

Projekt Vojaške žrtve prve svetovne vojne na Slovenskem

Andrej Pančur,* Neja Blaj Hribar,* Mihael Ojsteršek,* Mojca Šorn*

* Inštitut za novejšo zgodovino
Privoz 11, 1000 Ljubljana

andrej.pancur@inz.si, neja.blaj-hribar@inz.si, mihael.ojstersek@inz.si, mojca.sorn@inz.si

1 Uvod

Različni dogodki, ki so bili povezani z obeleževanjem stote obletnice prve svetovne vojne, so v zadnjih letih še dodatno povečale zanimanje za zgodovinske vire in raziskave o prvi svetovni vojni (1914-1918). Nekateri uspešni projekti so se odvijali na mednarodni ravni, npr. Europeana Collections 1914-1918,¹ 1914-1918-online² in CENDARI,³ še več takšnih projektov je bilo na nacionalni ravni, med drugim tudi v Sloveniji.⁴ Večina teh projektov se posveča omogočanju dostopa do digitalizirane kulturne dediščine o prvi svetovni vojni, ki jo hranijo ustanove s področja varstva kulturne dediščine. Nekatere ustanove pa so objavile tudi baze podatkov o (padlih) vojakih prve svetovne vojne [WW1]. (Clavert, 2018a, 2018b) V prispevku bomo predstavili digitalno bazo podatkov (vojaških) žrtev prve svetovne vojne iz Slovenije,⁵ ki je nastala na podlagi sodelovanja med različnimi raziskovalnimi in kulturnimi ustanovami ter nekaterimi posamezniki. V konzorciju je sodelovalo 10 muzejev,⁶ 2 arhiva⁷ in 2 raziskovalni ustanovi⁸ in posameznik.⁹ Projekt je koordiniral Inštitut za novejšo zgodovino, ki ima dobro razvito raziskovalno skupino za digitalno humanistiko in je tudi nacionalna koordinacijska ustanova za Digital Research Infrastructure for the Arts and Humanities (DARIAH) v Sloveniji.

V nadaljevanju bomo najprej predstavili zasnovo projekta zbiranja podatkov o vojaških žrtvah prve svetovne vojne iz Slovenije ter njegove prednosti in pomanjkljivosti. V naslednjem poglavju bomo opisali postopke, orodja in tehnologije, ki smo jih uporabili pri odpravljanju teh pomanjkljivosti projekta. Na koncu bomo na kratko predstavili glavni rezultat projekta - bazo vojaških žrtev ter naše načrte za bodočnost.

2 Projekt zbiranja podatkov

Projekt zbiranja podatkov o žrtvah prve svetovne vojne iz Slovenije (ZV1) je dober primer uspešnega institucionalnega povezovanja med lokalnimi ustanovami s področja varstva kulturne dediščine, raziskovalnimi ustanovami in digitalno raziskovalno infrastrukturo. Ker je projekt nastal na pobudo arhivov, muzejev, knjižnic in galerij (GLAM ustanov), je tudi dober primer uspešnega reševanja izzivov (Seillier, 2018), s katerimi se soočajo pobudniki listine o ponovni uporabi podatkov kulturne dediščine: DARIAH, APEF, CLARIN, Europeana, E-RIHS, Iperion-CH, PARTHENOS.

Delovanje konzorcijskih partnerjev je temeljilo na sledečih načelih:

- **Ločeno financiranje:** Projekt smo financirali iz obstoječih dejavnosti posameznih partnerjev.
- **Uporaba obstoječih baz podatkov:** Partnerji so prispevali podatke iz svojih že obstoječih baz podatkov.
- **Ločeno zbiranje podatkov:** Partnerji so zbirali podatke prvenstveno za namene dejavnosti matičnih ustanov.
- **Odprti dostop:** Zbrane podatke so partnerji dali na voljo pod pogoji odprtih licenc Creative Commons.
- **Enotno skrbništvo podatkov:** Partnerji so zbrane podatke shranili v verificiranem repozitoriju portala Zgodovina Slovenije – SIstory.
- **Enotna baza podatkov:** Zbrani podatki vseh partnerjev so bili poenoteni in uvoženi v relacijsko MySQL bazo podatkov.

