IM 44 (1-4) 2013.
IZ MUZEJSKE TEORIJE I PRAKSE
MUSEUM THEORY AND PRACTICE

PROJEKT DIGITALIZACIJE HEMEROTEKE HRVATSKOGA POVIJESNOG MUZEJA

JELENA BALOG VOJAK □ Hrvatski povijesni muzej, Zagreb

ZDENKA ŠINKIĆ □ Hrvatski povijesni muzej, Zagreb

sl.1. Hemeroteka (grč.ἡμέρα -dan + -teka), zbirka novina i časopisa; zbirka izrezaka iz tekućih novina, časopisa ili koje druge tiskane građe o određenim, unaprijed utvrđenim temama ili predmetima, obično složena prema abecednom ili kronološkom slijedu natuknica. Primjenjuje se u knjižnicama ili drugim informacijskim ustanovama kao brz izvor informacija i kao dopuna podataka, osobito o aktualnim temama i događajima.
Preuzeto sa: http://www.enciklopedija.hr/Natuknica.aspx?ID=24926

1. Uvod. Jedan od ciljeva ovog rada jest analizirati proces digitalizacije hemeroteke Hrvatskoga povijesnog muzeja, s posebnim osvrtom na metodološke, razvojne i druge izazove. Ideja primjene optičkog prepoznavanja znakova na povijesnim dokumentima nametnula se tijekom vremena kao moguće rješenje zaštite i lakše dostupnosti takvog materijala. Glavni izazovi projekta bili su što odabrati (koju građu) i kako (kojim alatom). Projekt je započet u siječnju 2014., a provodi ga dokumentacijska služba Hrvatskoga povijesnog muzeja. Prvi korak projekta bio je odabir materijala za digitalizaciju.

2. Hemeroteka u Hrvatskom povijesnom muzeju

Hemeroteka je jedan od fondova sekundarne dokumentacije te uz izdavačku i izložbenu djelatnost pripada najopsežnijim fondovima sekundarne dokumentacije Hrvatskoga povijesnog muzeja. Hrvatski povijesni muzej osnovan je 1991. spajanjem Muzeja revolucije naroda Hrvatske (osnovanoga 1945.) i Povijesnog muzeja Hrvatske (osnovanoga 1846.) te ima povijest djelovanja dulju od 150 godina i fundus s više od 300 000 predmeta. Bogatu djelatnost Muzeja pratile su brojne tiskovine, a članke su počeli skupljati djelatnici Muzeja prije više od 60 godina i danas hemeroteka broji oko 2 000 jedinica, djelomično skeniranih i pohranjenih u PDF formatu. Cjelokupna hemeroteka obrađena je u aplikaciji za obradu muzejske dokumentacije koja omogućuje ispis i pretraživanje po željenim kriterijima. Usto je povezana i s ostalim fondovima sekundarne dokumentacije.

Hemeroteka Hrvatskoga povijesnog muzeja podijeljena je na nekoliko cjelina: izložbe, povijesne teme, muzeološke teme, Hrvatski povijesni muzej i Domovinski rat. Povijesne i muzeološke teme pripadaju cjelinama koje obuhvaćaju pojedinačne članke vezane za neke

općenite povijesne ili muzejske teme. Posebnu cjelinu čine članci vezani za najave preseljenja Hrvatskoga povijesnog muzeja, koji su objavljivani još potkraj 1960-ih godina, te članci o nekim nemuzejskim aktivnostima. Primjerice, *Sajam cvijeća* (prethodnik današnjeg *Floraarta*) održavao se 1960-ih godina u tadašnjemu Povijesnom muzeju Hrvatske. Najbrojnija je cjelina ona o izložbenoj djelatnosti Muzeja, koja daje uvid u njegovu bogatu izložbenu djelatnost. Iz navedenoga je vidljivo da hemeroteka HPM-a sadržava zanimljive izvore ne samo za proučavanje djelatnosti Muzeja, već i za istraživanje društvenoga i kulturnog razvoja Zagreba i Hrvatske tijekom većeg dijela 20. st. pa do danas. Jedna od posebnosti fonda jesu jedinice koje se odnose na zbivanja za vrijeme Domovinskog rata, posebno na ratna razaranja kulturnih dobara.

sl.2. Epson V750 PRO (A4) i Epson GT 20000 (A3)

3. Projekt digitalizacije hemeroteke Hrvatskoga povijesnog muzeja

U višegodišnjem radu pokazalo se da je muzejska hemeroteka bogat izvor informacija za različite teme,pa se nametnula i potreba njezina šireg korištenja. Istraživači različitih profila (posebno povjesničari) često posežu za novinskim člancima kao izvorom informacija i sastavnicama društvene slike nekog vremena. Imajući na umu osjetljivost papirnog materijala, a time i njegovu ograničenu dostupnost, započet je projekt digitalizacije cjelokupne hemeroteke Muzeja.

