NADGRADNJA ZGODOVINARSKEGA INDEKSA CITIRANOSTI

Katja MEDEN
Jožef Stefan Institut; Inštitut za novejšo zgodovino

Ana CVEK
Inštitut za novejšo zgodovino; Filozofska fakulteta, Univerza v Ljubljani

Meden, K., Cvek, A. (2021): Nadgradnja Zgodovinarskega indeksa citiranosti. Slovenščina 2.0, 9(1): 216–235.

DOI: https://doi.org/10.4312/slo2.0.2021.1.216-235

Začetki Zgodovinarskega indeksa citiranja segajo v leto 2003, ko so raziskovalci Inštituta za novejšo zgodovino začeli spremljati in sistematično popisovati citate za prijave projektov in programov na ARRS. Citatni indeks je doživel nekaj nadgradenj, poskusov harmonizacije podatkov in prečiščevanja relacijskih baz, vendar je bilo v zadnjih letih ugotovljeno, da sistem ne zadostuje potrebam indeksatorjev in uporabnikov. Pred nadgradnjo smo izvedli analizo podatkov, kjer so se identificirale največje težave. Nadgradnja je potekala v dveh delih; v prvem delu smo nadgradili administrativni del, v drugem delu pa spletno aplikacijo. Zgodovinarski indeks citiranja je bil med nadgradnjo tehnično posodobljen in s tem oblikovan tako, da je intuitiven za indeksatorje in uporabnike.

Ključne besede: Zgodovinarski indeks citiranosti, ZIC, nadgradnja, citatni indeksi

1 UVOD

Ocenjevanje uspešnosti raziskovalcev v humanistiki je v primerjavi z drugimi raziskovalnimi področji, predvsem naravoslovnimi, že od samih začetkov precej prikrajšano. Med drugim ocenjevanje temelji na frekvenci citiranosti, te podatke pa pridobimo iz različnih citatnih indeksov, kot sta na primer Web of Science (v nadaljevanju WOS) in Scopus. Monografije so primarni produkt raziskovalnega dela v humanistiki in družboslovju (Glänzel in Schoepflin, 1999; Hicks, 2004; Huang in Chang, 2008; Nederhof, 2006). V nasprotju z vrednotenjem raziskovalne uspešnosti v naravoslovju se ta področja teže vrednotijo, predvsem zaradi dejstva, da so monografije po večini bolj obsežne kot znanstveni članki (Kousha idr., 2011), in visokih kriterijev vključevanja publikacij v obstoječe indekse citiranja, na primer WOS in Scopus. Med pomembnejše kriterije spadajo redno izhajanje serijske publikacije, jezik publikacije, recenziranost, spoštovanje mednarodnih standardov (kot so informativni naslov, povzetek, popolna bibliografska informacija za vse citirane reference), poleg pogojev pa težavo predstavlja tudi indeksiranje monografij. Obstoječi citatni indeksi se namreč bolj osredotočajo na serijske publikacije. Neenakosti pri vključevanju publikacij v citatne indekse so na Inštitutu za novejšo zgodovino skušali zamejiti že v letu 2003. Raziskovalci so začutili potrebo po spremljanju in sistematičnem popisovanju citatov za prijave projektov in programov, kar predstavlja zametek Zgodovinarskega indeksa citiranja (v nadaljevanju ZIC). Osnovni namen je bil ustvariti bazo citatov iz slovenskih zgodovinskih monografij, osrednjih znanstvenih časopisov in revij (Lazarević in Zemljič, 2003). Začetna shema baze, ki je bila precej enostavna, je ob nastanku dobro zadovoljevala potrebe raziskovalcev, vendar so se sčasoma pokazale pomanjkljivosti (Pančur idr., 2014), ki so vodile v nadaljnje nadgradnje, poskuse harmonizacije podatkov in prečiščevanja relacijskih baz. ZIC trenutno vsebuje 4.837 vseh vnosov, od tega 2.901 vnos serijskih publikacij in 1.936 vnosov monografij in poglavij iz monografij, kar predstavlja razmerje 59,9 % serijskih publikacij ter 39,1 % monografij in poglavij iz monografij.

Zadnja nadgradnja je potekala leta 2012 in predstavlja osnovo in temelj nadgradnje, ki je predstavljena v nadaljnjem besedilu članka.

2 CITATNI INDEKSI IN HUMANISTIKA

Kot omenjeno, sta humanistika in družboslovje pri vrednotenju znanstvene uspešnosti v nasprotju z naravoslovnimi vedami nekoliko prikrajšana pri vključevanju raziskovalne produkcije v mednarodne citatne indekse, kot sta Web of Science (WOS) in Scopus. V Sloveniji vrednotenje raziskovalne uspešnosti poteka prek Informacijskega sistema o raziskovalni dejavnosti (SICRIS), v katerem je popisana celotna slovenska raziskovalna produkcija in je povezan s prej omenjenima mednarodnima citatnima indeksoma WOS in Scopus (Curk idr., 2006). Pomembno je poudariti, da so točke, pridobljene prek SICRIS, osnovno merilo za točkovanje raziskovalne uspešnosti in so neposredno povezane s procesom financiranja raziskovalnih projektov in programov prek Agencije za raziskovalno dejavnost Republike Slovenije (ARRS).

Z vprašanjem vključenosti humanistike in družboslovja v WOS in Scopus se je ukvarjalo več raziskav (Ball in Tunger, 2006; Bartol idr., 2014), kjer obstaja konsenz o tem, da je za vključevanje humanistike in družboslovja Scopus občutno bolj primeren kot pa WOS. Vendar kot omenjeno, je monografija primarna oblika znanstvene produkcije v humanistiki, ki pa ji citatni indeksi niso najbolj naklonjeni. Podatki kažejo, da WOS zajema okoli 12.000 znanstvenih revij in samo okoli 50.000 monografij, medtem ko Scopus zajema več kot 21.500 znanstvenih revij in 113.000 znanstvenih monografij. Število monografij v indeksu Scopus odraža večji obseg monografij v primerjavi z WOS, pa vendar monografije v primerjavi s številom znanstvenih člankov v revijah predstavljajo zgolj zanemarljiv del citatnega indeksa (Južnič, 2017).

