

UVOZ PODATAKA IZ PDF-a U EXCEL

Ako ste mislili da je uvoz podataka iz HTML dokumenata bio problematičan, niste bili u pravu. Evo još većeg izazova. Uvoz podataka iz PDF fajla ume da bude komplikovan, ali se trud isplati jer je se sve više izvještaja i sličnih dokumenata sa korisnim podacima na sajtovima vladinih institucija, univerziteta, nevladinih i velikih svjetskih organizacija, objavljuje upravo u ovom formatu.

Na primjeru uvoza podataka iz PDF-a u *Excel* u ovom tekstu nastojimo da objasnimo neke opšte principe sa kojima ćete se sretati. Međutim, očekujte da će rad sa nekim drugim dokumentom donijeti nove izazove, pa ćete gotovo svaki put morati da se dovijate pri uvozu i čišćenju podataka. Ako naiđete na naizgled nerešiv problem, javite nam se i pokušaćemo da ga riješimo zajedno.

Za rad sa PDF fajlovima potreban vam je *Acrobat Reader* (ako nemate kompletan *Adobe Acrobat* paket), koji se besplatno downloaduje sa sajta firme *Adobe* (www.adobe.com).

Scenario prvi:

1. Na lokaciji www.census.gov/ipc/www/wp98.html nađite link na dokument *World Population Profile from the U.S. Census Bureau*. To je dokument sa kojim ćemo raditi.

2. Da biste radili sa PDF fajlom, on NE SMIJE biti otvoren unutar brauzera. Da biste ga spasili kao nezavistan dokument, kliknite desnim klikom na PDF fajl, odaberite opciju *Save Target As* i spasite ga na *Desktopu*.

3. Otvorite *Adobe Acrobat*. Iz menija *File* odaberite *Open* i otvorite željeni fajl.

4. Zanimaju nas podaci na stranama 113-117. (Možete kliknuti na *Thumbnails*, skrolovati se do strane 113 i kliknuti na nju.) Podaci se tiču prosječnog životnog vijeka u raznim zemljama, po regionima.

5. Da biste selektovali dijelove dokumenta, kliknite na *Text Select Tool* (to je dugme sa slovima *abc* ili sa velikim *T*). Selektirajte tekst koji želite da kopirate. (Napomena: da biste kasnije efikasno čistili podatke, često ćete morati da selektujete dijelove PDF fajla, što je slučaj u ovom primjeru. Da biste selektovali samo dio tabele, možete to uraditi držeći tipku *CTRL*.)



Možda je najbolje da uvozimo stranu po stranu podataka. Prvo ćemo selektovati podatke na strani 113. Da nas pri čišćenju podataka zaglavlje ne bi dodatno zbunjivalo,

Table A-10.
Life Expectancy at Birth by Region, Country, and Area

Region and country or area	1990	1995	2000	2005	2010
WORLD	63	67	70	71	72
Less Developed Countries	51	55	58	60	62
More Developed Countries	75	77	79	80	81

sada ga možemo preskočiti. (Pazite da ne selektujete i rečenicu na samom dnu strane jer bi nas i ona kasnije mogla zbuniti).

6. Otvorite *Notepad* (*Start* → *Programs* → *Accessories*). Komandom *Paste* "zalijepite" selektovani tekst. Spasite fajl kao text (*.txt*) fajl (nazovimo ga 113). Otvorite prozor *Notepad-a* na punu veličinu i pogledajte brojeve. Ako pokušamo da ih ovako uvezemo u *Excel* (možete probati),

```

. . . . . 63 70 61 67 65 72
. . . . . 62 69 60 66 63 71
. . . . . 75 79 71 76 79 82
. . . . . 51 58 50 56 52 60
. . . . . 49 57 47 55 50 59
. . . . . 48 63 46 59 50 66
. . . . . 54 65 52 62 56 69
. . . . . 40 49 39 48 41 50
. . . . . 46 54 45 53 47 55
. . . . . 46 54 44 52 47 56
  
```

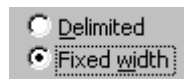
nećemo daleko stići. *Notepad* nam upravo služi da podatke pročistimo. Bitno je da sve brojeve poravnamo, za šta koristimo razmaknicu i tipku *backspace*. Sad se već pitate 'zar moram da prolazim kroz ove muke...! Ima i gorih, a za utjehu pomislite samo kako bi izgledalo "pješačko" ukucavanje svakog pojedinačnog podatka iz ovog PDF-a u *Excel*. Za poravnavanje nam je trebalo oko 3 minuta, što ipak nije strašno.

```

. . . . . 63 70 61 67 65 72
. . . . . 62 69 60 66 63 71
. . . . . 75 79 71 76 79 82
. . . . . 51 58 50 56 52 60
. . . . . 49 57 47 55 50 59
. . . . . 48 63 46 59 50 66
. . . . . 54 65 52 62 56 69
. . . . . 40 49 39 48 41 50
  