¹ <http://www.europeana-collections-1914-1918.eu/>.

² <http://www.1914-1918-online.net/>.

³ <http://www.cendari.eu/>.

⁴ <http://www.rtvsl.si/1sv>, <http://www.100letprve.si>.

⁵ <http://zv1.sistory.si/>.

⁶ Dolenjski muzej Novo mesto, Gorenjski muzej Kranj, Mestni muzej Idrija, Muzej novejšje zgodovine Slovenije, Muzej Velenje, Muzejsko društvo Vrhnika, Posavski muzej Brežice, Tolminski muzej, Ustanova "Fundacija poti miru v Posočju", Vojaški muzej Slovenske vojske.

⁷ Arhiv Republike Slovenije, Nadškofjski arhiv Ljubljana.

⁸ Inštitut za novejšo zgodovino; ZRC SAZU, Zgodovinski Inštitut Milka Kosa.

⁹ Renato Podbersič.

Za takšen način delovanja smo se odločili iz sledečih razlogov:

- Za namene zbiranja podatkov o vseh žrtvah prve svetovne vojne na ozemlju sedanje Slovenije ni bilo mogoče pridobiti ustreznega projektnega financiranja.
- Čeprav je bilo v okviru projekta opravljeno veliko prostovoljnega dela, so glede na svoje želje in pričakovanja intenzivnost delovanja v projektu določali partnerji sami.
- Konzorcij je združeval raziskovalce iz raznolikih raziskovalnih ustanov, z različnimi stopnjami tehničnega znanja in izkušenj. Zato so projektni partnerji sami izbrali ustrezna orodja in programe. Večina je uporabljala programe za preglednice, v prvi vrsti Microsoft Excel.

Zato pa je tako zasnovani projekt imel več pomanjkljivosti:

- **Manjkajoči podatki:** Ker so projektni partnerji večinoma zbirali podatke za samo določeno regijo, konzorciju pa se niso pridružile vse ustrezne ustanove iz Slovenije, se za nekatere regije podatkov sploh ni zbiralo.
- **Dvojniki:** Ker so podatke zbirali različni partnerji, vsak za svoje namene, imajo lahko v skupni bazi iste osebe več zapisov.
- **Neenotni podatki:** Ker so projektni partnerji podatke zbirali iz različnih razlogov, spremenljivke zbranih podatkov med seboj niso bile vedno primerljive.

3 Odpravljanje pomankljivosti projekta

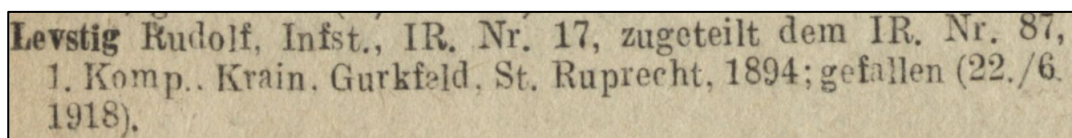
Infrastruktura skupina iz Inštituta za novejšo zgodovino si je zadala za nalogo, da s pomočjo digitalno humanističnih metod in orodij čim bolj učinkovito razreši pomankljivosti projekta zbiranja podatkov o vojaških žrtvah prve svetovne vojne, ki so izhajale iz same zasnove projekta:

3.1 Manjkajoči podatki

Po vzoru na nemške¹⁰ in avstrijske¹¹ geneološke projekte je uporabila digitalizirane Seznamne izgub (Verlustliste) za celotno Avstro-Ogrsko.¹² Ti tiskani sezname, ki so redno izhajali med prvo svetovno vojno in v mesecih neposredno po njej (12. 8. 1914 – 29. 1. 1919) vsebujejo osnovne podatke o ranjenih, zajetih, mrtvih in pogrešanih avstro-ogrskih vojakih (predvidoma skoraj 3 milijone zapisov). Podatki o žrtvi so zapisani na sistematičen način (vrstni red, vejice, podpičja, oklepaji), kar nam je omogočilo polavtomatsko označevanje teh zapisov:

- ime in priimek
- čin
- vojaški oddelek in pododdelek
- dežela, okraj in kraj
- leto rojstva
- kdaj in kaj se je vojaku zgodilo: ranjen, umrl, vojni ujetnik.