Jedan od osnovnih ciljeva projekta jest učiniti hemeroteku pristupačnom širem krugu korisnika ne samo za čitanje, već i za pretraživanje (unutar samog teksta). Stoga je cilj projekta, osim skeniranja, i OCR obrade, spremanje jedinica u nekom od formata koji omogućava pretraživanje.

Projekt obuhvaća:

- pripremu članaka za skeniranje
- skeniranje članaka
- automatsku obradu (OCR)
- unos teksta u aplikaciju za obradu sekundarne dokumentacije
- računalno pohranjivanje u PDF formatu prema inventarnoj oznaci fonda sekundarne dokumentacije.

4. Optičko prepoznavanje znakova

Optičko prepoznavanje znakova (engl. *Optical Character Recognition*, OCR) računalna je tehnologija koja omogućuje pretvorbu skeniranih papirnatih dokumenata, PDF dokumenata i slikovnih dokumenata snimljenih digitalnim fotoaparatom ili skeniranih u formate koji se mogu uređivati. Sustav OCR čine oprema (skener ili fotoaparat) i sofisticirani program.[1]

Postupak optičkog prepoznavanja znakova može se podijeliti na četiri glavne faze:

- digitalizaciju (skeniranje ili digitalno fotografiranje) materijala
- računalnu obradu teksta
- korekturu obrađenog teksta
- spremanje teksta u željeni računalni format.

4.1. Digitalizacija

Tekstualno se gradivo može digitalizirati na tri načina: prepisivanjem na računalu, skeniranjem i snimanjem digitalnim fotoaparatom. Prepisivanje je najdugotrajniji postupak te se malokad primjenjuje. S obzirom na vrstu građe, mi smo se odlučili za skeniranje. Opremu čine dva skenera Epson V750 PRO (A4) i Epson GT 20000 (A3) koje Muzej posjeduje za digitalizaciju građe, što djelatnici obavljaju sami unutar Muzeja. Skeniramo u rezoluciji 600 dpi kako bi se postigla što bolja kvaliteta.

Već je spomenuto da je dio hemeroteke HPM-a već spremljen u digitalnom obliku, većim dijelom u PDF-u, a manjim dijelom u JPG formatu. Ostala se građa najprije skenira, a potom se računalno obrađuje.

4.2. Računalna obrada teksta

Ono što odlikuje većinu današnjih OCR programa jest mogućnost obrade izgleda stranice. Današnji dokumenti i knjige nisu pisane isključivo tekstom, slijeva nadesno od vrha do dna, već imaju umetnute slike, tablice, dijagrame i ostale grafičke prikaze, a tekst na stranici može biti podijeljen u dva ili više stupaca. Pritom se najviše očituju razlike među programima. Zato je bilo potrebno odabrati program s najvjernijom mogućnosti reproduciranja dokumenta u digitalni format. Također, taj digitalni format mora biti prikladan za daljnju upotrebu.

Da bismo odabrali optimalno rješenje, trebali smo isprobati više dostupnih programa za optičko prepoznavanje znakova. Pri odabiru smo se fokusirali samo na besplatne programe, kako bi cijeli proces ostao unutar kuće i bez dodatnih financijskih troškova imajući na umu da nam je dio građe već prije skeniran Ovdje ćemo prezentirati neke od programa koje smo isprobali, pazeći pritom na faze OCR-a.

