Popisi prebivalstva Slovenije 1830–1931
Orodje za transkribiranje historičnih demografskih podatkov

Andrej Pančur*

* Inštitut za novejšo zgodovino
Kongresni trg 1, 1000 Ljubljana
andrej.pancur@inz.si

Povzetek

Avtor v prispevku predstavi orodje za transkribiranje historičnih demografskih podatkov. To je bilo razvito v sklopu projekta Popisi prebivalstva Slovenije 1830–1931, ki se odvija na Inštitutu za novejšo zgodovino v okviru raziskovalne infrastrukture Slovenskega zgodovinopisja. Projekt ima dvojen namen, saj želi zadovoljiti tako potrebe ustanov s področja varstva kulturne dediščine po čim lažjemu dostopu širše javnosti (prvenstveno rodoslovcev) do njihovega gradiva, kot tudi potrebe raziskovalcev, ki se ukvarjajo s historično demografijo, po čim širšem naboru relevantnih raziskovalnih podatkov.

Slovenian Population Censuses 1830–1931
Tool for the Transcription of Historical Demographic Information

In his contribution the author presents the tool for the transcription of historical demographic information, developed during the project entitled "Popisi prebivalstva Slovenije 1830-1931" (Slovenian Population Censuses 1830–1931), taking place at the Institute of Contemporary History in the context of the Research Infrastructure of Slovenian Historiography. The purpose of the project is twofold, as it attempts to satisfy the need of the institutions working in the field of cultural heritage protection to ensure that the wider public (especially genealogists) can benefit from easy access to these institutions' materials; as well as the need of the researchers, dealing with historical demographics, to have the widest possible collection of relevant research information at their disposal.

1 Uvod

Uporaba digitalnih orodij je sestavni del vsakršne raziskovalne metode v digitalni humanistiki. Orodja za transkribiranje lahko klasificiramo kot posebno skupino digitalnih orodij (Puhl et al., 2015, 22-23),[1] ki se jih relativno pogosto uporablja v projektih iz digitalne humanistike. Uporaba teh orodij se je zlasti razširila v zadnjih letih, ko se vse več projektov pri prepisovanju (historičnih) podatkov odloča za uporabo spletnih sistemov za izkoriščanje moči množic (crowdsourcing). Digitalna orodja za transkribiranje lahko pri tem razvrstimo v tri večje skupine (Noll, 2013, 10-11):

- vpisovanje v prost obrazec,
- XML ali HTML označevanje,
- vpisovanje v podatkovna polja.

Velika večina projektov pri tem uporablja orodja iz prve skupine, najmanj pa iz druge skupine. Digitalna orodja, pri katerih se podatke vpisuje v različna podatkovna polja se prvenstveno uporablja pri projektih, ki so povezani z rodoslovjem ali z raziskavami s področja demografije. Takšni projekti praviloma najprej poskrbijo za digitalizacijo relevantnih historičnih virov (matične knjige, popisi prebivalstva, davčni registri, domovinske knjige itd.), čemur sledi transkripcija podatkov. Ker so podatki v teh historičnih dokumentih ponavadi pisani z roko, poteka tudi njihovo prepisovanje ročno. Trenutno prepoznavanje rokopisov (Handwritten Text Recognition - HTR) pri historičnih besedil še ni zadosti razvito za množično uporabo, (Fornés et al., 2014) čeprav so se že pojavljali prvi projekti, ki so jo implementirali.[2] V rodoslovnih projektih se ponavadi prepisuje zgolj osnovne podatke, ki so potrebni za indeksacijo. Uporabnik s pomočjo indeksiranih podatkov poišče digitalizirano sliko, s katere si prebere ostale (neprepisane) podatke. V raziskovalnih projektih se ponavadi prepiše vse podatke, ki se jih nato klasificira v skladu z raziskovalnimi potrebami. Pri tem ni nujno, da se ohrani povezava med prepisanimi podatki in digitalizirano sliko.

Projekt Popisi prebivalstva Slovenije 1830–1931, ki se na Inštitutu za novejšo zgodovino (INZ) že pet let odvija v okviru raziskovalne infrastrukture Slovenskega zgodovinopisja, je primer projekta, ki združuje tako elemente rodoslovnih kot raziskovalnih projektov.

2 Popisi prebivalstva Slovenije 1830–1931

Digitalizirani historični popisi prebivalstva Slovenije so javno dostopni na portalu Zgodovina Slovenije – SIstory.[3] Projekt se izvaja v tesnem sodelovanju z Zgodovinskim arhivom Ljubljana (ZAL), ki hrani večjo število popisov prebivalstva, kateri vsebujejo mikropodatke o takratnem prebivalstvu. Tako so v celoti ohranjeni popisi prebivalstva Ljubljane (1830, 1857, 1869, 1880, 1890, 1900, 1910, 1921, 1928 in 1931). Relativno zelo dobro so ohranjeni še popisi za Idrijo, Novo mesto, Škofjo Loko in Vrhniko, delno pa tudi za različne podeželske občine. ZAL se je projektu pridružil, ker želi to gradivo dati preko spleta na voljo svojim uporabnikov, v prvi vrsti rodoslovcem.

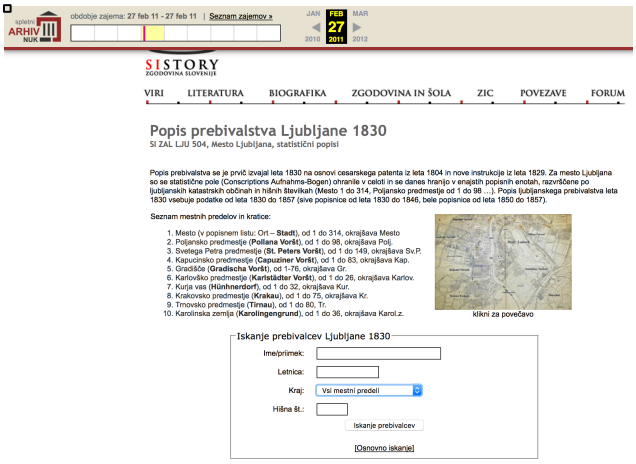
Sprva je ZAL sam digitaliziral popis prebivalstva Ljubljane 1830–1857. Ker je želel, da bi bil popis čim bolj dostopen širši javnosti, se je odločil, da bo popise objavil na portalu SIstory. V ta namen so bili v okviru dejavnosti raziskovalne infrastrukture s pomočjo optičnega prepoznavanja znakov (OCR) iz analognega indeksa (tipkopis iz leta 1934) pridobljeni osnovni podatki o prebivalcih Ljubljane iz popisa 1830–1857. Na podlagi

[1] DIRT: Digital Research Tools, http://dirtdirectory.org/tadirah/transcription.

[2] Transkribus, https://transkribus.eu/Transkribus/.

