Slovenski indeks citiranosti (SICI)

Načrt izgradnje in delovanja

Andrej Pančur, Mojca Šorn in Jurij Hadalin

2014-09-16

Raziskovalna infrastruktura slovenskega zgodovinopisja Inštituta za novejšo zgodovino

Uvod

Leta 2007 smo na Inštitutu za novejšo zgodovino v okviru IP začeli projekt Zgodovinarski indeks citiranosti (ZIC). V bazo smo vpisovali citate iz slovenskih zgodovinskih monografij, zbornikov in serijskih publikacij, saj smo želeli omogočiti pregled nad zgodovinopisno znanstveno produkcijo oziroma seznanitev z vsebino citiranih del. Popis citatov ZIC je prosto dostopen na spletnem portalu Zgodovina Slovenije – SIstory; preko iskalnika (http://sistory.si/zic) lahko uporabniki s pomočjo iskalnih nizov dostopajo do podatkov o delih, ki so citirala njihova dela.

Prvotna enostavna shema relacijske baze je ob svojem nastanku dobro služila prvotnim potrebam, toda z leti so se vzporedno s hitrim naraščanjem števila vpisanih podatkov pokazale nekatere večje pomanjkljivosti obstoječe sheme relacijske baze in celotnega delovnega procesa, zaradi česar je ZIC nujno potrebno nadgraditi.

Trenutno potekajo intenzivna dela pri čiščenju in povezovanju podatkov iz ZIC baze v novo bazo citatov. Na podlagi na novo vzpostavljenih relacij je že mogoče opraviti osnovne bibliometrične izračune o citiranosti slovenskih zgodovinarjev.

Nova baza Slovenski indeks citiranosti – Zgodovinopsije (SICI-Zgod) je implementacija podatkovnih baz in orodij SICI, ki procesira podatke iz publikacij iz področja humanistike (po vrstilcu UDK), katerih avtorji so raziskovalci iz področja zgodovinopisja (glede na raziskovalno šifro).

SICI-Zgod bi bil optimalen model tudi za:

- celotno humanistiko: Slovenski indeks citiranosti – Humanistika (SICI-Hum). Pomenil bi implementacijo podatkovnih baz in orodij SICI, ki bi procesiral podatke iz publikacij s področja humanistike (po vrstilcu UDK), katerih avtorji so raziskovalci iz humanističnih ved (glede na raziskovalno šifro),

- celotno raziskovalno polje v RS: Slovenski indeks citiranosti (SICI). Predstavljal bi skupek podatkovnih baz ter programskih orodij za luščenje (extraction), obdelavo, vnašanje, analizo, prikaz in vizualizacijo podatkov o citatih iz znanstvenih publikacij, ki so bile izdane v Republiki Sloveniji, oziroma so njihovi avtorji raziskovalci, registrirani v Sloveniji.

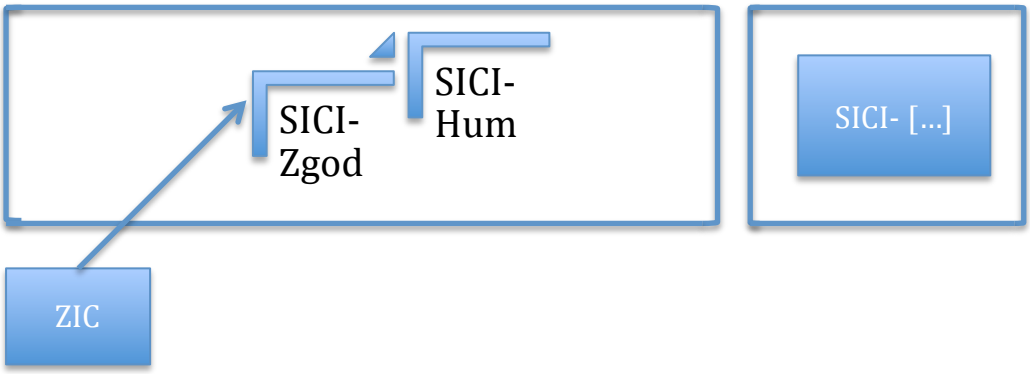
Osnovne komponente Slovenskega indeksa citiranosti (SICI)

Slovenski indeks citiranosti (SICI), angleško Sloveninan Citation Index (SICI), je skupek podatkovnih baz ter programskih orodij za luščenje (extraction), obdelavo, vnašanje, analizo, prikaz in vizualizacijo podatkov o citatih iz znanstvenih publikacij, izdanih v Republiki Sloveniji, oziroma so njihovi avtorji raziskovalci, registrirani v RS.

Slovenski indeks citiranosti – Humanistika (SICI-Hum), angleško Slovenian Citation Index – Arts & Humanities (SICI-A&H), je implementacija podatkovnih baz in orodij SICI, ki procesira podatke iz publikacij s področja humanistike (po vrstilcu UDK), katerih avtorji so raziskovalci humanističnih ved (glede na raziskovalno šifro).

Slovenski indeks citiranosti – Zgodovinopsije (SICI-Zgod), angleško Slovenian Citation Index – Historiography (SICI-Hist) je implementacija podatkovnih baz in orodij SICI, ki procesira podatke iz publikacij s področja humanistike (po vrstilcu UDK), katerih avtorji so raziskovalci s področja zgodovinopisja (glede na raziskovalno šifro).

Shema 1: Slovenski indeks citiranosti (SICI)



Osnova za izgradnjo podatkovnih baz SICI-Zgod bodo podatki, pridobljeni v okviru projekta Zgodovinarski indeks citiranosti (ZIC) – 81.500 citatov iz 3542 del. S postopnim vključevanjem podatkov iz publikacij iz ostalih področij humanistične vede, lahko SICI-Zgod organsko preraste v SICI-Hum. Poleg humanistične pa lahko tudi ostale vede uporabijo podatkovne baze in orodja Slovenskega indeksa citiranosti (SICI) pri izgradnji lastnih področnih indeksov citiranosti.

Vzroki za nadgradnjo ZIC v SICI-Zgod

V ZIC MySQL relacijsko podatkovno bazo se podatki o citatih vnašajo ročno z vpisovanjem v polja uporabniškega vmesnika. V polja za vnašanje podatkov se lahko podatke tudi kopira (copy-paste) iz besedila pridobljenega s pomočjo optičnega prepoznavanja znakov (OCR). V relacijski bazi dobi vsaka publikacija,

iz katere se izpisujejo citati, svojo unikatno identifikacijsko številko, na katero so vezani vpisani metapodatki o tej publikaciji. Citati iz te publikacije pri vpisu dobijo svojo identifikacijsko številko, ki je vedno vezana na publikacijo, iz katere je bil citat prepisan. Vsak citat ima svoje metapodatke.

Takšna enostavna shema relacijske baze ZIC je ob svojem nastanku leta 2007 dobro služila prvotnim potrebam. Preko iskalnika (http://sistory.si/zic) lahko raziskovalci s pomočjo iskalnih nizov dostopajo do podatkov o delih, ki so citirala njihova dela. Toda z leti so se vzporedno s hitrim naraščanjem števila vpisanih podatkov pokazale nekatere večje pomanjkljivosti obstoječe sheme relacijske baze in celotnega delovnega procesa, zaradi česar je ZIC nujno potrebno nadgraditi.