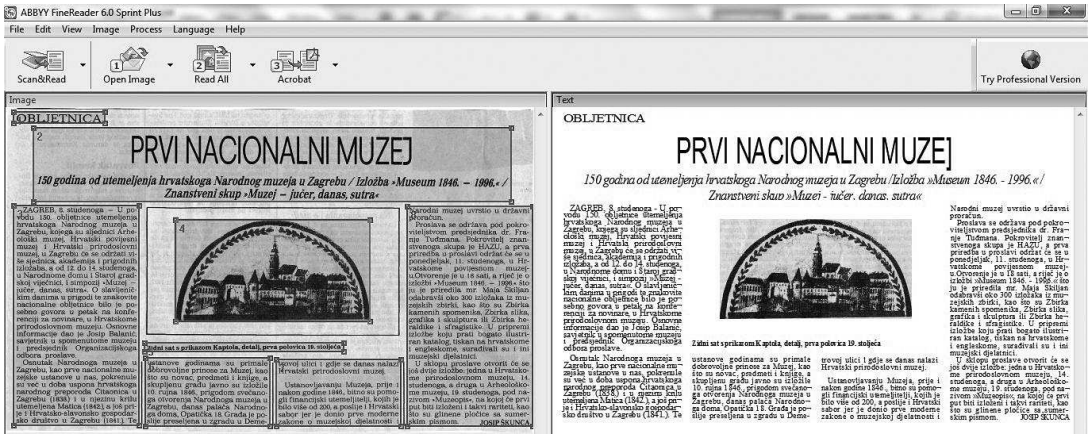
Na internetu je moguće naći velik broj besplatnih OCR programa: Free-Ocr, Simple OCR, i2OCR, FreeOCR, Microsoft Office Document Imaging (MODI), OnlineOCR, TopOCR, CuneiForm, OCROnline, Free Online OCR, NewOCR (da ne nabrajamo sve), od kojih smo dio i mi isprobali. Većina njih ima problem s prepoznavanjem hrvatskog jezika, odnosno s dijakritičkim znakovima, a ovdje ćemo na nekoliko izdvojenih primjera navesti koje su zapravo razlike među njima.

4.2.1. FREE OCR

Program se može besplatno preuzeti[2] i instalirati na računalo. Omogućuje skeniranje i OCR dokumenta

1 Letak objavljen na portalu URL: http://www.digitalmedia.hr/PDF/Business_use.pdf (30. lipnja 2014.).

2 http://www.softpedia.com/get/Others/Miscellaneous/FreeOCR.shtml



sl. 3. Proces OCR-a u ABBYY Fine Readeru

u jednom potezu. Uz slikovne, čita i PDF format te je pogodan za građu koja je unaprijed skenirana, kao i za onu koja to nije. Omogućuje odabir jezika, ali ne nudi hrvatski. Dijakritički se znakovi mogu pročitati odabirom poljskog jezika. Nema ograničenja broja dokumenata koji se skeniraju.

Nedostatak Free OCR-a jest to što program ne prepoznaje hrvatski jezik, pa ćete nakon skeniranja slike morati ručno ispraviti pojedine pogrešno napisane riječi, a i sam izlazni rezultat može biti dosta nečitljiv.

4.2.2. NEW OCR[3] i I2OCR[4]

Za dva slična programa u kojima se radi online nije potrebna registracija, no preduvjet je da je građa unaprijed skenirana i pohranjena u računalu. Čitaju sve formate (PDF, TIFF, JPG,..), bez ograničenja veličine. Najprije je potrebno odabrati jezik dokumenta, a postoji i mogućnost odabira hrvatskog jezika. Potom se pronalazi datoteka i šalje s računala. Slijedi OCR te program izbacuje konačni rezultat koji je moguće pohraniti u računalo. Programi su se pokazali lošima zbog nečitljivoga izlaznog rezultata.

4.2.3. ONLINE OCR[5]

Program se također primjenjuje online i podrazumijeva da je građa unaprijed skenirana i pohranjena u računalu. Čita sve formate (PDF, TIFF, JPG...). Postoji mogućnost odabira hrvatskog jezika. Potrebno je kreirati korisnika, no postoji i mogućnost rada u gost-režimu (bez registracije). Na taj je način moguće obraditi do 15 dokumenata po satu veličine do 4 MB. Ali tada su mogućnosti programa smanjene, npr. ima manje izlaznih formata za gosta (DOC, XLS, TXT), a time su izlazni rezultati nečitljiviji. Za korisnika nema ograničenja veličine kao za gosta. Konačni je rezultat za korisnika u PDF formatu dobar, ali najveća zamjerka programu jest ograničenje od samo 25 dokumenata koje je moguće besplatno obraditi.

4.2.4. ABBYY FineReader

Moramo napomenuti kako taj program nije besplatan. No kao što smo već naveli, u Muzeju trenutačno imamo dva skenera, Epson V750 PRO i Epson GT 20000. Pridružen im je softver u koji je uključen i standardni ABBYY FineReader OCR softver (za prepoznavanje teksta) koji omogućuje praktično uređivanje skeniranog teksta u jednom potezu. Ima mogućnost odabira jezika, uključujući i hrvatski, i izlaznog formata (RTF, DOC, XLS, PDF) te je pogodan za građu koja nije unaprijed skenirana. Nema ograničenje u broju skeniranih dokumenata, a besplatan je utoliko što se isporučuje zajedno sa skenerima koje smo nabavili radi digitalizacije građe. Program je odabran kao alat za digitalizaciju hemeroteke HPM-a jer je imao najbolji konačni rezultat, iako mu je nedostatak bio to što ne čita PDF format, a to je nama u konkretnom slučaju također bilo važno.