[3] Popisi prebivalstva Slovenije 1830-1931, Zgodovina Slovenije – Sistory, http://sistory.si/publikacije/?menu=510.

teh podatkov je bil nato izdelan iskalnik po popisih prebivalstva (glej Sliko 1).[4]



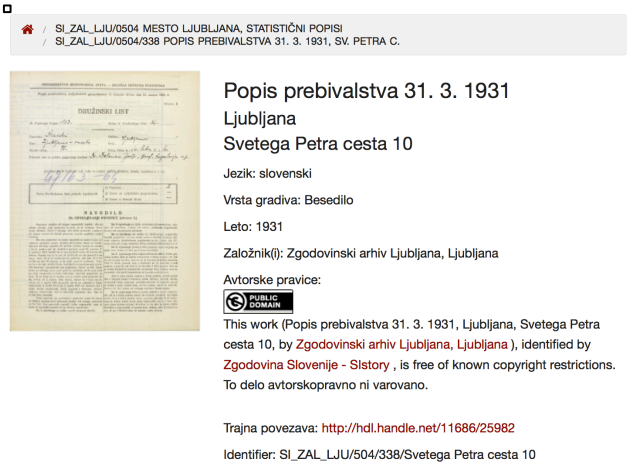
Slika1: Popis prebivalstva Ljubljane 1830: napredni iskalnik, Spletni arhiv NUK, 27. 2. 2011, http://nukrobi2.nuk.uni-lj.si:8080/wayback/20110227114943/http://www.sistory.si/popis_prebivalstva_1830_napredno-iskanje.html.

Že kmalu pa so se pokazale določene pomanjkljivosti takšnega pristopa pri objavljanju popisov prebivalstva:

- le za manjši del popisov so obstajali analogni indeksi, ki bi omogočili enostavno izdelavo elektronskih indeksov;
- obstoječi analogni indeksi niso zajemali vseh popisanih oseb (v glavnem le glave družine);
- ker so analogni indeksi nastali pred drugo svetovno vojno, so zajemali tudi osebe iz arhivskega gradiva, ki danes ni več ohranjeno.

Zato smo se odločili, da bomo po zgledu na rodoslovne projekte indeksacijo izvajali s prepisovanjem osebnih podatkov. Trenutno so na portalu SIstory objavljeni popisi prebivalstva Ljubljane (1830–57, 1869, 1921 - 1. del, 1931), občine Vrhnika (1870, 1880, 1890, 1900, 1910) in nekaterih občin okrajnega glavarstva Novo mesto iz leta 1869 (Dobrnič, Mirna, Velika Loka, Bela Cerkev, Črmošnjice, Kočevske Poljane, Prečna, Trebnje in novomeško predmestje Kandija). Za objavo se pripravlja še popis prebivalstva Ljubljane 1857. V sklopu digitalizacije vseh teh popisov je bilo narejenih 84000 slik in prepisani podatki za več kot 142000 oseb. V sodelovanju s FamilySearch[5] so bili leta 2015 digitalizirani še vsi ostali historični popisi, ki jih hrani Zgodovinski arhiv Ljubljana, skupaj več kot 270000 slik.

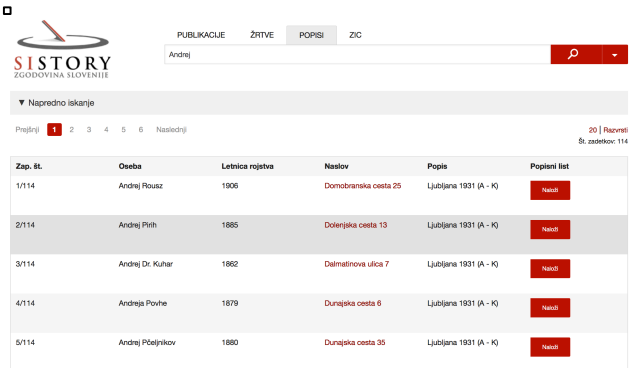
Na portalu SIstory so kot PDF publikacije objavljene popisnice za posamezne hiše. Metapodatke teh publikacij je mogoče iskati s pomočjo splošnega SIstory iskalnika. Metapodatki ne vključujejo informacij o osebah, temveč samo o kraju (naselje, ulica in hišna številka) in času popisa.



Slika 2: Popis prebivalstva kot publikacija na portalu SIstory.

Za iskanje podatkov o osebah, ki so bile popisane v okviru posameznih hiš, je predviden poseben iskalnik.[6] Sredi leti 2016 smo za iskalnik začeli uporabljati Elasticsearch, kateri temelji na Apache Lucene.[7] Iskalnik smo konfigurirali v skladu s potrebami rodoslovnih projektov:

- iskanje po osnovnih podatkih: ime in priimek, leto rojstva, bivališče (naslov) in hišna številka;
- uporaba n-gram algoritma za iskalne nize, ki se ne ujemajo povsem z zapisi v bazi;
- filtriranje rezultatov po popisih;
- povezava na sliko popisnega lista z ostalimi podatki o iskani osebi;
- povezava na PDF publikacije hiše (naslov hiše), v okviru katere je bila popisana iskana oseba.



Slika 3: Iskalnik po indeksu oseb iz popisov prebivalstva.

Iskalnik podatke o osebah zajema iz orodja za transkribiranje podatkov. V tabeli 1 so prikazani vsi digitalizirani popisi prebivalstva, katerih slike so bile že uvožene v to orodje (1), prepisani podatki o osebah pa omogočajo iskanje teh oseb na portalu SIstory. V tabeli je prikazano tudi število slik tistih digitaliziranih popisov, katere nameravamo v naslednjih letih postopoma uvoziti v orodje za transkribiranje podatkov in hkrati kot PDF publikacije tudi na SIstory.

[4] Podatki iz prvotne baze so še vedno dostopni kot Popis prebivalstva Ljubljane 1830: Tabelarni prikaz podatkov z iskalnikom in povezavami na digitalizirano gradivo, Zgodovina Slovenije – Sistory, http://sistory.si/SISTORY:ID:26731.

[5] FamilySearch, https://familysearch.org/.

[6] Popisi prebivalstva, Iskanje, Zgodovina Slovenije – SIstory, http://sistory.si/popis.

[7] https://www.elastic.co/products/elasticsearch.

popis	št. slik	št. oseb	orodje
Mesto Ljubljana 1830-1857	5069	16758	1
Mesto Ljubljana 1857	5260	15719	0
Mesto Ljubljana 1869	14627	23088	1
Mesto Ljubljana 1880	12934		0
Mesto Ljubljana 1890	22362		0
Mesto Ljubljana 1900	28198		0
Mesto Ljubljana 1910	36806		0
Mesto Ljubljana 1921	12841	2919	1
Mesto Ljubljana 1921	14559		0
Mesto Ljubljana 1928	36568		0
Mesto Ljubljana 1931	29561	59737	1
Občina Vrhnika 1870	576		1
Občina Vrhnika 1880	257		1
Občina Vrhnika 1890	992	2793	1
Občina Vrhnika 1900	657	4066	1
Občina Vrhnika 1910	705		1
Mesto Novo mesto 1870	284		0
Okr. glav. Novo mesto 1857	675		0
Okr. glav. Novo mesto 1869	10917	17609	1
Okr. glav. Novo mesto 1880	2869		0
Okr. glav. Novo mesto 1890	551		0
Okr. glav. Novo mesto 1900	506		0
Okr. glav. Novo mesto 1910	3026		0
Okr. glav. Novo mesto 1931	2947		0
Občina Čekovnik 1900	94		0
Občina Čekovnik 1910	138		0
Občina Čekovnik 1921	115		0
Občina Dole 1890	298		0
Občina Dole 1900	468		0
Občina Dole 1910	470		0
Občina Dole 1921	883		0
Občina Dole 1924	183		0
Občina Idrija 1870	1791		0
Občina Idrija 1880	1289		0
Občina Idrija 1890	1431		0
Občina Idrija 1900	247		0
Občina Idrija 1910	210		0
Občina Idrija 1921	2743		0
Občina Idrija 1931	5407		0
Občina Idrija 1936	5428		0
Občina Zminec 1880	502		0
Občina Zminec 1900	426		0
Občina Zminec 1931	1226		0
Občina Škofja Loka 1869	1133		0
Občina Škofja Loka 1880	793		0
Občina Škofja Loka 1890	640		0
Občina Škofja Loka 1900	300		0
Občina Škofja Loka 1931	951		0
	270913	142689	