Pomanjkljivosti obstoječega ZIC so:

- Metapodatki o publikaciji, ki so že bili vneseni v ZIC bazo, se v primeru, ko je bila ista publikacija ponovno citirana v drugi publikaciji, še enkrat vnesejo v ZIC bazo. Ker v bazi ni relacij (preko unikatnih identifikacijskih oznak) med isto publikacijo, ki je bila citirana v različnih publikacijah, je iskanje teh relacij v prvi vrsti odvisno od čim bolj poenotenega zapisa metapodatkov o tej publikaciji.
- Toda pri vpisovanju metapodatkov lahko pride do različnih napak in nedoslednosti:
    - različne publikacije lahko druge publikacije navajajo na različne načine in z različnimi podatki;
    - prvotni podatki o citiranih publikacijah so lahko napačni ali slovnično nepravilno zapisani;
    - OCR digitaliziranih podatkov ni brez napak, pri čemer se pri slabše digitaliziranih publikacijah nepravilnost zapisa samo še proporcionalno stopnjuje;
    - vpisovalec v bazo lahko podatke o citirani publikaciji napačno prepiše ali vnese v napačno vpisno polje.
- Zaradi neobstoječih relacij med zapisi iste publikacije in zaradi ne dovolj kvalitetnih zapisov so obstoječi podatki v bazi ZIC premalo natančni za izvajanje zanesljivih bibliometričnih raziskav.

Temeljne rešitve za odpravo pomanjkljivosti:

- Izgradnja nove SICI relacijske baze citatov, v kateri bo vsako citirano delo enotno identificirano samo enkrat.
- Shema relacijske baze bo prilagojena dejstvu, da so vhodni podatki lahko tudi napačni.
- Glavni namen SICI baze ne bo vnašanje čim bolj pravilnih opisnih metapodatkov o citiranih publikacijah, temveč vpisovanje čim bolj natančnih unikatnih identifikacijskih oznak in vzpostavljanje čim bolj natančnih relacij med njimi.
- Preverjanje pravilnosti vhodnih podatkov o publikacijah in relacijah med njimi bo temeljilo na širokem naboru kontrolnih podatkov.

Trenutno že potekajo intenzivna dela pri čiščenju in povezovanju podatkov iz ZIC baze v novo SICI bazo citatov. Na podlagi na novo vzpostavljenih relacij je že

mogoče opraviti osnovne bibliometrične izračune o citiranosti slovenskih zgodovinarjev.

Potencialni vhodni podatki

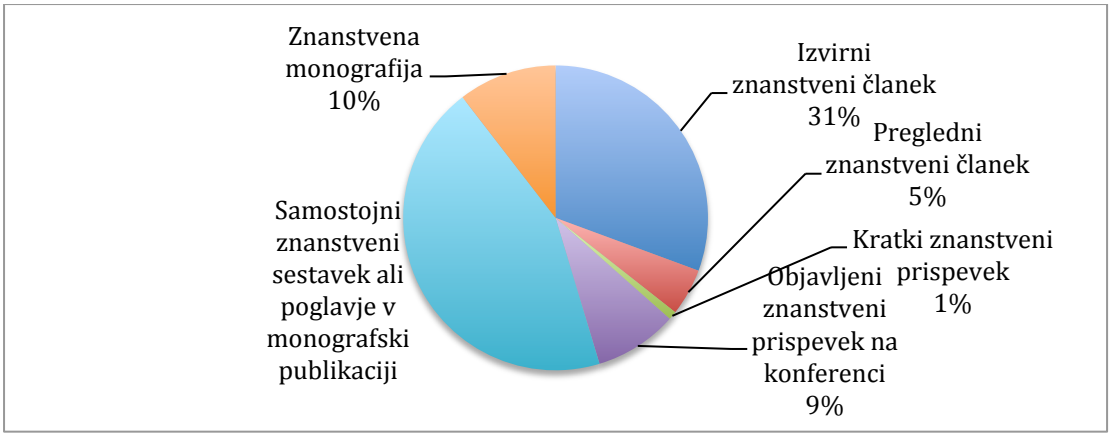
SICI-Zgod in SICI-Hum bosta vhodne podatke o citatih zajemala iz znanstvenih del, ki jih Bibliografska merila znanstvene in strokovne uspešnosti upoštevajo pri kvantitetnem ocenjevanju znanstvene uspešnosti:

- izvirni znanstveni članki (COBISS tip 1.01),
- pregledni znanstveni članki (COBISS tip 1.02),
- kratki znanstveni prispevki (COBISS tip 1.03),
- objavljeni znanstveni prispevki na konferenci (COBISS tipa 1.06 in 1.08),
- samostojni znanstveni sestavki ali poglavja v monografskih publikacijah (COBISS tip 1.16),
- znanstvene monografije (COBISS tip 2.01).

Vhodne podatke, ki pridejo v poštev pri izgradnji podatkovne baze citatov SICI-Zgod, bi bilo tako v primeru desetletnega obdobja 2004-2013 potrebno zajemati iz 4025 znanstvenih del, od katerih je znanstvenih člankov zgolj dobro tretjino (gl. Grafikon 1). Podatki iz približno treh četrtih[1] teh publikacij so že vpisani v bazo ZIC, podatki iz preostalih publikacij pa bodo po nadgradnji ZIC v SICI-Zgod prioritetno vpisani v novo relacijsko SICI bazo citatov.

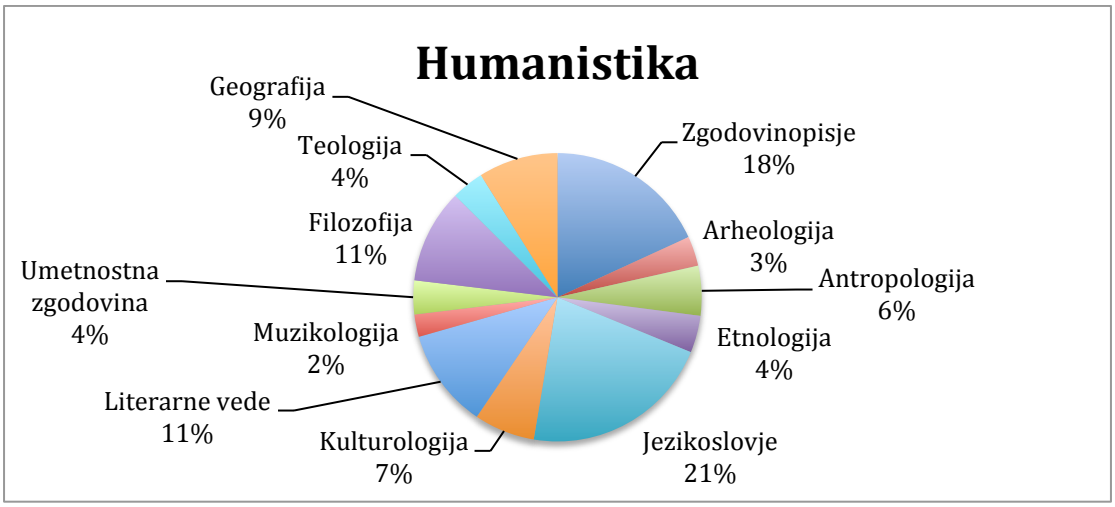
[1] Pred dokončnim prečiščenjem podatkov iz baze ZIC je možno podati zgolj približno oceno. Čeprav so trenutno v bazi ZIC vpisani citati iz 3542 publikacij, kar je kvantitativno gledano 88 % vseh zgodovinskih znanstvenih publikacij, ki so izšle v letih 2004-2013, pa ZIC baza vključuje tudi citate iz nezgodovinskih znanstvenih publikacij ter citate iz starejših publikacij. Med leti 2011 in 2013 so se projektu ZIC pridružili Inštitut za narodnostna vprašanja, Arhiv Republike Slovenije in Oddelek za muzikologijo Filozofske fakultete Univerze v Ljubljani, tako so njihovi knjižničarji v bazo ZIC vpisovali tudi podatke o citatih iz publikacij s širšega področja humanistike in družboslovja. Del baze ZIC so tudi podatki selektivnih izpisov citiranih del sodelavcev Inštituta za novejšo zgodovino (INZ) iz domačih, starejših in celo tujih publikacij. Ker iz teh publikacij niso bili izpisani vsi citati, predstavljajo izpisani citati iz teh publikacij motnjo pri izračunavanju znanstvene uspešnosti raziskovalcev, saj prefererajo raziskovalce z INZ. Zato bodo podatki o teh citatih izločeni oziroma ne bodo uvoženi v novo bazo SICI. Trenutno so v bazi ZIC vpisani citati iz 223 takšnih publikacij, ki so posebej označeni kot BazaINZ. Ker je bilo v procesu čiščenja in povezovanja obstoječih podatkov baze ZIC ugotovljeno, da je zaradi pogostih človeških napak (vpisovalec podatkov je lahko publikacijo nepravilno označil, da je BazaINZ, oziroma je še pogosteje pozabil označiti, da je BazaINZ) ta klasifikacija povsem neuporabna, bo potrebno še enkrat ročno pregledati večino spornih primerov.