Podobno stanje je tudi pri vključevanju slovenske raziskovalne produkcije v humanistiki. Južnič in Čadej (2016) v svoji raziskavi ugotavljata, da baza Scopus bistveno bolje zajema slovensko humanistično in družboslovno znanstveno publikacijo v primerjavi z WOS. Razlogi za to so različni: od dejstva, da je Scopus neprimerno bolj naklonjen vključevanju neangleških revij slabše razvitih in manjših držav vzhodne Evrope, do milejših meril vključevanja publikacij (Pajić, 2015).

Ne glede na dejstvo, da je Scopus bolj primeren za vključevanje slovenskih znanstvenih revij in monografij v humanistiki, pa še vedno obstaja vrzel pri vključevanju teh publikacij v Scopus. To pa poskušamo zamejiti s citatnimi indeksi, kot je npr. ZIC, ki so prilagojeni specifičnim lastnostim področja, ki ga

pokrivajo (v primeru humanistike je torej največje odstopanje v vključevanju monografskih publikacij).

3 CILJI IN POTEK NADGRADNJE

Pri postopku nadgradnje smo z uporabo sodobnih tehnologij in estetsko privlačne grafične podobe želeli preoblikovati administratorski spletni vmesnik in indeksatorju omogočiti prijazno in pregledno izkušnjo pri urejanju podatkov. Najpomembnejši cilj nadgradnje je bila postavitev ZIC kot ločene aplikacije. Ker je baza MySQL trenutno integralni del portala SIstory in se upravlja s pomočjo skupne administracije, je treba podatkovno bazo ZIC postaviti kot ločeno aplikacijo na poddomeni portala SIstory. Razlog za to je načrtovana postavitev nove digitalne knjižnice portala SIstory kot samostojnega repozitorija z ločeno administracijo. Poleg ločene baze in administracije smo pri nadgradnji upoštevali naslednje sklope problemov. V prejšnji nadgradnji uvoz in izvoz podatkov nista bila mogoča, zato smo želeli to omogočiti. Prav tako smo želeli, da je spletna aplikacija narejena modularno, kar bo omogočalo dodajanje novih funkcionalnih rešitev. Pri uporabniškem vmesniku smo želeli, da je stran prijazna za mobilne obiskovalce, pri iskalniku pa smo želeli doseči hitro in pregledno iskanje po podatkih. Nadgrajeni administracijski modul naj bi omogočal enostavnejši dostop in upravljanje vseh podatkov ter z geslom zaščiten dostop do administracije. Izbrani osnovni podatki morajo biti z ustreznim vmesnikom prosto dostopni strojnemu zajemu podatkov (Pančur, 2019b).

Pri postavljanju ciljev in procesu nadgradnje smo izhajali iz temeljnih načel Raziskovalne infrastrukture slovenskega zgodovinopisja (v nadaljevanju RI INZ), ki vključujejo uporabo uveljavljenih in razširjenih tehnologij, ki jih člani infrastrukture dobro poznajo in obvladajo (načeli enostavnosti in poznavanja), modularno nadgrajevanje obstoječih tehnologij (načelo fleksibilnosti) in uporabo odprtih ali lastniških standardov (načelo odprtosti) (Pančur in Šorn, 2019). V procesu nadgradnje smo tako uporabljali tehnologije, ki jih priporoča RI INZ (Pančur, 2019a) in upoštevajo načeli enostavnosti in poznavanja HTML5 in CSS3, najnovejše verzije PHP, MySQL, ElasticSearch engine, JavaScript in JavaScript knjižnice. Pomemben vidik nadgradnje je tudi vidik interoperabilnosti, ki se v svojem pomenu prepleta z načelom fleksibilnosti. Fleksibilnost in interoperabilnost sistema želimo doseči z implementacijo

aplikacijskega profila MODS za uvoz in izvoz metapodatkov v različnih formatih, ki podpirajo nadaljnjo diseminacijo in izmenjavo podatkov z drugimi informacijskimi sistemi. Nadgradnja je potekala v posameznih sklopih, ki so opisani v nadaljevanju besedila.

4 REZULTATI NADGRADNJE

Nadgradnja je potekala v dveh delih: prvi del se nanaša na administrativni sistem SIstory. Nadgradnja v tem delu zajema preoblikovanje mask in njihovih polj, postavitev nove sheme XML po standardu MODS za uvoz in izvoz podatkov, iskalnik, ki temelji na tehnologiji ElasticSearch, ter migracije vrednosti ločenih polj Avtor(ji). Drugi del se osredotoča na nadgradnjo spletne aplikacije in uporabniškega vmesnika. Pri programski nadgradnji smo sodelovali z zunanjimi sodelavci Infrastrukture.

4.1 Administrativni sistem Sistory

4.1.1 Maske za vnos podatkov

Glavna sprememba v administracijskem sistemu (admin) je prehod s prej enotne maske na dve ločeni. Enotna maska je vsebovala tri razdelke: *Splošni podatki, Podatki o viru* in *Vsebinska obdelava*. Vnos podatkov v maske poteka ročno, podatkovna polja v enotni maski pa so bila nejasna (npr. ponavljanje polja za vnos id številke COBISS, imena avtorja idr.), nekatera tudi brez pomena za potrebe citatnega indeksa. Tako je bil na primer razdelek Vsebinska obdelava za citatni indeks povsem neuporaben, saj vsak zapis vsebuje identifikatorje s povezavami na zapise publikacij (COBISS, SIstory) s polnim metapodatkovnim opisom.

