NESTANDARDIZIRANI PODACI U STANDARDIZIRANOM KONTEKSTU: IZRADA IZVJEŠTAJA POMOĆU PROGRAMSKOG JEZIKA AWK U ZBIRCI RUKOPISA I STARIH KNJIGA NACIONALNE I SVEUČILIŠNE KNJIŽNICE U ZAGREBU

NON-STANDARDIZED DATA IN THE STANDARDIZED CONTEXT: CREATING REPORTS USING THE AWK PROGRAMMING LANGUAGE IN THE MANUSCRIPTS AND OLD BOOKS COLLECTION AT THE NATIONAL AND UNIVERSITY LIBRARY IN ZAGREB

Ivan Kapec
Nacionalna i sveučilišna knjižnica u Zagrebu
ikapec@nsk.hr

Krunoslav Peter
Nastavni zavod za javno zdravstvo "Dr. Andrija Štampar"
kruno.peter@stampar.hr

UDK / UDC [025.17:09(497.5 Zagreb)]: 004.43
Stručni rad / Professional paper
Primljeno / Received: 30. 3. 2018.
Prihvaćeno / Accepted: 21.6.2018.

Sažetak

Cilj. Ovaj rad izlaže postojeći proces obrade u Zbirci rukopisa i starih knjiga u Nacionalnoj i sveučilišnoj knjižnici (NSK) u Zagrebu koji se odvija od rujna 2016. s ciljem da se takvi nestandardizirani podaci prevedu u neki od standardnih formata mrežnog okruženja te da se na temelju njih generiraju specifični izvještaji.

Vjesnik bibliotekara Hrvatske 61, 1(2018), 593-610
ISSN 0507-1925

Pristup. Upotrebom programskog jezika Awk prikazat će se mogućnosti prevođenja nestandardiziranih strukturiranih podataka u standardizirani format.

Rezultati. Zbirke građe posebne vrste uključuju građu koja u pojedinim slučajevima zbog svoje specifičnosti zahtijeva izradu posebnih opisa izvan knjižničnog standarda. Rukopisne ostavštine i sitni tisak u knjižničnom okruženju prema standardu obrađuju se na razini zbirke; pritom je detaljni opis dokumenata na razini komada, tj. najmanje moguće jedinice građe, često nedostupan korisniku koji pregledava središnji knjižnični *online* katalog. Detaljni opisi takvih dokumenata izrađeni su i sačuvani u tiskanim katalozima, posebnim repozitorijima u formi popisa na radnim listovima proračunskih tablica, tekstnim datotekama ili drugim formatima. Izvještaji kreirani Awkovim skriptama ubrzali su i unaprijedili proces obrade političkih letaka i ostavštine uredništva časopisa „Nova Hrvatska".

Vrijednost. Rezultati i opisani postupak moći će poslužiti kao korisno rješenje svim sličnim manjim organizacijskim jedinicama s manjim i specifičnim knjižničarskim bazama podataka.

Ključne riječi: nestandardizirani podaci, programski jezik Awk, rukopis, sintaktička interoperabilnost, zbirka građe posebne vrste

Abstract

Purpose. The article presents processing in the Manuscripts and Old Books Collections at the National and University Library in Zagreb. The process has been active since September 2016. Its aim is to translate non-standardized data into the standardized format and create various reports.

Approach. The authors will present the various possibilities of translating non-standardized data into the standardized format by using the programming language Awk.

Findings. Diverse private collections, as well as special collections, include materials which in some cases require specific detailed descriptions in non-standardized format due to their specificity. There are, for example, manuscripts, legacies, ephemeral printed materials and archive material in the collection, but the library catalogue stores only a collection record. The user of a library catalogue cannot access the description of an individual document of the collection. A separate repository can store the descriptions of the collection's documents. It can be in the form of a textual document, spreadsheet or a text file, etc. The reports created with Awk's scripts quickened and advanced processing of political pamphlets and legacies of the editorial of the "New Croatia" journal.

Value. The results and the described process can be useful for smaller organizational units with smaller and specific library databases.

Keywords: manuscript, special collection, syntax interoperability, the Awk programming language, unstandardized data

1. Uvod

Različite privatne zbirke, kao i zbirke građe posebne vrste, uključuju građu koja u nekim slučajevima zbog svoje specifičnosti zahtijeva izradu detaljnih opisa izvan standarda koji je trenutno na snazi u NSK-u. Primjerice rukopisi iz ostavština, pisma, stari sitni tisak ili skice članaka predstavljeni su zbirnim zapisom zbirke u središnjem katalogu knjižnice. Pritom je detaljni opis svakoga pojedinog dokumenta obično nedostupan korisniku koji pregledava središnji katalog. Detaljni opisi takvih dokumenata pohranjeni su u posebnim repozitorijima u formi popisa na radnim listovima proračunskih tablica, u tekstnim datotekama ili drugim formatima. Ovaj rad predstavlja dosadašnje rezultate i iskustva u obradi stare građe u Zbirci rukopisa i starih knjiga Nacionalne i sveučilišne knjižnice prilikom prevođenja takvih nestandardiziranih podataka u neki od standardiziranih formata mrežnog okruženja i generiranja raznih izvještaja na temelju spomenutih podataka u uvjetima oskudnih računalnih i ljudskih resursa te uz poštivanje mjera sigurnosti podataka. Standardizacija podataka temeljni je preduvjet svake semantičke interoperabilnosti koja omogućuje povezivanje i dijeljenje podataka u globalno umreženom kontekstu. Upotrebom programskog jezika Awk prikazat će se rezultati prevođenja nestandardiziranih strukturiranih podataka u standardizirani format i generiranja tekstnih i formatiranih izvještaja na temelju spomenutih podataka. Članak je posvećen konkretnom slučaju primjene (engl. *case study*) jednoga standardnog skriptnog jezika za rješenje problema u postupanju s knjižničnom građom (obrada političkih letaka i ostavštine uredništva časopisa „Nova Hrvatska“).