Tabela 1: Popisi prebivalstva Slovenije 1830-1931 (število digitaliziranih slik; število transkribiranih oseb; popis je (1) oziroma še ni (0) uvožen v orodje za transkribiranje podatkov).

Na primeru popisa Mesto Ljubljana 1857 se vidi, da imajo lahko nekateri popisi že prepisane osnovne podatke o osebah, čeprav ti popisi še niso bili uvoženi v orodje za transkribiranje podatkov.

Za lažjo iskanje po gradivu so uslužbenci arhiva tekom let za nekatere popise že naredili indekse, v katerih so bili zajeti osnovni podatki o popisanih osebah. Indeksi niso bili izdelani po enotnem podatkovnem modelu, temveč se glede na nabor spremenljivk med seboj lahko precej razlikujejo. V glavnem vsebujejo le ime ime in priimek, leto rojstva in bivališče. Praviloma vsebujejo še podatek o tem, v katerem arhivskem dokumentu oziroma na kateri digitalizirani sliki se nahaja originalni arhivski zapis o indeksirani osebi. Nekateri arhivarji so v izdelavo indeksa vložili še dodaten trud in prepisali nekatere dodatne podatke kot so družinski stan, poklic, družinska razmerja, kraj rojstva in domovinska pravica. Zaradi lažjega iskanja oseb so stare oziroma ponemčene zapise slovenskih priimkov pogosto normalizirali v sodoben slovenski zapis.

Praviloma torej ti indeksi vsebujejo le osnovne podatke o osebah, s pomočjo katerih je nato mogoče najti še dodatne podatke o indeksirani osebi v analognem ali digitaliziranem arhivskemu gradivu. Na ta način so bili zbrani podatki za 15700 oseb popisa Ljubljane 1857 (v urejanju za uvoz v orodje) in za 10400 oseb popisov okrajnega glavarstva Novo mesto 1869 (uvoženo v orodje) ter še za ostale popise okrajnega glavarstva Novo mesto (še nismo dobili).

Na drugi strani spektra zbiranja podatkov iz historičnih popisov prebivalstva pa so raziskovalci in ostali zgodovinarji, ki prepišejo vse podatke iz popisov prebivalstva. Na ta način smo od Muzejskega društva Vrhnika dobili podatke za 6800 oseb, kateri so bili prepisani iz popisov občine Vrhnika 1890 in 1900 (Anžič, 2004).[8]

Indeksi, ki so jih naredili arhivisti, so bili pretvorjeni v XML, primeren za uvoz v relacijsko MySQL bazo orodja za transkribiranje. Te osnovne podatke o osebah se nato v orodju dopolni še z ostalimi, pred tem neprepisanimi podatki.

3 Orodje za transkribiranje

3.1 Osnovni namen

Na Inštitutu za novejšo zgodovino razvito orodje za transkribiranje ima v skladu z različnimi interesi partnerjev (Zgodovinski arhiv Ljubljana) v projektu Popisi prebivalstva dvojen namen:

- Zadovoljiti potrebe ustanov s področja varstva kulturne dediščine po čim lažjemu dostopu širše javnosti (prvenstveno rodoslovcev) do njihovega gradiva.
- Zadovoljiti potrebe raziskovalcev, ki se ukvarjajo s historično demografijo, po čim širšem naboru relevantnih raziskovalnih podatkov.

Zaradi zelo različnih interesov rodoslovnih (prepisuje se osnovne podatke vseh oseb) in raziskovalnih projektov (prepisuje se vse podatke reprezentativnega vzorca oseb) so bila do sedaj razvita orodja za transkribiranje prilagojena potrebam samo ene od teh dveh skupin projektov. Z razvojem novega orodja za transkribiranje smo to razdvojenost učinkovito presegli na način, ki maksimalno koristi obema partnerjema v projektu. Arhivisti tako prispevajo digitalizirano gradivo in podatke o indeksiranih osebah, raziskovalci prepišejo manjkajoče podatke o že indeksiranih osebah, hkrati pa prepisujejo

[8] SI_ZAL_VRH/0016 Matični urad Vrhnika, 1870–1959, Zgodovina Slovenije – Sistory, http://sistory.si/publikacije/?menu=737.

tudi podatke o osebah, ki jih arhivisti še niso indeksirali. Posledično se s tem širi tudi nabor indeksiranih oseb, do katerih preko iskalnika dostopajo uporabniki rodoslovnih projektov.

3.2 Skupine uporabnikov

Orodje za transkribiranje podatkov je prosto dostopno za vse raziskovalce, vendar se morajo zainteresirani raziskovalci (in ostali zainteresirani uporabniki) najprej registrirati.[9] Glede na pravice in dolžnosti se registrirani uporabniki delijo na tri skupine:

- navadni uporabniki,
- napredni uporabniki,
- uredniki popisov.

Z registracijo uporabniki najprej dobijo status navadnega uporabnika. Navadni uporabniki v orodju ne vidijo podatkov o že prepisanih osebah, temveč le tiste podatke, ki so jih sami prepisali. Posledično lahko prepisujejo podatke o osebah samo iz tistih hiš, iz katerih ni prepisoval še nihče drug. Podatke, ki so jih prepisali, lahko izvozijo v XLS datotekah. Šele ko (pravilno) prepišejo podatke za 300 oseb, jim urednik popisov lahko status nadgradi v naprednega uporabnika.

Napredni uporabniki v orodju vidijo ne le podatke, ki so jih sami prepisovali, temveč vse podatke, ki so bile pred tem vneseni v bazo tudi s strani ostalih uporabnikov. Hkrati pridobijo pravico do korigiranja vseh že obstoječih zapisov in pravico do izvoza celotne baze podatkov.

Urednike popisov lahko določi le administrator orodja. Glede dostopa do že prepisanih raziskovalnih podatkov ima enake pravice kot napredni uporabnik. Urednik popisov ima pravico, da ustvari XML datoteko za uvoz novega popisa, da novemu popisu preko administracije določi dodatna polja za prepisovanje in da nadzira pravilnost prepisov navadnega uporabnika.