Ta del projekta je potekal v treh delovnih fazah. Najprej smo na spletni strani ANNO (Historische Österreichische Zeitungen und Zeitschriften) iz besedilnih datotek (format txt, Unicode) izluščili OCR vseh dostopnih seznamov izgub (515 datotek, 4.356.826 vrstic). Zaradi relativno slabega tiska, ki se je zaradi vojnih razmer in vedno slabšega papirja z leti samo še slabšal, je temu primerno tudi OCR vseboval napake. Na Sliki 1 je prikazan primer izseka digitalizirane slike zapisa podatkov o umrlem vojaku, na Sliki 2 pa originalni OCR tega zapisa.

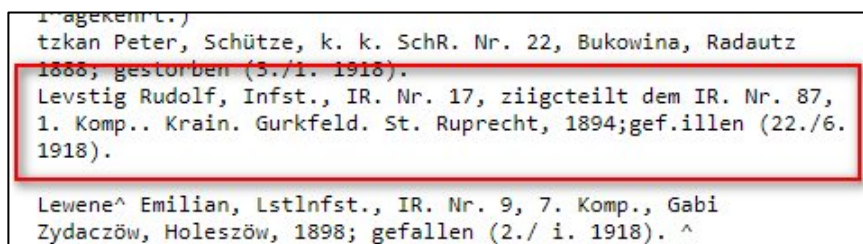


Slika 1: Izsek skena.

¹⁰ http://wiki-de.genealogy.net/Verlustlisten_Erster_Weltkrieg/Projekt.

¹¹ http://wiki-de.genealogy.net/Verlustlisten_%C3%96sterreich-Ungarns_1914-1919.

¹² <http://anno.onb.ac.at/cgi-content/anno?aid=vll>.



Slika 2: OCR zapis skena zgornje slike 1.

V drugem koraku smo besedilne datoteke pretvorili v datoteke XML. Pri tem smo se naslonili na smernice konzorcija TEI (Text Encoding Initiative). Prvotne datoteke TEI so v <teiHeader> imele zapisane osnovne metapodatke, vsaka vrstica prvotnega OCR besedila pa je bila označena kot odstavek <p>. S pomočjo regularnih izrazov smo najprej označili dele besedil z navedbo habsburških dežel (Kranjska, Štajerska, Koroška, Primorska; Goriška, Istra), katerih deli ozemlja so danes sestavni del slovenskega državnega ozemlja. Potem so sledili popravki OCR napak.

V tretjem koraku smo na podlagi niza pretvorb XSLT avtomatsko označili strukturo zapisa o žrtvah v skladu s TEI modulom Nemes, Dates, People, and Places (TEI Consortium, 2018). Končne datoteke TEI vsebujejo seznam oseb (<listPerson>), znotraj katere je vsaka žrtev iz slovenskih dežel Avstro-Ogrske označena znotraj elementa <person>. Na Sliki 3 je primer takšnega zapisa v skladu s smernicami TEI. Pri tem je čin označen z elementom <socecStatus> (socialno ekonomski status), pripadnost vojaškemu oddelku pa kot <affiliation>. Vojakovo usodo smo označili z elementom <event>, pri katerem vrsto usode določa atribut @type (death, pow, wounded, burial).

```
<person xml:id="vll.19181204_703.9082"
  facs="http://anno.onb.ac.at/cgi-content/anno?aid=vll&datum=19181204&seite=34&zoom=33">
  <persName>
    <surname>Levstig</surname>
    <forename>Rudolf</forename>
  </persName>
  <socecStatus>Infst.</socecStatus>
  <affiliation>IR. Nr. 17, zugeteilt dem IR. Nr. 87, 1. Komp.</affiliation>
  <birth>
    <placeName key="SI" type="municipality" corresp="krn:c.040304" cert="high">
      <country>Krain</country>
      <region>Gurkfeld</region>
      <settlement>St. Ruprecht</settlement>
    </placeName>
    <date>1894</date>
  </birth>
  <event type="death">
    <label>gefallen</label>
    <desc>(<date when="1918-06-22">22./6. 1918</date>)</desc>
  </event>
</person>
```

Slika 3: Kodirane podatkov o žrtvi 1. svetovne vojne iz seznama izgub Avstro-Ogrske v skladu s Smernicami TEI.