2. Građa posebne vrste

U građu posebne vrste uvrštava se građa koja je smještena u specijalne zbirke kao što su zbirka rukopisa i starih knjiga, grafička zbirka, kartografska zbirka i zbirka muzikalija, tj. građa koja većim dijelom uključuje rijetke primjerke ili unikate. Takva građa ima svoje specifičnosti i zbog toga ju je teško standardizirati onako kako je to moguće u slučaju nove građe. Proces obrade fokusiran je na neknjižnu tekstnu građu. Ona uključuje pisma, neobjavljene članke, letke, skice ili koncepte raznih radova, dnevnike, novinske izvještaje i sl. Sva ta građa pojavljuje se u obliku rukopisa, strojopisa, fotokopija, faksova, tiskopisa i, u novije vrijeme, računalnih ispisa, tj. *isprinta*. Takva vrsta građe opisiva je arhivski ili bibliografski. Arhivski opis polazi od općega, pa različitim razinama do pojedinačnog. Opisuje se prvo cjelina (fond), a potom se, po potrebi, stvaraju različite razine (podfond, serija itd.) koje su hijerarhijski povezane.[1] Nasuprot tomu, bibliografski opis određen je me-

[1] Hurem, M.; J. Kolanović; S. Zgorelec. ISAD(G): opća međunarodna norma za opis arhivskoga gradiva, 2. izdanje. Zagreb: Hrvatski državni arhiv, 2001.

đunarodnom bibliografskom normom – ISBD[2] te polazi od pojedine jedinice građe; može se odnositi i na dio i na cjelinu. Veze između tih dijelova mogu se uspostaviti naknadno.

3. Obrada ostavština i rukopisne građe u NSK-u

U Nacionalnoj i sveučilišnoj knjižnici (NSK) u Zagrebu rukopisi i ostavštine obrađivani su prema pruskim i njemačkim kataložnim i knjižničnim tradicijama.[3] To znači da se takva vrsta građe opisuje na skupnoj razini (razini zbirke) koja se razlikuje od opisa na razini komada. U tiskanom Katalogu[4] br. 4 rukopisa Nacionalne i sveučilišne knjižnice čiji je prvi svezak objavljen 1991. godine, a posljednji, šesti svezak, 2000. godine, opisani su rukopisi i arhivi pojedinih ostavština jednim dijelom kao zbirni zapisi; drugim su dijelom pojedini rukopisi opisani kao zasebni i njima je dodijeljena posebna signatura, tj. izvučeni su iz cjeline pojedinih ostavština. U tom katalogu navode se naslov, podatak o odgovornosti, vrijeme nastanka rukopisa, materijalni opis i napomene. Na kataložnim listićima nisu uvijek navedeni svi takvi podaci, a u vezi s rukopisnim ostavštinama novijeg datuma na listiću je navedena samo signatura.

U strojno čitljivom *online* katalogu, koji je u formatu MARC21, kao zbirke je uneseno samo šest zapisa i cjelokupna rukopisna ostavština Miroslava Krleže. Ostali rukopisi iz ostavština uneseni su kao zasebne jedinice građe. U tim zapisima navedeni su vrsta građe, autor, vrijeme nastanka zbirke, materijalni opis, napomena o organizaciji i rasporedu, biografski i povijesni podaci i sažetak. Međutim u radnim knjigama proračunske tablice Microsoft Excel, nestandardizirano, obrađeno je 18 novih signatura, tj. 18 rukopisnih ostavština, a povrh toga je obrađena zbirka političkih letaka. U tim Excelovim radnim knjigama navedeni su podaci za autora, naslov, mjesto i godinu, materijalni opis i napomena. Prema posljednjim smjernicama za obradu rukopisne građe, takva građa trebala bi se obrađivati u zbirnim zapisima[5], a u takvim zapisima nedostaju podaci za pojedine jedinice građe ili su oni nepotpuni. Isto se odnosi i na letke koji prema uputama za katalogizaciju u NSK-u spadaju u sitni tisak i katalogiziraju se u obliku zbirnog zapisa. Zbog toga se u središnji ka-

---

[2] Barbarić, A. ISBD: međunarodni standardni bibliografski opis: objedinjeno izdanje. Zagreb : Hrvatsko knjižničarsko društvo, 2014.

[3] Galić Bešker, I. Arhiv, fond i zbirka: rukopisne ostavštine Nacionalne i sveučilišne knjižnice u Zagrebu. // Rukopisne ostavštine kao dio hrvatske baštine: zbornik radova: znanstveno stručni skup, Zagreb, 9. listopada 2014. / uredile Melina Lučić, Marina Škalić. Zagreb : Hrvatski državni arhiv, 2015. Str. 39–62.

[4] Katalog rukopisa Nacionalne i sveučilišne biblioteke u Zagrebu. Zagreb: Nacionalna i sveučilišna biblioteka, 1991.

[5] Buzina, T; D. Salaj Pušić. Sitni tisak: upute za katalogizaciju u bibliografskom formatu MARC 21. Zagreb : Nacionalna i sveučilišna knjižnica, 2012. Str. 7. [citirano: 2018-03-28]. Dostupno na: http://www.nsk.hr/sitni-tisak-upute-za-katalogizaciju-u-bibliografskom-formatu-marc-21/.

talog unose samo zbirni zapisi u kojima se nalaze podaci o broju jedinica i načinu njihova smještaja u određeni broj mapa. U takvom zbirnom zapisu izostaju podaci o pojedinim jedinicama građe i korisniku je prepušteno da pregledava cjelokupnu ostavštinu kako bi mogao pronaći pojedini dokument koji ga zanima. Od takva načina obrade jedini je izuzetak rukopisna ostavština Miroslava Krleže koja je u cijelosti unesena u središnji *online* katalog i gdje je vidljiv opis svake pojedine jedinice.