Grafikon 1: Znanstvena dela s področja zgodovinopisja (glede na raziskovalno šifro prvega avtorja) izdana v letih 2004-2013; % glede na COBISS tip



V primeru širitve nabora vhodnih podatkov iz publikacij, kateri avtorji so zgodovinarji (SICI-Zgod) na vse pri ARRS registrirane avtorje s področja humanistike (SICI-Hum), se temu ustrezno poveča tudi sama količina vhodnih podatkov, primernih za vnos v bazo SICI. Podatki vzorčnega leta[2] nam tako kažejo, da so v letu 2013 od skupno 1951 znanstvenih del le 18 % vseh del s področja humanistike izdali zgodovinarji (gl. Grafikon 2). V tabeli Znanstvena dela po področjih humanistike (upoštevan prvi avtor) izdana v letu 2013; število izdanih del, so ti podatki bolj nadrobno prikazani.

Grafikon 2: Znanstvena dela s področja humanistike (glede na raziskovalno šifro prvega avtorja) izdana leta 2013; % glede na področja humanistike



[2] Da bi dobili boljšo predstavo o tem, za koliko bi se povečala količina vhodnih podatkov, če bi SIC-Zgod nadgradili v SIC-Hum, smo zbrali podatke o vseh publikacijah, ki so jih v letu 2013 izdali avtorji s področja humanistike. Pri tem smo upoštevali javno dostopne podatke o raziskovalcih in znanstvenih delih iz SICRIS-a. Da ne bi prišlo do podvajanj pri štetju istih publikacij, smo vedno upoštevali samo prvega avtorja.

Table 1: Znanstvena dela po področjih humanistike (upoštevan prvi avtor) izdana v letu 2013; število izdanih del

	Zgodovinopisje	Arheologija	Antropologija	Etnologija	Jezikoslovje	Kulturologija	Literarne vede	Muzikologija	Umetnostna zgodovina	Filozofija	Teologija	Geografija	skupaj
Izvirni znanstveni članek (1.01)	103	24	45	30	149	64	87	19	26	91	26	50	714
Pregledni znanstveni članek (1.02)	9	1	3	5	7	6	19	2	3	2	9	24	90
Kratki znanstveni prispevek (1.03)	3	1	5	0	2	3	1	0	1	1	0	5	22
Objavljeni znanstveni prispevek na konferenci (1.06, 1.08)	39	5	10	5	65	6	25	18	4	20	7	14	218
Samostojni znanstveni sestavek ali poglavje v monografski publikaciji (1.16)	149	23	36	23	180	38	65	9	25	56	26	56	686
Znanstvena monografija (2.01)	51	11	10	18	15	16	20	1	15	36	3	25	221
Skupaj	354	65	109	81	418	133	217	49	74	206	71	174	1951

Še bolj natančno predstavo o predvideni količini vhodnih podatkov za SICI-Hum v enem letu pa dobimo, če upoštevamo skupno število citatov iz teh skoraj dva tisoč publikacij. V povprečju namreč posamezna znanstvena monografija vsebuje veliko več citatov kot pa znanstveni članek ali poglavje. Glede na podatke iz baze ZIC povprečna zgodovinska monografija citira 153 različnih virov literature, članek 23 in poglavje 20. Glede na te podatke bi bilo torej pri prehodu iz SICI-Zgod na SICI-Hum na letni ravni potrebno v bazo SICI vnesti podatke in urediti ustrezne relacije za dobrih 70.000 v citatih navedenih virov literature (gl. Table 2).

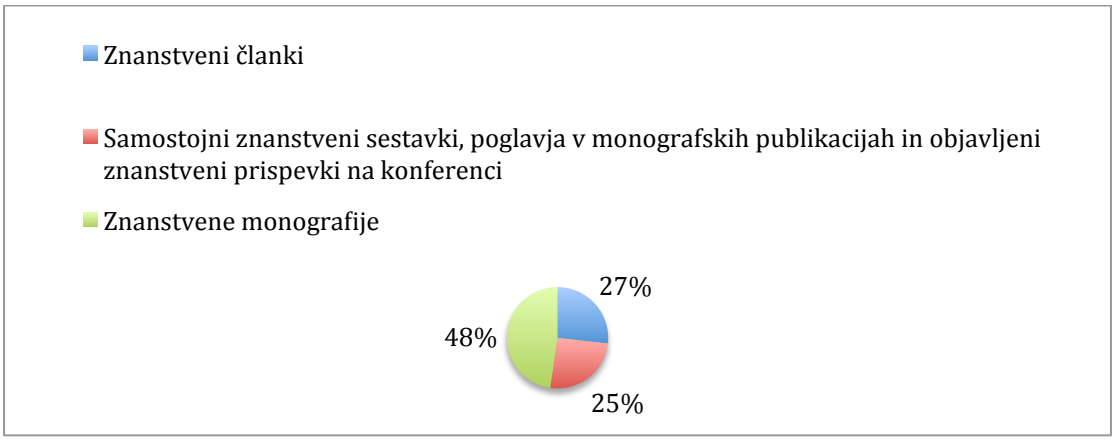
Table 2: Ocena števila citiranih virov literature v znanstvenih delih s področja humanistike, izdanih v letu 2013.