4.3. Korektura obrađenog teksta

Brzina OCR procesa impresivna je - jedna se stranica može obraditi 100 puta brže od profesionalnog daktilografa, uz manje pogrešaka. Nijedan OCR program ipak nema stopostotnu učinkovitost. Razne studije govore da učinkovitost optičkog prepoznavanja teksta varira između 95 i 99%, što nije dovoljno za apsolutno vjernu kopiju nekog dokumenta. Stoga nakon optičkog prepoznavanja ručnu korekturu mora obaviti čovjek. Opseg korekture ovisi i o formatu u koji se sprema rezultat procesa.

4.4. Spremanje u odabrani računalni format

Ako su pravilno izrađeni, formati poput PDF-a vizualno izgledaju poput originalnog dokumenta. PDF ima mogućnost da pri pretraživanju označi željenu riječ ili rečenicu u svim dijelovima. Isprobavajući različite formate, odlučili smo se upravo za PDF jer je rezultat bio učinkovitiji nego u Wordu. Potrebnih korektura u Wordu je znatno više nego u PDF-u, gdje ih je gotovo zanemariv broj.

3 http://www.newocr.com

4 http://www.i2ocr.com

5 http://www.gla.ac.uk/services/archives/guardian/about/

MUZEJSKO-EDUKATIVNA NAGRADNA IGRA '2000'

Obidite deset muzeja!

I ove godine u povodu Svjetskog dana muzeja (18. svibnja) organizira se obrazovna nagradna igra za školsku mladež nazvana jednostavno »2000«. Muzejski pedagozi u 24 zagrebačka muzeja i muzeja u Velikoj Gorici i Krapinsko-zagorskoj županiji priredili su niz programa i otvorili svoja vrata radi poticanja djece na individualne i obiteljske posjete muzejima. Muzeji predstavljaju raznovrsnost svoje građe i omogućuju sudjelovanje i kreiranju muzeja za novo tisućljeće. Naglasak je na izgradnji i djelovanju muzeja na promicanju harmonije i mira u društvu, prema preporuci Međunarodnoga savjeta za muzeje (ICOM-a).

Prema riječima Željke Jelavić, muzejske pedagoginje iz Etnografskoga muzeja koji je koordinator ovogodišnjega projekta, učenici trebaju obići deset muzeja (djeca s posebnim potrebama pet) i u Muzejsku knjižicu prikupiti deset pečata da bi ostvarili pravo na nagradne na završnoj priredbi 18. svibnja. Ujedno, knjižica kupljena za 10 kuna služi kao ulaznica za sve muzeje, a u svakom muzeju treba zamoliti kustosa ili ovlaštenu osobu da pečatom ovjere stranicu posjećenog muzeja. Maskota igre »2000« je mali svemirac, a organizaciju potpisuje Sekcija za muzejsku pedagogiju i kulturnu akciju Hrvatskoga muzejskoga društva. (dd)

Željka Jelavić, muzejska pedagoginja Etnografskoga muzeja

Snimio T. BANIĆ

NAGRADNA IGRA

deset

2000

Željka Jelavić

sl. 4. Primjer rezultata istog članka u PDF formatu i Wordu

Usto, ostaje sačuvan izvorni oblik dokumenta, što je važan faktor, posebno kada je riječ o povijesnim izvorima (dokumentima ili tiskovinama). Naime, već i sam izgled nekog dokumenta ili tiskovine može reći nešto o vremenu kada je nastao.

5. Izazovi na koje smo naišli

Jedan od prvih izazova tijekom projekta bio je dio fonda hemeroteke koji je otprije digitaliziran. Naime, program koji se pokazao najučinkovitijim alatom nije mogao pročitati te jedinice jer su spremljene kao PDF datoteke. Doduše, postoje i u fizičkome, tiskanom obliku, pa je postojala mogućnost ponovne digitalizacije i spremanja u obliku slikovnih datoteka. Istodobno, veći dio novije hemeroteke spremljen je upravo kao PDF i ne postoji u drugom obliku. Riječ je o člancima pronađenima na različitim internetskim stranicama, odakle su izravno "printani" u PDF radi uštede prostora i novca te ekoloških razloga. Takvo stanje nametnulo je potrebu rješavanja tog izazova na drugačiji način. Postojeće PDF datoteke pretvorili smo uz pomoć skripte *Image Processor* u Photoshopu (koji također imamo u Muzeju radi obrade fotografija) u slikovne datoteke. To je način kojim možemo brzo i kvalitetno mijenjati različite parametre fotografija - dimenzije, format spremanja i kvalitetu obrade. Nakon obrade u Photoshopu dobili smo datoteke koje je bilo moguće pročitati programom ABBYY FineReader i provesti OCR.