Međutim pojedinačne jedinice takvih ostavština popisane su samo u tablicama koje su ispisane te su dostupne samo korisnicima koji dolaze u radni prostor Zbirke rukopisa i starih knjiga; jedino tamo oni mogu dobiti detaljan uvid u opis pojedine jedinice.

4. Proces obrade je i proces organizacije

U procesu obrade ovakve vrste građe, koja nije standardizirana, postupno se nameće način njezine organizacije, tj. kako ju je najbolje (najpraktičnije) organizirati i obraditi. Da bi konsolidiranje načina obrade bilo lakše i preciznije, potrebno je informacije unositi u fleksibilni sustav koji se tijekom procesa obrade može modelirati prema promjenjivim zahtjevima, a koji nastaju zbog raznovrsnosti oblika i sadržaja građe. U obradi letaka određenoga razdoblja, pregledom njihove veće količine uočava se tip informacija koji je relevantan za letke, npr. podatak o potpisniku letka, adresatu, tiskaru ili političkoj grupaciji koja je dala izraditi takav letak. U procesu obrade informacija uočava se da su neke važnije, a neke manje važne. Procjena podataka za objavu i prikaz korisnicima odvija se prema potrebi. Zbog toga je potreban fleksibilan i interoperabilan sustav za manipuliranje podacima i stvaranje izvještaja na temelju njih. Da bi se takvi podaci mogli učitati u repozitorij knjižnice ili repozitorij virtualne izložbe te da bi se integrirali u neki veći sustav, potrebno ih je izlučiti iz posebnih repozitorija (proračunskih tablica ili tekstnih datoteka).

5. Izrada izvještaja pomoću programskog jezika Awk u Zbirci rukopisa i starih knjiga Nacionalne i sveučilišne knjižnice u Zagrebu

Krajem 2016. godine započela je obrada zbirke Političkih letaka u NSK-u. Zbirka broji 2201 letak raznih političkih grupacija u razdoblju od 1847. do 1988. godine. Kako je već pisano u posljednjim službenim uputama za obradu rukopisne građe, leci se svrstavaju u sitni tisak i obrađuju se u zbirnom zapisu. To znači da će podaci o svakom pojedinom letku, tj. točan materijalni opis, naslov letka, potpis itd. ostati nevidljivi ili nepotpuni onomu koji pretražuje *online* katalog NSK-a. Dugogodišnja praksa u Zbirci rukopisa i rijetkih knjiga jest da se građa iz ostavština obrađuje prema arhivskim načelima i zbog toga se svi podaci ne unose u glavni mrežni katalog, nego u Excelove radne knjige koje su pohranjene na memorijskim medijima računala djelatnika zbirke. Da bi se ti podaci iz popisa u Excelovim radnim knjiga-

ma mogli najbolje prezentirati, realizirana je tijekom 2017. godine ideja o uporabi programskoga jezika Awk s ciljem lakog i brzog generiranja raznih vrsta korisnih izvještaja. U slučaju letaka, pomoću Awka mogu se generirati ispisi u obliku kataložnih listića ili skraćeni zapisi koji bi se upotrijebili u izgradnji digitalne zbirke. Tako je krajem 2017. godine, nakon što su podaci za svaki pojedini letak bili upisani u Excelove radne knjige, kreiran izvještaj u obliku kataložnih listića za cijelu zbirku političkih letaka. Taj izvještaj pohranjen je kasnije u PDF-format te je dostupan korisnicima koji dolaze u Zbirku za pretraživanje i istraživanje.

Zbog specifičnosti građe i njezina korištenja bilo je važno ostvariti fleksibilan način manipuliranja njezinim metapodacima. Npr. ako se pojavi potreba da se radi nekog određenog istraživanja izluče samo podaci o tiskarima letaka u jedan popis ili samo oni leci koji imaju točan potpis ili bilo koji drugi parametar, tada programski jezik Awk omogućava brz i jednostavan način formiranja takvih vrsta ispisa podataka.

Nakon obrade letaka pristupilo se korištenju programskog jezika Awk za generiranje izvještaja ostavštine uredništva časopisa „Nova Hrvatska“. Struktura njihove građe ponešto je drukčija nego ona političkih letaka, pa se stoga i njihov opis razlikuje od opisa letaka. Radi se uglavnom o člancima iz novina koji su u fazi uredničke korekture, pa zbog toga dolaze u obliku rukopisa, strojopisa, strojopisa s rukopisnim bilješkama, tiskopisa, fotokopija itd. Zbog toga je izvještaj za tu građu trebao biti promijenjen; osim navođenja naslova, podataka o odgovornosti, vremena i mjesta nastanka, trebalo je dodati broj i godinu tiskanog časopisa na koju se odnosi ta građa, a potom i podatke o izvorniku, jer je velik broj članaka zapravo prerada ili prijevod već objavljenih članaka iz nekog drugog tiska. Opet je programskim jezikom Awk napisana skripta koja je ispunila te zahtjeve da bi se generirao ispis prema izmijenjenim parametrima. Upravo zbog te jednostavnosti i preglednosti izabran je programski jezik Awk koji se potvrdio kao uporabljiv alat za lako i brzo manipuliranje podacima (koji su prethodno pohranjeni u CSV-format) za prezentaciju u potrebnim oblikovanjima. Obrada ostavštine uredništva „Nove Hrvatske“ još je u tijeku, a od rujna 2017. godine, kada je počela obrada, do ožujka 2018. godine obrađeni su članci i generiran je izvještaj u obliku kataložnog listića za više od 50 brojeva „Nove Hrvatske“, što ukupno broji više od 2000 članaka.

Osim u svrhu generiranja izvještaja, plan je da se programski jezik Awk iskoristi za pripremu metapodataka koji se unose u digitalnu zbirku i izradu izvještaja u formi *web*-stranice.

U slučaju da se u skorijoj budućnosti na višoj razini donese odluka da se svaki pojedini letak, pismo ili članak mora unijeti u *online* katalog NSK, kao što je to bio slučaj s ostavštinom Miroslava Krleže, sva obrađena građa čiji su podaci pohranjeni u Excelove radne knjige mogla bi se pomoću programskoga jezika Awk preoblikovati u format MARCXML i time brže učitati u mrežni katalog.