	število del	povprečno število citiranih virov na delo	skupno število citiranih virov
Znanstveni članki	826	23	18.998
Samostojni znanstveni sestavki, poglavja v monografskih publikacijah in objavljeni znanstveni prispevki na konferenci	904	20	18.080
Znanstvene monografije	221	153	33.813
Skupaj	1951		70.891

V to število niso všteti samo viri literature, katerih avtorji so v Sloveniji registrirani raziskovalci s področja humanistike, temveč tudi dela vseh ostalih citiranih slovenskih raziskovalcev ter dela vseh ostalih slovenskih in tujih citiranih avtorjev. Pri vnašanju vhodnih podatkov v SICI bazo citatov se ne bo ugotavljajo, katero od citiranih del je znanstveno in še manj, katero od citiranih del je napisal registrirani raziskovalcev s področja humanistike, temveč se bo neselektivno vnašalo vso v citatih navedeno literaturo. Selekcija med citati bo narejena naknadno s pomočjo podatkov, pridobljenih iz drugih baz podatkov, predvsem iz SICRIS baz. Povezovalna identifikacijska oznako med SICI bazo citatov in SICRIS bo COBISS:ID.

Na podlagi teh podatkov lahko torej trdimo, , da je za večjo zanesljivost merjenja znanstvene uspešnosti v humanistiki poleg citatov iz znanstvenih revij nujno potrebno upoštevati še citate iz ostalih znanstvenih del. Članki v znanstvenih revijah namreč zajemajo samo dobro četrtino vseh citatov, medtem ko monografije zajemajo skoraj polovico vseh citatov. Povsem enako težo kot znanstveni članki pa imajo samostojni znanstveni sestavki, poglavja v monografskih publikacijah in objavljeni znanstveni prispevki na konferencah (gl. Grafikon 3), ki jih tuji indeksi (z delno izjemo prispevkov na konferencah) ne upoštevajo.

Grafikon 3: Ocena števila citiranih virov literature v znanstvenih delih s področja humanistike, izdanih v letu 2013



Računalniško podprto procesiranje podatkov

Zaradi relativno zelo velike količine podatkov o citiranosti, ki jih vsebujejo znanstvene publikacije s področja humanistike, bi samo ročno prepisovanje podatkov o citatih v SICI bazo citatov zahtevalo zelo veliko količino dela in s tem povezane (pre)visoke stroške delovanja Slovenskega indeksa citiranosti za Humanistiko (SICI-Hum). Če pri prehodu iz Zgodovinarskega indeksa citiranosti na SICI-Hum ne bi opustili dosedanjih postopkov ročnega vnašanja podatkov, bi se potrebe po delovni sili od dosedanje ene osebe (zunanji sodelavec) povečale za najmanj petkrat. Zato bo postopek vnašanja podatkov v bazo v prihodnosti nujno potrebno podpreti z uporabo razpoložljivih programskih orodij. V primeru, da ta orodja še ne obstajajo, oziroma njihovi algoritmi ne podpirajo slovenščine, jih bo potrebno razviti.

Prednosti računalniško podprtih delovnih postopkov pri pridobivanju in procesiranju podatkov za SICI bazo citatov so v primerjavi z ročnim vnašanjem več kot očitne:

- hitrejše vnašanje podatkov;
- natančnejše vnašanje podatkov;
- nižji stroški delovne sile/vnašalcev podatkov;
- razvoj novih tehnologij in postopkov.

Hkrati pa ima ta sistem seveda tudi nekatere pomanjkljivosti:

- Predvsem v prvih letih, ko programska orodja še ne bodo dovolj dobro razvita, bo podatke potrebno vnašati delno ročno. Ker bo ročno vnašanje potekalo v okolju, za katerega je nujno osnovno poznavanje programskih jezikov za delo z besedili in XML zapisi (regularne izraze, Python ali Perl, XSLT, xQuery, xPath), sodelovanje z (občasnimi) vpisovalci podatkov iz sorodnih ustanov ne bo več ustrezno. Za optimalne rezultate vnašanja podatkov v bazo SICI bi potrebovali eno (novo) osebo in polovični delovni čas druge (že realizirano v času trajanja projekta ZIC) na matični ustanovi.

- Potrebna bo zaposlitev kvalificirane osebe z računalniško izobrazbo, s katero INZ trenutno ne razpolaga.
- Večino programskih orodij bo potrebno šele razviti.

Dolgoročno pa bodo te pomanjkljivosti imele veliko pozitivnih učinkov za razvoj raziskovalne infrastrukture slovenskega zgodovinopisja, širše infrastrukture za humanistiko in celo širšo interdisciplinarno povezovanje humanistike z ostalimi vedami (strojno učenje, jezikovne tehnologije, vizualizacija podatkov):

- Razvojni sodelavec z računalniško izobrazbo bo za Raziskovalno infrastrukturo INZ pomenil večjo interdisciplinarno vpetost.
- Potreba po ročnem vnašanju podatkov se bo z leti zaradi razvoja programske opreme hitro zmanjševala. Zato se bodo ta sredstva v vedno večji meri lahko usmerjala v dodatni tehnološki razvoj.
- Ker bo celoten projekt usmerjen prvenstveno v razvoj programskih orodij, ko bodo uporabna tudi za širšo evropsko skupnost digitalne humanistike, bo IP INZ večino potreb po razvojnem denarju lahko pokril s sredstvi, namenjenemi za razvoj v okviru DARIAH (Digital Research Infrastructure for the Arts and Humanities).
- V okviru projekta SICI bodo nastali obsežni korpusi ročno označenih besedil. Te se bo uporabilo pri razvoju programov, ki bodo temeljili na strojnem učenju.
- Ti korpusi bodo uporabni tudi pri razvoju jezikovnih tehnologij za slovenski jezik.
- Podatki iz baze citatov bodo v kombinaciji z ostalimi podatkovnimi bazami omogočili nove raziskave na področju analize in vizualizacije podatkov socialnih omrežij.

Vrste vhodnih podatkov

Predpogoj za vzpostavitev uspešnega delovnega postopka z računalniško podprtim procesiranjem vhodnih podatkov je čim bolj enostavna in zanesljiva pridobitev teh podatkov v elektronski obliki. Do sedaj smo veliko večino literature, iz katere so se podatki o citatih prepisovali v ZIC, dobili v relativno zelo dobro založeni knjižnici Inštituta za novejšo zgodovino. V prihodnje pa bo potrebno publikacije za ostala področja humanistike pridobili v drugih zunanjih knjižnicah ali pa občutno povečati sredstva za nakup te dodatne literature. Zaradi čim nižjih stroškov delovanja Slovenskega indeksa citiranosti za Humanistiko bi bilo zato dobrodošlo, da bi se sistemsko uredil začasen dostop do humanističnih publikacij, ki ne bi bil povezan z večjimi dodatnimi stroški. Obenem bi bilo zaželeno, da bi čim več teh publikacij dobili že v elektronski obliki, pri čemer naj bi bilo tudi že izvedeno optično prepoznavanje besedila (OCR). Še bolj optimalno bi bilo, da bi v primeru, ko so bile znanstvene publikacije narejene na podlagi osnovnega XML dokumenta, ki označuje tudi citate in vire literature,[3] od založnikov lahko dobili ta XML.

