, GT) koristi se za testiranje točnosti automatiziranog procesa analize slike (digitisation.eu navedeno u Kettunen, Kervinen i Koistinen, 2018: 2). U kontekstu OCR-a, to je tekstualni sadržaj slike, tj. cjeloviti i točan zapis svakoga znaka i riječi sa slike. Koristi se kako bi se procijenila točnost optički prepoznatog teksta te kako bi se procijenila važnost bilo kakvih zastranjivanja od izvornika (digitisation.eu u Reynaert, 2014: 160).

Temeljni tekst može se opisati kao idealni ishod savršenog procesa OCR-a. Iz tog razloga, vrlo je važan kao alat za evaluaciju. Njegova izrada uglavnom je ručni ili u najboljem slučaju napola automatizirani zadatak (Clausner et al., 2015).

Temeljni tekst može se koristiti u treniranju modela prepoznavanja, fokusirajući se na posebne setove dokumenata. Poboljšanja u OCR softveru ABBYY FineReader omogućuju da se tekstovi izrađeni unazad do sredine 18. stoljeća mogu prepoznati, no mogu biti nestrukturirani i često sa slabom točnošću teksta. Ipak, mogu se pokazati korisnima za mnoge znanstvene upotrebe. Digitalizacija punoga teksta (*full text digitization*) u djelima nastalima u 17. stoljeću i prije kompliciranija je zbog toga što je kvaliteta prepoznavanja još uvijek nezadovoljavajuća. Suvremeni napreci u prepoznavanju teksta koji se temelje na neuronskim mrežama (engl.*neural networks*) omogućuju da se i ovakva građa može digitalizirati s visokom razinom točnosti (Boenig et al., 2019). Ovakve metode zahtijevaju izradu temeljnog teksta za treniranje modela za prepoznavanje (Boenig et al., 2019).

Temeljni tekst može se izraditi bez upotrebe specifičnih alata (Boenig et al., 2019). Mogu se koristiti i posebni programi zvani uređivači temeljnog teksta (*Ground Truth Editor*) kao što su Transkribus i Aletheia. Ova dva programa imaju jednake osnovne funkcije, a to su transkripcija i ručna segmentacija, no imaju različite prioritete. Aletheia je bolja za segmentaciju i evaluaciju jer se često koristila u ocjenjivanju analize dokumenata u

natjecanjima. Transkribus, s druge strane, ima dobru podršku za uvoz iz digitalnih knjižnica jer se fokusira na knjižničnu i arhivsku zajednicu (Boenig et al., 2019).

4. Vrste građe

U ovom dijelu predstavit će se podjela građe važna za proces optičkog prepoznavanja teksta i definicije pojmova koji se koriste u radu, kao što su serijska publikacija i novine.

Za proces optičkog prepoznavanja znakova korisna je podjela građe prema tehnologiji nastanka, tj. s obzirom na način na koji je tekst zapisan na medij. U knjižnicama se može pronaći i druge vrste građe, no ovaj se rad bavi papirnatom građom (Trbušić, 2022). S obzirom na to, građu možemo podijeliti na (Trbušić, 2022):

1. rukopisnu,
2. strojopisnu,
3. tiskanu i
4. građu nastalu kombiniranim pristupom.

Rukopisna građa obuhvaća tekstove napisane rukom. Svaka osoba ima svoj stil pisanja, koji je nepredvidiv i može se mijenjati tijekom njezina života. Za prepoznavanje ovakve građe, koristi se primjerice prethodno spomenuta HWR i HTR tehnologija. U razvoju ovih tehnologija pomaže razvoj umjetne inteligencije i strojnog učenja. Primjer sustava koji se koristi za optičko prepoznavanje znakova rukopisne građe je platforma Transkribus. Platforma uključuje treniranje već izrađenih modela prepoznavanja i izradu novih, čime je moguća prilagodba vrsti građe i stilu pisanja (Trbušić, 2022).

Strojopisna građa izrađena je na pisaćem stroju. Moguća je upotreba optičkog prepoznavanja znakova. Strojopisna se građa razlikuje od tiskane jer su razlike između otiska znakova veće. Kod tiskane građe one su vidljive na razini od nekoliko tisuća stranica, dok kod strojopisne građe razlike u točnosti optičkog prepoznavanja ovise o jačini kojem je osoba pritisnula tipku na stroju jer jačina pritiska utječe na batić koji udara na površinu papira pa se tako može dogoditi da postoji nejednakost otiska na istoj stranici, što je problem u optičkom prepoznavanju (Trbušić, 2022).

Tiskana građa započinje izumom Gutenbergova tiskarskog stroja 1450. godine. Ovim izumom započinje brža i jeftinija reprodukcija tekstova. Tiskani tekstovi imali su veliku ulogu u obrazovanju stanovništva. Tiskanoj građi povećan je značaj upotrebom računala i softvera za stvaranje teksta te raznih vrsta pisača koji postoje, primjerice laserski i tintni (Trbušić, 2022).

Građa nastala kombiniranim pristupom odnosi se na građu nastalu korištenjem više od jedne prethodno opisane metode. Primjerice, strojopisni i tiskani tekstovi mogu sadržavati rukopis na marginama. Moguća je kombinacija na jednoj stranici, u dokumentu ili u skupu dokumenata. Ovakva je građa izdvojena jer kombinacija metoda znači teže optičko prepoznavanje teksta. Rukopisni dijelovi mogu se digitalizirati drugim metodama, npr. ručnim prepisivanjem. Tekstovi koji uključuju strojopisne i tiskane dijelove ne predstavljaju toliki problem jer je kod suvremenih OCR programa moguće istodobno prepoznavanje različitih oblika slova (Trbušić, 2022).

Publikacija je "tiskano ili drugom tehnikom umnoženo djelo, obično proizvedeno u više primjeraka i namijenjeno raspačavanju u javnosti. S obzirom na opseg sadržaja publikacija može biti omeđena (knjiga u jednom ili u više svezaka) ili neomeđena" (Publikacija, Hrvatska enciklopedija), u kojem slučaju se radi o serijskoj publikaciji. Kad se radi o tiskanim djelima prenesenima u elektronički oblik ili onima koja su izvorno objavljena u elektroničkom obliku nazivaju se elektroničkim publikacijama (Publikacija, Hrvatska enciklopedija). Serijska publikacija, pak, nema unaprijed planiran završetak objavljivanja. Izlazi u uzastopnim zasebnim sveščićima ili dijelovima, koji se uglavnom označuju brojčano ili kronološki. U periodičke publikacije ubrajaju se časopisi, revije, magazini, novine, godišnjaci i nizovi knjiga.

Serijske publikacije prvi se put javljaju u obliku časopisa u 17. stoljeću. U 18. stoljeću u Francuskoj pojavljuju se prve novine, a u 18. stoljeću magazini. Na početku knjižnice nisu ulagale napor u njihovo očuvanje jer ih se smatralo prolaznom građom. Njihovo očuvanje započelo je u 19. stoljeću s odjelima za periodiku. Važan dokument bio je međunarodni bibliografski standard za serijske publikacije (ISBD/S/), koji je kasnije prerađen i poznat kao ISBD/CR i danas uključuje postupak opisa elektroničkih serijskih publikacija. ISDS ili Međunarodni sustav podataka o serijskim publikacijama (International Serials Data System) utemeljio je UNESCO, koji je danas baza podataka koja uključuje sve serijske publikacije. U sklopu ovog sustava, pojavio se ISSN, sustav za identifikaciju serijskih publikacija (Serijska publikacija, Hrvatska enciklopedija).

Novine su svako periodično izdanje koje izlazi dnevno, tjedno, petnaestodnevno ili mjesečno. Njihov je cilj informiranje o događajima, obrazovanje i dr. Njihova se važnost u društvu povećala nakon prve industrijske revolucije. Nakon Francuske revolucije, dobile su na značaju kao alat političke borbe (Novine, Hrvatska enciklopedija).

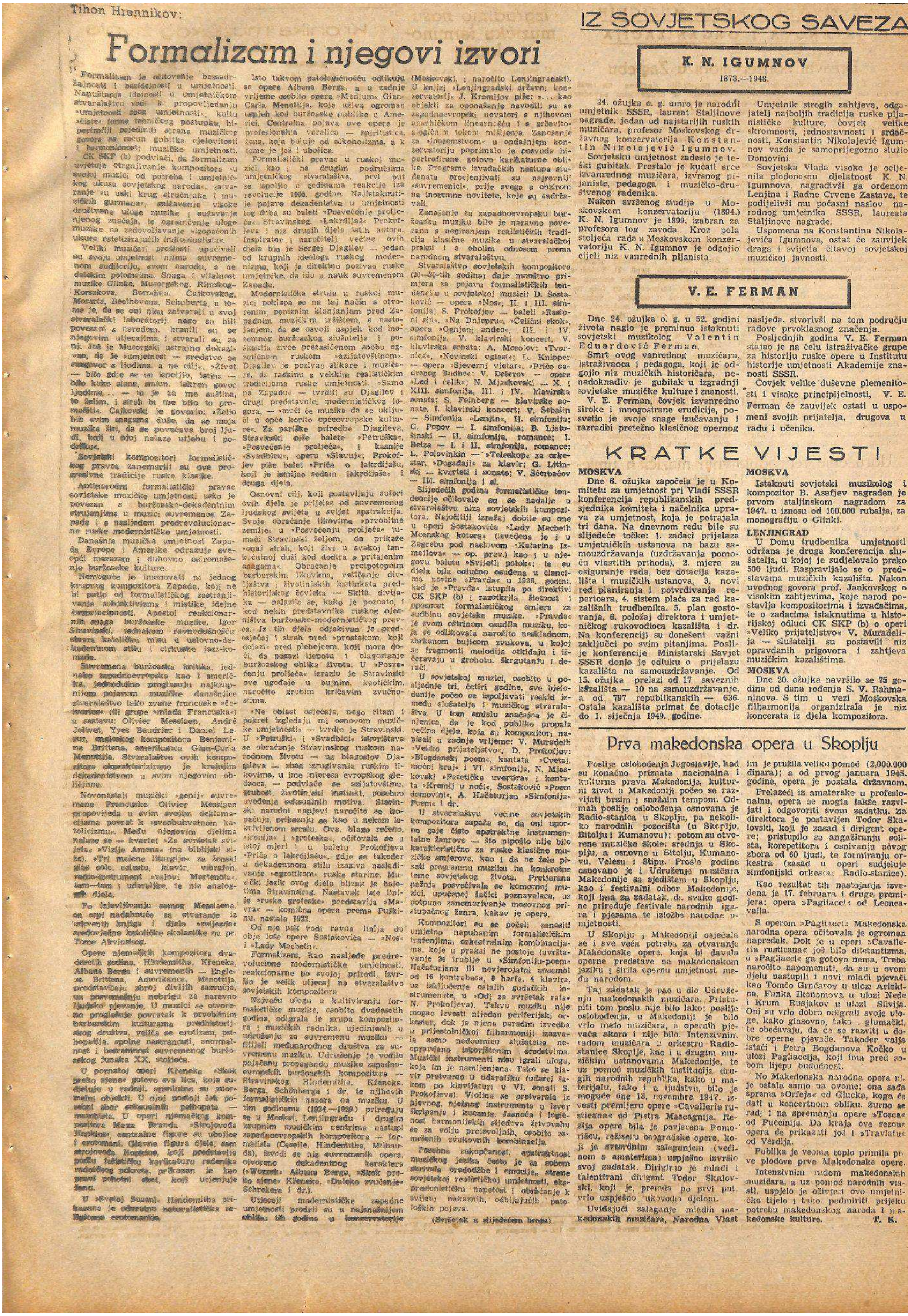
5. Istraživanje

Ovaj dio rada donosi informacije o procesu istraživanja koji uključuje obradu i skeniranje građe, izradu datoteka s optički prepoznatim tekstom, izradu temeljnog teksta i testiranje uspješnosti prepoznavanja. Predstavlja se analiza procesa s uključenim optimizacijskim parametrima, a to su vrijeme u radnim satima.