6. Primjena Awka s ciljem postizanja interoperabilnosti knjižničnih podataka

Popis knjižnične građe moguće je ostvariti na radnom listu proračunske tablice sa svojim zaglavljem (metapodacima) i podatkovnim recima u kojima svaki redak odgovara jednom dokumentu knjižnične građe te uključuje ćelije (polja) u skladu sa zaglavljem. Nad takvim popisom izvediva je validacija podataka (engl. *data validation*) tijekom upisa radi njihove konzistentnosti. Na zaglavlju popisa može se postaviti podatkovni filtar radi izdvajanja redaka popisa prema zadanom kriteriju.

Na temelju popisa knjižnične građe neki korisnik proračunske tablice može izraditi jednostavne izvještaje koji će iskazivati sve ili filtrirane retke popisa s njegovim zaglavljem. Jedan od problema u kreiranju izvještaja jest kako podatke jednoga retka toga popisa prikazati u nekoliko redaka izvještaja, a kakvi mogu biti potrebni za tiskane kataloge ili kataložne listiće. Drugi problem jest generiranje izvještaja u HTML-formatu ili XML-formatu. Ovaj članak ima polazište u korisnikovu stavu da on može generirati izvještaj o knjižničnoj građi na temelju njezina popisa prema svojim potrebama te donosi prijedlog primjene programskoga jezika Awk za kreiranje takvih izvještaja na temelju popisa knjižnične građe na radnim listovima proračunske tablice.

Programski jezik Awk standardni je programski jezik operacijskoga sustava Unix, a njegov interpretator dostupan je i za druge operacijske sustave poput Microsoftova sustava Windows i Appleova MacOS-a.[6] Namijenjen je pretraživanju i obradi tekstnih datoteka.[7] Izabran je kao alat za rješenje spomenutih problema u izvještavanju o knjižničnoj građi zbog jednostavnosti svoje instalacije i primjene, a ujedno je besplatan za primjenu. Programska rješenja uporabom Awka ostvariva su uz minimum izvornoga koda.

Awkov program traži zadane uzorke u recima (ili zapisima unutar) tekstnih datoteka.[8] Ako je u retku (zapisu) pronađen traženi uzorak, tada program nad njim izvodi obradu. Otvaranje i zatvaranje tekstne datoteke te čitanje njezinih zapisa automatizirano je. Tako Awkovi programi uključuju jedan ili više parova uzoraka i obrada u formatu[9]:

uzorak { obrada }
uzorak { obrada }
...

[6] Robbins, A. Effective Awk programming: universal text processing and pattern matching, 4th ed. Sebastopol : O'Reilly Media, 2014. Str. XX.

[7] Aho, A.; B. Kernighan; P. Weinberger. The Awk programming language. Reading: Addison-Wesley Publishing Company, 1988., str. III.

[8] Robbins, A. Nav. dj., str. 3.

[9] Aho, A.; B. Kernighan; P. Weinberger. Nav. dj., str. 2.

Awkovi programi zadaju se interpretatoru u okruženju ljuske (engl. *shell*) ili retka za upis naredbi[10] (engl. *command line*), a koji korisniku omogućuju da operacijskom sustavu zadaje tekstne naredbe. Tako se Awkovu interpretatoru kraći program zadaje u samom retku za upis naredbi. Program s više redaka izvornoga koda može se zapisati tekstnim *editorom* u tekstnu datoteku te potom zadati Awku na izvršenje.[11]

Awk je podrobno dokumentiran u knjizi autora programskoga jezika[12], a njegov je naziv kratica njihovih prezimena: Aho, Weinberger i Kernighan. Tako će u radu biti sažeto pojašnjeni ključni dijelovi izvornoga koda programskih primjera.

Da bi se podaci popisa na radnom listu proračunske tablice mogli podvrgnuti obradi uz pomoć Awka, prethodno je potrebno pohraniti sadržaj radnoga lista u tekstnu datoteku, i to u CSV-format, što je uobičajena mogućnost aplikacija za proračunske tablice.[13]

Umjesto programskoga jezika Awk, za obradu knjižničnih podataka i kreiranje izvještaja o njima mogu se upotrijebiti programski jezici Python, Perl, Java, C# i drugi. Programska rješenja implementirana tim jezicima mogu otvarati Excelove datoteke, pa njihove podatke nije potrebno pohraniti u CSV-format. U odnosu na Awk takve su mogućnosti prednost, no potrebno ih je realizirati izvornim kodom za otvaranje i zatvaranje Excelove datoteke te čitanje njezinih podataka. Awk donosi automatizaciju otvaranja i zatvaranja te čitanja redaka tekstne datoteke, što pridonosi minimalizaciji njegovih programskih rješenja.

7. Tekstni format podataka

U prethodnom dijelu članka spomenuto je da se sadržaj radnoga lista proračunske tablice pohranjuje u tekstnu datoteku prema CSV-formatu. CSV je skraćenica engleskog izraza *Comma-separated values* (hrv. *vrijednosti razdvojene zarezom*). Taj podatkovni format podrazumijeva pohranu podatkovnih zapisa (slogova) u zasebne retke tekstne datoteke, a podatkovna polja unutar zapisa razdvojena su zarezom[14], odnosno točkom sa zarezom u slučaju postavaka operacijskoga sustava za hrvatski jezik. Preporuka je da se taj znak ne pojavljuje ni u jednoj ćeliji popisa, a može se nadomjestiti zarezom.

Primjer jednoga popisa knjižnične građe sa zaglavljem i četiri retka na radnom listu proračunske tablice prikazan je slici 1.