Iz enotne maske sta nastali dve neodvisni maski za vnos podatkov v ZIC V2. Iz maske za vnos publikacije sta nastali dve: *maska za vnos monografij* in *maska za vnos serijskih publikacij*, ki dovoljujeta natančnejši opis glede na publikacijo, ki jo indeksiramo. Vsaka izmed mask, tako kot v prejšnji verziji, vsebuje tudi masko za vnos citatov. Maske so bile oblikovane na podlagi zaznanih težav v prejšnjem administracijskem sistemu, o katerih so poročali indeksatorji, ter na podlagi potreb za opis določene publikacije in citatnega indeksa. Spodnja preglednica (Preglednica 1) prikazuje polja oziroma metapodatke za opis posameznih del in citatov.

Preglednica 1: *Metapodatki mask za vnos podatkov*

Metapodatek	min/max. št	Podatkovni tip	Maska (Mono, Serijska, Citat)	Primer
Cobiss ID	0,1	ID	M, S, C	3278924
Sistory ID	0,1	ID	M, S, C	handle. net/11686/4320
ISBN	0,1	ID	M	987-961-3421-43
ISSN	0,1	ID	S	0353-0329
Jezik	1,1	ISO639-2b	M, S	slv - slovenski
Tipologija	1,1	COBISS tipologija	M, S	1.16 – Samostojni znan. sestavek
Tip	0,1	interni seznam	M	Poglavje v monografiji
Avtorji	1,neomejeno	niz	M, S, C	Marko Zajc
Naslov	1,1	niz	M, S, C	Slovenski intelektualci in …
Vzporedni naslov	0,1	niz	M, S	Slovenian Intellectuals …
Naslov zbornika	0,1	niz	M	Slovenija v Jugoslaviji
Naslov vira	0,1	niz	S	Prispevki za novejšo zgodovino
Uredniki	0,neomejeno	niz	M	Zdenko Čepič (ur.)
Kraj	0,1	niz	M, S, C	Ljubljana
Založba	0,1	niz	M, S, C	Založba INZ
Leto	0,1	številčna vrednost	M, S, C	2015
Letnik	0,1	številčna vrednost	S, C	57
Številka	0,1	številčna vrednost	S, C	1
Zbirka	0,1	niz; št. vrednost	M	Vpogledi; 10
Stran	0,1	št. vrednost	M, S, C	241–256
DOI	0,1	ID	S, C	10.1090/019339135
Baza citatov INZ	0,1	gumb	M, S	DA
Citat na strani	1,1	št. vrednost	C	34
Vir	0,1	niz	C	Prispevki za novejšo zgodovino

Večina elementov, potrebnih za opis publikacij, je ostala nespremenjena. Po opravljeni analizi elementov mask smo izpostavili ključna polja za potrebe

opisa publikacij in njihovih citatov. Večina polj je splošne narave (npr. avtor, naslov, leto, kraj itd.), publikacije, ki jih vnašamo (monografije in serijske publikacije), pa se med seboj razlikujejo v določenih vidikih. Ločeni maski s prilagojenimi polji omogočata (z indeksatorskega vidika) kakovostnejšo indeksacijo publikacije. Elementi so bili spremenjeni ali prilagojeni, saj določeni niso bili ažurirani (na primer element Tipologija) ali niso omogočali dovolj natančnega opisa (element Avtor). Pri poljih Avtor in Urednik smo metapodatkovno polje ločili na dve polji: Ime in Priimek. S tem smo zagotovili natančnejši, bolj strukturiran opis in posledično boljše prikazovanje podatkov. Zaradi nove strukture polja je bilo za povezovanje vrednosti polj treba opraviti migracijo vrednosti iz starih, neločenih polj v nova, strukturno ločena polja v obliki Priimek, Ime (za namen prikaza). Nekaterih elementov iz stare maske v novih maskah nismo vključili, npr. Ključne besede ali Država, saj so bili za opis publikacij v citatnem indeksu nepotrebni. Dodani so bili tudi novi elementi, ki jih starejša maska za vnos podatkov ni vsebovala, ker ti podatki še niso bili potrebni. Tu govorimo predvsem o maski za vnos serijskih publikacij in citatov, kjer smo dodali polji DOI in URL, ki omogočata enoznačno, trajno identifikacijo, prav tako pa poleg polja Sistory ID uporabniku omogočata hiter dostop do publikacije.

Pri analizi obstoječih zapisov se je izkazalo, da so pomanjkljivi in neenotni. Do takšnih napak je prihajalo predvsem zato, ker indeksatorji niso imeli nobenih konkretnih navodil in so publikacije v maski (glavni vnos in citat) vpisovali po lastni presoji. Zato smo se pri nadgradnji odločili, da indeksatorjem ponudimo pomoč, ki jim bo olajšala vnos podatkov, še bolj pomembno pa je, da bi s temi navodili oz. pomočjo radi zagotovili čim bolj enotno indeksacijo ter pravilnejše in natančnejše zapise v indeksu. Ob vsakem polju je pri vseh treh maskah opis polja z navodili za vnos in primeri, ki naj bi bili indeksatorju v pomoč oz. oporo pri vpisovanju podatkov. Tu velja poudariti, da se zavedamo, da se bodo napake kljub pomoči še vedno pojavljale, saj se podatki vpisujejo ročno. S tem, da dajemo navodila za vnos, poskušamo zmanjšati število pogostih napak.