Cilj istraživanja jest istražiti utjecaj rezolucije na brzinu digitalizacije i njezinu isplativost u radnim satima. Drugi utjecaj povezan je uz različite verzije korištenog temeljnog teksta. U prvoj verziji ispravljena je struktura, dakle vodilo se računa da stupci teksta slijede svoj originalni raspored. U drugoj verziji, pak, ispravljene su samo pogreške u prepoznavanju znakova. U istraživanju se mjerilo vrijeme skeniranja i vrijeme prepoznavanja.

Uzorak koji se koristio u istraživanju sastoji se od deset stranica publikacije Muzičke novine. Uzorak se sastoji od dva broja novina. Prvi broj uključuje četiri stranice, a drugi šest. Tekst na stranicama tiskan je strojopisno i raspoređen u neravnomjerne stupce koji otežavaju proces optičkog prepoznavanja teksta.

Muzičke novine Hrvatskog državnog konzervatorija publikacija je koja je izlazila mjesečno od početka 1946. do kraja 1948. godine (Katalog knjižnica grada Zagreba). Novine donose informacije o radu muzičkih ustanova i organizacija (npr. Muzičkih škola), muzičkim priredbama i festivalima, kongresima kompozitora, događanjima u inozemstvu vezanima uz glazbu, muzičkim programima, nagradama. Obrađuju, nadalje, biografije glazbenika i neke probleme koji se javljaju u području glazbe. Primjer stranice Muzičkih novina može se vidjeti na slici 1.



Tihon Hrennikov:

Formalizam i njegovi izvori

IZ SOVJETSKOG SAVEZA

K. N. IGUMNOV

1873.—1948.

V. E. FERMAN

KRATKE VIJESTI

Prva makedonska opera u Skoplju

Slika 1. Primjer stranice Muzičkih novina.

5.1. Odabir i obrada građe

Muzičke novine odabrane su jer knjižnica Muzičke akademije u Zagrebu želi digitalizirati stare novine koje posjeduje. Kako knjižnica posjeduje tri primjerka, za potrebe digitalizacije razvezan je jedan primjerak radi bolje kvalitete skenova. Svaki primjerak je uvez svih brojeva iz tri godišta novina. Postupak razvezivanja proveden je skalpelom. Odstranile su se korice i razvezali pojedinačni brojevi novina.

Kao što je spomenuto, softver za skeniranje može omogućiti automatsko uklanjanje točke na skenu nastale zbog probušenih stranica i popraviti orijentaciju te ukošenost (Stančić, 2009). Kod građe u ovom istraživanju, softver skenera nije automatski prepoznao i uklonio točke nastale zbog probušenih stranica, već se to moralo uraditi ručno obrezivanjem slike skena. Softver uređaja ScanSnap ima mogućnost automatskog popravljanja orijentacije i ukošenosti stranica, no nije bio precizan pa je bilo potrebno ručno ispravljanje.

U procesu digitalizacije preporuča se planiranje konvencije za imenovanje datoteka, kojom se može spojiti datoteku s originalom (Stančić, 2009). U ovom slučaju, izrađene su tri datoteke s nazivima IMAGE, OCR i GT. Datoteka IMAGE sadržavala je datoteke svih rezolucija pojedinačno, u kojima su se nalazile datoteke s imenima uređaja za skeniranje. Skenovi su nazvani na sljedeći način: S_3_4_200_s1. "S" označava da se radi o ScanSnap uređaju, broj je 3 je godište, 4 je broj, broj 200 je rezolucija, a "s1" označava broj stranice. Ovaj način imenovanja datoteka omogućio je lakše snalaženje u istraživanju i može se lako povezati s izvornikom. Datoteka OCR sadrži OCR datoteke, a datoteka GT temeljne tekstove u dvije datoteke nazvane GT1 i GT2. Dokumenti su nazvani prema sljedećem primjeru: GT1_3_4_s1. Naziv ne sadrži informaciju o rezoluciji jer nije relevantna kod temeljnog teksta. Preporuča se imati pričuvne kopije u slučaju požara, elementarnih nepogoda itd. (Stančić, 2009). Sve datoteke istraživanja čuvale su se na Google Driveu.

Pohrana digitalizirane građe koju opisuje ovaj rad provedena je na Digitalnom repozitoriju Muzičke akademije, koji je dio sustava Dabar, tj. Digitalni akademski arhivi i repozitoriji. Knjižnica u repozitoriju okuplja završne i diplomske radove, disertacije, *pre-print* radove, znanstvene i stručne radove, podatke istraživanja, knjige, nastavne materijale, slike, video i audiozapise, prezentacije i digitaliziranu građu (Digitalni repozitorij Muzičke akademije). Sustav Dabar podržava nacionalnu e-infrastrukturu Hrvatske koja institucijama i sustavu znanosti i visokog obrazovanja omogućava podršku digitalnoj imovini. Radi se o tematskim i institucijskim digitalnim repozitorijima i arhivima. Ima funkcionalnost prikupljanja, trajne

pohrane i diseminacije. Dabar korisnicima pruža kontrolu prava pristupa i korištenja, podržava otvoreni pristup, koji vodi k povećanju vidljivosti građe i institucije te dugoročnu pohranu u standardnim, preporučenim formatima. Građi, tj. digitalnim objektima dodaju se metapodaci radi lakšeg pronalaženja (Digitalni akademski arhivi i repozitoriji). Što se tiče zaštite i osiguranja identiteta, građa koja je objavljena u repozitoriju nalazi se u otvorenom pristupu, po odluci Knjižnice Muzičke akademije. Određena građa u repozitoriju dostupna je npr. samo korisnicima matične ustanove, no to nije slučaj kod Muzičkih novina. Građu može preuzeti bilo tko, bez upotrebe lozinke.

5.2. Skeniranje

U ovom potpoglavlju opisat će se postupak skeniranja građe, za što su korištena dva različita uređaja. To su Fujitsu ScanSnap SV600 skener i uredski kopirni uređaj. U ovom dijelu istraživanja mjereno je vrijeme potrebno za skeniranje svake stranice na svakoj od dostupnih rezolucija na oba uređaja. Kako dva uređaja ne podržavaju iste rezolucije, spomenut će se i ovaj problem. Na kraju, navest će se programe koji su se koristili u istraživanju.

Proces digitalizacije koji rad opisuje proveden je za potrebe istraživanja, ali i za potrebe knjižnice Muzičke akademije. Nakon odabira građe za digitalizaciju, zaposlenici knjižnice donijeli su odluku da se građa skenira u boji, iako su tekstovi i slike crno-bijeli, a samo je pozadina žuta zbog starenja papira. Korisnicima se željelo predstaviti građu onako kako je izgledala u trenutku digitalizacije, ali s potpuno pretraživim tekstom. Kao što je rečeno, vrijedna i povijesna građa digitalizira se fotografskim aparatima, uz posebno osvjetljenje i komore (Digitalizacija, Hrvatska enciklopedija, 2021). To se u ovom slučaju nije primijenilo jer se nije smatralo da je građa krhka za digitalizaciju skenerima. Bila je dostupna u više primjeraka, a knjižnica nije raspolagala drugim uređajima za digitalizaciju. Kako je proces digitalizacije u knjižnici završen prije dovršetka pisanja ovoga rada, za objavljivanje u repozitoriju odabrana je rezolucija od 300 dpi jer knjižnica inače odabire ovu rezoluciju i jer je viđena kao zadovoljavajuća za predviđeno korištenje. Provedeno istraživanje u obzir je uzelo omjer vremena i točnosti OCR-a na svakoj od rezolucija. Istražuje se koja rezolucija i koji od dva skenera omogućuju brži i jeftiniji proces digitalizacije. Slika 2 prikazuje Fujitsu ScanSnap skener upotrijebljen u ovome istraživanju.

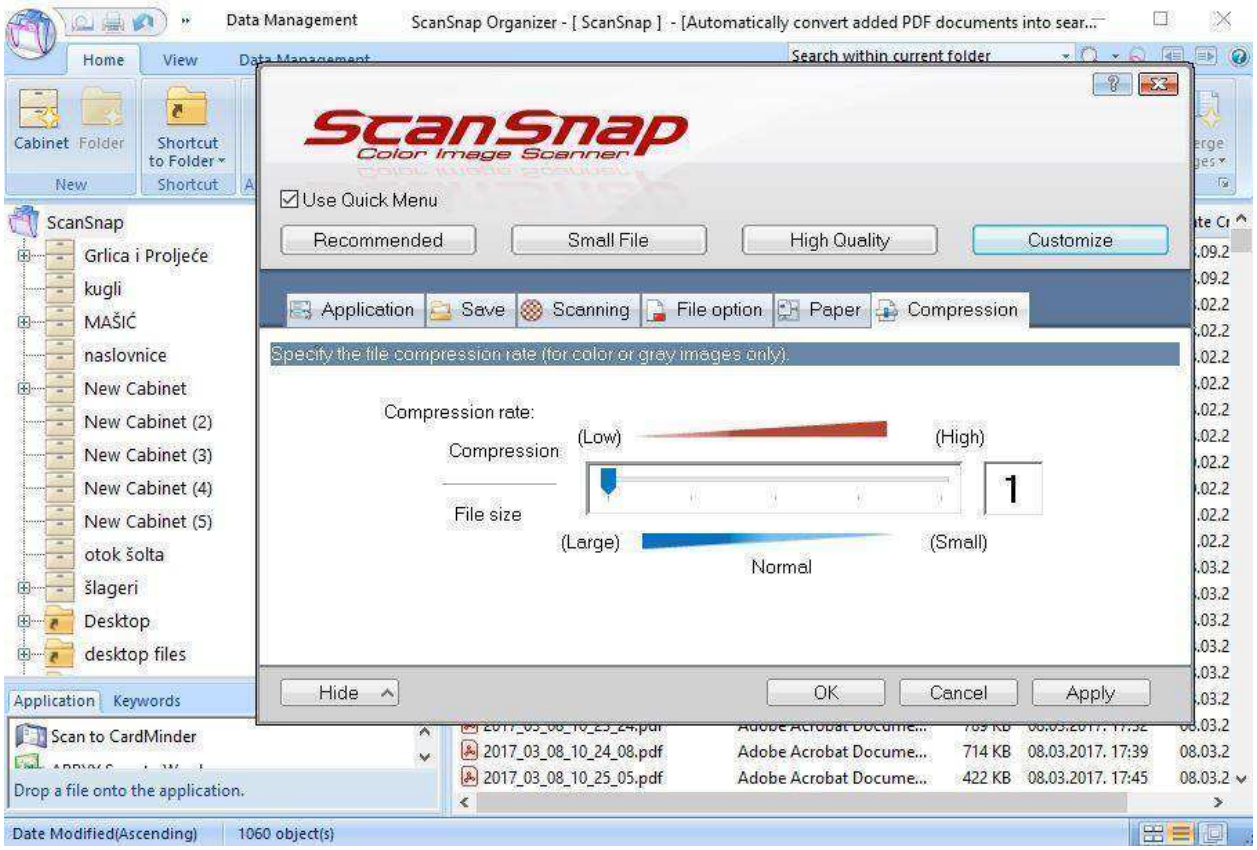
Slika 2. Skener Fujitsu ScanSnap SV600 s podmetačem i pokretnom glavom za skeniranje

Izvor: FUJITSU Image Scanner ScanSnap SV600. Fujitsu.

https://www.fujitsu.com/ph/products/computing/peripheral/scanners/soho/sv600/.