---

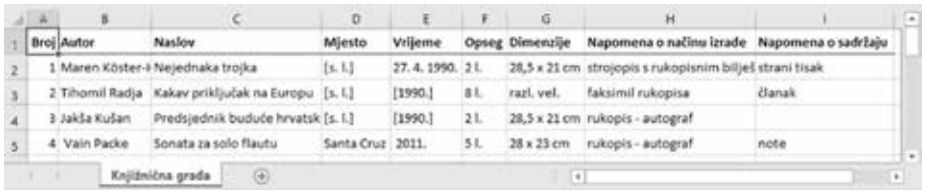
[10] Haas, J. How to write AWK commands and scripts. lifewire, 2018. [citirano: 2018-06-12]. Dostupno na: https://www.lifewire.com/write-awk-commands-and-scripts-2200573.

[11] Robbins, A. Nav. dj., str. 4.

[12] Aho, A.; B. Kernighan; P. Weinberger. Nav. dj.

[13] Robbins, A. Nav. dj., str. 73.

[14] Isto.



	A	B	C	D	E	F	G	H	I
1	Broj	Autor	Naslov	Mjesto	Vrijeme	Opseg	Dimenzije	Napomena o načinu izrade	Napomena o sadržaju
2	1	Maren Köster-H	Nejednaka trojka	[s. l.]	27. 4. 1990.	2 l.	28,5 x 21 cm	strojopis s rukopisnim bilješ	strani tisak
3	2	Tihomil Radja	Kakav priključak na Europu	[s. l.]	[1990.]	8 l.	razl. vel.	faksimil rukopisa	članak
4	3	Jakša Kušan	Predsjednik buduće hrvatsk	[s. l.]	[1990.]	2 l.	28,5 x 21 cm	rukopis - autograf	
5	4	Vain Packe	Sonata za solo flautu	Santa Cruz	2011.	5 l.	28 x 23 cm	rukopis - autograf	note

Knjižnična građa

Slika 1. Popis knjižnične građe sa zaglavljem na radnom listu proračunske tablice

CSV-datoteka u koju je pohranjen sadržaj radnoga lista proračunske tablice ima pet redaka, a neka joj je naziv *gradja.csv*.

Broj;Autor;Naslov;Mjesto;Vrijeme;Opseg;Dimenzije;Napomena o načinu izrade;Napomena o sadržaju

1;Maren Köster-Hetzendorf;Nejednaka trojka;[s. l.];27. 4. 1990.;2 l.;28,5 x 21 cm;strojopis s rukopisnim bilješkama;strani tisak

2;Tihomil Radja;Kakav priključak na Europu;[s. l.];[1990.];8 l.;razl. vel.;faksimil rukopisa;članak

3;Jakša Kušan;Predsjednik buduće hrvatske vlade;[s. l.];[1990.];2 l.;28,5 x 21 cm;rukopis - autograf;

4; Vain Packe ;Sonata za solo flautu;Santa Cruz; 2011.;5 l.;28 x 23 cm;rukopis - autograf;note

Programskim jezikom Awk moguće je čitati, analizirati i obrađivati CSV-datoteke. Evo primjera programa za prikaz sadržaja drugoga polja zapisa u CSV-datoteci zadanog u retku za upis naredbi bilo kojeg od operacijskih sustava Windows (znak '#' u ovom primjeru jest *prompt* – znak spremnosti; on se ne upisuje[15]):

 # awk -F; "{ print $2 }" gradja.csv

Autor

Maren Köster-Hetzendorf

Tihomil Radja

Jakša Kušan

Vain Packe

Prekidačem *-F* Awkove naredbe određuje se znak za razdvajanje polja unutar zapisa.[16] U slučaju CSV-formata znak za razdvajanje polja jest točka sa zarezom, pa se prekidaču *-F* dopisuje taj znak (*-F;*).

[15] Aho, A.; B. Kernighan; P. Weinberger. Nav. dj., str. 2.

[16] Robbins, A. Nav. dj., str. 22.

Ključna riječ *print* prikazuje znakovni niz ili sadržaj nekoga polja (ili zapisa) na standardnom izlazu – zaslonu; u prethodnom primjeru prikazuje se sadržaj drugoga polja. Awk sadržaj učitana retka razdvaja prema zadanom separatoru polja redom, prvo polje u varijablu *$1*, drugo u *$2* itd[17].

Awkov program jest *{ print $2 }*. On prikazuje drugo polje učitanoga zapisa. Ako je u paru *uzorak { obrada }* izostavljen *uzorak*, tada se *{ obrada }* izvodi nad svakim zapisom ulazne datoteke. Ako se pak izostavi *{ obrada }*, tada Awkov program prikazuje svaki zapis koji odgovara zadanom uzorku. Tako sljedeći Awkov program prikazuje sve zapise CSV-datoteke koji u petom polju imaju broj (ili znakovni niz) *2011*:

 # awk -F; "$5 ~ /2011/" gradja.csv

4; Vain Packe ;Sonata za solo flautu;Santa Cruz; 2011.;5 l.;28 x 23 cm;rukopis - autograf;note

8. Generiranje tekstnog izvještaja Awkom

Da bi se generirao izvještaj koji će za svaki redak u CSV-datoteci u izvještaju ispisati tri retka, (1.) redak s rednim brojem, (2.) naslovni redak s podacima i (3.) opisni redak, pripremljena je ova Awkova skripta (izvorni kod 1):

Izvorni kod 1. Awkova skripta za generiranje tekstnoga izvještaja

```
BEGIN  { FS = ";" }

NR > 1 {
   print ""
   print "Br. " $1 "."
   print "    " $3 " / " $2 ", " $4 ", " $5 "; " $6 ", " $7
   napomena = ", " $9
   if ($9 == "") napomena = ""
   print "    " "(" $8 napomena ")"
}
```

Prvi par uzorak-obrada u Awkovoj skripti uključuje uzorak *BEGIN*; njegova se obrada izvodi prije učitavanja bilo kojega zapisa ulazne datoteke.[18] U toj obradi određuje se znak za razdvajanje polja – točka sa zarezom.