Za uspešno izvedbo projekta so bili zelo pomembni podatki o kraju rojstva oziroma bivanja (<placeName>), saj smo v primeru nekdanjih avstrijskih dežel, ki so danes hkrati del Slovenije in sosednjih držav, potrebovali natančne geografske podatke o izvoru teh oseb. Pri tem smo se naslonili (atribut @corresp elementa <placeName>) na podatke iz avstrijskih krajevnih repertorijev (Pančur et al., 2017), ki smo jih prav tako kodirali v skladu s smernicami TEI. Od prvotnih 66.442 oseb v seznamu žrtev je na koncu ostalo 35.899 oseb iz slovenskih dežel, od tega 18.536 ranjenih, 9535 ujetih in 8790 mrtvih.

3.2 Dvojniki

Zaradi takšnega načina zbiranja manjkajočih podatkov, se je še dodatno povečalo število dvojnikov v bodoči skupni bazi. Ta problem smo rešili na dva načina:

- V administracijo skupne baze smo vgradili osnovno orodje za povezovanje podatkov in deduplikacijo: ujemanje po imenu, priimku, rojstvu (datum in/ali kraj) in smrti (datum in/ali kraj).
- Za bolj napredno deduplikacijo zapisov o osebah uporabljamo:
 - R paket RecordLinkage
 - Fine-grained Record Integration and Linkage Tool (FRIL).

3.3 Neenotni podatki

Urejanje neenotnih podatkov preko administracije je ponavadi zelo zamudno in dolgotrajno, saj lahko na enkrat urejamo le podatke za nekaj oseb. Zato smo se odločili, da bomo podatke naknadno urejali in poenotili zunaj obstoječe baze. To urejanje tako poteka s pomočjo uvoza in izvoza vseh ali le dela podatkov iz baze v CSV formatu in naknadna obdelava teh podatkov v orodjih in okoljih (predvsem R), ki so specializirani za takšno urejanje podatkov.

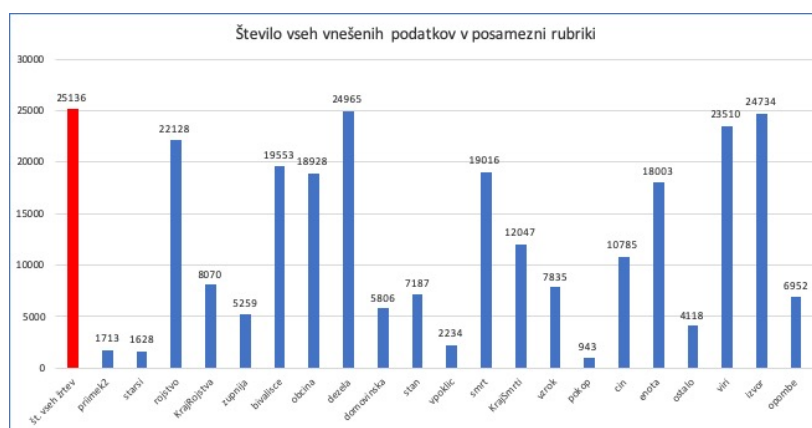
4 Baza žrtev

Skupna baza podatkov je dostopna na poddomeni portala Zgodovina Slovenije – Sistory kot samostojna spletna aplikacija.¹³ Digitalno bazo žrtev prve svetovne vojne smo zasnovali s pomočjo klasičnih spletnih tehnologij: HTML5, CSS, MySQL, PHP7 (Laravel framework) in JavaScript. Za iskanje po bazi uporabljamo ElasticSearch.¹⁴ Za tabelarni prikaz iskalnih rezultatov uporabljamo DataTables,¹⁵ ki omogoča nadaljnjo filtriranje podatkov po izbranih stolpcih (glej Slika 4). Vsi podatki o žrtvi se izpišejo na ločeni strani, uporabnik si lahko te podatke naloži v formatu PDF.