Velike razlike med pravicami navadnih in naprednih uporabnikov glede prostega dostopa do vseh raziskovalnih podatkov so bile določene iz dveh razlogov:

- Ker uporaba orodja ni omejena le na ožjo projektno skupino, temveč je odprta za vse zainteresirane raziskovalce in študente, smo se z zahtevo, da mora uporabnik najprej pravilno prepisati podatke za 300 oseb, preden dobi pravico do dostopa do celotne baze podatkov, hoteli izogniti problemu prostega strelca (free rider problem).
- Hkrati smo presodili, da so posamezni popisi prebivalstva kot zgodovinski vir lahko tako zelo specifični, da je za čim bolj pravilno interpretacijo spremenljivk potrebno prepisati vsaj nekaj originalnih podatkov.

3.3 Osnovna načela

Relacijska baza podatkov, v katero se preko orodja za transkribiranje prepisuje podatke iz popisov, je bila zgrajena v skladu s sledečimi načeli:

1. Struktura popisov prebivalstva se je tekom časa zelo spreminjala, zato ni mogoče za vse popise vnaprej točno določiti vsa podatkovna polja.
2. Poleg popisov prebivalstva mora orodje omogočiti prepisovanje podatkov še iz matičnih knjig in drugih podobnih tabelarnih historičnih osebnih podatkov (domovnice, vojaške konskripcije ipd.).
3. Vsi popisi imajo skupna samo sledeča osnovna polja, podatke katerih zajema tudi iskalnik: ime in priimek, leto rojstva in širša popisna enota, znotraj katere je bila oseba popisana.
4. Osebe so med seboj povezane v razmerju, pri katerem se na eno izhodiščno osebo veže nič ali več odvisnih oseb.
5. Osebe morajo biti popisane v okviru ene enote, enota pa lahko ima nič ali več podenot, na katero je vezana ena ali več izhodiščnih oseb.
6. Prepisuje se lahko podatke za enote, podenote in osebe.
7. Enote morajo imeti povezavo na eno ali več slik (digitaliziranega arhivskega gradiva), osebe morajo imeti povezavo na eno sliko.
8. Popisi, enote, podenote, osebe in slike imajo unikatne identifikacijske oznake
9. Enote, podenote in osebe ni mogoče brisati.
10. Prepisanih podatkov enot, podenot in oseb ni mogoče brisati. Popravljen prepis je shranjen kot nova verzija. Za vsake shranjeno verzijo prepisanih podatkov se shrani tudi podatek o uporabniku in časovna znamka.

3.4 Izgradnja orodja

V skladu s temi načeli je bilo izdelano orodje za transkribiranje historičnih demografskih podatkov, ki omogoča čim hitrejše prepisovanje iz slik v relacijsko MySQL bazo podatkov. Prva, poskusna verzija orodja (Popisi 1.0) je bila izdelana leta 2011. Ta verzija orodja je bila v glavnem namenjena le dodatni indeksaciji oseb. Na podlagi pridobljenih izkušenj z delom na prvi verziji orodja se je leta 2012 začelo izdelovati novo verzijo (Popisi 2.0), v kateri so bila upoštevana vsa prej našteta načela. Ta verzija je v naslednjih letih doživela le manjše modifikacije. Lastnik kode je Inštitut za novejšo zgodovino, kodo pa so napisali zunanji izvajalci. Orodje je bilo zavestno zgrajeno s pomočjo nekaterih temeljnih tehnologij: PHP, MySQL, JavaScript, CSS in HTML. Ker je veliko programerjev, ki obvlada te tehnologije, je vzdrževanje orodja relativno poceni. V naslednjih letih načrtujemo večje posodobitve (Popisi 3.0), predvsem glede uporabe obstoječih JavaScript knjižnic (mdr. YUI library). Ob tej priložnosti načrtujemo objavo pod odprtokodno licenco.

3.5 Upravljanje orodja

Velikost popisov v orodju ni vnaprej določena, temveč je odvisna od konkretnih potreb uporabnikov in posameznih parcialnih projektov. V praksi sta se pri tem izoblikovali dve pravili:

- posamezen popis v orodju naj ustreza sestavi analognega arhivskega gradiva;
- večje popise (več kot 30000 oseb) naj se razdeli na manjše, bolj obvladljive dele, ki omogočajo hitrejši izvoz podatkov.

Ker vsak popis praviloma vsebuje na stotine slik in enot smo se zavestno določili, da ne potrebujemo uporabniškega vmesnika za shranjevanje slik in ustvarjanje novih enot. Že ob digitalizaciji arhivskega gradiva se uredi osnovne podatke o enotah (minimalno naslov in število enote) in katere slike so vezane na

[9] http://sistory.si/admin.

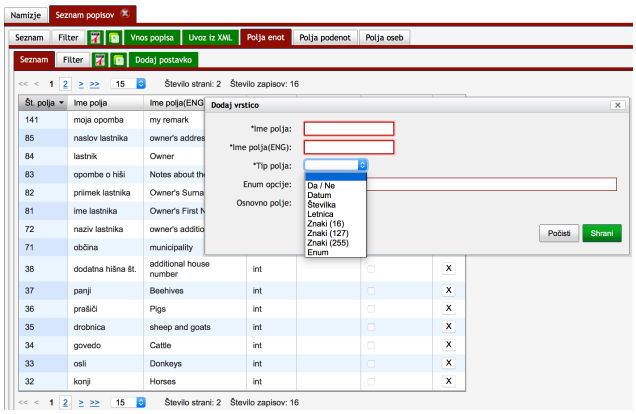
posamezno enoto. Nato pa je te slike in osnovne podatke nujno potrebno množično uvoziti v orodje. Pri popisih prebivalstva so enote posamezne hiše (naslov je ulica, število je hišna številka), znotraj katerih se je popisovalo njihove prebivalce. V skladu z osnovnim načelom 2 pa lahko urednik za enoto določi katero koli poljubno enoto, ki vsebuje podatke o osebah. Ker je iz praktičnih razlogov težje prepisovati podatke iz hiše, ki ima več kot 20 slik, se lahko posamezno hišo razdeli na več smiselnih delov, pri katerih ima sicer vsak del svojo unikatno identifikacijo, vendar imajo vsi deli (enote) iste podatke o hiši. Pri matičnih knjigah ali domovnicah je enota lahko samo ena stran, na kateri se nahaja obrazec o osebi, vse enote pa imajo posledično isti naslov in samo drugo število (stran v matični knjigi).

Glede na načelo 4 mora biti za vsako prepisano osebo določeno:

- oseba je izhodiščna oseba;
- oseba je vezana na točno določeno posamezno izhodiščno osebo.

Če med osebami ni nobene relacije, je vsaka oseba izhodiščna oseba. Pri popisih prebivalstva je izhodiščna oseba t. i. glava družine, pri krstni matični knjigi je izhodiščna oseba krščenec, pri poročni matični knjigi je izhodiščna osebe mož, pri mrliški pa umrli. Na izhodiščno osebo so vezane osebe v točno določenem razmerju (glede na šifrant: žena, mož, sin, hči, mati, oče, brat, sestra) ali v poljubnem razmerju (polje ostala razmerja).

Podenote se uporablja samo v primerih, ko osebe niso bile popisane samo znotraj enote, temveč tudi znotraj njegove podenote. Podenote se v praksi uporablja predvsem za prepisovanje podatkov o stanovanjih (podenota) v posameznih hišah (enota). Ker je bila znotraj ene podenote lahko popisana ena ali več družin, je za vsako podenoto potrebno določiti pripadajoče izhodiščne osebe.