Ograničenje istraživanja je rezolucija jer dva uređaja koja su se koristila za skeniranje nisu podržavala jednaku rezoluciju. Na uređaju ScanSnap dostupne rezolucije su 150 dpi, 200 dpi, 300 dpi i 600 dpi. Na kopirnom uređaju dostupne rezolucije su 100 dpi, 200 dpi, 300 dpi, 400 dpi i 600 dpi. Time dobivamo da su usporedive rezolucije 200 dpi, 300 dpi i 600 dpi.

Prije skeniranja odabrale su se postavke rada. Kao izlazni format procesa digitalizacije odabran je JPG format na oba raspoloživa uređaja. Kopirni uređaj podržava spremanje u TIFF format, ali ta opcija ne postoji kod ScanSnap skenera. Pritom se vodilo računa o tome da se skenovi spremaju u najvećoj mogućoj kvaliteti, što se u programu uređaja ScanSnap odabire kao razina kompresije koja je postavljena na nulu, Slika 3. Na kopirnom uređaju odabrala se najveća moguća kvaliteta skena.



Slika 3. Odabir razine kompresije u programu ScanSnap uređaja.

Razvezane stranice postavilo se jednu po jednu na površinu skenera i skeniralo pokretnom glavom na svim dostupnim rezolucijama. Pokretna glava dio je skenera koji se nalazi na vrhu uređaja, kao što se može vidjeti na Slici 2. Taj se dio pokrene prema naprijed provede skeniranje te se nakon toga vraća u prvotni položaj. Na kopirnom uređaju odabrana je opcija automatskog određivanja veličine stranice. ScanSnap uređaj automatski prepoznaje rubove stranica građe, nakon čega su u programu skenovi ručno uređeni i popravljene pogreške. Problem se može dogoditi ako program netočno obreže stranicu, uključujući crnu površinu podmetača ili obrezujući rubove ili sami tekst. Provjerilo se koliko je dobro program automatski odredio rubove, nakon čega su ručno promijenjene granice obrezivanja stranice. Mjereći vrijeme skeniranja na svim dostupnim rezolucijama, dobiveni su podaci u tablicama koje slijede. Mjereno je vrijeme između početka skeniranja uređaja i trenutka kad je uređaj bio spreman za sljedeći sken. U Tablici 1 može se vidjeti vrijeme potrebno za skeniranje na uređaju ScanSnap. Nakon toga, postupak se ponovio na kopirnom uređaju, a rezultati su vidljivi u Tablici 2. Vrijeme u tablicama izraženo je u sekundama.

Tablica 1. Vrijeme potrebno za skeniranje na uređaju ScanSnap.

Vrijeme potrebno za skeniranje korištenjem uređaja ScanSnap	150 dpi	200 dpi	300 dpi	600 dpi
Muzičke novine 3/4, str. 1	8:38	8:94	9:83	11:20
Muzičke novine 3/4, str. 2	9:02	9:38	10:35	12:20
Muzičke novine 3/4, str. 3	8:97	9:49	11:74	13:20
Muzičke novine 3/4, str. 4	9:96	10:28	11:82	12:92
Muzičke novine 3-5/6, str. 1	9:65	10:27	11:16	12:44
Muzičke novine 3-5/6, str. 2	10:11	10:85	11:48	13:02
Muzičke novine 3-5/6, str. 3	9:88	10:31	11:80	12:87
Muzičke novine 3-5/6, str. 4	9:02	10:35	11:48	12:73
Muzičke novine 3-5/6, str. 5	9:80	10:33	11:61	12:77
Muzičke novine 3-5/6, str. 6	10:08	10:30	11:10	12:83

Tablica 2. Vrijeme potrebno za skeniranje na kopirnom uređaju.

Vrijeme potrebno za skeniranje korištenjem kopirnog uređaja	100 dpi	200 dpi	300 dpi	400 dpi	600 dpi
Muzičke novine 3/4, str. 1	3:61	4:99	8:09	12:33	23:53
Muzičke novine 3/4, str. 2	3:77	5:12	8:09	12:23	23:98
Muzičke novine 3/4, str. 3	3:90	5:08	8:18	12:15	23:64
Muzičke novine 3/4, str. 4	3:82	5:02	8:15	12:33	23:72
Muzičke novine 3-5/6, str. 1	3:94	5:05	8:05	12:32	23:65
Muzičke novine 3-5/6, str. 2	3:79	4:99	8:36	12:36	23:74
Muzičke novine 3-5/6, str. 3	3:80	5:01	8:27	12:38	23:79
Muzičke novine 3-5/6, str. 4	3:81	5:14	8:16	12:17	23:88
Muzičke novine 3-5/6, str. 5	3:60	5:07	8:29	12:36	23:67
Muzičke novine 3-5/6, str. 6	3:73	4:97	8:11	12:33	23:69

Može se vidjeti da se vrijeme između najniže i najviše rezolucije znatno razlikuje na kopirnom uređaju. Razlika je i do dvadeset sekundi. Na primjer, skeniranje prve stranice na najnižoj rezoluciji trajalo je 3:61 sekundi. Na najvišoj rezoluciji trajanje je iznosilo 23:53 sekundi. Na ScanSnap skeneru razlika između najniže i najviše rezolucije je otprilike tri sekunde. Na primjer, skeniranje iste stranice na najnižoj rezoluciji trajalo je 8:38 sekundi, a

na najvišoj 11:20. Skeniranje prve stranice na najnižoj usporedivoj rezoluciji od 200 dpi na ScanSnapu trajalo je 8:94 sekundi, dok je na kopirnom uređaju tek 4:99. Na rezoluciji od 300 dpi, kopirni je uređaj opet bio brži s 8:09 sekundi naprema 9:83 sekundi na ScanSnap uređaju. Na maksimalnoj usporedivoj rezoluciji od 600 dpi, pak, ScanSnap je bio otprilike 22 sekunde brži. Kopirnom je uređaju trebalo 23:53 sekundi, a ScanSnapu 11:20.

Kako bismo vidjeli koliko radnih sati bi nam bilo potrebno ako se radi o projektu digitalizacije koji obuhvaća vrlo veliki broj stranica, kao primjer navest će se projekt Europeana Newspapers, koji nastoji dodati osamnaest milijuna stranica povijesnih novina Europeani i Europskoj knjižnici (The European Library), a deset milijuna stranica novina konvertirat će se u puni tekst (Europeana Newspapers). Upotreba ovakvih procjena vrlo je važna u procesu planiranja projekta digitalizacije. Rad zaposlenika koji obavlja skeniranje najviše utječe na trošak projekta (Trbušić, 2022). Jedan sat rada jest otprilike 12 američkih dolara (Blostein i Nagy, 2012).

Kako bismo dobili podatke u tablici 3, izračunale su se prosječne vrijednosti za svaku od usporedivih rezolucija i pomnožile s deset milijuna, koliko projekt uključuje.

Tablica 3. Broj radnih sati potrebnih za skeniranje deset milijuna stranica na oba uređaja i usporedivim rezolucijama.

	ScanSnap skener	Kopirni uređaj
200 dpi	27.916,67h	14.011,11h
300 dpi	31.213,89h	22.708,33h
600 dpi	35.050,00h	65.913,89h

Skeniranje na najvišoj rezoluciji na ScanSnap skeneru 20,35% je sporije nego na najnižoj usporedivoj rezoluciji. Na kopirnom uređaju ta brojka iznosi 78,74%. Možemo zamijetiti da je na najnižoj usporedivoj rezoluciji od 200 dpi na kopirnom uređaju potrebno dvostruko manje vremena za razliku od ScanSnap uređaja. Na rezoluciji od 300 dpi kopirni je uređaj 27,22% brži. Na najvećoj usporedivoj rezoluciji ScanSnap skener gotovo je dvostruko brži od kopirnog uređaja.

5.3. OCR

Slijedi opis postupka OCR-a, navode se korišteni program i predstavlja se vrijeme potrebno za izradu.

Spomenuto je da se OCR program može se upotrijebiti tijekom skeniranja ili nakon (Stančić, 2009). U ovom istraživanju proveden je nakon što su izrađeni svi uzorci. Za obavljanje optičkog prepoznavanja znakova dobivenih skenova koristio se program ABBYY FineReader 12 Professional. Način kodiranja znakova postavljen je na UTF-8. UTF-8 najpopularnija je metoda kodiranja znakova (engl. *character encoding method*), koja se koristi na internetu i u aplikacijama (Jari, 2014 u Teahan i Alhawiti, 2015). Predstavlja latinične znakove upotrebom jednog bajta, a koristi do četiri bajta za ostale abecede, iako većina abeceda zahtijeva samo dva bajta po znaku (Teahan i Alhawiti, 2015). Tablice koje slijede prikazuju vrijeme izraženo u sekundama potrebno za provođenje postupka OCR-a na dva uređaja i dostupnim rezolucijama.

Tablica 4. Vrijeme potrebno za obavljanje optičkog prepoznavanja znakova na uređaju ScanSnap.

Vrijeme potrebno za obavljanje optičkog prepoznavanja znakova na uređaju ScanSnap	150 dpi	200 dpi	300 dpi	600 dpi
Muzičke novine 3/4, str. 1	21:79	26:12	19:68	35:96
Muzičke novine 3/4, str. 2	29:40	35:02	33:24	55:87
Muzičke novine 3/4, str. 3	79:07	96:15	53:31	101:27
Muzičke novine 3/4, str. 4	53:36	36:80	33:74	37:49
Muzičke novine 5/6, str. 1	37:06	37:32	34:93	44:25
Muzičke novine 5/6, str. 2	56:66	45:26	78:08	78:42
Muzičke novine 5/6, str. 3	14:19	18:98	17:96	46:00
Muzičke novine 5/6, str. 4	17:87	18:81	17:66	27:07
Muzičke novine 5/6, str. 5	19:97	20:48	21:85	33:04
Muzičke novine 5/6, str. 6	28:44	26:48	25:82	46:21

Tablica 5. Vrijeme potrebno za obavljanje optičkog prepoznavanja znakova na kopirnom uređaju.