4.1.2 ElasticSearch iskalnik in filtriranje

Iskalnik ElasticSearch je distribucijsko, odprtokodno in analitično orodje za vse vrste podatkov, skupaj z besedilnimi, številčnimi, geoprostorskimi, strukturiranimi in nestrukturiranimi podatki (What is ElasticSearch, b.d.). ElasticSearch temelji na knjižnici Lucene Apache, ki je odprtokodna Java knjižnica za besedilno iskanje. ElasticSearch ponuja najrazličnejše možnosti, kot so prilagodljiva mapiranja podatkovnih polj, shranjevanje vrednosti ključev (ang. Key Value Store) itd., sam delovni tok pa je sestavljen iz petih korakov (What is ElasticSearch, b.d.; Divya in Goyal, 2013):

- **Zajem podatkov** (ang. *Data ingestion*): Postopek zajema vrednosti se začne s tako imenovanim *data ingestion*, v katerem so surovi podatki zajeti v iskalnik iz različnih virov. Podatki, ki jih zajamemo, so lahko v kateremkoli formatu in kakršnekoli velikosti.
- **Pretvorba v format JSON**: Zajete podatke pretvorimo v format JSON *JavaScript Object Notation*), ki omogoča interoperabilnost podatkov med različnimi sistemi.
- **Tokenizacija**: Zajete podatke je potrebno ločiti na posamezne besede, kar dosežemo z uporabo funkcije *Tokenizer*.
- **Indeksacija**: V naslednjem delu se oblikuje ElasticSearch index, ki je zbirka med seboj povezanih dokumentov. Vsak izmed dokumentov je povezan s ključi (imena, podatkovna polja ali lastnosti) in njihovimi vrednostmi (niz, številke, Boolovi operatorji, nabor vrednosti ...).
- **Parsiranje podatkov** (*Data parsing*): Parser bo procesiral iskalno poizvedbo (ang. search query), preiskal indeksirani dokument in poiskal morebitne ustrezne zadetke.

Za implementacijo iskalnika ElasticSearch za ZIC v administrativnem sistemu podatke zajamemo iz relacijske baze, ki temelji na tehnologiji MySQL (*What is ElasticSearch, b.d.*). Indeksirani ključi so v tem primeru podatkovna polja, ki bodo namenjena iskalnim poizvedbam, in njihove vrednosti (ki so večinoma besedilni nizi ali številčne vrednosti). Iskalnik ponuja izvajanje kompleksnih iskalnih poizvedb, ZIC uporablja funkcijo *simple string query*:

```
GET /_search
{
   »query«: {
      »simple_query_string« : {
      »«query«: »Mojca + Šorn +
      \«Življenje Ljubljančanov
      med drugo svetovno vojno\««
      »fields«: [»title^5«, »body«],
      »default_operator«: »and«
      }
   }
}
```

Funkcija uporablja preprosto sintakso za besedilne iskalne poizvedbe, na podlagi katere vrača iskalne rezultate z uporabo parserja.

Za iskalnik v spletni aplikaciji indeksiramo zgolj polji *Avtor* in *Naslov*, filtri v spletni aplikaciji pa imajo indeksirana polja (in njihove vrednosti) *Identifikator*, *Avtor*, *Naslov*, *Tipologija*, *Leto*, *Kraj* in *Št. citatov*. V administrativnem sistemu je bil filter nadgrajen. Prej je omogočal filtriranje po naslednjih parametrih: *Avtor*, *Leto*, *Naslov*, *Vir* in *Kraj*. Ti po mnenju indeksatorjev niso omogočali učinkovitega in natančnega iskanja zapisov znotraj baze. Novi filtri vsebujejo večje število parametrov: *Tip (monografija/serijska publikacija)*, *ID*, *Avtor*, *Naslov*, *Leto* in *Vir*. Iskalnik ElasticSearch podpira tudi funkcijo samodokončanja iskalne poizvedbe, poznano tudi pod imenom *Autocomplete* ali *Completion suggester*. Funkcija je optimizirana za hitrost tipkanja, saj se prilagaja hitrosti tipkanja iskalne poizvedbe, ki jo uporabnik vnese. Podpira izključno funkcijo *type as you go* in ni mišljena za samodejno korekcijo iskalne poizvedbe ali funkcije *Ali ste mislili* (What is ElasticSearch, b.d.). V našem primeru se na funkcijo samodokončanja, enako kot pri osnovnem iskalniku, vežeta zgolj polji Avtor in Naslov.

4.1.3 Uvoz in izvoz metapodatkov – MODS aplikacijski profil

XML ali eXtensible Markup Format prihaja iz družine označevalnih jezikov, kot sta SGML in HTML. Vendar pa se od omenjenih formatov razlikuje predvsem po fleksibilnosti – v primerjavi s HTML omogoča oblikovanje lastnih označevalcev oz. elementov (angl. tag) in s tem predstavlja enega izmed najpogosteje uporabljenih standardov za izmenjavo podatkov v digitalni humanistiki (Extensible markup language (XML) 1.0 (fifth edition), b. d.). Že v

prejšnjih verzijah baze je izvoz podatkov bil mogoč v formatu XML. Shema je predpostavljala lastne elemente (npr. *OpTipBiblEnote* za označevanje tipologije vpisanega vnosa ali *OpSistoryUrnId* za vnos SIstory identifikatorja) in ni upoštevala kateregakoli metapodatkovnega standarda, kot je na primer Dublin Core. Kot je bilo že omenjeno, to pomeni zmanjšano stopnjo interoperabilnosti podatkov, saj gre za unikatne elemente oz. označevalce, ki jih drugi (informacijski) sistemi ne uporabljajo. Pri prenosu podatkov lahko zaradi neujemajočih shem (oziroma elementov) prihaja do izgube določenega dela podatkov ali celo do izgube konteksta, v katerem so podatki. Čeprav je med metapodatkovnimi standardi najbolj razširjen in uporabljen standard Dublin Core ali njegova razširjena različica, DCTERMS, pa imata oba standarda precej omejen nabor elementov, ki ne zadostuje našim potrebam. Čeprav bi z implementacijo enega izmed omenjenih standardov dosegli višjo stopnjo interoperabilnosti, pa smo se zaradi omejitev nabora elementov odločili za metapodatkovni standard MODS.