```

7. Otvorite *Excel*. Nađite i otvorite sačuvani tekst fajl 113 (kod *Files of Type* odaberite *All*

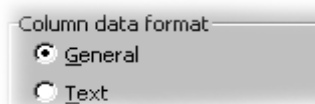
Files ili *Text Files*). Pokrenuće se *Import Wizard* (programčić koji vas korak po korak vodi kroz uvoz podataka). Trebalo bi da vaše podatke prepozna kao *Fixed Width* (fiksirane širine) - ako ste ih dobro



poredali u *Notepad-u*. Pogledajte kako ih prikazuje *Wizard* - da li vertikalne strelice odvajaju brojeve bez greške? Ako ne, vratite se

	60	70	80
63	70	61	67 65 72
62	69	60	66 63 71
75	79	71	76 79 82
51	58	50	56 52 60
49	57	47	55 50 59

u *Notepad* i dovršite čišćenje podataka. Sada kliknite na *Next*. (Ako vam *Import Wizard* sam postavi strelice između svake tačkice, uklonite ih dvoklikom.) Pratite korake kroz koje vas vodi *Wizard*. Za *Column Data Format* odaberite *General* (što znači da će brojeve



razumjeti kao brojeve, a tekst kao tekst). OBAVEZNO sačuvajte fajl kao *Microsoft Excel Workbook* (*Save As Type*).

8. Raširite kolonu A. Hm... Prvu kolonu brojeva nismo dobro sredili u *Wizardu*. Zatvorite ovaj spreadsheet i ponovite korak 8. U drugom prozoru *Wizarda* jednim klikom dodajte vertikalnu strelicu ispred prve kolone

	60	70	80
63	70	61	67 65 72
62	69	60	66 63 71
75	79	71	76 79 82
51	58	50	56 52 60
49	57	47	55 50 59

. 63	70	61	67	65	72
. 62	69	60	66	63	71
. 75	79	71	76	79	82
. 51	58	50	56	52	60
. 49	57	47	55	50	59

brojki - ona će ih odvojiti od teksta s lijeve strane kad podatke uvezemo u *Excel*. Dovršite korake u *Wizardu*.

9. Podaci u *Excelu* izgledaju odlično, samo nam smetaju one tačkice poslije naziva zemalja. Idite na *Edit* → *Replace*, pa u

	A	B
D	63	70
Developed Countries	62	69
Developed Countries	75	79
A	51	58
African Africa	49	57
	48	63
	54	65

gornjem okviru ukucajte jednu tačku, a u donjem ništa. Kliknite na *Replace All*. Zar nije dobro? Sve tačke smo zamijenili praznim prostorom. Opet sačuvajte *Excel* fajl kao *Microsoft Excel Workbook*.

10. Sada na isti način kopirajte podatke sa strane 114 (nemojte selektovati zaglavlje i tekst na dnu strane, a ako to ipak uradite, obrišite ih ili u *Notepad-u* ili u *Excelu*), "zaljepite" ih u *Notepad*, i nakon čišćenja spasite fajl kao *txt* fajl pod imenom 114. Savjet: da vam oznake za polja gdje nema podataka (NA) ne bi kasnije pravile probleme, obrišite zagrade i tako ih poravnajte sa brojkama.

11. Otvorite novi *Excel* dokument, i slijedite korak 8. Obavezno ga spasite kao *Microsoft Excel Workbook*, pod imenom 114.

12. Isto uradite i sa podacima sa strana 115, 116 i 117. Sada imamo pet *Excel* dokumenata koje treba da sastavimo u jedan.

13. Otvorite *Excel* dokument 114 i selektujte sve podatke (kliknite na prvu ćeliju u gornjem lijevom uglu i držeći *Shift* označite sve podatke koristeći strelice na tastaturi). Idite na *Edit* → *Copy*, a zatim otvorite *Excel* dokument 113 i kliknite u prvu praznu ćeliju ispod posljednjeg imena zemlje, pa idite na *Edit* → *Paste*.