Vrijeme potrebno za obavljanje optičkog prepoznavanja znakova na kopirnom uređaju	100 dpi	200 dpi	300 dpi	400 dpi	600 dpi
Muzičke novine 3/4, str. 1	29:62	23:18	20:44	28:60	28:94
Muzičke novine 3/4, str. 2	39:18	30:30	25:62	32:75	36:46
Muzičke novine 3/4, str. 3	86:86	56:27	90:41	62:47	84:93
Muzičke novine 3/4, str. 4	29:83	24:35	49:82	21:15	25:02
Muzičke novine 3-5/6, str. 1	39:93	35:65	27:13	30:69	37:38
Muzičke novine 3-5/6, str. 2	74:98	52:68	63:03	41:95	72:59
Muzičke novine 3-5/6, str. 3	30:85	20:68	21:51	21:79	29:02
Muzičke novine 3-5/6, str. 4	21:73	35:48	16:79	17:05	32:30
Muzičke novine 3-5/6, str. 5	37:42	24:56	21:12	28:96	34:71
Muzičke novine 3-5/6, str. 6	55:28	26:87	30:3	34:56	37:68

Za projekt Europeane bilo bi korisno znati koliko radnih sati bi nam bilo potrebno ako bismo koristili uređaje koje istraživanje opisuje.

Tablica 6. Usporedba vremena potrebnog za obavljanje optičkog prepoznavanja znakova u projektu Europeane.

	ScanSnap skener	Kopirni uređaj
200 dpi	100.394,44h	91.672,22h
300 dpi	93.408,33h	101.713,89h
600 dpi	140.438,89h	116.397,22h

Optičko prepoznavanje na ScanSnap skeneru bilo je najbrže pri rezoluciji od 300 dpi. Bilo je 6,96% brže od optičkog prepoznavanja znakova na najnižoj i 33,49% brže od optičkog prepoznavanja znakova na najvišoj rezoluciji. Na kopirnom uređaju, rezolucija od 200 dpi rezultirala je najbržim vremenom. Bila je 9,87% brža od sljedeće i 21,24% brža od najviše rezolucije. Pri rezoluciji od 200 dpi, kopirni je uređaj 8,69% brže proveo postupak optičkog prepoznavanja znakova. Pri 300 dpi, 8,17% brže optičko prepoznavanje znakova primijećeno je kod ScanSnap skenera. Na 600 dpi, 17,12% brže optičko prepoznavanje znakova dobilo se upotrebom kopirnog uređaja.

U procesu digitalizacije, kao i bilo kojem projektu, kao što je spomenuto, vrlo su važni troškovi realizacije. U tablici 7 vidimo zbroj radnih sati potrebnih za izradu skenova i OCR-a na oba uređaja i usporedivim rezolucijama te približni trošak procesa. Spomenuto je da je jedan sat rada otprilike 12 američkih dolara (Blostein i Nagy, 2012).

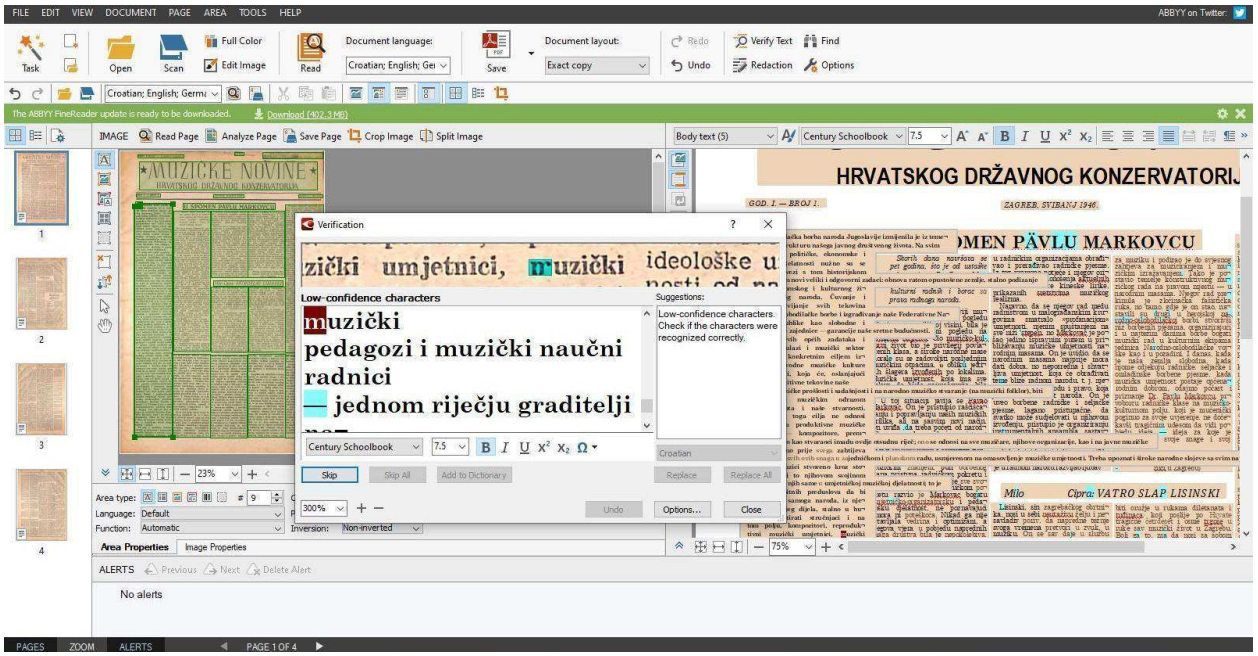
Tablica 7. Usporedba ukupnog vremena potrebnog za skeniranje i OCR na oba uređaja i usporedivim rezolucijama te procjena ukupnih troškova radnih sati zaposlenika.

	ScanSnap - ukupan zbroj radnih sati za skeniranje i OCR	ScanSnap - procjena ukupnih troškova radnih sati zaposlenika	Kopirni uređaj - ukupan zbroj radnih sati za skeniranje i OCR	Kopirni uređaj - procjena ukupnih troškova radnih sati zaposlenika
200 dpi	128.311,11h	1.539.733,32 USD	105.683,33h	1.268.199,96 USD
300 dpi	124.622,22h	1.495.466,64 USD	124.422,22h	1.493.066,64 USD
600 dpi	175.488,89h	2.105.866,68 USD	182.311,11h	2.187.733,32 USD

Iz tablice 7 možemo vidjeti da je na najnižoj usporedivoj rezoluciji od 200 dpi, skeniranje i optičko prepoznavanje znakova s kopirnog uređaja 17,64% brže i za 271.533,36 USD jeftinije. Na 300 dpi, upotreba kopirnog uređaja jeftinija je za 2.400,00 USD. Na 600 dpi bržim i jeftinijim uređajem pokazao se ScanSnap skener s razlikom u cijeni od 81.866,64 USD te 3,74% bržim vremenom. Trošak skeniranja deset milijuna stranica najniži je na kopirnom uređaju pri 200 dpi, a najviši na istom uređaju pri 600 dpi.

5.3.1. Testiranje OCR-a

Ovaj dio donosi opis postupka izrade temeljnog teksta i provjere OCR-a.



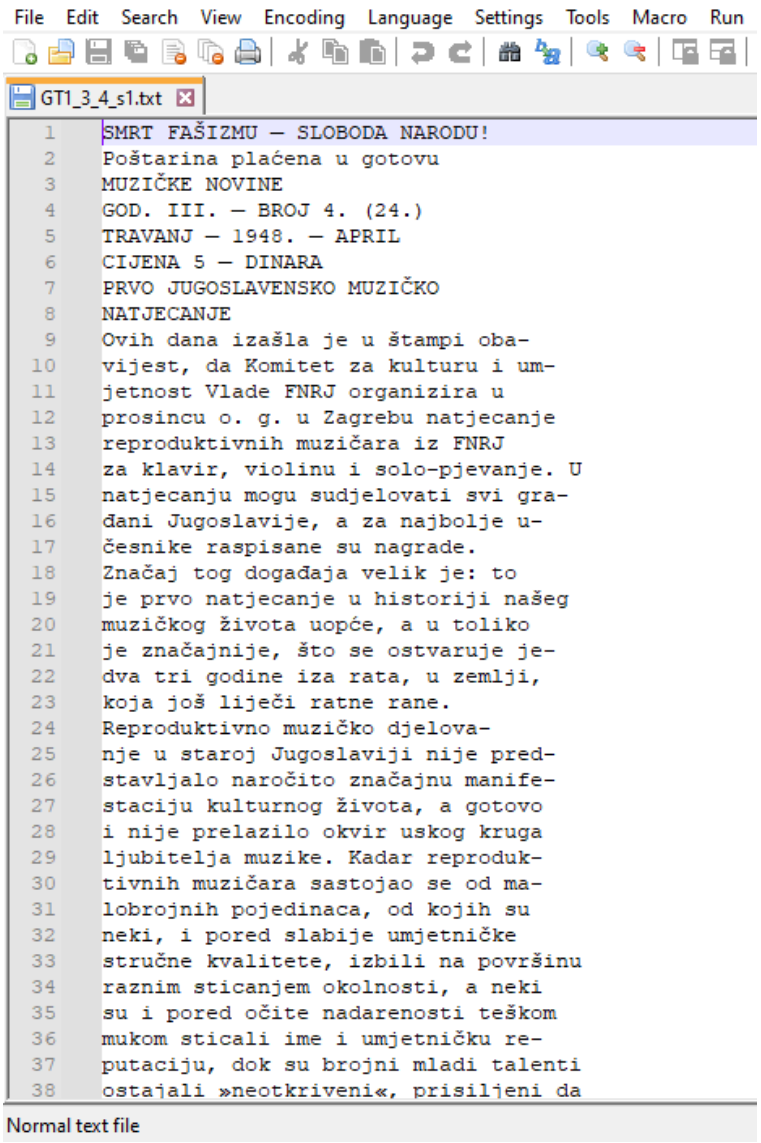
Slika 4. Ispravljanje OCR-a u prozorčiću u programu ABBYY FineReader.

Kako bi se testirala uspješnost optičkog prepoznavanja, izrađen je sto posto točan tekst koji je identičan originalnom. Za ispravljanje znakova korišten je program ABBYY FineReader, u kojem je i proveden OCR, a program Notepad++ upotrijebljen je kako bi se ispravila struktura dokumenta. Naime, tekst se u novinama nalazi u stupcima i OCR program nije sa stopostotnom točnošću utvrdio njihov redoslijed. Kako stupci koji nisu jednake duljine niti je svaka stranica raspoređena na isti način, cilj istraživanja bio je i vidjeti koliko pogrešaka se pojavljuje u OCR-u kad se ručno isprave struktura teksta i pogreške. Naime, tekst u novinama organiziran je u više stupaca. Na više mjesta je jedan stupac odvojen crtom od stupca ispod, koji je dio drugoga članka. Pod ispravljanjem strukture teksta mislilo se na ručno organiziranje stupaca na način da ispravno prate redoslijed, a ispravljalo se i crtice na krajevima redaka koje označuju da se riječ nastavlja u sljedećem retku. S druge strane, ispravljale su se samo pogreške, dok je struktura ostavljena onakvom kakvom ju je automatski prepoznao program ABBYY FineReader. Dakle, sve je uređeno na način da je istovjetno izvorniku, ali je struktura stupaca zadržana u samo jednoj verziji kako bismo

vidjeli koliko ručno ispravljanje strukture stupaca utječe na točnost optičkog prepoznavanja znakova. Na ovaj način istražena je potreba za ručnim ispravljanjem strukture optički prepoznatog teksta koja zahtijeva mnogo vremena. Problem sa strukturom pojavio se već kod ručnog ispravljanja pogrešaka u programu ABBYY FineReader, gdje bi npr. program prepoznao tekst u istom retku, ali iz sljedećeg stupca kao dio iste rečenice. Kako izgleda ispravljanje OCR-a u programu ABBYY FineReader, možemo vidjeti na slici 4. Izrada temeljnog teksta odrađivala se u programu Notepad++. OCR datoteke eksportirane su kao TXT datoteke jer program za analizu podržava samo taj format.

