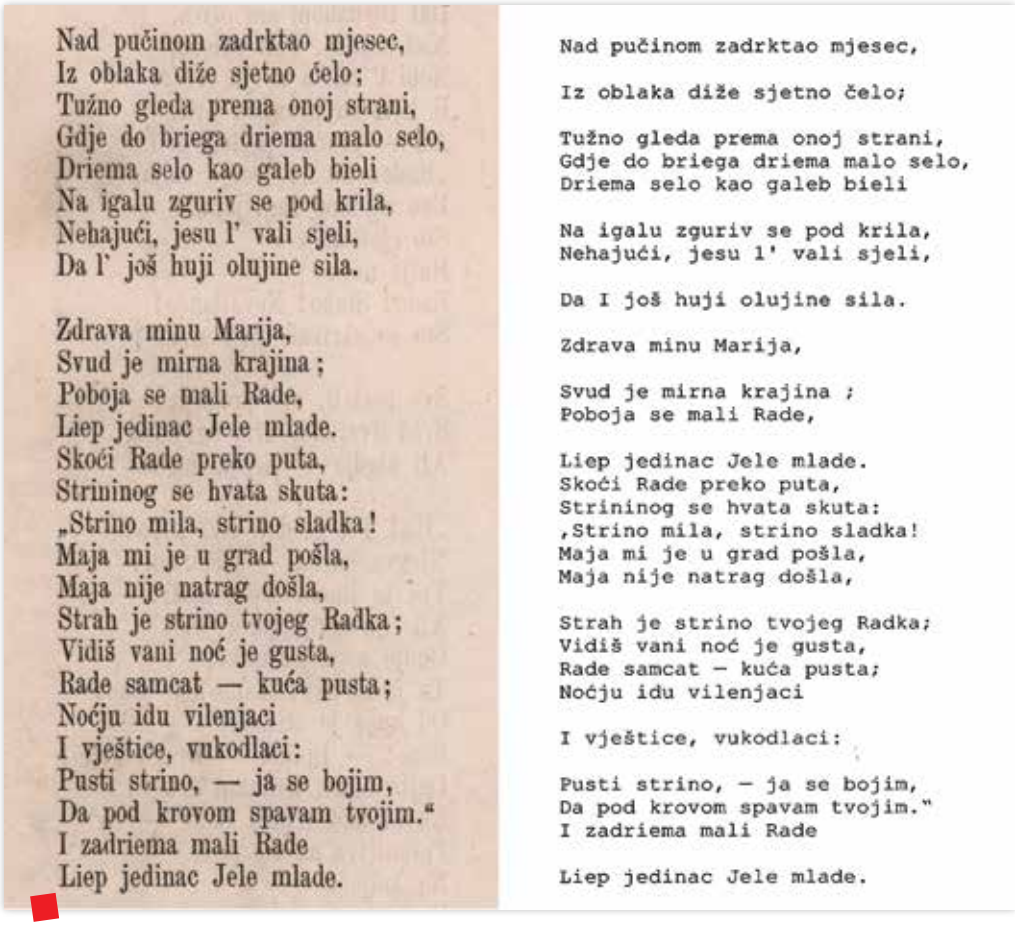
Zašto je arhivima potreban sustav za optičko prepoznavanje znakova?

Željko Trbušić

Što je OCR?

Optičko prepoznavanje znakova (engl. *Optical Character Recognition*) tehnologija je kojom se slika teksta pretvara u oblik razumljiv računalima i strojevima. Ovaj proces omogućuje pretraživanje, indeksiranje (automatsko katalogiziranje) i drugu manipulaciju skeniranog ili fotografiranog teksta na način kao da je izvorno nastao u elektroničkom okruženju. Praktična primjena ove tehnologije može se, primjerice, vidjeti pri automatskom očitavanju registarskih oznaka pri ulasku automobilom na parkiralište (engl. *smart parking*, pametno parkiranje), no njen značaj seže puno dalje i dublje od toga. Bitno je razlikovati tehnologije koje služe za prepoznavanje rukopisa (HWR/HTR – *Handwriting recognition / Handwritten text recognition* (Prepoznavanje rukopisa / Prepoznavanje rukopisnog teksta), ICR – *Intelligent Character Recognition* (Inteligentno prepoznavanje znakova)) od OCR-a koji se koristi za prepoznavanje tiskanog teksta kojeg odlikuje konzistentnost otiska. Da bi OCR sustav prepoznao slovo a mora taj proces za svaki oblik slova naučiti samo jednom, dok je svako rukopisno a na svoj način individualno.