Metadata Object Description Schema (MODS) je shema XML z bibliografskimi elementi (oziroma naborom elementov), ki jo lahko uporabljamo za najrazličnejše potrebe. Shema izhaja iz standarda za bibliografske zapise MARC21, vendar za svoje elemente namesto številčnega zapisa (na primer polje 222 za glavni naslov (ang. *Key Title*) in 210 za skrajšan naslov (ang. *Abbreviated Title*) uporablja besedilne označevalce oziroma elemente (ang. *language-based tags*) (MODS User Guidelines, Version 3 (Metadata Object Description Schema), b.d.).

MODS namreč vsebuje dovolj obsežen nabor elementov, ki ustreza našim potrebam, hkrati pa je še vedno dovolj razširjen in zato omogoča zaželeno stopnjo interoperabilnosti naših podatkov z minimalno izgubo konteksta.

Postopek prenosa podatkov iz interne sheme v metapodatkovno shemo MODS je vključeval tri faze:

- Pregled elementov stare sheme, ki je za svoje elemente upoštevala imena, kot so *OpTipBiblEnote* ali *OpSistoryUrnId;* del elementa 'Op' se nanaša na publikacijo, ki jo opisujemo (Op = original publication), 'Pv' pa označuje podatke za vir publikacije, sledi interno poimenovanje polja (ki ustreza imenu polja, iz katerega vzamemo podatke).

- Preslikava internih polj (poimenovanje po meri) v metapodatkovni standard MODS in komentiranje kode (navodila za programerja, iz katerih polj v stari metapodatkovni shemi se vežejo vrednosti v nove elemente). Iz ene sheme sta nastali dve novi, upoštevali smo novo strukturo mask za vnos podatkov, tako kot smo predhodno enotno masko razdelili na masko za monografije in serijske publikacije. V aplikacijskem profilu v skupnem metapodatkovnem zapisu v formatu XML sta ločena zapisa mask definirana z elementom mods in identifikatorjem ID=pub za oznako zapisa za monografijo ali serijsko publikacijo (na primer *mods ID=pub.224*) ali elementom relatedItem in identifikatorjem za oznako navedenih del, na primer *relatedItem type=referencesID=ref.1*.

- Prenos vrednosti iz starih internih polj v polja MODS ima svoje prednosti; poleg dejstva, da tako povečamo interoperabilnost svojih podatkov z drugimi sistemi, s tem pridobimo večjo strukturiranost in pogosto

```
<?xml version="1.0"?>
<root>
  <delo>
      <field name="ID" >4916</field>
      <field name="OpTipBiblEnote" >1</field>
      <field name="OpTipologija" >1</field>
      <field name="OpZvrst" >1</field>
      <field name="OpJezik" >21</field>
      <field name="OpDrzava" >17</field>
      <field name="OpStAvtorjev" >1</field>
      <field name="OpCobId" >3955316</field>
      <field name="OpSistoryUrnId" >0</field>
      <field name="OpAvtor0" >Hadalin Jurij</field>
      <field name="OpAvtor1" ></field>
      <field name="OpAvtor2" ></field>
      <field name="OpUrednik" ></field>
      <field name="OpNaslov" >Jugoslovanska malica</field>
      <field name="OpVzpNaslov" ></field>
      <field name="OpPodnaslov" ></field>
      <field name="PvCobId" >302353920</field>
      <field name="PvISSN" >978-961-6386-99-9</field>
      <field name="PvTip" >2</field>
      <field name="PvAvtor" ></field>
      <field name="PvNaslov" >Mimohod blaga : materialna kultura potrošniške družbe na Slovenskem</field>
      <field name="PvNaslovKratki" ></field>
      <field name="PvPodnaslov" ></field>
      <field name="PvVzporedniNaslov" ></field>
      <field name="PvZbirka" >Vpogledi ; 22</field>
      <field name="PvKraj" >Ljubljana</field>
      <field name="PvZalozba" >Inštitut za novejšo zgodovino</field>
      <field name="PvLetnik" ></field>
      <field name="PvLeto" >2019</field>
      <field name="PvSt" ></field>
      <field name="PvStran" >151-165</field>
      <field name="KwDeskriptor1" ></field>
      <field name="KwDeskriptor2" ></field>
```

Slika 1: Metapodatkovna polja maske za vnos podatkov pred nadgradnjo.

tudi dodatne podatke, ki jih v stari shemi ne bi mogli implementirati. Element OpJezik ima za svojo vrednost na primer le številčno vrednost »21«, kar se navezuje na interni nekontroliran seznam jezikovnih vrednosti, novi element pa v svoji strukturi dovoljuje navedbo avtoritete in tipa poimenovanja. Tako poleg jezikovne kode pridobimo tudi podatek o standardu oziroma kontroliranem seznamu, ki je bil uporabljen, s tem pa tudi standardiziramo vrednost zapisa. Slika 1 prikazuje strukturo in del elementov stare, interne metapodatkovne sheme.

Spodaj so prikazani stari in novi način poimenovanja ter primerjava strukture posameznega zapisa:

Interna shema ZIC (element Avtor):

```
<field name="OpAvtor0">Hadalin Jurij</field>
```

Aplikacijski profil v XML:

```
<name type="personal">
      <namePart>Priimek Ime avtorja</namePart>
   <role>
      <roleTerm type="code">cre</roleTerm>
      <roleTerm>Avtor</roleTerm>
   </role>
      <namePart type="family">Priimek</namePart>
      <namePart type="given">Ime</namePart>
</name>
```

Interna shema ZIC (element Jezik)

```
<field name="OpJezik" >21</field>
```

Aplikacijski profil v XML:

```
<language>
   <languageTerm type="code">slv</languageTerm>
   <scriptTerm type="code">Latin</scriptTerm>
</language>
```

Interna shema ZIC (element Tipologija):

```
<field name="OpTipologija" >1</field>
```

Aplikacijski profil:

```
<classification authorityURI="https://www.izum.si/">101</classification>
```

Z novim aplikacijskim profilom, ki izhaja iz metapodatkovnega standarda MODS, smo namesto internih metapodatkovnih elementov v shemi uporabili obstoječi in razširjeni metapodatkovni standard MODS. S tem smo naslovili dve izmed temeljnih načel: poznavanje oziroma uporabo poznanih in razširjenih tehnologij ter načelo interoperabilnosti. Format XML nam namreč zagotavlja lažje izmenjevanje in diseminacijo podatkov z drugimi sistemi.