14. Uradite isto sa *Excel* dokumentom 115, ali obratite pažnju na naslov *Latin America and the Caribbean* - ako pogledate stranu 115 PDF-

a, vidjećete da on već postoji na strani 114, a da se na strani 115 ponavlja da bi naznačio da se nastavlja sa pominjanjem latinoameričkih

LATIN AMERICA AND THE CARIBBEAN—Con.

Aruba	70
Bahamas, The	69
Barbados	79
Belize	58
Bolivia	57

zemaalja. Preskočićemo ga da nas ne bi kasnije zbunjivalo pri čišćenju podataka, i ostatak podataka selektovati na prethodno opisan način i "zaljepiti" u *Excel* dokument 113.

15. Uradićemo isto i sa *Excel* dokumentom 116, ali pri selektovanju treba zaobići ponovljenu oznaku za Evropu - sve do *Guernsey*. Ako je ipak kopirate, možete je obrisati nakon "ljepljenja" u *Excel* fajl 113, s tim da ćete morati obrisati par redova koji su ostali prazni nakon brisanja. Isto uradite i sa 117.

16. Ostaje da novodobijenoj tabeli dodamo zaglavlje. Možemo ga prekopirati iz PDF dokumenta (pri čemu u *Excel* dokumentu 113 dodamo četiri reda - klik na oznaku za red br. 1, pa *Insert* → *Row*), ali je jednostavnije da iznad kolona sa brojkama ukucamo naizmenično oznake za godinu - 1998 i 2025. Ako vam *Excel* ubaci decimalu u oznaku za godinu, formatirajte samo te ćelije (nikako cijeli spreadsheet) kao *Text* (*Format* → *Cells* → *Text*).

Iznad njih ćemo staviti oznake kolona - prvo *Both sexes (oba pola)* iznad prve kolone brojeva, a zatim označimo tu i ćeliju do nje, pa idemo na *Format* → *Cells* → *Alignment* i kliknemo na *Merge cells*, pa na OK. Ovim smo povezali dvije ćelije u jednu, pa ovaj natpis stoji iznad prve dvije kolone brojeva, kao i u PDF dokumentu. Isto uradimo i sa natpisima *Males* i *Females*. I to bi bilo to. Sada podatke možete da obrađujete kako vam drago.

Scenario drugi:

Kada u PDF dokumentu imate posla sa tabelama (kao što je ovdje slučaj), a imate instaliran *Adobe Acrobat* paket (dakle, ne

samo besplatni *Acrobat Reader*), podatke možete selektovati *Table Text Select Tool-om*

(klik na strelicu pored dugmeta sa slovom T). Izgleda da n a j b o l j e "radi" kada njome selektujemo pojedinačne stranice. Zahvaljujući ovoj moćnoj alatki, kopirane podatke možete zaljepiti DIREKTNO u *Excel* i izgledaće odlično uz vrlo malo čišćenja.

Scenario treći:

Firma *Adobe* nudi vam besplatnu uslugu pretvaranja PDF dokumenta u HTML ili tekst fajl (detaljno na <http://access.adobe.com/online-tools.html>).

Naime, u formular upišete URL (web adresu) PDF dokumenta i za par sekundi *Adobe* vam ga prikazuje kao HTML stranicu. Najbolje je da odete na stranicu sa naprednim formularom na adresi http://www.adobe.com/products/acrobat/access_adv_form.html jer tu možete naznačiti samo dio PDF dokumenta koji želite da pretvorite u HTML tako što navedete prvu i posljednju stranu tog dijela (*First page* i *Last page*). Sa HTML strane podatke kopirate u *Notepad*, pa ih sređujete kako je opisano u *Scenariju dva*.

Adobe vam nudi i da PDF dokument pošaljete e-mailom kao atačment (ako ga imate sačuvanog) ili samo njegov URL (ako je na Internetu), a kao odgovor ćete dobiti njegovu HTML ili text verziju - po izboru (detaljno se informišite na adresi http://www.adobe.com/products/acrobat/access_email.html). Savjet: ako imate *Adobe Acrobat*, a iz velikog dokumenta kao atačment hoćete da pošaljete samo par strana, idite na *Thumbnails*, pa kliknite na desni taster miša i odaberite opciju *Extract Pages*. Ukucajte redne brojeve stranica, kliknite na OK i odmah ćete se naći u dokumentu koji sadrži samo izdvojene stranice. Obavezno ga spasite.

I ne zaboravite da nam se uvijek možete obratiti sa pitanjima.

Nevena RŠUMVIĆ
nevena@netnovinar.org