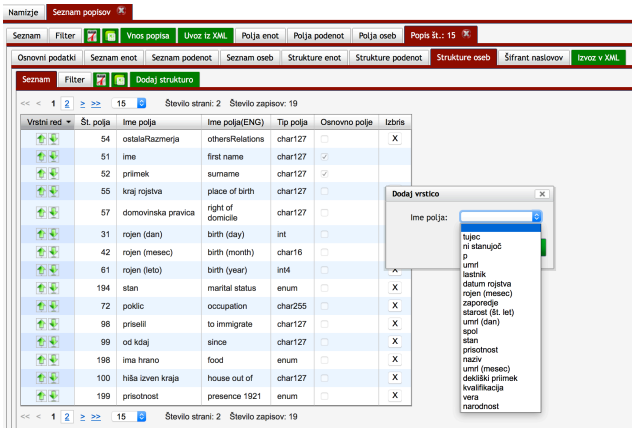


Slika 4: Uporabniški vmesnik za ustvarjenje dodatnih (extra) polj za vnos podatkov.

V skladu z načeli 1, 3 in 6 lahko urednik popisov poleg obveznih osnovnih podatkov za enote (naslov in št.), podenote (izhodiščna oseba) in osebe (ime in priimek, izhodiščna oseba, razmerje) preko uporabniškega vmesnika (Polja enot, Polja podenot, Polja oseb) kreira poljubna dodatna (extra) polja za vnašanje podatkov (glej Sliko 4). Vsako polje mora dobiti ustrezno slovensko in angleško ime. Ker se ime polja izpiše v obrazcu za prepisovanje podatkov, naj bo smiselno in čim krajše. Za vsako polje je potrebno določiti njegov podatkovni tip:

- Booleova spremenljivka (bool; da / ne),
- celo število (int; številka),
- štiri cela števila (int 4; letnica),
- datum (date),
- največ 16 znakov (char 16; znaki 16),
- največ 127 znakov (char 127; znaki 127),
- največ 255 znakov (char 255; znaki 255),
- naštevni (enum).

Urednik popisov nato preko uporabniškega vmesnika (Struktura enot, Struktura podenot, Struktura oseb) za vsak nov popis določi dodatna (extra) popisna polja in njihov vrstni red (glej Sliko 5).



Slika 5: Uporabniški vmesnik za določanje strukture polj posameznega popisa.

Pri tem je priporočljivo, da urednik pred tem skrbno analizira vsebino popisa in šele na podlagi te analize izbere najbolj primerna polja. Če bo npr. za prepisovanje rojstnega datuma izbral tip polja datum, se mora dobro zavedati, da je v to polje poleg letnice obvezno potrebno vpisovati še dan in mesec rojstva. Zlasti skrbno mora izbrati enega od podatkovnih tipov znaki. Če bo izbral tip s premajhnim številom znakov, višek znakov ne bo shranjen v bazo podatkov. Če bo izbral tip s prevelikim številom znakov, bo vizualno preveč razširil obrazec za prepisovanje podatkov in s tem otežil preglednost in hitrost prepisovanja. Polje s podatkovnim tipom znaki 255 je namreč šestkrat daljše od polja s tipom znaki 16.

Pri prepisovanju historičnih demografskih podatkov poznamo tri načine prepisovanja:

1. Uporabnik prepisuje podatke natančno takšne kot so bili prvotno zapisani.
2. Uporabnik prepisuje podatke natančno takšne kot so bili prvotno zapisani, vendar pri tem sproti popravlja očitne napake.
3. Uporabnik pri prepisovanju podatke sproti normalizira na način, ki ustreza njegovi zastavljeni raziskovalni nalogi.

Prvi in drugi način se večinoma uporabljata pri rodoslovnih in dolgoročno naravnanih raziskovalnih projektih, tretji način se večinoma uporablja pri enkratnih raziskovalnih projektih. Pri Popisih prebivalstva Slovenije smo se odločili za prvi način prepisovanja. Z našega stališča ima ta način prepisovanja dve veliki prednosti:

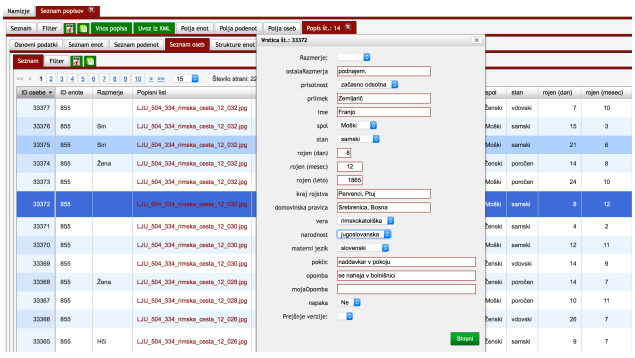
- Ker uporabniku med prepisovanjem ni potrebno izvajati normalizacijo zapisa, poteka prepisovanje sorazmerno veliko hitreje.
- Ker projekt Popisi prebivalstva predvideva ponovno uporabo raziskovalnih podatkov je podatke potrebno prepisati samo enkrat. Prepisane

podatke lahko nato raziskovalci vedno znova drugače interpretirajo in klasificirajo.

Toda v nekaterih primerih se je ta način prepisovanja izkazal za povsem kontraproduktivno, saj je imel za posledico manjšo hitrost prepisovanja in nobene dodatne interpretativne vrednosti. V primeru popisa prebivalstva Ljubljane 1931 je bilo pri opisni spremenljivki stan vrednost poročen zapisana na 53 različnih načinov in vrednost samski na 42 načinov, pri spremenljivki vera je bila vrednost rimskokatoliška zapisana na 46 načinov, pri spremenljivki narodnost pa vrednost jugoslovanska na 45 načinov. Še večja raznolikost načina zapisa istih vrednosti spremenljivke je bila v primeru večjezičnih popisov (slovenski in nemški). Pri popisu prebivalstva Ljubljane 1869 je bilo tako pri opisni spremenljivki stan vrednost poročen zapisana na 134 različnih načinov in vrednost samski na 51 načinov, pri spremenljivki vera pa je bila vrednost rimskokatoliška zapisana na kar 190 načinov. Zato smo sprejeli nova pravilna za način prepisovanja podatkov:

- Opisne spremenljivke, ki vsebujejo veliko število različnih vrednosti (imena, priimki, poklici, kraji ipd.), se prepisuje nespremenjene.
- Opisne spremenljivke, ki vsebujejo malo število različnih vrednosti in kateri so bili že prvotno ustrezno klasificirani (spol, vera, družinski stan, narodnost, jezik ipd.) se pri prepisovanju sproti klasificira na način, ki ustreza prvotni klasifikaciji.