Drugi par uzorak-obrada ima uzorak *NR > 1* i svoju obradu. Varijabla *NR* služi za pohranu broja trenutno učitanoga zapisa. Tako uzorak *NR > 1* izdvaja iz CSV-da-

[17] Aho, A.; B. Kernighan; P. Weinberger. Nav. dj., str. 5.

[18] Robbins, A. Nav. dj., str. 4.

toteke sve retke osim početnoga. U njegovoj se obradi naredbama "print" prikazuju sadržaji polja zapisa[19]. Prva naredba – *print ""* – ispisuje prazni redak. Druga niže znakovni niz „Br.", sadržaj prvoga polja CSV-datoteke i točku.

Opisana Awkova skripta ima nekoliko redaka izvornoga koda, pa ju je uputno pohraniti u tekstnu datoteku[20], a datoteku imenovati i pridružiti joj nastavak *.awk*. Awkova skripta u datoteci poziva se na izvršenje naredbom (navodi se u nastavku prekidača *-f*), a rezultat njezina izvršenja preusmjerava se u tekstnu datoteku *izvjestaj.txt* (što je mogućnost koju pruža redak za upis naredbi):

awk -f skripta.awk gradja.csv > izvjestaj.txt

Rezultat izvršenja skripte nad podacima CSV-datoteke *gradja.csv* jest:

Br. 1.
Nejednaka trojka / Maren Köster-Hetzendorf, [s. l.], 27. 4. 1990.; 2 l., 28,5 x 21 cm
(strojopis s rukopisnim bilješkama, strani tisak)

Br. 2.
Kakav priključak na Europu / Tihomil Radja, [s. l.], [1990.]; 8 l., razl. vel.
(faksimil rukopisa, članak)

Br. 3.
Predsjednik buduće hrvatske vlade / Jakša Kušan, [s. l.], [1990.]; 2 l., 28,5 x 21 cm
(rukopis - autograf)

Br. 4.
Sonata za solo flautu / Vain Packe , Santa Cruz, 2011.; 5 l., 28 x 23 cm
(rukopis - autograf, note)

Tekstnu datoteku *izvjestaj.txt* korisnik može otvoriti tekstnim *editorom* ili programom za obradu teksta te dodatno urediti ili ispisati.

U pisanje Awkove skripte za kreiranje tekstnoga izvještaja ili promjenu njezina izvornoga koda mogao bi se uputiti sam bibliotekar, no za sljedeće dvije skripte u ovome radu on bi ipak mogao potražiti potporu programera, odnosno prepustiti mu razvoj programskih rješenja programskim jezikom Awk.

[19] Aho, A.; B. Kernighan; P. Weinberger. Nav. dj., str. 5.

[20] Robbins, A. Nav. dj., str. 4.

9. Generiranje izvještaja u formi *web*-stranice

Ako je u izvještaju o knjižničnoj građi potrebno izgledom naglasiti neke podatke ili ga prikazati određenim fontom, tada je uputno generirati ga u formi *web*-stranice. Polazište za takvo programsko rješenje jest Awkova skripta za generiranje tekstnoga izvještaja, a njezina nadgradnja uključuje generiranje oznaka jezikom *HTML* kojima će biti opisana struktura i sadržaj *web*-stranice te će u nju biti umetnute postavke njezina izgleda iskazima jezika *CSS*.[21]

Evo primjera jedne Awkove skripte za generiranje takva izvještaja (izvorni kod 2) u kojem će slova tamnoplave boje biti prikazana fontom bez ukrasa (engl. *serif*), a redni broj naglašen (engl. *strong*):

Izvorni kod 2. Awkova skripta za generiranje izvještaja u formi *web*-stranice

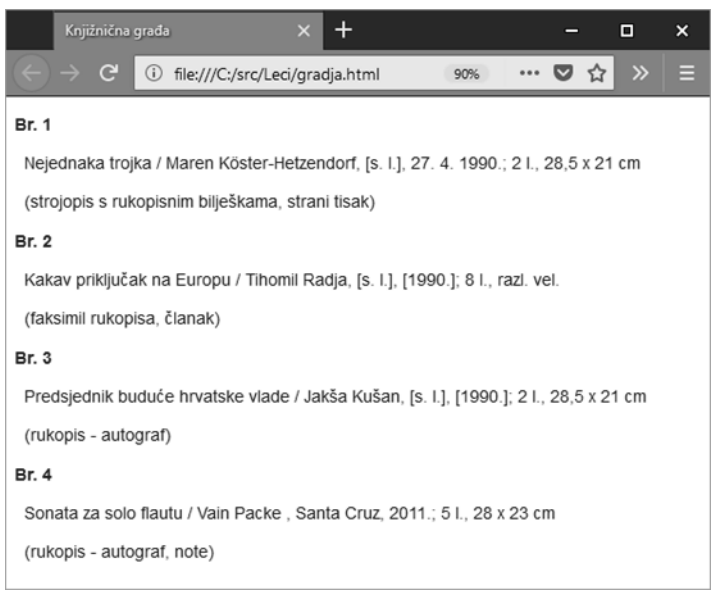
```
BEGIN {
   FS = ";"
   print "<!doctype html>"
   print "<html>"
   print "<head>"
   print "  <title>Knjižnična građa</title>"
   print "  <meta charset=\"utf-8\">"
   print "  <style>html { color: DarkBlue; font-family: sans-serif }</style>"
   print "</head>"
   print "<body>"
}

NR > 1 {
   print "<p>" "<strong>" "Br. " $1 "." "</strong>" "</p>"
   print "<p>" "  " $3 " / " $2 ", " $4 ", " $5 "; " $6 ", " $7 "</p>"
   napomena = ", " $9
   if ($9 == "") napomena = ""
   print "<p>" "  " "(" $8 napomena ")" "</p>"
}

END {
   print "</body>"
   print "</html>"
}
```

[21] West, M. HTML5 Foundations. Chichester: John Wiley & Sons, 2013., str. 8.