4.1.4 Migracija vrednosti polj avtorji

Enega izmed večjih problemov, ki nam ga je delno uspelo rešiti med nadgradnjo, predstavlja migracija vrednosti polja Avtor(ji) iz skupnega polja v dve ločeni. Problem je nastal zaradi neenotnega zapisa oziroma različnih oblik vrednosti Priimek in Ime (oblike: *Priimek, Ime; Ime in Priimek, Ime, Priimek* …) ter naštevanja več avtorjev v enem polju (*Avtor1; Avtor2* …), ki so bili med seboj ločeni z različnimi ločili. Ta problem nam je uspelo rešiti zgolj delno: migracija, ki je potekala strojno, je bila uspešna na poljih, ki so se med seboj ujemala, pri določenih zapisih pa to ni bilo mogoče (primer *Ime Ime, Priimek*), zato zahteva ročne popravke. Te napake bomo lahko odpravili po začetku procesa prečiščevanja baze, ki pa za zdaj še ni predviden.

4.2 Spletna aplikacija in uporabniški vmesnik

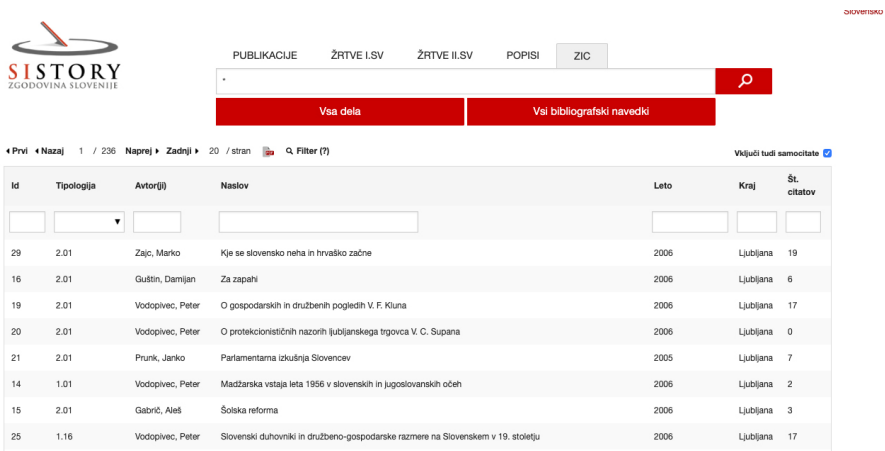
4.2.1 Podatkovna baza Vseh del in podatkovna baza Vseh bibliografskih navedb

Spletna aplikacija vsebuje dve podatkovni bazi: bazo *Vsa dela* in podatkovno bazo *Vse bibliografske navedbe*. Razlog za dve medsebojno ločeni bazi je v prikazu rezultatov, še natančneje v prikazu števila prejetih citatov pri določenem zapisu. Pri izpisu rezultatov je na voljo število citatov, ki jih je določeno delo prejelo, vendar ti podatki morda niso pravilni, ker se število prejetih citatov določenega dela veže na ujemanje naslova pri glavnem vnosu (maska za vnos glavnega zapisa) in pri citatu (maska za vnos citata). Kot pa smo omenili že zgoraj, nemalokrat pride do napak. Zaradi tega je potrebna druga baza *Vse bibliografske navedbe*, po kateri je omogočeno brskanje z uporabo filtrov. Ta baza dovoljuje uporabniku dodaten in bolj natančen vpogled v citate, saj tu dejansko vidimo vse vnesene citate, indeksatorjem pa predstavlja dodatno orodje za lažje popravke že obstoječih zapisov (preglednejše iskanje zapisov slabše kakovosti).

4.2.2 Prikaz iskalnih rezultatov

Iskalni rezultati so prikazani v obliki tabel, ki uporabnikom ponujajo tudi filtriranje rezultatov oziroma omogočajo oženje iskalne poizvedbe znotraj tabele. Rezultate je mogoče tudi razvrščati. Poleg filtriranja je uporabniku omogočen izvoz zadetkov na seznamu rezultatov in posameznega zadetka v formatu PDF. Za uporabnike sta prav tako pripravljeni tudi dve vrsti pomoči: osnovna razlaga uporabe citatnega indeksa na prvi strani ZIC (iskanje/brskanje) in manjši namig pri uporabi filtrov s primeri uporabe ločil. Prikaz posameznega zapisa uporabniku dovoljuje vpogled v osnovne podatke (metapodatke dela), osnovne podatke vseh del, v katerih je bil citiran, in avtorjev seznam literature. Podatki so prikazani v dveh ločenih tabelah, *Citirano v* in *Seznam literature*, zapisi so med seboj povezani.