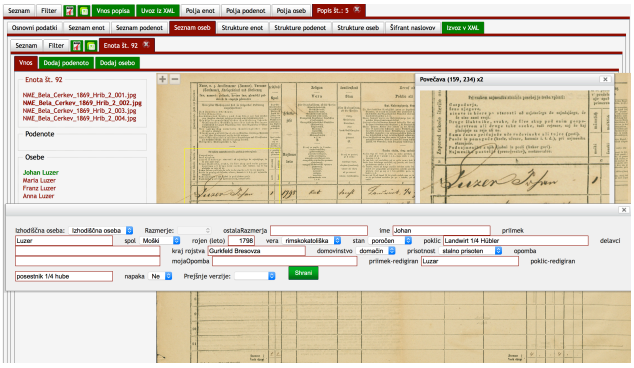
Za slednje primere smo uvedli naštevni (enum) tip podatkov. Urednik popisov predhodno preko vmesnika za ustvarjanje dodatnih polj določi vrednosti (med seboj ločene z vejico), ki jih bo uporabnik preko spustnega menija lahko vnašal v bazo. Ker se v MySQL bazi te vrednosti zapisujejo v varchar tip podatkov, jih lahko urednik popisov vedno naknadno spreminja, dodaja ali briše glede na potrebe projekta.



Slika 6: Urejanje zapisov v seznamu oseb.

Do že prepisanih podatkov o enoti, podenoti ali osebi uporabnik dostopa preko Seznama enot, Seznama podenot in Seznama oseb. Uporabnik lahko ureja in dopolnjuje obstoječi seznam (glej Sliko 6).

Uporabnik začne prepisovati nove podatke preko Seznama enot (gumb Podatki). Podatke prepisuje v obrazec s polji za prepisovanje podatkov, ki lebdi nad digitaliziranо sliko. Obrazec lahko uporabnik prosto premika na način, da si z njim podčrtuje podatke, katere prepisuje. Sliko je mogoče povečati ali pomanjšati, podatke je možno brati tudi s pomočjo povečevalne lupe.



Slika 7: Obrazec s polji za prepisovanje podatkov o osebi.

4 Uvoz in izvoz podatkov

Uporabnik lahko sezname enot, sezname podenot in sezname oseb izvozi v XLS format in izvožene podatke nato izven orodja za transkribiranje podatkov dalje ureja v skladu s svojimi raziskovalnimi potrebami.

Vse podatke posameznega popisa lahko napredni uporabnik izvozi tudi v XML zapisu (glej Sliko 8). Za format XML je izdelana posebna shema, ki se čim bolj dosledno naslanja na relacije v SQL bazi.[10] Podatki enot, podenot in oseb so zapisani v okviru ločenih elementov. Vsak od njih ima svojo identifikacijsko številko (id), verzija shranitve podatkov (version), identifikacijska številka osebe, ki je shranila to verzijo zapisa (user_id_added) in kdaj je bil zapis shranjen (date_added). Podatki enot, podenot in oseb imajo nekatere za njih specifične elemente. Podatek enot ima identifikacijsko številko mesta in številko mesta, podatek podenot ima identifikacijsko številko poglavarja (če jih je več, so med seboj ločene z znakom |), podatek oseb ima identifikacijsko številko razmerja, slike (file) in poglavarja. Vsi pa imajo še dodatna (extra) polja. Vsako dodatno (extra) polje ima v atributu id zapisano identifikacijsko oznako tega polja. Lastnosti dodatnih (extra) polj so na enoten način zapisani v okviru elementov polja_enot, polja_podenot in polja_oseb. V sklopu elementa povezave so zabeležene relacije med identifikacijskimi številkami enot in popisnih listov, enot in oseb ter naposled še enot in podenot. Identifikacijske številke in nazivi popisnih listov (slik), mest (naslov enote) in razmerij (med izhodiščno in odvisno osebo) se nahajajo v sklopu elementa sifranti.

To XML shemo se uporablja tudi za uvoz podatkov v orodje za transkribiranje. Na ta način poteka ne samo uvoz podatkov za slike in enote, temveč tudi za osebe, za katere so arhivarji in ostali zgodovinarji že izdelali indekse. Ti so te podatke prepisovali v XLS datoteke ali celo v tabele in odstavke DOC datotek. Zato je potrebno pred uvozom v orodje za transkribiranje te podatke še dodatno urediti in pretvoriti v XML. Zaradi specifičnosti vsakega popisa je pred vsakim uvozom novih podatkov le-te potrebno pretvoriti z na novo napisanimi XSLT stili.

[10] Imena XML elementov so večinoma izbrana glede na zasnovo prve verzije orodja Popisi 1.0 in jih kasneje nismo več spreminjali. Zato nekatera od njih ne ustrezajo več njihovemu novemu pomenu. Mapiranje: poglavar = izhodiščna oseba, popisni list = slika, mesto = naslov enote, mesto_st = št. enote.

```
<popis>
    <id>3</id>
    <naziv>Ljubljana 1869</naziv>
    <polja_enot>
        <polje>
            <id_extra>84</id_extra>
            <order>3</order>
            <ime_polja>lastnik</ime_polja>
            <tip_polja>char127</tip_polja>
        </polje> ...
    </polja_enot>
    <podatki_enot>
        <podatek>
            <id>1</id>
            <version>1</version>
            <user_id_added>11</user_id_added>
            <date_added>2014-12-11 09:14:30</date_added>
            <mesto>1</mesto>
            <mesto_st>1</mesto_st>
            <extra_polja>
                <polje id="84">Magist. Laibach</polje> ...
            </extra_polja>
        </podatek> ...
    </podatki_enot>
    <polja_podenot>
        <polje> ... </polje> ...
    </polja_podenot>
    <podatki_podenot>
        <podatek> ...
            <poglavar_id>1234|4567</poglavar_id> ...
        </podatek> ...
    </podatki_podenot>
    <polja_oseb>
        <polje> ... </polje> ...
    </polja_oseb>
    <podatki_oseb>
        <podatek> ...
            <razmerje>2</razmerje>
            <file>209</file>
            <poglavar_id>245</poglavar_id> ...
        </podatek>
    </podatki_oseb>
    <povezave>
        <enote_popisni_listi>
            <povezava>
                <enota_id>1</enota_id>
                <file_id>1</file_id>
            </povezava> ...
        </enote_popisni_listi>
        <enote_osebe>
            <povezava>
                <enota_id>1</enota_id>
                <oseba_id>3013</oseba_id>
            </povezava> ...
        </enote_osebe>
        <enote_podenote>
            <povezava>
                <enota_id>6</enota_id>
                <podenota_id>1</podenota_id>
            </povezava> ...
        </enote_podenote>
    </povezave>
    <sifranti>
        <popisni_listi>
            <popisni_list>
                <id>1</id>
                <file>Mesto_1_001.JPG</file>
            </popisni_list> ...
        </popisni_listi>
        <mesta>
            <mesto>
                <id>2</id>
                <naziv>Gradišče</naziv>
            </mesto> ...
        </mesta>
        <razmerja>
            <razmerje>
                <id>2</id>
                <naziv>Mož</naziv>
            </razmerje> ...
        </razmerja>
    </sifranti>
</popis>
```