[3] Prim. npr.: http://sistory.eu/TEI_publikacije/monografije/Gasparic_Parlamentaria1/bibliogr.html

Obstajata dva prevladujoča načina navajanja uporabljenih virov v znanstvenih publikacijah, ki bosta prišla v poštev pri vnašanju vhodnih podatkov v SICI bazo citatov:

1) Sprotne opombe: Podatke o viru se v besedilu sproti citira v opombi pod črto ali v končni opombi. Pri prvem citatu vira se navede vse zahtevane (s strani izdajatelja) podatke o viru, pri morebitnih naslednjih citiranjih istega vira pa se nato navede okrnjene podatke o tem viru. Način navajanja virov v sprotnih opombah se deli na dve večji skupini:
   a) Na koncu publikacije so vsi uporabljeni viri še enkrat navedeni v seznamu bibliografije.
   b) V publikaciji vsi uporabljeni viri niso posebej navedeni v skupnem seznamu.
2) Seznam citiranih virov (referenc): Na koncu publikacije je naveden seznam vseh navajanih virov, ki vsebuje vse zahtevane (s strani izdajatelja) podatke o teh virih. V besedilu (in-text citation) pa se ta vir citira z vnaprej določeno identifikacijsko oznako, ki se sklicuje na oznako tega vira v seznamu virov:
   a) Citiran priimek avtorja in letnica izdaje, od besedila ločena z oklepajem. V humanistiki se v redkih primerih ta način lahko kombinira tudi z zgoraj navedenim načinom navajanja v sprotnih opombah, pri čemer se v opombah navaja samo okrajšane oznake virov, ki se navezujejo na oznake v seznamu virov.
   b) Citirana zaporedna številka vira v seznamu, od besedila ločena z oglatim oklepajem.
   c) Citirane druge unikatne identifikacijske oznake.

Tako v zgodovinopisju kot v humanistiki so uveljavljeni prav vsi ti načini citiranja, kar bo potrebno upoštevati pri izdelavi programskih orodij za vnos podatkov.

Obenem obstajajo tudi zelo različni načini zapisa podatkov o navedenem viru. Pri tem sicer obstajajo nekatera splošno sprejeta širša pravila, toda v podrobnostih se lahko posamezni načini navajanj podatkov o viru med seboj razlikujejo. V praksi vsako uredništvo določi svoja natančna pravila.

Glede na delovni postopek v sistemu SICI bodo ti različni načini citiranja deljeni na dve skupini:

1. Sprotne opombe, ki imajo na koncu bibliografijo in na sezname citiranih virov: Digitalizirane bibliografije in seznami bodo pretvorjeni v XML in posamezni metapodatki o publikacijah najprej avtomatsko čim bolj natančno označeni (avtorji, naslovi, letnice in kraji izdaj, založbe, urednike, strani, izdaje itd.). Metapodatki v seznamih navedenih publikacij bodo sprva glede na kontekst in ločila označeni s pomočjo regularnih izrazov. Pri tem se bo sproti izgrajevala knjižnica regularni izrazov za označevanje navedene literature: SICI Regular Expression Library for Annotation of References (RegexLib-ARef). Ker bodo tako označena besedila sproti vedno ročno pregledana in dopolnjena, bodo v bazo citatov uvoženi le skoraj povsem zanesljivi podatki. Na ta način se bo sproti oblikoval vedno večji korpus označenega besedila, ki ga bo mogoče uporabiti kot učne podatke za algoritme strojnega učenja: SICI Machine

Learning for Automatic Annotation of References (ML-AARef). Ko se bo to orodje za avtomatično označevanje seznamov citirane literature s pomočjo strojnega učenja izkazalo za dovolj zanesljivo, bo ta postopek lahko povsem avtomatiziran.

2. Sprotne opombe brez seznama citirane literature: Pri današnjem razvoju tehnologije na področju strojnega učenja in v primerjavi z angleščino predvsem veliko slabše razvitih jezikovnih tehnologij, bo delovni postopek v teh primerih še vedno potekal v večji meri ročno. Šele ko se bo v okviru ML-AARef sčasoma nabral dovolj velik korpus označenega besedila, ki ga bo mogoče učinkovito kombinirati s širokim naborom podatkov o publikacijah iz baze citatov, bo mogoče izdelati tudi prve prototipe za avtomatično označevanje citatov znotraj besedila: SICI Machine Learning for Automatic Annotation of In-Text Citatons (ML-AACit). Poleg strojnega učenja bo ta sistem kombiniral še eksplicitno predznanje iz vseh baz podatkov sistema SICI.

Količina ročnega dela, potrebnega pri vnašanju vhodnih podatkov v SICI bazo citatov, bi se lahko občutno zmanjšala, če bi vse znanstvene publikacije, izdane v Sloveniji,vsebovale seznam citirane literature. To navodilo bi bilo potrebno upoštevati tudi pri samostojni znanstvenih sestavkih ali poglavjih v znanstvenih monografijah, ki imajo neredko skupne sezname citirane literature, zaradi česar je oteženo enotno avtomatsko procesiranje posameznih poglavij ali sestavkov.

Ker ne pričakujemo, da bo te pomanjkljivosti mogoče nemudoma odpraviti, smo delovni postopek vnašanja podatkov v SICI bazo citatov prilagodili do te mere, da bo nujno potrebno ročno delo pri vnašanju podatkov dolgoročno čim bolj učinkovito izkoriščeno:

1. Načrtno pridobivanje vzorcev, na podlagi katerih ne bo mogoče meriti le skupnega števila citiranih znanstvenih publikacij (vsak citiran članek je štet samo enkrat), temveč tudi frekvenco citiranja vsake posamezne znanstvene publikacije (šteje se, kolikokrat je bilo posamezno delo citirano v posamezni drugi publikaciji). To bi dolgoročno lahko omogočilo izdelavo novih in potencialno bolj natančnih algoritmov za merjenje znanstvene uspešnosti.[4]
2. Če bi se ugotovilo, da je frekvenca citiranja posameznega dela zelo pomemben podatek za merjenje znanstvene uspešnosti, bi bilo na podlagi dodatnega ročnega označevanja besedila (pridobivanje zanesljivih vzorcev) mogoče relativno hitro izdelati tudi nova orodja za avtomatično označevanje (strojno učenje). Takšno avtomatsko označevanje bi bilo mogoče najhitreje vpeljati pri publikacijah, ki imajo na koncu seznam citiranih virov, v besedilu pa točno določene identifikacijske oznake, ki se sklicujejo na ta seznam. V primeru citiranja virov v opombah bi potrebovali veliko večji korpus učnega besedila.

[4] Prim. Wen-Ru Hou, Ming Li in Deng-Ke Niu: Counting citations in texts rather than reference lists to improve the accuracy of assessing scientific contribution. V: Bioessays, 33, 2011, str. 724-727.

Kontrolni podatki

Podatki o citiranih publikacijah vsebujejo veliko manjkajočih ali napačnih podatkov, obenem pa so določeni tipi podatkov v različnih seznamih literature drugače in celo nekonsistentno zabeleženi. Zato lahko pričakujemo, da se bodo tudi pri vnašanju teh podatkov v SICI bazo vnašali nekonsistentni, pomanjkljivi in napačni podatki. Da pa se bo v SICI bazo vnašalo čim manj takšnih podatkov, smo se odločili vpeljati dvojno vrsto kontrole:

- V času vnašanja vhodnih podatkov s pomočjo primerjave in dopolnjevanja vhodnih podatkov s kontrolnimi podatki.
- Po vnosu vhodnih podatkov še z občasnim dodatnim iskanjem duplikatov oziroma z njihovim povezovanjem.