Nad pučinom zadrktao mjesec,
Iz oblaka diže sjetno čelo;
Tužno gleda prema onoj strani,
Gdje do briega driema malo selo,
Driema selo kao galeb bieli
Na igalu zguriv se pod krila,
Nehajući, jesu l' vali sjeli,
Da l' još huji olujine sila.

Zdrava minu Marija,
Svud je mirna krajina;
Poboja se mali Rade,
Liep jedinac Jele mlade.
Skoći Rade preko puta,
Strininog se hvata skuta:
„Strino mila, strino sladka!
Maja mi je u grad pošla,
Maja nije natrag došla,
Strah je strino tvojeg Radka;
Vidiš vani noć je gusta,
Rade samcat — kuća pusta;
Nоćju idu vilenjaci
I vještice, vukodlaci:
Pusti strino, — ja se bojim,
Da pod krovom spavam tvojim.“
I zadriema mali Rade
Liep jedinac Jele mlade.

Nad pučinom zadrktao mjesec,

Iz oblaka diže sjetno čelo;

Tužno gleda prema onoj strani,
Gdje do briega driema malo selo,
Driema selo kao galeb bieli

Na igalu zguriv se pod krila,
Nehajući, jesu l' vali sjeli,

Da I još huji olujine sila.

Zdrava minu Marija,

Svud je mirna krajina ;
Poboja se mali Rade,

Liep jedinac Jele mlade.
Skoći Rade preko puta,
Strininog se hvata skuta:
‚Strino mila, strino sladka!
Maja mi je u grad pošla,
Maja nije natrag došla,

Strah je strino tvojeg Radka;
Vidiš vani noć je gusta,
Rade samcat — kuća pusta;
Nоćju idu vilenjaci

I vještice, vukodlaci:

Pusti strino, — ja se bojim,
Da pod krovom spavam tvojim."
I zadriema mali Rade

Liep jedinac Jele mlade.

Ulomak Kugine kuće Augusta Šenoe objavljen u Viencu broj 19 iz 1869. i optički prepoznat Tesseract OCR sustavom

Za prve korake u razvoju OCR-a zaslužni su, između ostalih, C. R. Carey (1870. – sustav prijenosa slika mozaikom fotoćelija), P. Nipkow (1884. – Nipkow disk, uređaj za prepoznavanje uzoraka) i ruski znanstvenik Tyuring (1900. – pokušao je izraditi sustav za pomoć slijepim i slabovidnim osobama), a Ray Kurzweil (američki izumitelj i nositelj dvadeset i jednog počasnog doktorata) prvi izrađuje OCR stroj sposoban pročitati više različitih oblika slova (1978. – *Kurzweil Reading Machine*). Godine 1989. osnovana je ruska tvrtka Abbyy čiji proizvod Abbyy Fine Reader postaje *de facto* zlatni standard za pretvaranje papirnatoga gradiva u digitalni oblik, a u 21. stoljeću OCR se širi u domenu mobilnih aplikacija te sustava u oblaku (engl. *cloud computing*), ali se i pridružuje *open source* (sustavi otvorenog koda) zajednici (2005. – Tesseract, softver za optičko prepoznavanje znakova, postaje *open source*).

Kako OCR primijeniti u arhivima?

Arhivski informacijski sustavi organizirani su na način da koristeći postojeće resurse (ljudske i sistemske) pružaju uslugu čuvanja, obrade i davanja na uvid gradiva koje im je predano (dobar primjer standarda za izgradnju arhivskog informacijskog sustava je OAIS referentni model). Digitalizacijom klasičnog, papirnatog gradiva te korištenjem sustava za njegovu pohranu i obradu ubrzava se proces pristupa, a često i eliminira potreba za čuvanjem originalnih primjeraka. Svaki dokument, nakon što je digitaliziran, mora biti logički pohranjen i označen metapodacima (kao što su autor, vrijeme, struktura, količina ili predmet) radi lakšeg pronalaska i migracije unutar sustava (no i dokazivanja vjerodostojnosti i autentičnosti). Metapodatke je moguće upisati ručno ili automatski generirati korištenjem OCR-a. Kod velike količine dokumenata ručni unos, iako pouzdaniji, nije vremenski efikasan i otvara se pitanje automatizacije procesa.