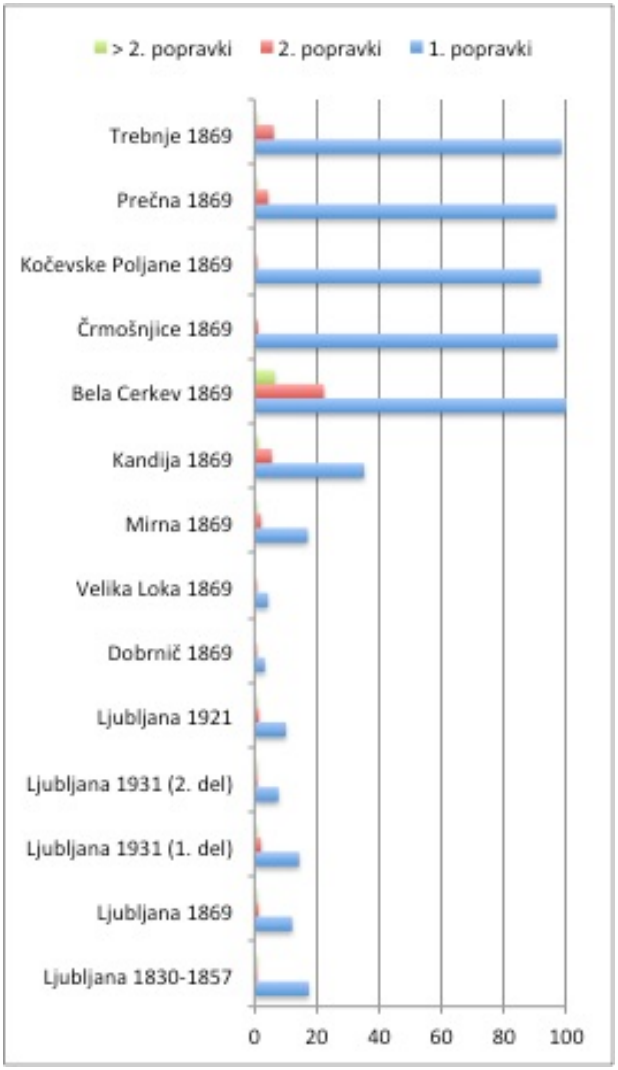
Slika 8: Zapis popisa v XML.

5 Analiza uporabe orodja

Poleg članov raziskovalne infrastrukture Slovenskega zgodovinopisja orodje uporabljajo tudi študentje Filozofske fakultete (in delno FDV) Univerze v Ljubljani. Čeprav projekt Popisi prebivalstva torej ni tipičen 'crowdsourcing' projekt, menim, da bi statistična analiza uporabe tega orodja pomagala ostalim raziskovalcem pri načrtovanju podobnih projektov. Takšnih analiz (Causer in Terras, 2014, 6-9) namreč do sedaj ni bilo veliko predstavljenih. Za analizo sem uporabil v XML izvožene podatke. Za razliko od XLS izvoza, ki vključuje le zadnjo verzijo zapisa podatkov, se v XML izvozi celotna baza podatkov: vse verzije zapisov podatkov ter kdo in kdaj jih je shranil. Z analizo bom zajel samo podatke o osebah, saj so uporabniki do sedaj le redko dopolnjevali podatke o enotah in podenotah. Ker smo v bazi dodali časovno znamko šele marca 2015, bom časovno analizo moral omejiti le na obdobje po tem datumu.

Za vsakogar, ki načrtuje raziskovalni projekt, pri katerem je nujno potrebno prepisati podatke o osebah, je najbolj zanimiv podatek o povprečni hitrosti prepisovanja. Pri tem se je potrebno zavedati, da bo ta hitrost odvisna tudi od jezika in vrste pisave, s katero je zapisan originalni popis prebivalstva. Popis prebivalstva Ljubljane 1869 (18 spremenljivk) je večinoma zapisan v nemščini in v nemški kurenti. Zato pri njem znaša srednja vrednost (mediana) vpisovanja novih oseb v bazo 1 minuta in 56 sekund. Nasprotno znaša mediana pri popisi prebivalstva Ljubljane 1931 (20 spremenljivk, slovenščina, latinica) samo 1 minuta in 15 sekund. Pri tem pa lahko hitro opazimo precejšnje razlike med začetnimi in izkušenimi uporabniki. Medtem, ko začetniki potrebujejo za vpis ene osebe 1 minuto in 44 sekund, jo izkušeni uporabniki prepišejo v 1 minuti in 13 sekundah. Zelo velike razlike so tudi med samimi izkušenimi uporabniki. Ena uporabnica je tako v povprečju za vpis ene oseba porabila samo 36 sekund.

Velik vpliv na končno hitrost prepisovanja podatkov imajo nujni popravki podatkov. V grafikonu na sliki 9 je prikazano, koliko odstotkom oseb so uporabniki naknadno popravljali podatke. Pri tem je tudi prikazano, koliko odstotkom so bili podatki popravljeni enkrat, koliko dvakrat in koliko več kot dvakrat. Pri tem je potrebno razlikovati med popisi, za katere so bili v orodje uvoženi tudi indeksi (Trebnje, Prečna, Kočevske Poljane, Črmošnjice in Bela Cerkev) in med ostalimi popisi. Pri prvi skupini popisov je bilo pri skoraj vsaki osebi potrebno osnovne podatke iz indeksov najprej dopolniti še z ostalimi podatki. Pri ostalih popisih pa so uporabniki praviloma že prvič prepisali vse podatke o osebah. Pri popisih Ljubljana 1830, 1868 in 1931 (1. del), kateri so bili najprej uvoženi v orodje za transkribiranje, je nekoliko večje število prvih popravkov tudi posledica razvoja orodja (iz verzije 1.0 na 2.0) in ustvarjanja novih podatkovnih polj. Hkrati moramo tudi vedeti, da sta popisa Bele Cerkve in Kandije edina popisa, ki sta bila naknadno tako temeljito dvojno pregledana, da v prihodnosti praktično ne bosta več potrebovala popravkov.



Slika 9: Odstotki popravljenih podatkov oseb glede na popis.

Če upoštevamo te dejavnike, lahko pri analizi grafikona pridemo do sledečih zaključkov:

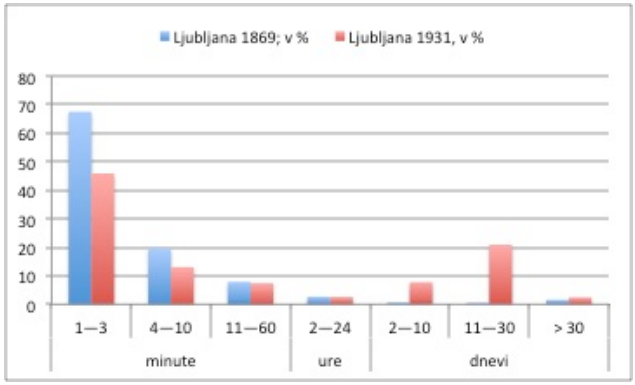
- Pri prepisovanju podatkov o osebah lahko pričakujemo, da bodo uporabniki morali popravljati podatke pri manj kot 10 % osebah.
- Ta odstotek popravkov je večji v primerih, ko mora uporabnik brati različne rokopise (ljubljanski popisi so zaradi različnih rokopisov veliko težje berljivi kot popisi okrajnega glavarstva Novo mesto).
- Preden bo osnovni prepis popisa postal uporaben za raziskovalne namene, ga bo uporabnik moral še enkrat skrbno pregledati in razrešiti vse morebitne dileme glede pravilnosti prepisa. Pri tem lahko pričakuje, da bo moral popraviti podatke za dosti več kot petino oseb.