U procesu OCR-a primijetili smo da datoteke koje su skenirane na minimalno 300 dpi prolaze bez teškoća, no za one koje su skenirane na nižoj rezoluciji (npr. 150-200 dpi) program upozorava da je riječ o nezadovoljavajućoj rezoluciji. Ipak, i te su datoteke nakon nešto više korektura dale dobar rezultat. Pokazalo se i to da program teže prepoznaje dokumente s tamnijom podlogom te ih čak okrene naopako.

Postavilo se pitanje na koji način maksimalno iskoristiti rezultate našeg projekta. Kako smo već na početku napomenuli, hemeroteka HPM-a obrađena je u računalnoj aplikaciji za obradu sekundarne dokumentacije. Tako smo dosadašnje rezultate projekta kopirali u polje "sadržaj" unutar aplikacije, koje je pretraživo po cijelom sadržaju. Time je postignuta mogućnost da se hemeroteka sadržajno pretražuje te je znatno unaprijeđena njezina dostupnost i iskoristivost. Pokazalo se da u navedenu aplikaciju u PDF formatu nije moguće implementirati podatke i dobiti pretraživa polja. Stoga je nužno nakon procesa optičkog prepoznavanja dokument otvoriti u Wordu, obaviti korekturu i potom ga kopirati u aplikaciju. Ti Word dokumenti samo su radni materijal i ne čuvaju se trajno. Pokušali smo PDF pretvoriti u Word, no pokazalo se da je količina pogrešaka u tekstu znatno manja ako se nakon OCR-a dokument pohranjuje izravno u Word.

6. Zaključak. Cilj je ovog projekta digitalizacijom hemeroteke Hrvatskoga povijesnog muzeja omogućiti kvalitetniji pristup njezinu sadržaju. Nakon završetka ovog projekta, sljedeća bi faza trebala biti primjena optičkog prepoznavanja znakova u povijesnim dokumentima. Iako je projekt digitalizacije hemeroteke HPM-a još u tijeku, pokušali smo prikazati koji se problemi pojavljuju pri primjeni optičkog prepoznavanja znakova u muzejskome dokumentacijskom fondu. Realizacija projekta zamišljena je i provodi se bez dodatnih financijskih sredstava. Možemo napomenuti kako je pri nabavi opreme za skeniranje potrebno obratiti pozornost i na alate koji se isporučuju u paketu s opremom. Konačan izbor programa za optičko prepoznavanje znakova ovisio je o njegovu izlaznom rezultatu, kao i o rješavanju naknadnog OCR-a već digitaliziranih dokumenata. Preporučena rezolucija za OCR iznosi minimalno 300 dpi, no mi smo se odlučili za 600 dpi kao za optimalno rješenje, iako smo uspješno proveli OCR i s dokumentima digitaliziranim na 150 dpi. Na kraju, rezultate implementiramo u aplikaciju za obradu sekundarne dokumentacije unutar koje su sadržajno pretraživi i bolje iskoristivi.

Primljeno: 7. srpnja 2014.

THE PROJECT FOR THE DIGITISATION OF THE PRESS CLIPPINGS LIBRARY OF THE CROATIAN HISTORY MUSEUM

The Croatian History Museum has collected materials for its press clippings library for more than 60 years; today it numbers about 2000 items, now available though a museum documentation processing application. Items that relate to events during the Homeland War are one of the special features of the holdings, especially those related to the destruction of cultural properties during that war. In work lasting numbers of years it has been shown that the museum press clippings library is a rich source of information for various themes, which gave rise to the idea that the use of it should be extended.

Bearing in mind the sensitivity of the material of paper and its restricted accessibility, a process for digitising the whole of the press clippings archives of the museum was started. One of the basic objectives of the project was to make the collection more accessible to a wider group of users, not only for reading, but also for searching, within the text. The paper shows how tools for digitisation and OCR are chosen as well as implementation in an application for the computer processing of secondary museum documentation. Listed are the challenges that appeared during the project, and recommendations that stemmed from them.