5.3.1.1. Prva verzija temeljnog teksta

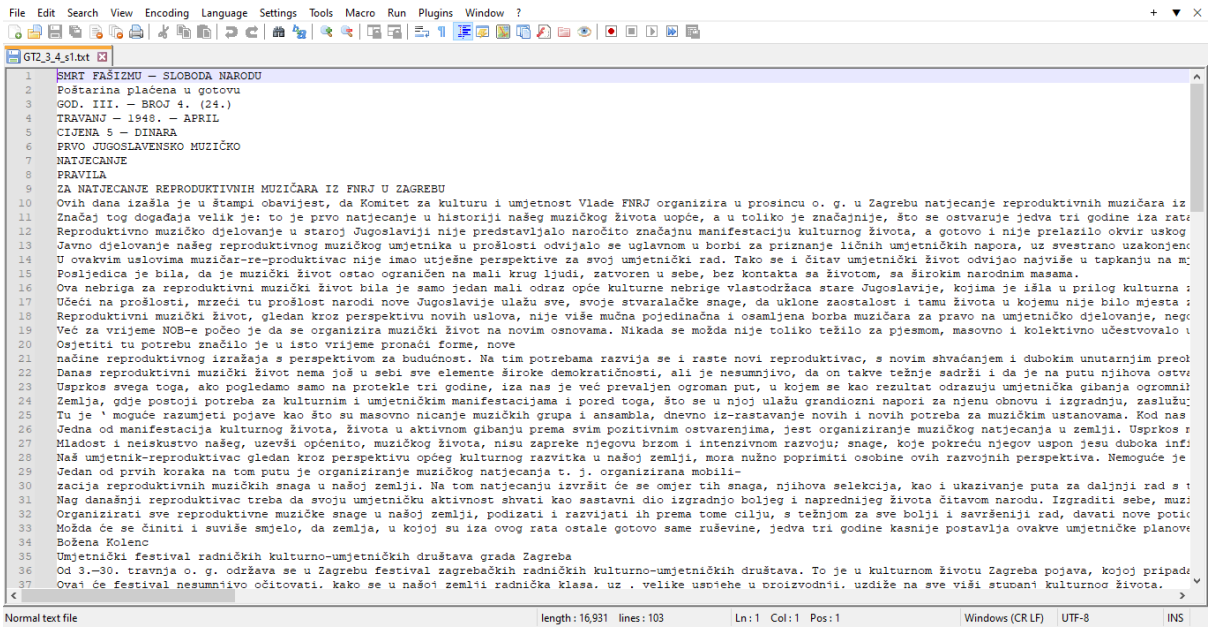
Kao što je spomenuto, u prvoj verziji temeljnog teksta struktura je ispravljena na način da se pratio izvornik i sve stupce teksta posložilo u izvorni redoslijed. Kako ova verzija temeljnog teksta izgleda možemo vidjeti na Slici 5.



Slika 5. Prva verzija temeljnog teksta sa ispravljenom strukturom stupaca.

5.3.1.2. Druga verzija temeljnog teksta

U drugoj verziji temeljnog teksta struktura je onakvom kakvom ju je prepoznao program ABBYY FineReader. Kako izgleda druga verzija temeljnog teksta vidljivo je na Slici 6.



Slika 6. Druga verzija temeljnog teksta sa zadržanom strukturom stupaca.

Temeljni tekst se tada usporedio s uzorcima, tj. deset stranica skeniranih na tri različite rezolucije. Za dobivanje podataka korišteni su ISRI (The Information Science Research Institute) analitički alati za evaluaciju OCR-a u verziji 5.1. (Rice i Nartker, 1996), tj. njihova ažurirana verzija pod imenom ocreval (GitHub). Od 1991. godine, ISRI je radio na razvoju mjera performansi sustava za prepoznavanje teksta. Njihove mjere omogućuju sveobuhvatnu evaluaciju. ISRI Analytic Tools dio je ISRI OCR Experimental Environmenta, a to je skup softverskih alata za provođenje automatiziranih testova prepoznavanja teksta širokih razmjera. Korisnik mora učitati tekstualne datoteke generirane OCR-om i tekstualne datoteke koje sadržavaju temeljni tekst. ISRI alati uključuju programe koji donose informacije o mjerama performansi, uspoređujući te dvije tekstualne datoteke

5.4. Analiza rezultata

Analizira se proces digitalizacije s uključenim optimizacijskim parametrima. To su vrijeme u radnim satima.

U istraživanju je korišten je program *accuracy*. Accuracy program uspoređuje temeljni tekst s tekstom generiranim putem OCR-a. Izvještaj programa sastoji se od šest dijelova. U prvome dijelu nalazi se ukupan broj znakova u temeljnom tekstu, broj pogrešaka i postotak točnosti (engl. *character accuracy*). Drugi dio donosi informacije o odbijenim znakovima, nesigurnim oznakama (engl. *suspect markers)* i netočnim oznakama (engl. *false marks)*. Također donose podatke o efikasnosti označenih znakova, što znači da ako korisnik pregleda označene znakove i ispravi ih, točnost će se povećati na iskazani broj. Pogreške se u programu vode kao radnje uređivanja (engl. *edit operations*) i mogu biti unos (engl. *insertion*), zamjena (engl. *substitution*) i brisanje (engl. *deletion*), a koje su potrebne kako bi se ispravio tekst generiran OCR-om. U trećem dijelu izvještaja nalaze se podaci o označenim pogreškama, neoznačenim pogreškama i ukupnim pogreškama po radnji uređivanja. Četvrti dio pokazuje točnost prema kategoriji znakova. Znakovi temeljnog teksta podijeljeni su u kategorije i predstavlja se izvještaj u postocima za prepoznate znakove u svakoj od kategorija. Peti dio analize navodi nedoumice, posložene prema broju pogrešaka koje su dodijeljene svakoj. Šesti dio izvještaja prikazuje kompletan popis znakova iz temeljnog teksta (Rice i Nartker, 1996). Za potrebe ovog postupka, u svakom direktoriju s OCR datotekama nalazile su se i GT datoteke, a u isti je direktorij sustav za evaluaciju generirao i datoteke s podacima o točnosti. Njihovi nazivi izgledali su ovako: acc_GT1_K_3_4_100_s1. Kratica “acc” označava da se radi o programu *accuracy.* GT1 označava verziju temeljnog teksta. Slovo “K” označava da se radi o kopirnom uređaju. Oznaka “3_4” odnosi se na godište i broj novina, a broj 100 označava rezoluciju skeniranja. Oznaka”s1” označava da se radi o prvoj stranici.Datoteke su spremljene u TXT formatu.

Prva četiri dijela izvještaja o točnosti programa ocreval možemo vidjeti na slici 37. Peti dio analize, koji navodi nedoumice (Rice i Nartker, 1996), prikazan je na slici 48. Šesti dio izvještaja s kompletnim popisom znakova iz temeljnog teksta (Rice i Nartker, 1996) može se vidjeti na slici 59.

```
ocreval Accuracy Report Version 7.0
-----------------------------------
   16761   Characters
    3180   Errors
   81.03%  Accuracy

       1   Reject Characters
       0   Suspect Markers
       0   False Marks
    0.01%  Characters Marked
   81.05%  Accuracy After Correction

     Ins    Subst      Del   Errors
       0        3        0        3   Marked
     758     2168      251     3177   Unmarked
     758     2171      251     3180   Total

   Count   Missed   %Right
    2591      583    77.50   ASCII Spacing Characters
     708      330    53.39   ASCII Special Symbols
     104       44    57.69   ASCII Digits
     501      132    73.65   ASCII Uppercase Letters
   12420     1754    85.88   ASCII Lowercase Letters
       6        2    66.67   Latin1 Special Symbols
     424       81    80.90   Latin Extended-A
       1        1     0.00   Cyrillic
       6        2    66.67   General Punctuation
   16761     2929    82.52   Total
```

Slika 7. Prva četiri dijela izvještaja o točnosti.

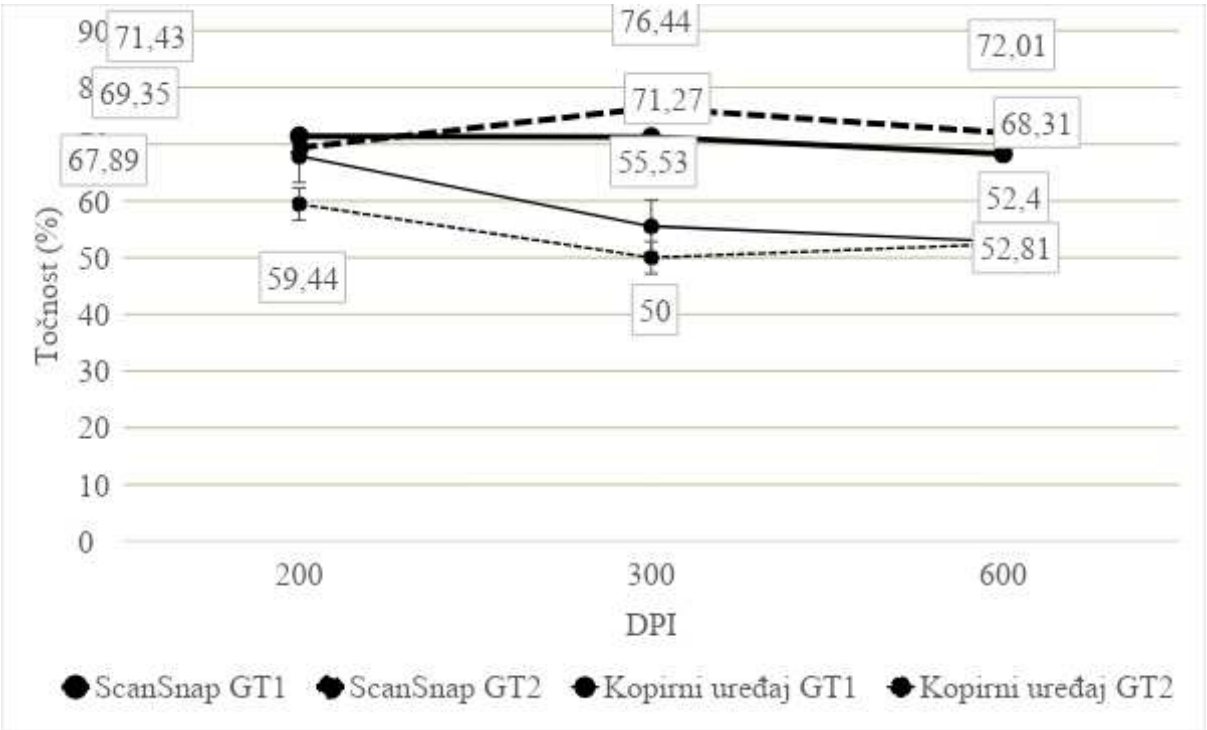
Errors	Marked	Correct-Generated
244	0	{-<\n>}-{}
156	0	{<\n>}-{ }
26	0	{}-{PRAVILA<\n>ZA NATJECANJE...}
25	0	{,}-{.}
25	0	{en-<\n>(Svršetak na 4. ...}-{}
20	0	{}-{DUKTIVNIH MUZIČARA<\n>O}
18	0	{MUZIČARA<\n>IZ FNRJ U}-{Olabrani kandidati}
16	0	{), Čehova<\n>(Prosi}-{cembra 19-18. go}
15	0	{tavom tur-<\n>nusu}-{, uzdiže na sve}
15	0	{u<\n>Zagrebu.<\n>2. U}-{c), dl, e) niže}
14	0	{REPRODUKTIVNIH}-{izlučni izbor.}
14	0	{ansambli.<\n>Ovaj}-{FNRJ te stalni}
14	0	{je-<\n>canja sači}-{a (Prst pred }
14	0	{nastu-<\n>pit će }-{žiri. Organi}
14	0	{prireda-<\n>ba; 9}-{na dan 3. XII.}
14	0	{ta.<\n>PRAVILA<\n>ZA}-{j etapi vršit}
13	0	{ kandidati iz}-{gor Bu-lyčo}

Slika 8. Peti dio izvještaja o točnosti.