Pridobitev ustreznih kontrolnih podatkov, s pomočjo katerih se bo preverjalo in dopolnjevalo vhodne podatke v SICI bazo citatov, bo eden ključnih dejavnikov pri uspešni izgradnji SICI. Pri tem je na srečo veliko želenih kontrolnih podatkov že dostopnih v elektronski obliki. Ker se bo v SICI bazo citatov vnašalo podatke o vseh citiranih delih, bodo v primeru pridobivanja kontrolnih podatkov prišle v poštev tako domače kot tuje baze podatkov. Spodnji Grafikon 4 shematsko prikazuje pomembnost in količino potencialnih baz s kontrolnimi podatki. Količinsko najmanj je podatkov iz slovenskega COBISS in na njegovi podlagi zgrajenega SICRIS, vendar so ti podatki veliko bolj pomebni pri izgradnji slovenskega indeksa citiranosti kot pa količinsko neprimerno bolj številčni podatki iz tujih baz. Do slednjih lahko brez težav dostopamo s pomočjo protokola OAI-PMH (Open Archives Initiative – Protocol for Metadata Harvesting), vedno več podatkov pa je dostopnih tudi na semantičnem spletu.

Grafikon 4: Potencialne baze podatkov s kontrolnimi podatki za SICI bazo citatov

Prav tako pomembna kot ustreznost, količina in kvaliteta teh podatkov je tudi njihova dostopnost in avtorsko-pravna ureditev. Kot kontrolni podatki SICI baze citatov namreč lahko pridejo v poštev samo podatki, ki so po eni strani tehnično

ustrezno dostopni in po drugi strani tudi avtorsko-pravno urejeni na način, da omogočajo prosti dostop:

- SICRIS: Podatki o osebah in publikacijah so prosto dostopni v XML, skupaj z XML shemami. Nekaj teh podatkov se že uporablja pri konverziji podatkov iz ZIC baze citatov v SICI-Zgod bazo citatov. Zaželeno bi bilo, da bi enkrat do dvakrat na leto pridobili vse te podatke (dump) na enkrat.
- COBISS: Podatki niso splošno dostopni. Preko ARRS bi se lahko zaprosilo za prosti dostop, vendar ni nujno, da bo prošnja odobrena; alternativa je luščenje podatkov (web scraping) celotnega COBISS-a ali sprotno luščenje podatkov vsake posamezne strani izbrane publikacije posebej.
- OAI-PMH: Veliko prosto dostopnih podatkov, vendar večinoma za digitalne oziroma digitalizirane vsebine (Europeana, European Library in še mnogi manjši repozitoriji).
- Semantični splet: Večje nacionalne in druge knjižnice (nemška, britanska, španska, švedska, ameriška kongresna, združenje britanskih raziskovalnih knjižnic predvidoma kmalu sledijo tudi druge) so začele svoje podatke dajati na voljo pod odprtodostopnimi licencami in v formatu RDF. Nekatere dajo na voljo samo podatke, druge so že vzpostavile preko endpointov dostopne triplestore baze.
- Svetovni splet (www): Še veliko več podatkov kot na semantičnem spletu je dostopnih ne svetovnem spletu. Pri tem bodo prišle v poštev predvsem različne specializirane spletne strani, kjer so podatki o publikacijah objavljeni v HTML, ki omogoča luščenje podatkov. Pridobivanje teh podatkov bo potekalo na način, ki je uveljavljen v okviru delovanja prosto dostopnega programa za upravljanje seznamov literature Zotero (https://github.com/zotero).

Avtorske pravice

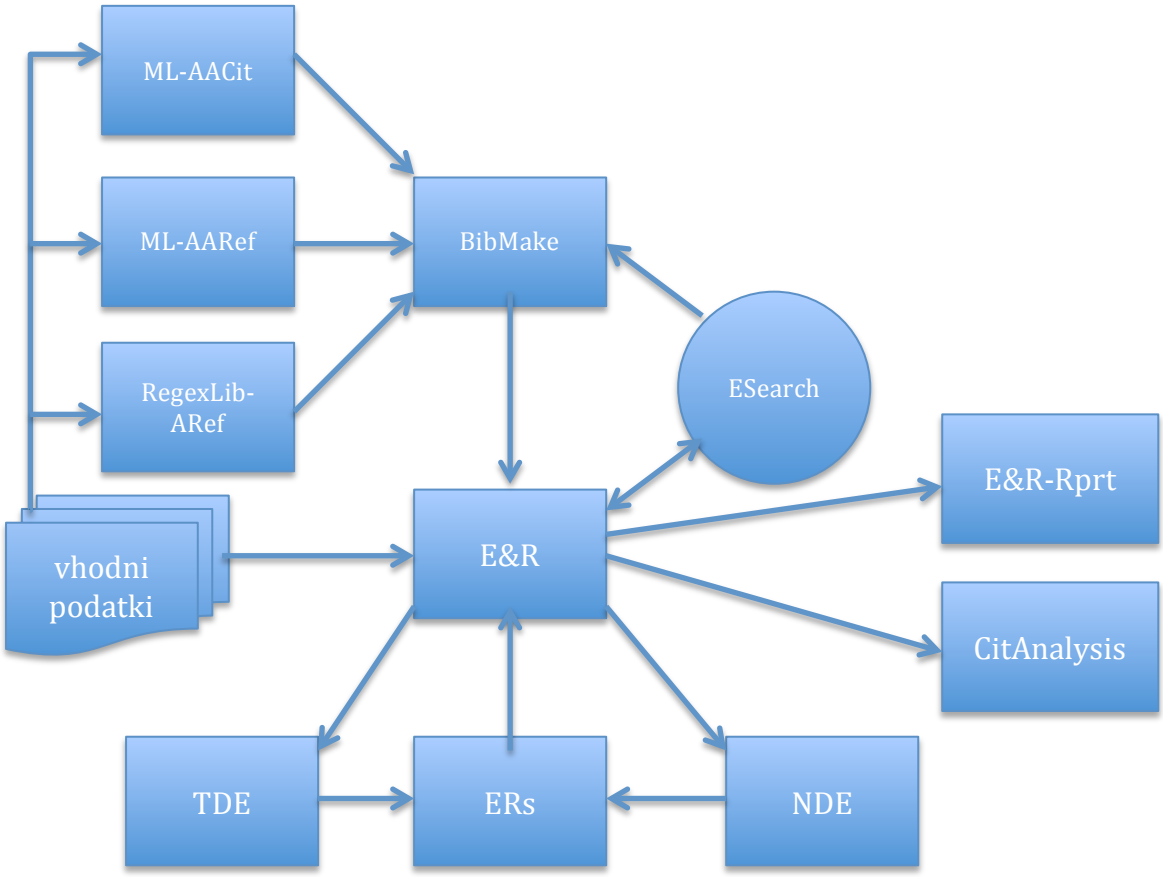
Slovenski indeks citiranosti (SICI) bo svoje podatke in programska orodja javnosti ponujal na razpolago pod odprtimi licencami:

- Vsi metapodatki o citatih publikacij bodo odprto dostopni (open access) za nekomercialno in komercialno uporabo, z možnostjo nadaljnje predelave in z upoštevanjem navajanja avtorstva-izvora. Najbolj primerna bo licenca Creative Commons CC-BY 4.0 (Priznanje avtorstva).
- V okviru SICI razvita programska oprema bo odprtokodna (open source). Koda bo komentirana v angleščini in objavljena na GitHub (https://github.com/). Pri izgradnji SICI bo cilj prevzeti čim več obstoječih odprtokodnih programskih orodij, pri čemer se bo morebitne predelave in izboljšave teh orodij dalo na voljo za nadaljnjo uporabo pod odprtokodno licenco.