Prvi par uzorak-obrada uključuje uzorak *BEGIN* i pripadajuću obradu. Njegovim se izvršenjem u izvornom kodu *web*-stranice generiraju HTML-oznake zaglavlja[22] sa stilom u jeziku CSS. Drugi par uzorak-obrada primjenjuje se nad svim zapisima ulazne datoteke osim na prvom (koji predstavlja zaglavlje i nije potreban u izvještaju). Treći par uzorak-obrada izvodi se nakon što se iščitaju svi zapisi CSV-datoteke te zapisuje završne HTML-oznake[23].

Naredba za pokretanje te Awkove skripte istovjetna je naredbi za pokretanje skripte za tekstni izvještaj, a uz iznimku da se rezultat izvršenja preusmjerava u datoteku s nastavkom *.html*.

Spomenuta skripta generira *web*-stranicu čiji je izgled prikazan u prozoru *web*-preglednika na slici 2.



Slika 2. Izvještaj o knjižničnoj građi u formi *web*-stranice

10. Pohrana podataka iz CSV-formata u XML-format (MARCXML)

Bibliotekar se u pohrani, obradi te pretraživanju i prikazu podataka o knjižničnoj građi posebne vrste svakako može poslužiti proračunskom tablicom i programskim jezikom Awk, a uz moguću potporu programera. Awkom se oni također mogu poslužiti da bi takve podatke pohranili (preveli) u format pogodan za učitavanje u bazu

[22] Isto, str. 17.

[23] Isto, str. 14.

podataka bibliotekarskoga informacijskog sustava. Jedan od standarda za pohranu bibliotekarskih podataka jest *MARC* – kratica engleskoga izraza *MAchine-Readable Cataloging*[24]; u okvirima toga standarda definirana je pohrana podataka u XML-formatu[25] – *MARCXML*.

XML je kratica engleskog izraza *Extensible Markup Language*; to je standardna sintaksa označavanja podataka tekstnim čitljivim oznakama.[26] XML-dokumenti čitljivi su tekstnim *editorom* te se mogu generirati programskim jezikom Awk.

Radi preglednosti programskoga primjera, Awkova skripta za prevođenje podataka iz formata CSV u MARCXML (izvorni kod 3) obuhvaća tri polja zapisa CSV-formata: naslov, opseg i dimenzije. Naslov se pohranjuje u polje MARCXML-a za bibliografske informacije broj *245*. U polje MARCXML-a broj *300* – fizički opis – pohranjuju se opseg i dimenzije, i to u dva potpolja: opseg u potpolje kodne oznake *a*, a dimenzije u potpolje *c*. Za svaki zapis generira se XML-kod unutar oznaka *<record>* i *</record>*. Awkova skripta za prevođenje podataka prema tim načelima iz CSV-a u MARCXML-format jest:

Izvorni kod 3. Awkova skripta za prevođenje podataka iz formata CSV u MARCXML

```
BEGIN {
   FS = ";"
   print "<?xml version=\"1.0\" encoding=\"UTF-8\"?>"
   print "<collection xmlns=\"http://www.loc.gov/MARC21/slim\">"
}

NR > 1 {
   print "  " "<record>"
   print "    " "<datafield tag=\"245\" ind1=\"1\" ind2=\"0\">"
   print "      " "<subfield code=\"a\">"
   print "        " $3
   print "      " "</subfield>"
   print "    " "</datafield>"
   print "    " "<datafield tag=\"300\" ind1=\" \" ind2=\" \">"
   print "      " "<subfield code=\"a\">"
```

[24] MARC Standards. [citirano: 2018-03-28]. Dostupno na: https://www.loc.gov/marc/.

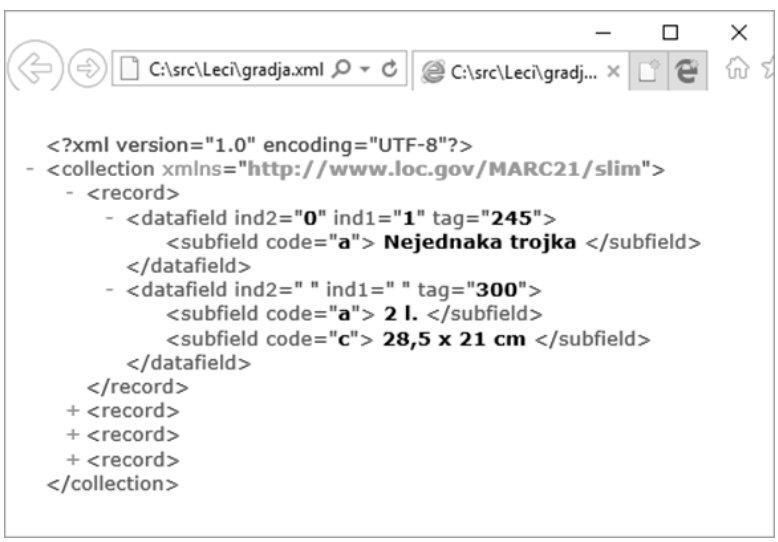
[25] MARCXML – MARC 21 XML schema. [citirano: 2018-03-28]. Dostupno na: http://www.loc.gov/standards/marcxml/.

[26] Harold, E.; S. Means. XML in a nutshell: a desktop quick reference, 3rd ed. Sebastopol: O'Reilly Media, 2004., str. 3.

```
    print “       “ $6
    print “     “ “</subfield>”
    print “     “ “<subfield code=\”c\”>”
    print “       “ $7
    print “     “ “</subfield>”
    print “   “ “</datafield>”
    print “  “ “</record>”
}

END {
    print “</collection>”
}
```

Kao i u slučaju skripte za generiranje izvještaja u formatu *web*-stranice, prvi par uzorak-obrada stvara zaglavlje XML-datoteke, a treći njezinu zaključnu oznaku. Drugi par uzorak-obrada primjenjuje se na svim zapisima ulazne datoteke osim na prvom.

Na slici 3 je *web*-preglednikom prikazan (renderiran) generirani XML-kod u formi grananja s jednim razvijenim zapisom.