Count	Missed	%Right	
528	428	18.94	{<\n>}
2063	155	92.49	{ }
2	1	50.00	{!}
1	0	100.00	{%}
12	7	41.67	{(}
38	18	52.63	{)}
193	56	70.98	{,}
232	186	19.83	{-}
163	52	68.10	{.}
20	8	60.00	{0}
27	9	66.67	{1}
7	2	71.43	{2}
10	4	60.00	{3}
10	7	30.00	{4}
10	4	60.00	{5}
3	2	33.33	{6}
3	0	100.00	{7}
7	4	42.86	{8}
7	4	42.86	{9}
33	6	81.82	{:}
34	4	88.24	{;}
27	9	66.67	{A}
16	2	87.50	{B}

Slika 9. Posljedni dio izvještaja o točnosti.

Grafikon 1 prikazuje podatke dobivene u istraživanju. Uspoređuje se rad na dva uređaja, rezolucija, dvije verzije temeljnog teksta te je prikazana dobivena točnost za svaku stranicu. Detaljni izračuni točnosti nalaze se u Prilogu 1, a grafikon prikazuje samo usporedive rezolucije.



Grafikon 1. Usporedba točnosti s obzirom na uređaj, rezoluciju i korištenje dvaju temeljnih tekstova.

Korištenjem ScanSnap uređaja na 150 dpi u OCR-u je postignuta točnost od 68,26% s prvom verzijom temeljnog teksta, a 67,08% s drugom verzijom. Na taj način možemo zamijetiti da je upotreba prvog temeljnog teksta dovela do procijenjene više točnosti, a razlika je otprilike 1%. Kopirni uređaj na 100 dpi postiže točnost od 52,61% s prvom verzijom temeljnog teksta, a 54,93% s drugom verzijom, čime se viša točnost postigla upotrebom drugog temeljnog teksta s razlikom od otprilike 2%. Iako se ne radi o usporedivim rezolucijama, ScanSnap je postigao puno bolje rezultate točnosti s oba temeljna teksta.

Na usporedivoj rezoluciji od 200 dpi, ScanSnap skener postigao je točnost od 71,43% s prvom verzijom temeljnog teksta, a 69,35% s drugom pa tako vidimo da je upotreba prvog temeljnog teksta dovela do više točnosti, a razlika iznosi otprilike 2%. Kopirni uređaj, pak,

postigao je rezultat od 67,89% s prvim temeljnim tekstom i 59,44% s drugim, gdje je došlo do veće razlike u točnosti između dva temeljna teksta te se boljim pokazao prvi temeljni tekst, s otprilike 8% razlike. Viša točnost zamjećuje se kod ScanSnap skenera.

Na rezoluciji od 300 dpi i u usporedbi s prvim temeljnim tekstom, skener je postigao točnost od 71,27%, a u usporedbi s drugim temeljnim tekstom ta vrijednost iznosi 76,43%. Možemo iščitati da je drugi temeljni tekst bio za oko 5% točniji. Kod kopirnog uređaja, usporedivši s prvim temeljnim tekstom, dobili smo točnost od 55,53%, dok je s drugim taj broj iznosio 50%. U ovom slučaju, prvi je temeljni tekst doprinio višoj točnosti i to za 5,53% više, a između dva uređaja, ScanSnap postigao je daleko višu točnost, s razlikom od otprilike 15% i 25%.

Kopirni je uređaj na rezoluciji od 400 dpi i s prvim temeljnim tekstom polučio rezultat od 65,39% točnih znakova, dok je s drugim temeljnim tekstom taj broj iznosio 59,06%. Time je prvi temeljni tekst dao bolje rezultate jer je razlika u točnosti iznosila otprilike 6%.

Na najvećoj rezoluciji od 600 dpi, ScanSnap uređaj omogućio je točnost koja je iznosila 68,31% koristeći prvi temeljni tekst, a 72,01% koristeći drugi. Možemo primijetiti da se rezultati razlikuju za otprilike 4% i time smo višu točnost postigli koristeći drugi temeljni tekst. Upotrebom kopirnog uređaja i prvog temeljnog teksta dobili smo 52,81% točnosti, a s drugim temeljnim tekstom taj broj iznosi 52,40%. Ovi su rezultati vrlo slični, no upotreba prvog temeljnog teksta dovela je do 0,41% više točnosti. ScanSnap uređaj i u ovom je slučaju omogućio postizanje više točnosti OCR-a i to s razlikom od otprilike 16% i 20%.

Najviša točnost postignuta je upotrebom ScanSnap skenera na rezoluciji od 300 dpi i drugog temeljnog teksta. Najniža točnost postignuta je kopirnim uređajem pri 600 dpi i drugim temeljnim tekstom. Viša točnost na svim usporedivim rezolucijama postignuta je upotrebom ScanSnap skenera te je razlika iznosila i do 25%. Prvi temeljni tekst u 66,67% slučajeva vodio je k višoj točnosti OCR-a. Dobiveni rezultati pokazuju da je i temeljni tekst, tj. njegovo oblikovanje, važno prilikom evaluacije točnosti.

6. Zaključak

Diplomski rad obradio je ciljeve postavljene u istraživanju. Istražen je utjecaj rezolucije na brzinu procesa digitalizacije pa time i njezine troškove. Rezultati su izraženi u radnim satima i kao točnost OCR-a korištenjem dviju verzija temeljnog teksta. Uzorak koji se upotrijebio u istraživanju sastojao se od dva broja, tj. deset stranica Muzičkih novina, a za skaliranje rezultata na stvarno okruženje korišten je primjer projekta Europeana Newspapers koji nastoji digitalizirati deset milijuna stranica. U istraživanju se pojavilo ograničenje, a to je činjenica da dva uređaja koja su korištena za skeniranje ne podržavaju jednake rezolucije, čime smo dobili tri usporedive rezolucije. Odabir postavki prilikom izvođenja procesa digitalizacije može dugoročno uvelike smanjiti troškove, a uz testiranje točnosti OCR-a i zadržati prihvatljive razine točnosti prepoznatog teksta. Savršenstvo u svim segmentima nemoguće je postići pa se traži određena "zlatna sredina" između troška digitalizacije i postignute razine točnosti OCR-a.

Na najnižoj rezoluciji od 200 dpi, kopirni se uređaj pokazao bržom i jeftinijom opcijom. Na 300 dpi je također brža i jeftinija upotreba kopirnog uređaja, no na maksimalnoj rezoluciji od 600 dpi ScanSnap uređaj je brži i jeftiniji. Trošak skeniranja pokazao se najnižim na kopirnom uređaju pri 200 dpi, a najviši na tom istom uređaju sa 600 dpi. Što se tiče točnosti, najviša točnost postignuta je upotrebom ScanSnap skenera s rezolucijom od 300 dpi i drugog temeljnog teksta. Najniža je, pak, točnost postignuta upotrebom kopirnog uređaja pri 600 dpi i drugog temeljnog teksta. Viša točnost na svim usporedivim rezolucijama omogućena je sa ScanSnap skenerom te je razlika iznosila i do 25%. Prvi temeljni tekst u 66,67% slučajeva vodio je k višoj točnosti OCR-a. Najviša točnost, kako je spomenuto, postignuta je na ScanSnap skenerom na rezoluciji od 300 dpi i upotrebom drugog temeljnog teksta. Ipak, opcija koja je imala najvišu točnost treća je po trošku. Opcija koja se pokazala najboljom u svim aspektima osim točnosti je kopirni uređaji s 200 dpi, što se ispostavilo kao najbrža opcija za skeniranje i OCR pa tako i prema trošku. Točnost korištenjem ove kvalitete nije bila najviša, međutim razlika s opcijom s najvišom točnošću je 8,54% (što je još uvijek relativno prihvatljivo).

Testiranjem prije početka projekta digitalizacije i procesa optičkog prepoznavanja znakova, štedi se vrijeme, novac, predviđaju se rezultati, moguće je dugoročno planiranje projekta, financiranje, prijaviti se za dodatno financiranje temeljem dobivenih rezultata itd. Nije uvijek moguće sve testirati, a postoje i drugi parametri koje je moguće uključiti (npr. drugi uređaji,

rezolucije, dubina boje itd.). Potrebno je pri donošenju odluke o tome biti pažljiv i donositi ju temeljem prethodnih istraživanja i proučavanjem relevantne literature.

7. Literatura

ABBYY FineReader PDF. Learning Center. Dostupno na: https://pdf.abbyy.com/learning-center/what-is-ocr/ [1.7.2023.]

Besser, H. (2000) Digital Longevity. U: Sitts, M. K. (ur.), Handbook for Digital Projects: A Management Tool for Preservation and Access. Andover: Northeast Document Conservation Center, str. 165-178.

Blostein, D. i Nagy, G. (2012) Asymptotic cost in document conversion. U Viard-Gaudin, C. i Zanibbi, R.. (ur.), *Proc. SPIE 8297, Document Recognition and Retrieval XIX, 82970N: San Francisco, 23. siječnja 2012.*

Boeing, M., Baierer, K., Hartmann, V., Federbusch, M. i Neudecker, C. (2019) Labelling OCR Ground Truth for Usage in Repositories. U: DATeCH2019: *Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage: Brussels, 8-10.5.2019*. New York: Association for Computing Machinery, str. 3-8.

Chapman, S. (2000). Working with Printed Text and Manuscripts. U: Sitts, M. K. (ur.), Handbook for Digital Projects: A Management Tool for Preservation and Access. Andover: Northeast Document Conservation Center, str. 114-122.

Clausner, C., Papadopoulos, C. Pletschacher, S. i Antonacopoulos, A. (2015) The ENP Image and Ground Truth Dataset of Historical Newspapers. U: *2015 13th International Conference on Document Analysis and Recognition (ICDAR): Nancy, 23-26.8.2015*. Tunis: IEEE, str. 931-935.

Commonwealth Consolidated Acts. Archives Act 1983. Dostupno na: http://www8.austlii.edu.au/cgi-bin/viewdb/au/legis/cth/consol_act/aa198398/ [10.5.2023]

Digitalizacija. (2021). Hrvatska enciklopedija, mrežno izdanje. Leksikografski zavod Miroslav Krleža. Dostupno na: http://www.enciklopedija.hr/Natuknica.aspx?ID=68025 [5.5.2023.]