The screenshot shows the Sistory application interface. At the top, there's a search bar and navigation tabs: PUBLIKACIJE, ŽRTVE I.S.V., ŽRTVE II.S.V., POPISI, ZIG. Below the search bar, there's a table with columns: Id, Priimek, Ime, Datum rojstva, Datum smrti/pogrebanja, Kraj bivanja, Dežela, and Vojaška enota, oddelak in pododdelak. The table contains several rows of data. Below the table, there's a footer section with the Sistory logo, a list of columns, and a list of project partners.

Slika 4: Iskanje in filtriranje osnovnih podatkov spletne aplikacije.

Trenutno so v bazi podatki o 25.136 osebah. Redakcijska skupina projekta je določila enoten podatkovni model: priimek, druga oblika priimka, ime, očetovo/materino ime, datum rojstva, kraj rojstva, rojstna župnija, kraj bivanja, občina/okraj bivanja, dežela, domovinska pravica, stan (poklic, otroci), datum vpoklica, datum smrti/pogrebanja, kraj smrti, vzrok smrti, kraj pokopa, čin, vojaška enota, drugi podatki, viri, moje opomb, izvor. Zaradi raznovrstnosti izvornih podatkov projektnih partnerjev so podatki tega podatkovnega modela zelo različno zastopani (glej spodnji grafikon).



5 Načrti

¹³ <https://zv1.sistory.si/>.

¹⁴ <https://www.elastic.co/elasticsearch/>.

¹⁵ DataTables: Table plug-in for jQuery, <https://datatables.net/>.

Podatki o žrtvah v zapisu TEI se trenutno prosto dostopni v GitHub repozitoriju.¹⁶ Zato kratkoročno načrtujemo, da bomo to in nekatere ostale izvorne baze projektnih partnerjev objavili kot ločene digitalne izdaje. Prvi poskusni primer takšne ločene postavitve je že dostopen v GitHub repozitoriju¹⁷ in spletni strani¹⁸ ter poleg podatkov o smrtnih žrtvah vsebuje še podatke o vojnih ujetnikih in ranjencih.

V projektu smo za označevanje historičnih krajev uporabljali GeoNames. Hkrati načrtujemo še klasifikacijo dogodkov in vojaških enot. Po vzoru uspešnih Linked Open Data (LOD) projektov, ki so uporabili podatke o prvi svetovni vojni (Mäkelä et al., 2017), je naš dolgoročni cilj, da tudi podatke o žrtvah prve svetovne vojne iz Slovenije objavimo na semantičnem spletu.

6 Zahvala

Raziskavo je sofinancirala Javna agencija za raziskovalno dejavnost Republike Slovenije skozi program Raziskovalne infrastrukture slovenskega zgodovinopisja (IO-0013) in slovenske raziskovalne infrastrukture DARIAH-SI.

Literatura

- Frédéric Clavert. 2018a. Traces de mémoires collectives et bases de données de la Première Guerre mondiale. V: *L'histoire contemporaine à l'ère numérique (blog)*. <https://histnum.hypotheses.org/2743>.
- Frédéric Clavert. 2018b. WW1 (fallen) soldiers on-line: databases as traces of collective memories. V: *Digital Humanities Benelux Conference*. International Institute for Social History, Amsterdam. http://2018.dhbenelux.org/wp-content/uploads/sites/8/2018/05/Clavert_FallenSoldiers_DHBenelux2018.pdf.
- Eetu Mäkelä, Juha Törnroos, Thea Lindquist in Eero Hyvönen. 2017. WW1LOD – An application of CIDOC-CRM to World War 1 Linked Data. *International Journal on Digital Libraries* 18(4): 333-343.
- Andrej Pančur, Neja Blaj Hribar in Barbara Györfi. 2017. *Krajevna imena na Slovenskem 1.0*. Zgodovina Slovenije – SIstory. <http://hdl.handle.net/11686/36052>.
- Dorian Seillier. 2018. Assessing the survey on the Charter's principles. V: *Heritage Data Reuse Charter: Information and development around the Heritage Reuse Charter (blog)*. <https://datacharter.hypotheses.org/179>.
- TEI Consortium (2018). *Guidelines for Electronic Text Encoding and Interchange*. <http://www.tei-c.org/P5/>.

¹⁶ <https://github.com/SIstory/Verlustliste>.

¹⁷ <https://github.com/sidih/verlustliste>.

¹⁸ <https://sidih.github.io/verlustliste>.