Delovni proces in programska orodja SICI

Grafikon 5 shematsko prikazuje celotno vizijo delovnega procesa in vseh programskih orodij, ki bodo omogočili čim bolj avtomatsko procesiranje podatkov. Ker bodo vsa ta orodja pod odprto kodo dostopna na GitHub širši zainteresirani svetovni javnosti kot slovenski prispevek k digitalni humanistiki v okviru projekta SIDIH, so njihova imena v angleščini.

Grafikon 5: Pretok podatkov med programskimi orodji Slovenskega indeksa citatov (SICI)



Orodja za predprocesiranje podatkov o publikacijah:

- SICI Regular Expression Library for Annotation of References (**RegexLib-ARef**): Vhodni podatki, za katere bo narejen OCR, bodo najprej procesirani s pomočjo knjižnice regularnih izrazov.
- SICI Machine Learning for Automatic Annotation of References (**ML-AARef**): Orodje za avtomatično označevanje metapodatkov o citiranih publikacijah, navedenih v seznamu literature.
- SICI Machine Learning for Automatic Annotation of In-Text Citatons (**ML-AACit**): Orodje za avtomatično označevanje citatov v besedilu.
- SICI Bibliography Maker (**BibMake**): Orodje za povezovanje bibliografskih enot, pridobljenih iz RegexLib-ARef, ML-AARef ali ML-AACit z že vnesenimi podatki iz baze citatov SICI.

Orodji za vzpostavljanje relacij med entitetami (publikacijami):

- SICI Entities & Relations Management (**E&R**): Osrednje orodje sistema SICI, preko katerega so posredno ali neposredno povezana vsa ostala orodja. To orodje je uporabniški vmesnik, ki omogoča vzpostavljanje relacij med publikacijami, katerih podatki so bili vneseni v baco citatov. V E&R se podatke lahko vnaša ročno, ali pa se jih uvozi iz BibMake.
- SICI Entities Search Module (**ESearch**): Iskalni modul, ki omogoča iskanje tako po bazi SICI kot po različnih bazah kontrolnih podatkov. Omogoča iskanje v okviru orodij BibMake in E&R.

Orodja za iskanje duplikatov in napačnih relacij med entitetami:

- SICI Tabular Data Explorer (**TDE**): Prikaz podatkov o entitetah v tabelarni obliki. To orodje bo imelo širok nabor vdelanih filtrov in iskalnih algoritmov, ki bodo omogočili čim enostavnejšo primerjavo podatkov o morebitnih duplikatih.
- SICI Network Data Explorer (**NDE**): Vizualizacija mreže razmerij med entitetami, ki bo olajšala iskanje nedoslednosti v podatkih.
- SICI Entity Resolution Management (**ERs**): Orodje za združevanje duplikatov entitet oziroma za razdruževanje napačno identificiranih entitet.

Orodja za analizo podatkov:

- SICI Entities & Relations Report (**E&R-Rprt**): Orodje za opisno statistično analizo podatkov iz baze citatov. Te analize se bo v glavnem uporabljalo v namen spremljanja učinkovitosti delovnega procesa in za poročila financerjem.
- SICI Citation Analysis (**CitAnalysis**): Orodje za bibliometrično izračunavanje raziskovalne uspešnosti. Poleg analiz, ki bi jih potreboval financer, bo to orodje zagotavljalo tudi analizo podatkov v povsem raziskovalne namene.

Dostop do podatkov SICI baze citatov

Ker bodo vsi podatki iz SICI baze citatov odprto dostopni, bodo ti podatki javno objavljeni na načine, ki bodo omogočali čim večjo fleksibilnost pri njihovi nadaljnji uporabili. Ker pa ne obstaja noben idealen način dostopa do podatkov, ki bi lahko zadovoljil vse potrebe uporabnikov teh podatkov, SICI predvideva paleto različnih načinov dostopa do teh podatkov (prim. pot podatkov od baze citatov do izhodnih podatkov, prikazano na Grafikon 6: Shematski prikaz poti podatkov med različnimi bazami v sistemu SICI):

- Izvoz XML: Ker bo delovni proces pri vnašanju podatkov v SICI temeljil tudi na uvažanju podatkov v XML formatu po SICI XML shemi, bo mogoče celotne ali delne podatke tudi izvoziti v XML datoteki. Zato bodo zunanji uporabniki na zahtevo lahko te podatke dobili kot izvoženo XML datoteko. Zaradi relativno velike količine teh podatkov bi v praksi izvoz

XML bolj kot pri izvozu celotne baze prišel v poštev pri izvozu delnih podatkov, primernih za nadaljnjo specialno analizo.

- Izvoz CSV: Delovni proces bo temeljil tudi na stalni kontroli kvalitete že vnesenih podatkov in njihovi delni statistični analizi. V ta namen se bodo uporabljala različna zunanja orodja (mdr. Open Refine, Microsoft Excel, programske knjižnice v jeziku Python in R). CSV (comma-separated values) je podatkovni format, ki podpira uvoz podatkov v vsa ta orodja.
- Aplikacijski programski vmesnik (API): Oddaljen dostop do podatkov SICI baze citatov bo mogoč po registraciji in pridobitvi API-ključa. Po vzoru na Europeano bo iskanje po SICI podatkih omogočil iskalnik Slovenske digitalne infrastrukture za humanistiko in umetnosti (SIDIH) (http://www.sidih.si/). Predvidevamo, da bi preko SICI API do podatkov, potrebnih za bibliometrične izračune, dostopal tudi SICRIS. Uporabnik, ki bo zaprosil za API-ključ, bo lahko do vseh podatkov dostopal na zelo hiter in pregleden način, zaradi česar bo lahko svojo aplikacijo povsem prilagodil svojim potrebam. Glavna pomanjkljivost tega sistema je, da se bo moral pred implementacijo API vmesnika temeljito seznaniti s SICI podatkovno shemo. Zato bo izdelan tudi Applet, ki bo omogočil hitro integracijo iskalnika po vnaprej določenih osnovnih metapodatkih v zunanje spletne strani.
- OAI protokol za avtomatično zajemanje metapodatkov (OAI-PMH): SICI baza podatkov bo tudi OAI-PMH strežnik, s čimer bo zainteresiranim harvesterjem (zajemalcem vsebin) omogočeno avtomatično zajemanje metapodatkov po metapodatkovni shemi Dublin Core. Ker je Dublin Core splošno razširjena in dobro poznana metapodatkovna shema, bo na ta način uporabnik lahko podatke iz SICI enostavno vključil med svoje metapodatke. Glavna pomanjkljivost Dublin Core sheme pa je, da ni najbolj primerna za označevanje citiranosti.
- Dostop preko SPARQL točke (SPARQL endpoint): Podatki iz relacijske SICI baze citatov se bodo z jezikom R2RML dinamično mapirali v RDF podatkovni model v SICI triplestore podatkovno bazo. SPARQL točka SICI triplestore baze ne bo omogočala le iskanja po tej, temveč tudi po vseh ostalih poljubnih svetovnih triplestore bazah. Pri objavi podatkov bodo upoštevane ontologije, ki so izrecno namenjene prikazovanju podatkov o citiranosti, predvsem Citation Typing Ontology (CiTO) in Citation Counting and Context Characterization Ontology (C4O). Edina slabost takšnega dostopa do podatkov je v tem, da so triple store podatkovne baze, SPARQL poizvedovalni jezik ter ontologije semantičnega spleta relativno nove tehnologije in bo potrebno ta sistem dostopa do podatkov vedno znova prilagajati novim rešitvam.