Digitalni akademski arhivi i repozitoriji. Što je Dabar?. Dostupno na: https://dabar.srce.hr/dabar [9.5.2023.]

Digitalni repozitorij Muzičke akademije. Dostupno na: https://drma.muza.unizg.hr/ [5.5.2023]

Eikvil, L. (1993) OCR: Optical Character Recognition. Dostupno na: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=5c1f384f2478efb17a83c2f5ade6da059c7dbbea [12.10.2023.]

Europeana Newspapers. About. Dostupno na: http://www.europeana-newspapers.eu/ [15.7.2023.]

Gifford, Fenton, E. (2000). An OCR Case Study. U: Sitts, M. K. (ur.), Handbook for Digital Projects: A Management Tool for Preservation and Access. Andover: Northeast Document Conservation Center, str. 131-134.

GitHub. eddieantonio/ocreval. Dostupno na: https://github.com/eddieantonio/ocreval [20.7.2023.]

Hamad, A. i Kaya, M. (2016) A Detailed Analysis of Optical Character Recognition Techonology. International Journal of Applied Mathematics, Electronics and Computers, 4, 244-249.

Isthiaq, A. i Saif, N. A. (2020) OCR for Printed Bangla Characters Using Neural Network. I. J. Modern Education and Computer Science, 2, 19-29.

Katalog Knjižnica grada Zagreba. Muzičke novine Hrvatskog državnog Konzervatorija. Dostupno na: https://katalog.kgz.hr/pagesresults/bibliografskiZapis.aspx?&currentPage=1&searchById=1&sort=0&fid0=2&fv0=novine&spid0=1&spv0=glazba&xm0=1&selectedId=2004770 [2.7.2023.]

Kettunen, K. T., Kervinen, J. i Koistinen, J. (2018) Creating and Using Ground Truth OCR Sample Data for Finnish Historical Newspapers and Journals. U: *DHN 2018 Digital Humanities in the Nordic Countries 3rd Conference: Helsinki, 7-9. ožujka 2018.* Str. 162-169.

Kordić, M. (2021) Digitalna tranzicija turizma: Digitalna rješenja u sektoru turizma koja otvaraju nove prilike poduzećima, potiču razvoj pouzdane tehnologije, podupiru otvoreno i demokratsko društvo, omogućuju dinamično i održivo gospodarstvo te pomažu u borbi protiv klimatskih promjena i zelenoj tranziciji. Ministarstvo turizma i sports. Dostupno na: https://mint.gov.hr/UserDocsImages/NPOO/M3_3_1_digitalna_tranzicija_turizma.pdf [25.9.2023.]

Kumar, A. i Pati, P. B. (2023) Offline HWR Accuracy Enhancement with Image Enhancement and Deep Learning Techniques. Procedia Computer Science, 218, 35-44.

Manjunath Aradhya, V. N., Hemantha Kumar, G. i Noushath, S. (2008) Multilingual OCR system for South Indian scripts and English documents: An approach based on Fourier transform and principal component analysis. *Engineering Applications of Artificial Intelligence,* 21, 658–668.

Mithe, R., Indalkar, S. i Divekar, N. (2013) Optical Character Recognition. *International Journal of Recent Technology and Engineering (IJRTE)*, 2(1), 72-75.

Novine. (2021). Hrvatska enciklopedija, mrežno izdanje. Leksikografski zavod Miroslav Krleža. Dostupno na: https://www.enciklopedija.hr/Natuknica.aspx?ID=44284 [2.7.2023.]

Nunamaker, B., Bukhari S. S., Borth, D. i Dengel, A. (2016) A Tesseract-Based OCR Framework for Historical Documents Lacking Ground-Truth Text. U: *2016 IEEE International Conference on Image Processing (ICIP): Phoenix, 25-28. rujna 2016*. Phoenix: The Institute of Electrical and Electronics Engineers Signal Processing Society, str. 3269-3273.

Nwokoma, F., O., Odii, J. N., Ayogu, I. I. i Ogbonna, J. C. (2021) Camera-based OCR scene text detection issues: A review. *World Journal of Advanced Research and Reviews*, 12(3), 484–489.

Oluakachukwu Nneji, K. (2018) Digitization of academic library resources: A case study of Donal E. U. Ekong Library. Dostupno na: https://core.ac.uk/download/pdf/189483989.pdf [12.10.2023.]

Pandey, P. i Misra, R. (2014) Digitization of Library Materials in Academic Libraries: Issues and Challenges. *Journal of Industrial and Intelligent Information,* 2(2), 136-141.

Parekh, H. (2001) Digitization: An Overview of Issues. Dostupno na: https://ir.inflibnet.ac.in/bitstream/1944/79/1/cali_1.pdf [12.10.2023.]

Publikacija. (2021). Hrvatska enciklopedija, mrežno izdanje. Leksikografski zavod Miroslav Krleža. Dostupno na: https://www.enciklopedija.hr/Natuknica.aspx?ID=50998 [2.7.2023.]

Reynaert, M. (2014) On OCR ground truths and OCR post-correction gold standards, tools and formats. U: *DATeCH '14: Proceedings of the First International Conference on Digital*

Access to Textual Cultural Heritage: Madrid, 19-20. svibnja 2014. New York: Association for Computing Machinery, str. 159-166.

Rice, S. V. i Nartker, T. A. (1996). The ISRI Analytic Tools for OCR Evaluation: Version 5.1. Las Vegas: Information Science Research Institute. Dostupno na: https://github.com/jmokoistinen/isri-ocr-evaluation-tools/blob/master/user-guide.pdf [20.7.2023.]

Schumacher, A., Sihn, W. i Erol, S. (2016). Automation, digitization and digitalization and their implications for manufacturing processes. U: *Innovation and Sustainability 2016: International Scientific Conference: Bukurešt, 28-29. listopada 2016*.

Serijska publikacija. (2021). Hrvatska enciklopedija, mrežno izdanje. Leksikografski zavod Miroslav Krleža. Dostupno na: https://www.enciklopedija.hr/Natuknica.aspx?ID=55493 [2.7.2023.]

Shinde, A. A. i Chougule, D. G. (2012) Text Pre-processing and Text Segmentation for OCR. *IJCSET*, 2(1), 810-812.

Shigwan, R. (2015). Restoration and Digitization of Library Archival Materials: Issues and Challenges. Bulletin of the Deccan College Research Institute, 75, 351-368.

Singh, V. S. i Pal, P. (2014) Survey of Different Types of CAPTCHA. *International Journal of Computer Science and Information Technologies*, 5(2), 2242-2245.

Smith, R. (2007) An Overview of the Tesseract OCR Engine. U: *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, 2, 629 - 633.

Stančić, H. 2009. Digitalizacija. Zagreb: Zavod za informacijske studije.

Ströbel, P. B., Clematide, S., Volk, M., Schwitter, R., Hodel, T. i Schoch, D. (2022). Evaluation of HTR Models without Ground Truth Material. U: *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), Marseille, 20-25. lipnja 2022*. European Language Resources Association, str. 4395-4404.

Strugačevac, P. (1999). Teorijska osnova imaging CT tehnike. Osijek: Klinička bolnica Osijek.

Touj, S., Amara, N. B. i Amiri, H. (2005) Generalized Hough Transform for Arabic Printed Optical Character Recognition. *The International Arab Journal of Information Technology*, 2,4, 326-333.

Trbušić, Ž. (2022) Metode analize i optimizacije procesa optičkog prepoznavanja znakova u arhivskim informacijskim sustavima (Doktorska disertacija). Filozofski fakultet, Zagreb, Sveučilište u Zagrebu.

Ulges, A., Lampert, C. H. i Breuel, T. M. Document Image Dewarping using Robust Estimation of Curled Text Lines. U: *Eighth International Conference on Document Analysis and Recognition (ICDAR'05), Seoul, 31. kolovoza - 1. rujna 2005*. Dostupno na: https://www.dfki.de/fileadmin/user_upload/import/2003_UlgesCHLTMBDocImageDewarping.pdf [25.9.2023.]

updatestar. Recognita Standard OCR 3.2. Dostupno na: https://recognita-standard-ocr.updatestar.com/en [12.10.2023.]

Popis slika

Popis tablica

Popis grafikona

Prilozi

Prilog 1 - Usporedba točnosti s obzirom na uređaj, rezoluciju i korištenje dvaju temeljnih tekstova

Uređaj	Rezolucija	GT	Str.	Točnost	Uređaj	Rezolucija	GT	Str.	Točnost
ScanSnap	150 dpi	GT1	1	93,96%	Kopirni uređaj	100 dpi	GT1	1	81,03%
ScanSnap	150 dpi	GT1	2	86,96%	Kopirni uređaj	100 dpi	GT1	2	43,66%
ScanSnap	150 dpi	GT1	3	32,03%	Kopirni uređaj	100 dpi	GT1	3	53,97%
ScanSnap	150 dpi	GT1	4	24,46%	Kopirni uređaj	100 dpi	GT1	4	20,79%
ScanSnap	150 dpi	GT1	5	65,87%	Kopirni uređaj	100 dpi	GT1	5	67,57%
ScanSnap	150 dpi	GT1	6	35,34%	Kopirni uređaj	100 dpi	GT1	6	24,79%
ScanSnap	150 dpi	GT1	7	95,62%	Kopirni uređaj	100 dpi	GT1	7	63,29%
ScanSnap	150 dpi	GT1	8	75,73%	Kopirni uređaj	100 dpi	GT1	8	63,54%
ScanSnap	150 dpi	GT1	9	80,11%	Kopirni uređaj	100 dpi	GT1	9	54,39%
ScanSnap	150 dpi	GT1	10	92,48%	Kopirni uređaj	100 dpi	GT1	10	53,11%
				68,26%					52,61%
ScanSnap	150 dpi	GT2	1	97,62%	Kopirni uređaj	100 dpi	GT2	1	84,11%
ScanSnap	150 dpi	GT2	2	61,70%	Kopirni uređaj	100 dpi	GT2	2	80,27%
ScanSnap	150 dpi	GT2	3	27,65%	Kopirni uređaj	100 dpi	GT2	3	38,79%
ScanSnap	150 dpi	GT2	4	25,31%	Kopirni uređaj	100 dpi	GT2	4	20,30%
ScanSnap	150 dpi	GT2	5	68,53%	Kopirni uređaj	100 dpi	GT2	5	70,30%
ScanSnap	150 dpi	GT2	6	32,53%	Kopirni uređaj	100 dpi	GT2	6	24,83%
ScanSnap	150 dpi	GT2	7	93,70%	Kopirni uređaj	100 dpi	GT2	7	62,37%
ScanSnap	150 dpi	GT2	8	81,51%	Kopirni uređaj	100 dpi	GT2	8	71,13%
ScanSnap	150 dpi	GT2	9	84,18%	Kopirni uređaj	100 dpi	GT2	9	43,57%