Za izgradnju sustava koji implementira optičko prepoznavanje znakova potrebna je detaljna analiza postojećih tehnoloških rješenja (softvera) te je poželjno, prije odabira, testirati OCR na gradivu koje se planira obraditi u većoj količini. Sam proces često se dijeli

Optičkim prepoznavanjem znakova stvaramo dokument čijim sadržajem možemo elektronički manipulirati

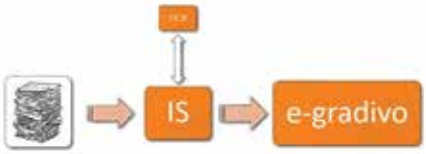
u nekoliko faza, a u svakoj je moguće ostvariti ispitivanje elemenata sustava što u konačnici rezultira većom točnošću prepoznatog teksta. Već u početnoj fazi digitalizacije proces je moguće prilagoditi OCR-u korištenjem pravilne rezolucije i orijentacije dokumenta na skeneru te odlukom o skeniranju u boji ili u nijansama sive. Osim toga, potrebno je voditi računa o kompresiji i veličini datoteka, jer se ti parametri odražavaju na vrijeme obrade (radne sate računala i zaposlenika arhiva), ali i na potrebnu količinu prostora za pohranu. Nakon digitalizacije slijedi proces obrade dobivenih datoteka (engl. *preprocessing*) s ciljem dobivanja još boljih rezultata optičkog prepoznavanja. Odabranim softverom za obradu slike najčešće se uklanjaju nepotrebni rubovi i slika se binarizira (postiže se kontrast između teksta i pozadine), a dodatno se prilagođava veličina, orijentacija i uklanjaju druge smetnje (mrlje ili tragovi korištenja). Svaki tip dokumenta zahtjeva drugačije postavke i stoga je teže u potpunosti automatizirati proces (bez ljudske kontrole) jer krajnji rezultat moraju biti predlošci na kojima su sva slova jasno vidljiva, koji imaju dobar kontrast između teksta i pozadine te nisu zakrenuti, niti sadrže bilo kakve anomalije koje izazivaju smetnje pri prepoznavanju teksta. Treća faza obuhvaća proces optičkog prepoznavanja nakon čega slijedi evaluacija rezultata i ispravljanje grešaka (ručno ili automatski). Odluke donesene u pojedinim fazama i odabir načina obrade rezultiraju određenim stupnjem točnosti i većom ili manjom uštedom vremena.

Nakon što je ovaj proces završen, dokument se vraća u sustav, a tekst koji je optički prepoznat postaje njegov sastavni dio i koristi se za dohvat i pretraživanje, koje bez OCR-a ne bi bilo moguće ili bi zahtijevalo dugotrajno i neisplativo ručno prepisivanje i ispunjavanje.

Budućnost

Osim što je ispravno očekivati veću brzinu obrade podataka, bolju implementaciju i automatizaciju cjelokupnog procesa, budućnost OCR i HWR/HWT tehnologija je i u pamćenju svih tekstualnih podataka koji nas okružuju, pa tako i onih arhivskih. Dolazi vrijeme kada nam neće biti potrebni glomazni skeneri kojima digitaliziramo gradivo, pa čak ni rukovanje fotografskim aparatom, već će pregled (uređajem u sklopu nosive tehnologije, engl. *wearable technology*) biti dovoljan kako bi sve relevantne informacije bile upisane i lako dostupne kroz postojeći informacijski sustav. To je svijet u kojem će se sve manje čitati, a sve više pretraživati. Arhivi će morati postati dio te stvarnosti kako bi ostali, ili postali, od važnosti za zajednicu, a ne samo spremišta gradiva koje nitko ne želi i nema vremena pretraživati.

Ray Kurzweil i njegov izum Kurzweil Reading Machine (1977.)



Pojednostavljeni prikaz uloge OCR-a u informacijskim sustavima za digitalizaciju dokumenata

SAZNAJTE VIŠE:

Više informacija o *Open Archival Information System* (OAIS, Otvoreni arhivski informacijski sustav) dostupno je na: http://www.oais.info/

O procesu obrade digitaliziranih datoteka saznajte više na: https://github.com/tesseract-ocr/tesseract/wiki/ImproveQuality

G. Nagy (2018) The Lifetime Reader. U: *IEEE Pervasive Computing*, sv. 17, br. 4, str. 86-95. Dostupno na: https://www.ecse.rpi.edu/~nagy/PDF_chrono/2018_The_Lifetime_Reader.pdf