Med oblikovanjem vmesnika so v vmesnih fazah sodelovali raziskovalci/uporabniki, s katerimi smo testirali odzive na novi vmesnik, novo podatkovno strukturo in nove funkcionalnosti. Največ težav je predstavljala terminologija, predvsem na podlagi dejstva, da se zgodovinarsko dojemanje terminov literature in virov precej razlikuje od pojmovanja na področju tehnologije. Nerodna poimenovanja iz prejšnje verzije vmesnika (*Avtor citira, Citiranost Avtorja*) je bilo treba nadomestiti s terminom, ki bo uporabnikom razumljiv. Kot že omenjeno, smo se na podlagi tega odločili za osnovno iskanje in dve ločeni bazi, ki sta po številnih preimenovanjih pridobili ime *Vsa dela* in *Vsi bibliografski*



Slika 2: Trenutni uporabniški vmesnik ZIC-a.

navedki. Čeprav sta imeni daljši, smo prednost namenili razlagi terminov, saj so uporabniki menili, da sta ti poimenovanji najbolj jasni in logični.

Poleg terminologije je problem predstavljala tudi postavitev elementov na spletni strani (predvsem gumbi). Tu se je izkazalo, da je uporabnike precej zmedla postavitev gumbov za obe bazi, saj so mislili, da s klikom na npr. *Vsa dela* dobijo vsa dela iskanega avtorja. Težavo smo odpravili tako, da smo ustvarili različne statične verzije uporabniškega vmesnika in s pomočjo uporabnikov določili tisto, ki je najbolj jasna in intuitivna.

4.2.3 Uporaba indeksa citiranosti

Primarni uporabniki citatnega indeksa so raziskovalci, ki lahko v sistemu enostavno preverijo št. prejetih citatov za posamezno avtorsko delo; če je to indeksirano v sistem. Poleg izpisa iz sistema SICRIS (Slovenian Current Research Information System), ki je osnova za vrednotenje znanstvene uspešnosti na posameznem raziskovalnem področju, lahko izpis iz ZIC predstavlja dodano vrednost pri prijavljanju projektov ali programov na področju humanistike in pri obnavljanju ali napredovanju v višje znanstvene nazive. Poleg raziskovalcev si z ZIC lahko pomagajo tudi uredniki revij, ki želijo preveriti, kolikokrat so bili posamezni članki citirani, in s tem upravičijo obstoj revije. Poleg primarne naloge, ki je zagotavljanje vpogleda v število prejetih citatov, pa indeks ponuja tudi druge možnosti, ki jih stari ZIC ni ponujal. Te naj bi uporabniku omogočile prijetnejšo interakcijo s sistemom. Ena izmed takšnih funkcionalnosti je npr. možnost *prijaznega kopiranja*, ki uporabniku omogoča lažje navajanje virov v svojih delih, saj ZIC ponuja skoraj popolne bibliografske podatke, ali npr. izpis števila citatov v formatu PDF ipd. Indeks ponuja tudi možnost dostopa do polnega besedila, če je le-to na voljo na sestrskem spletnem portalu Zgodovina Slovenije – SIstory.

5 SKLEP

Sistem je bil že v začetni zasnovi izjemno ambiciozen in zaradi načina objavljanja v zgodovinopisju izjemno potreben. Vendar je Zgodovinarski indeks citiranja zadnja leta nekoliko stagniral. Po pregledu in analizi podatkov smo ugotovili, da je nadgradnja potrebna, saj sistem ne zadostuje potrebam indeksatorjev in uporabnikov. Začeli smo nadgradnjo administrativnega dela,

kjer smo preoblikovali oz. nadgradili nove maske, nadgradili metapodatkovno shemo oziroma ustvarili nov aplikacijski profil na podlagi metapodatkovnega standarda MODS, filtre in dodali pomoč indeksatorjem, ki naj bi pripomogla k poenotenim zapisom. Poleg administrativnega dela smo nadgradili tudi uporabniški vmesnik z občasnim testiranjem baze in njenih komponent z raziskovalci. Z omenjeno nadgradnjo smo rešili večino zaznanih problemov, od nejasnih in nepotrebnih polj vnosa podatkov in razčlenitve mask, ki indeksatorju omogočajo lažje in natančnejše oblikovanje zapisov, oblikovanja aplikacijskega profila MODS, ki omogoča lažji uvoz in izvoz podatkov, do uporabniku prijaznejšega vmesnika itd. Vseh težav pa zaradi omejitev, povezanih z ročnim vnosom podatkov, ni bilo mogoče v celoti rešiti. To velja predvsem za postopek migracije polja Avtorji, kjer bo problem v celoti rešen šele po prečiščenju celotne baze podatkov. Postopek prečiščenja bo pripomogel tudi k poenotenju zapisov, kar bo omogočalo, da uporabniki v sistemu pridobijo zanesljive in kakovostne informacije. Pri nadgradnji Zgodovinarskega citatnega indeksa smo dosegli zastavljene cilje. Sistem smo tehnično posodobili in ZIC postavili kot ločeno spletno aplikacijo na poddomeni portala SIstory. Spletna aplikacija je narejena modularno, zato je mogoče dodajati nove funkcionalne rešitve, iskalnik s tehnologijo ElasticSearch pa omogoča natančnejše in preglednejše iskanje po podatkih.

V prihodnosti želimo poleg že obstoječih funkcionalnosti dodati še druge možnosti, ki bi olajšale delo indeksatorjem, uporabnikom pa omogočile prijetnejšo uporabniško izkušnjo. Te možnosti so npr. avtomatizirano vnašanje osnovnih podatkov iz vnosov, ki so povezani in dostopni na portalu SIstory, ter možnost samodejnega generiranja citatov po različnih citatnih stilih (npr. APA, Chicago idr.). Z nadgradnjo Zgodovinarskega indeksa citiranosti smo tako oblikovali sistem, ki je intuitiven za indeksatorje in uporabnike, s tem pa zagotovili, da ZIC izpolni svoj namen.

Zahvala

Raziskavo je sofinancirala Javna agencija za raziskovalno dejavnost Republike Slovenije v okviru programa Raziskovalne infrastrukture slovenskega zgodovinopisja (I0-0013) in slovenske raziskovalne infrastrukture DARIAH SI.