Slika 3. XML-format (MARCXML) (prikaz grananja s jednim razvijenim zapisom)

11. Prednosti primjene Awka u obradi knjižničnih podataka u tekstnom formatu

Opisani pristup obradi knjižničnih podataka pohranjenih u CSV-format jedan je od mnogih mogućih. Polazišta su mu korisnikov stav da on može generirati izvještaj o knjižničnoj građi na temelju njezina popisa te ograničenja dostupnih računalnih resursa i vremena. Zato je za implementaciju programskoga rješenja izabran Awk, koji je besplatan i dostupan za razne operacijske sustave te se može lako i brzo instalirati (njegov je interpretator doslovce u jednoj izvršnoj datoteci). Sintaksa Awkovih programa jednostavna je, a programska rješenja ostvariva su uz minimum izvornoga koda. Zato je on pogodan za rapidni razvoj programskih rješenja i prototipova, čime se može prevladati vremensko ograničenje obrade specifične knjižnične građe. Uzevši u obzir kompleksnost programiranja, bibliotekar bi mogao razviti jednostavne Awkove skripte za generiranje tekstnih izvještaja o knjižničnoj građi, no razvoj skripti za generiranje formatiranih izvještaja u formi *web*-stranice ili za prevođenje knjižničnih podataka iz CSV-formata u MARCXML ipak bi bio posao za programera koji poznaje Awk.

12. Primjena drugih programskih jezika i alata s ciljem postizanja interoperabilnosti knjižničnih podataka

Ovaj članak opisuje konkretan slučaj obrade knjižnične građe uporabom proračunske tablice za evidenciju knjižničnih podataka, programskih rješenja implementiranih Awkom za generiranje podatkovnih izvještaja i raznih pomagala operacijskoga sustava (redak za upis naredbi, tekstni *editor* i *web*-preglednik). Za obradu knjižničnih podataka moguće je upotrijebiti i druge programske jezike kao što su Python, Perl, Java, C# i Node.js. U usporedbi sa spomenutim jezicima Awk bi imao prednost u rapidnoj implementaciji programskog rješenja za jednokratnu primjenu ili prototipa, a ostali programski jezici bili bi primjereniji za razvoj aplikacije s korisničkim sučeljem za redovitu primjenu u obradi knjižničnih podataka.

Za generiranje izvještaja na temelju knjižničnih podataka u formi CSV-datoteke ili Excelove datoteke također se mogu upotrijebiti razni generatori izvještaja na računalu. Takvi alati dostupni su i na *webu*, no prije njihove uporabe svakako je potrebno informirati se o sigurnosnoj politici ustanove ili tvrtke u čijemu su vlasništvu knjižnični podaci.

13. Zaključak

Zbirke građe posebne vrste u knjižnicama i sličnim manjim organizacijskim jedinicama koje obrađuju i čuvaju neku specifičnu i često nestandardiziranu građu u nekim slučajevima imaju potrebu za posebnim načinom obrade građe. Takva obra-

da može izlaziti izvan knjižničarskih standarda i normi, no može biti primjerena takvom slučaju. U određenim okolnostima naknadno se pojavljuje potreba za standardizacijom takvih podataka i njihovim prihvatom u baze podataka knjižničnih informacijskih sustava. Primjena programskog jezika Awk radi postizanja sintaktičke operabilnosti jedan je od mogućih koraka u učinkovitom integriranju podataka iz nestandardizirane okoline u standardnu prema postojećim standardima za integraciju podataka iz različitih izvora. Pritom treba naglasiti jednostavnost i fleksibilnost takva pristupa te mnogostrukost njegovih primjena u obradi bibliotekarskih podataka.

LITERATURA

Aho, A.; B. Kernighan; P. Weinberger. The Awk programming language. Reading: Addison-Wesley Publishing Company, 1988.

Barbarić, A. ISBD: međunarodni standardni bibliografski opis: objedinjeno izdanje. Zagreb: Hrvatsko knjižničarsko društvo, 2014.

Buzina, T.; D. Salaj Pušić. Sitni tisak: upute za katalogizaciju u bibliografskom formatu MARC 21. Zagreb: Nacionalna i sveučilišna knjižnica, 2012. [citirano: 2018-03-28]. Dostupno na: http://www.nsk.hr/sitni-tisak-upute-za-katalogizaciju-u-bibliografskom-formatu-marc-21/.

Galić Bešker, I. Arhiv, fond i zbirka: rukopisne ostavštine Nacionalne i sveučilišne knjižnice u Zagrebu. // Rukopisne ostavštine kao dio hrvatske baštine: zbornik radova: znanstveno stručni skup, Zagreb, 9. listopada 2014. / uredile Melina Lučić, Marina Škalić. Zagreb : Hrvatski državni arhiv, 2015. Str. 39–62.

Haas, J. How to write AWK commands and scripts. Lifewire, 2018. [citirano: 2018-06-12]. Dostupno na: https://www.lifewire.com/write-awk-commands-and-scripts-2200573.

Harold, E.; S. Means. XML in a nutshell: a desktop quick reference, 3rd ed. Sebastopol: O'Reilly Media, 2004.

Hurem, M.; J. Kolanović; S. Zgorelec. ISAD(G): opća međunarodna norma za opis arhivskoga gradiva. 2. izd. Zagreb: Hrvatski državni arhiv, 2001.

Katalog rukopisa Nacionalne i sveučilišne biblioteke u Zagrebu. Zagreb: Nacionalna i sveučilišna biblioteka, 1991.

MARC standards. [citirano: 2018-03-28]. Dostupno na: https://www.loc.gov/marc/.

MARCXML – MARC 21 XML schema. [citirano: 2018-03-28]. Dostupno na: http://www.loc.gov/standards/marcxml/.

Robbins, A. Effective Awk programming: universal text processing and pattern matching, 4th ed. Sebastopol: O'Reilly Media, 2014.

West, M. HTML5 foundations. Chichester: John Wiley & Sons, 2013.