ScanSnap	150 dpi	GT2	10	98,05%	Kopirni uređaj	100 dpi	GT2	10	53,62%
				67,08%					54,93%
ScanSnap	200 dpi	GT1	1	93,97%	Kopirni uređaj	200 dpi	GT1	1	93,48%
ScanSnap	200 dpi	GT1	2	87,31%	Kopirni uređaj	200 dpi	GT1	2	87,91%
ScanSnap	200 dpi	GT1	3	33,72%	Kopirni uređaj	200 dpi	GT1	3	75,96%
ScanSnap	200 dpi	GT1	4	59,91%	Kopirni uređaj	200 dpi	GT1	4	20,88%
ScanSnap	200 dpi	GT1	5	65,69%	Kopirni uređaj	200 dpi	GT1	5	65,74%
ScanSnap	200 dpi	GT1	6	37,38%	Kopirni uređaj	200 dpi	GT1	6	54,53%
ScanSnap	200 dpi	GT1	7	95,71%	Kopirni uređaj	200 dpi	GT1	7	51,63%
ScanSnap	200 dpi	GT1	8	66,05%	Kopirni uređaj	200 dpi	GT1	8	76,59%
ScanSnap	200 dpi	GT1	9	80,46%	Kopirni uređaj	200 dpi	GT1	9	94,98%
ScanSnap	200 dpi	GT1	10	94,11%	Kopirni uređaj	200 dpi	GT1	10	57,18%
				71,43%					67,89%
ScanSnap	200 dpi	GT2	1	97,88%	Kopirni uređaj	200 dpi	GT2	1	97,41%
ScanSnap	200 dpi	GT2	2	61,59%	Kopirni uređaj	200 dpi	GT2	2	62,01%
ScanSnap	200 dpi	GT2	3	30,43%	Kopirni uređaj	200 dpi	GT2	3	68,59%
ScanSnap	200 dpi	GT2	4	54,13%	Kopirni uređaj	200 dpi	GT2	4	20,20%
ScanSnap	200 dpi	GT2	5	68,34%	Kopirni uređaj	200 dpi	GT2	5	68,39%
ScanSnap	200 dpi	GT2	6	34,52%	Kopirni uređaj	200 dpi	GT2	6	36,28%
ScanSnap	200 dpi	GT2	7	93,75%	Kopirni uređaj	200 dpi	GT2	7	50,92%
ScanSnap	200 dpi	GT2	8	71,81%	Kopirni uređaj	200 dpi	GT2	8	61,39%
ScanSnap	200 dpi	GT2	9	84,47%	Kopirni uređaj	200 dpi	GT2	9	72,42%
ScanSnap	200 dpi	GT2	10	96,55%	Kopirni uređaj	200 dpi	GT2	10	56,77%
				69,35%					59,44%

ScanSnap	300 dpi	GT1	1	94,75%	Kopirni uređaj	300 dpi	GT1	1	79,59%
ScanSnap	300 dpi	GT1	2	54,54%	Kopirni uređaj	300 dpi	GT1	2	88,27%
ScanSnap	300 dpi	GT1	3	88,14%	Kopirni uređaj	300 dpi	GT1	3	23,15%
ScanSnap	300 dpi	GT1	4	47,58%	Kopirni uređaj	300 dpi	GT1	4	20,43%
ScanSnap	300 dpi	GT1	5	65,50%	Kopirni uređaj	300 dpi	GT1	5	60,72%
ScanSnap	300 dpi	GT1	6	17,28%	Kopirni uređaj	300 dpi	GT1	6	35,64%
ScanSnap	300 dpi	GT1	7	95,91%	Kopirni uređaj	300 dpi	GT1	7	51,87%
ScanSnap	300 dpi	GT1	8	75,93%	Kopirni uređaj	300 dpi	GT1	8	43,55%
ScanSnap	300 dpi	GT1	9	80,38%	Kopirni uređaj	300 dpi	GT1	9	94,88%
ScanSnap	300 dpi	GT1	10	92,69%	Kopirni uređaj	300 dpi	GT1	10	57,23%
				71,27%					55,53%
ScanSnap	300 dpi	GT2	1	98,72%	Kopirni uređaj	300 dpi	GT2	1	83,12%
ScanSnap	300 dpi	GT2	2	97,19%	Kopirni uređaj	300 dpi	GT2	2	62,36%
ScanSnap	300 dpi	GT2	3	77,95%	Kopirni uređaj	300 dpi	GT2	3	22,96%
ScanSnap	300 dpi	GT2	4	48,17%	Kopirni uređaj	300 dpi	GT2	4	19,99%
ScanSnap	300 dpi	GT2	5	68,14%	Kopirni uređaj	300 dpi	GT2	5	63,17%
ScanSnap	300 dpi	GT2	6	16,65%	Kopirni uređaj	300 dpi	GT2	6	38,32%
ScanSnap	300 dpi	GT2	7	94,04%	Kopirni uređaj	300 dpi	GT2	7	51,31%
ScanSnap	300 dpi	GT2	8	80,32%	Kopirni uređaj	300 dpi	GT2	8	29,64%
ScanSnap	300 dpi	GT2	9	84,63%	Kopirni uređaj	300 dpi	GT2	9	72,29%
ScanSnap	300 dpi	GT2	10	98,53%	Kopirni uređaj	300 dpi	GT2	10	56,88%
				76,43%					50,00%
ScanSnap	600 dpi	GT1	1	94,16%	Kopirni uređaj	400 dpi	GT1	1	73,90%
ScanSnap	600 dpi	GT1	2	45,03%	Kopirni uređaj	400 dpi	GT1	2	81,06%

ScanSnap	600 dpi	GT1	3	33,20%	Kopirni uređaj	400 dpi	GT1	3	90,49%
ScanSnap	600 dpi	GT1	4	89,93%	Kopirni uređaj	400 dpi	GT1	4	20,89%
ScanSnap	600 dpi	GT1	5	65,82%	Kopirni uređaj	400 dpi	GT1	5	66,06%
ScanSnap	600 dpi	GT1	6	16,60%	Kopirni uređaj	400 dpi	GT1	6	35,89%
ScanSnap	600 dpi	GT1	7	93,35%	Kopirni uređaj	400 dpi	GT1	7	51,88%
ScanSnap	600 dpi	GT1	8	75,48%	Kopirni uređaj	400 dpi	GT1	8	81,62%
ScanSnap	600 dpi	GT1	9	77,00%	Kopirni uređaj	400 dpi	GT1	9	94,90%
ScanSnap	600 dpi	GT1	10	92,49%	Kopirni uređaj	400 dpi	GT1	10	57,23%
				68,31%					65,39%
ScanSnap	600 dpi	GT2	1	98,18%	Kopirni uređaj	400 dpi	GT2	1	77,24%
ScanSnap	600 dpi	GT2	2	83,60%	Kopirni uređaj	400 dpi	GT2	2	48,41%
ScanSnap	600 dpi	GT2	3	29,80%	Kopirni uređaj	400 dpi	GT2	3	76,90%
ScanSnap	600 dpi	GT2	4	71,87%	Kopirni uređaj	400 dpi	GT2	4	20,33%
ScanSnap	600 dpi	GT2	5	68,48%	Kopirni uređaj	400 dpi	GT2	5	68,72%
ScanSnap	600 dpi	GT2	6	16,32%	Kopirni uređaj	400 dpi	GT2	6	39,30%
ScanSnap	600 dpi	GT2	7	91,54%	Kopirni uređaj	400 dpi	GT2	7	51,13%
ScanSnap	600 dpi	GT2	8	80,63%	Kopirni uređaj	400 dpi	GT2	8	79,52%
ScanSnap	600 dpi	GT2	9	81,54%	Kopirni uređaj	400 dpi	GT2	9	72,33%
ScanSnap	600 dpi	GT2	10	98,18%	Kopirni uređaj	400 dpi	GT2	10	56,71%
				72,01%					59,06%
					Kopirni uređaj	600 dpi	GT1	1	82,05%
					Kopirni uređaj	600 dpi	GT1	2	51,45%
					Kopirni uređaj	600 dpi	GT1	3	28,63%
					Kopirni uređaj	600 dpi	GT1	4	21,03%

					Kopirni uređaj	600 dpi	GT1	5	61,02%
					Kopirni uređaj	600 dpi	GT1	6	34,45%
					Kopirni uređaj	600 dpi	GT1	7	51,84%
					Kopirni uređaj	600 dpi	GT1	8	44,60%
					Kopirni uređaj	600 dpi	GT1	9	94,78%
					Kopirni uređaj	600 dpi	GT1	10	58,29%
									52,81%
					Kopirni uređaj	600 dpi	GT2	1	85,69%
					Kopirni uređaj	600 dpi	GT2	2	77,21%
					Kopirni uređaj	600 dpi	GT2	3	25,50%
					Kopirni uređaj	600 dpi	GT2	4	20,46%
					Kopirni uređaj	600 dpi	GT2	5	63,48%
					Kopirni uređaj	600 dpi	GT2	6	37,42%
					Kopirni uređaj	600 dpi	GT2	7	51,06%
					Kopirni uređaj	600 dpi	GT2	8	33,12%
					Kopirni uređaj	600 dpi	GT2	9	72,44%
					Kopirni uređaj	600 dpi	GT2	10	57,64%
									52,40%

Digitalizacija arhivskoga gradiva s uključenim procesom optičkog prepoznavanja znakova na primjeru Muzičkih novina Hrvatskog državnog konzervatorija

Sažetak

Diplomski rad donosi temu procesa digitalizacije u knjižnici Muzičke akademije u Zagrebu s uključenim procesom optičkog prepoznavanja znakova na primjeru Muzičkih novina Hrvatskog državnog konzervatorija. Istražen je utjecaj rezolucije na brzinu digitalizacije, a time i troškove. Rezultate se izrazilo u radnim satima i kao točnost OCR-a, a korištene su dvije verzije temeljnog teksta. Uzorak se sastojao od dva broja novina, tj. deset stranica. Odabir postavki digitalizacije može dugoročno smanjiti troškove, a uz testiranje točnosti OCR-a i zadržati prihvatljive razine vjerodostojnosti. Važna je ravnoteža između troška digitalizacije i postignute razine točnosti. Najviša je točnost postignuta sa ScanSnap skenerom na 300 dpi i drugim temeljnim tekstom. Najniža je postignuta upotrebom kopirnog uređaja pri 600 dpi i drugog temeljnog teksta. Sveukupno je viša točnost omogućena sa ScanSnapom. Prvi temeljni tekst u 66,67% slučajeva vodio je k višoj točnosti OCR-a. Najbolja opcija u svim aspektima osim točnosti je kopirni uređaji na 200 dpi.

Ključne riječi: digitalizacija, OCR, temeljni tekst, Muzičke novine, ocreval, ISRI alati

Digitization of archival material with the included process of optical character recognition on the example of Muzičke novine

Summary

The thesis presents the digitization process in the library of the Academy of Music in Zagreb with OCR on "Muzičke novine". The impact of the resolution on the speed and the costs was investigated. The results were expressed in working hours and OCR accuracy. Two versions of GT were used. The sample consisted of two issues, i.e. ten pages. Choosing digitization settings can reduce costs in the long run, and with OCR accuracy testing, maintain acceptable levels of accuracy. The balance between the cost of and the level of accuracy is important. The highest accuracy was achieved with the ScanSnap scanner at 300 dpi and the second version of GT. The lowest was achieved using a copier at 600 dpi and the second version of GT. Overall higher accuracy was observed with the ScanSnap scanner. The first version of GT led to higher accuracy in 66.67% of cases. The best option in all aspects except accuracy is the copier at 200 dpi.

Key words: digitization, OCR, Ground Truth, Muzičke novine, ocreval, ISRI Tools