LITERATURA

Ball, R., & Tunger, D. (2006). Science indicators revisited-Science Citation Index versus SCOPUS: A bibliometric comparison of both citation databases. *Information Services and Use*, *26*(4), 293–301.

Bartol, T., Budimir, G., Dekleva-Smrekar, D., Pušnik, M., & Južnič, P. (2014). Assessment of research fields in Scopus and Web of Science in the view of national research evaluation in Slovenia. *Scientometrics*, *98*(2), 1491–1504.

Curk, L., Budimir, G., Seljak, T., & Gerkes, M. (2006). Linking the SICRIS-COBISS.SI-Web of Science systems. *Organizacija znanja*, *11*(4), 230–235.

Divya, M. S., & Goyal, S. K. (2013). ElasticSearch: An advanced and quick search technique to handle voluminous data. *Compusoft*, *2*(6), 171.

Extensible markup language (XML) 1.0 (fifth edition). Pridobljeno https://www.w3.org/TR/xml/

Glänzel, W., & Schoepflin, U. (1999). A bibliometric study of reference literature in the sciences and social sciences. V *Information Processing & Management* (str. 31–44).

Hicks, D. (2004). The four literatures of social science. V *Handbook of quantitative science and technology research* (str. 473–496).

Huang, M. H., & Chang, Y. W. (2008). Characteristics of research output in social sciences and humanities: From a research evaluation perspective. *Journal of the American Society for Information Science and Technology*, *59*(11), 1819–1828.

Južnič, P. (2017). *Bibliometrijski indikatorji*. Pridobljeno s https://www.youtube.com/watch?v=l9W5glZl97I&feature=youtu.be

Kousha, K., Thelwall, M., & Rezaie, S. (2011). Assessing the citation impact of books: The role of Google Books, Google Scholar, and Scopus. *Journal of the American Society for information science and technology*, *62*(11), 2147–2164.

Lazarević, Ž., & Zemljič, I. (2003). *Slovenski zgodovinarski indeks citiranosti – izhodišča in pomisleki*. [Neobjavljena dokumentacija.]. Ljubljana: Inštitut za novejšo zgodovino.

MODS User Guidelines, Version 3 (Metadata Object Description Schema). Pridobljeno s https://www.loc.gov/standards/mods/userguide/introduction.html

Nederhof, A. (2006). Bibliometric monitoring of research performance in the social sciences and the humanities: A review. *Scientometrics*, *66*(1), 81–100.

Pajić, D. (2015). Globalization of the social sciences in Eastern Europe: genuine breakthrough or a slippery slope of the research evaluation practice? *Scientometrics*, *102*(3), 2131–2150.

Pančur, A. (2019a). *Preprosta raziskovalna infrastruktura za kompleksne raziskovalne podatke v humanistiki – si4 (Simple research Infrastructure FOR complex research data in digital humanities)*. [Neobjavljena dokumentacija.]

Pančur, A. (2019b). *Specifikacije za izvedbo naročila izdelave Zgodovinarskega indeksa citiranosti (ZIC)*. [Neobjavljena dokumentacija.]

Pančur, A., & Šorn, M. (2019). Na začetku je bil SIstory: raziskovalna infrastruktura slovenskega zgodovinopisja. V J. Hadalin in Ž. Lazarević (ur.), *Inštitut za novejšo zgodovino: 60 let mislimo preteklost* (str. 47–58). Ljubljana: Inštitut za novejšo zgodovino.

Pančur, A., Šorn, M., & Hadalin, J. (2014). Slovenski indeks citiranosti (SICI): Načrt izgradnje in delovanja. Tehnično poročilo. Pridobljeno s https://www.sistory.si/11686/36153

What is ElasticSearch. Pridobljeno s https://www.elastic.co/what-is/elasticsearch

THE HISTORIOGRAPHY CITATION INDEX UPGRADE

The fields of humanities and social sciences are often deprived of inclusion within the international citation indexes such as Scopus and Web of Science (WOS). The reason for this offshift in the indexes are commonly associated with the format of published works, e.g. the most common type of published works in humanities are monographs (though the scientific journals are on the rise), which are not typically included in WOS and Scopus. Even though Scopus is far more inclusive of such types and fields in comparison to WOS, there is still a gap to be filled. As a response to this predicament the Institute of Contemporary History developed its own citation index – the Historiography Citation Index (HCI), which was first meant to only track the research production within the institution, but has since been expanded to cover the production of the whole field of Slovene historiography. Over the years HCI was a subject of several upgrades and data harmonization attempts. Even with the upgrades, several shortcomings of the systems were apparent, and therefore, another upgrade was taken into consideration, and after the extensive analysis was performed, we identified the most problematic aspects of the index and began working on another upgrade.

The upgrade was performed in two parts – in the first one, we took upon ourselves to improve the administrative system in which we implemented the ElasticSearch technology to improve our search engine and filtration system, as well as improving the data masks to increase the precision and accuracy of the data input into the index. As a part of the administrative system upgrade we also modeled the MODS application profile to increase the interoperability of our data and therefore, enabling the exchange of our data between different information systems without losing data and its context. In the second part, we upgraded the user interface of the citation index to be more user friendly. In order to increase the coherence of the data display, we implemented a table-like design of the search result, equipped with filters in each column. To increase the visibility of the most important factor of the citation index, number of citations the work has received, we included additional column just for that information. The index aims to enable researchers access to the information on the number of citations, cited works ect. It is also recognised by the Slovenian Research Agency (ARRS) as a valid source of citations and could be used to provide proof

of the researchers achievements and scientific excellency, though it is still not recognised as equal to the SICRIS information system.

With the upgrade we increased the efficiency of the citation index, as well as its usability, and with it ensured a more intuitive system to its indexators and users.

Keywords: the Historiography Citation Index, HCI, upgrade, citation indexes