V spodnji tabeli 3 je prikazana analiza načina popravljana besedila. Pri tem sem popravke razvrstil v tri večje skupine: besedilo je bilo delno popravljeno, v prej prazno polje so bili vpisani podatki, besedilo je bilo v celoti izbrisano. Dobljeni rezultati analize so lahko pri različnih popisih povsem različnih. Te razlike so posledice samo enega dejavnika: v primeru, da uporabnik ni prepisal vseh podatkov o osebi, je veliko večja verjetnost, da bo te podatke dopolnila druga oseba.

	popravljeno besedilo	vpisano besedilo v prazno polje	povsem izbrisano besedilo	popravek naredila druga oseba
Bela Cerkev 1869	5,1	94,9	0,0	98,3
Ljubljana 1931 (1)	7,8	91,4	0,9	87,8
Ljubljana 1869	45,6	48,6	5,7	31,8
Ljubljana 1931 (2)	41,2	28,8	30,0	18,4

Tabela 3: Načini popravljanja besedila glede na izbrane popise; v %.

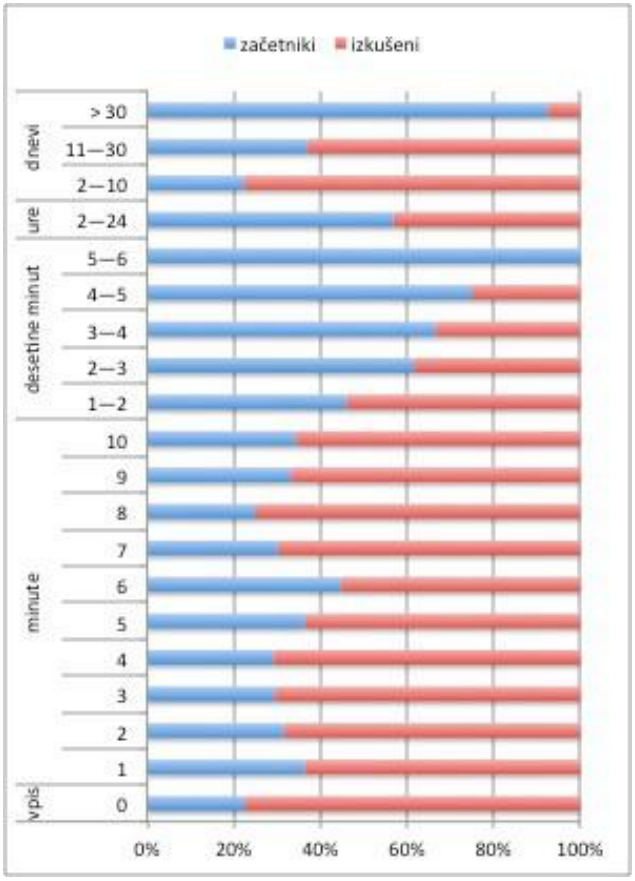
Iz teh podatkov je tudi razvidno, da uporabniki pogosto sproti popravljajo svoje prepisane podatke. Da bi lahko bolje razumeli to prakso, sem pripravil še časovno analizo popravkov. Na žalost pri tem razpolagamo samo s podatki od marca 2015, zaradi česar ti podatki niso primerni za časovno analizo popravkov, ki so se zgodili po daljšem časovnem obdobju.



Slika10: Koliko časa je preteklo od shranitve izvorne in popravljene verzije podatkov; popisi prebivalstva in časovne skupine v %.

Iz zgornjega grafikona (slika 10) je tako jasno razvidno, da lahko prakso popravljanja zapisanih podatkov razdelimo na dve večji skupini. Naknadno popravljanje in dopolnjevanje besedil po preteku 10 in več dni, ter na sprotno popravljanje napačnih zapisov. V slednjem primeru je tako velika večina popravkov opravljena že v prvih treh minutah, večinoma kar v prvi minuti.

Pri tem so zanimive precejšnje razlike med izkušenimi uporabniki in začetniki (slika 11). Glede na število opravljenih vpisov namreč začetniki pogosteje popravljajo (svoje) vpisane podatke kot pa izkušeni uporabniki. Sorazmerno največ je teh popravkov zlasti po prvih desetih minutah. Razlog za to tendenco je lahko samo eden. Zaradi svoje neizkušenosti se začetni uporabniki pri prepisovanju podatkov relativno bolj pogosto kot izkušeni srečajo z dilemo kako prebrati ali kako zapisati kakšen podatek. Pri tem reševanju te dileme namenijo dosti več časa kot njihovi izkušeni kolegi. V primeru morebitne večje vključenosti študentov v bodoče projekte bo vsekakor potrebno večjo pozornost nameniti reševanju teh problemov (npr. intenzivnejše in dolgotrajnejše delavnice).



Slika 11: Razlike med izkušenimi in začetnimi uporabniki glede popravljanja podatkov.

6 Zaključek

V prispevku sem predstavil orodje za transkribiranje historičnih demografskih podatkov, ki ga na Inštitutu za novejšo zgodovino uporabljamo v sodelovanju z arhivi, društvi in univerzo. Ker bi si želeli, da bi k uporabi orodja pritegnili še nove uporabnike, nameravamo temu primerno orodje razvijati tudi v prihodnje. Tako si bodo parcialni raziskovalni projekti v bodoče lahko rezervirali želeni sklop digitaliziranega gradiva (kateri še ni uvožen v orodje) in dobili zanj izključno pravico za obdelavo. Šele po preteku projekta bodo ti podatki v skladu s politiko odprtega dostopa do raziskovalnih podatkov dani na razpolago tudi ostalim raziskovalcem.

7 Literatura

Sonja Anžič. 2004. Prebivalstvo občine Vrhnika na prelomu 19. in 20. stoletja. *Vrhniški razgledi*, 5: 95-100. http://www.dlib.si/details/URN:NBN:SI:spr-LQTKRUDL.

Tim Causer in Melissa Terras. 2014. Crowdsourcing Bentham: beyond the traditional boundaries of academic history. *International Journal of Humanities and Arts Computing*, 8(1): 46-64. http://dx.doi.org/10.3366/ijhac.2014.0119.

Alicia Fornés, Josep Lladós, Joan Mas, Joana Maria Pujades in Anna Cabré. 2014. A Bimodal Crowdsourcing Platform for Demographic Historical Manuscript. V: *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, str. 103-108, New York, NY. DOI: 10.1145/2595188.2595199.

Aaron G. Noll. 2013. Crowdsourcing Transcription of Archival Materials. V: *Graduate History Conference: Interdisciplinary Approaches to Historical Inquiry*, Boston, MA. http://scholarworks.umb.edu/ghc/2013/panel6/4/.

Johanna Puhl, Peter Andorfer, Mareike Höckendorff, Stefan Schmunk, Juliane Stiller in Klaus Thoden. 2015. *Diskussion und Definition eines Research Data LifeCycle für die digitalen Geisteswissenschaften*. Göttingen: GOEDOC, Dokumenten- und Publikationsserver der Georg-August-Universität. http://nbn-resolving.de/urn:nbn:de:gbv:7-dariah-2015-4-4.