Podatkovne baze SICI

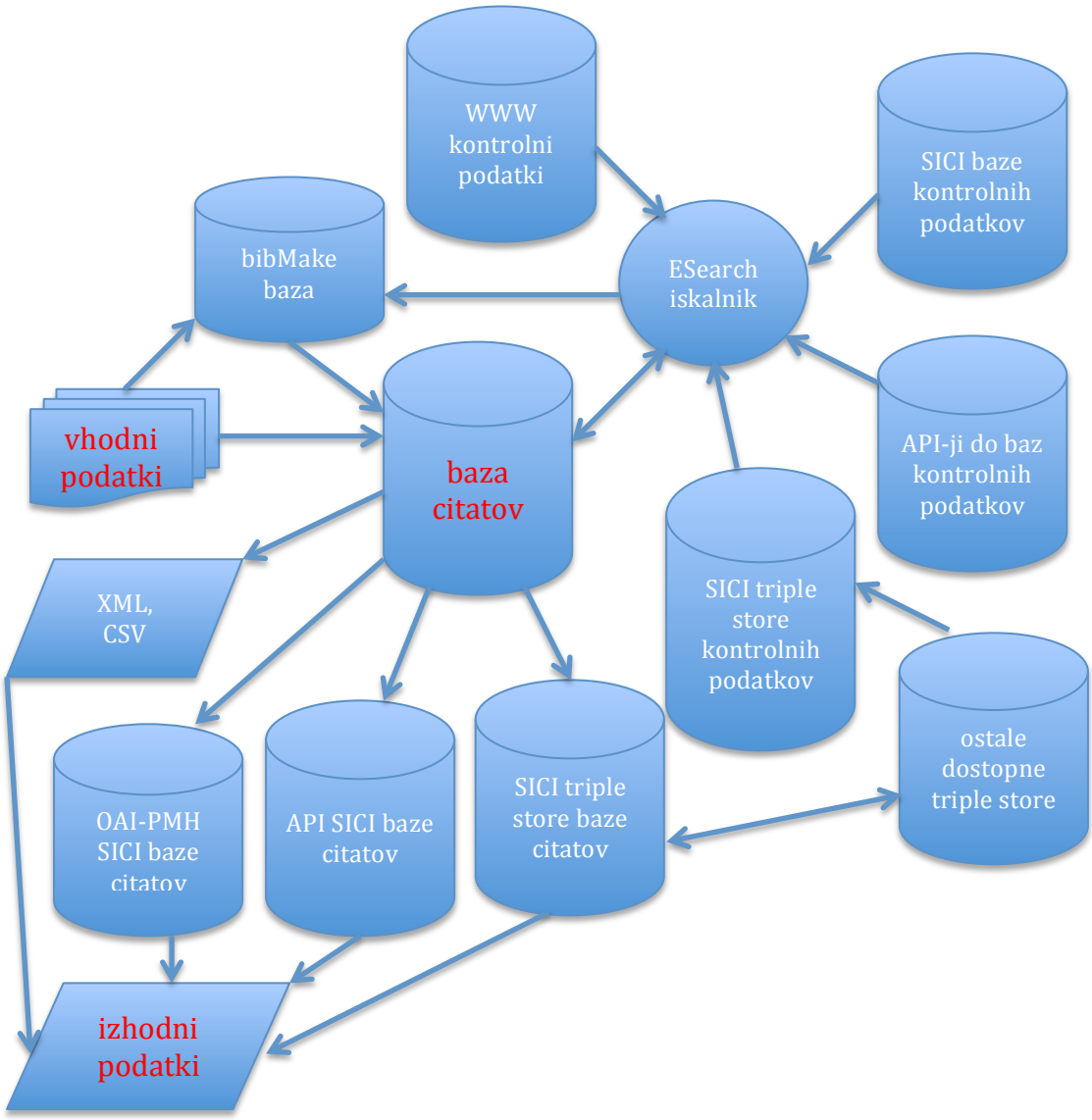
Zaradi zelo različnih vhodnih in kontrolnih podatkov bo SICI združeval več baz podatkov (prim. pot podatkov od začetka poti vhodnih podatkov in od baz podatkov preko ESearch iskalnika do baze citatov, ki ga shematsko prikazuje

Grafikon 6: Shematski prikaz poti podatkov med različnimi bazami v sistemu SICI).

Osrednja baza podatkov je SICI baza citatov. SICI baza citatov je relacijska baza, ki vsebuje podatke o entitetah, ki so med seboj nujno povezane z relacijami. V SICI bazo citatov se ne vnaša entitet brez povezav z ostalimi entitetami.

Baza podatkov orodja Bibliography Maker (BibMake) je prav tako relacijska baza podatkov, ki vsebuje podatke o publikacijah iz seznamov literature, katerih metapodatki so bili pred tem avtomatsko označeni s pomočjo regularnih izrazov ali algoritmov strojnega učenja. Preko iskalnika ESearch se bo tem podatkom o publikaciji dodalo SICI unikatno identifikacijsko oznako, ki bodo omogočile prenos relacij o citiranosti iz BibMake baze v SICI bazo citatov.

Grafikon 6: Shematski prikaz poti podatkov med različnimi bazami v sistemu SICI



Osrednja SICI baza bo kontrolne podatke preko Esearch iskalnika pridobivala iz različnih baz:

- Relacijske SICI podatkovne baze kontrolnih podatkov, v katere se bo uvažalo podatke, pridobljene iz drugih baz podatkov. V te baze se bo uvažalo samo podatke, za katere se bo presodilo, da so po eni strani zelo pomembni za izvajanje kontrole kvalitete vhodnih podatkov in po drugi strani vsebujejo nujno potrebne podatke, predvsem različne identifikacijske oznake, katerih ni med vhodnimi podatki. Najpomembnejši so podatki iz SICRIS in COBISS. Uvoz teh podatkov v SICI baze kontrolnih podatkov bo omogočil izvajanje veliko bolj zahtevnih iskalnih ukazov s Solr iskalnikom (mdr. tudi iskanja po Lavenshteinovem algoritmu) ter hitrejše ročno vpisovanje podatkov s pomočjo avtomatičnega dopolnjevanja besed (autocomplete). To bo mdr. pohitrilo

vnašanje podatkov v bazo citatov. Po eventualni vzpostavitvi sodelovanja, bi podatke iz SICRIS-a dobivali na podlagi OAI-PMH protokola. V nasprotnem primeru bi podatke dobili v XML formatu, jih nato pretvori v SICI XML shemo in nato enkrat do dvakrat na leto uvozi v SICI bazo kontrolnih podatkov

- Dostop do kontrolnih podatkov preko različnih aplikacijskih programskih vmesnikov. V prvi vrsti predvidevamo vključitev API - European Library.
- Triple store. Glede na izvor podatkov sta predvideni dve vrsti triple store podatkovnih baz:
    - Mapiranje podatkov iz relacijske SICI baze citatov z jezikom R2RML v RDF podatkovni model. Namenjena zunanjemu dostopu do podatkov iz SICI baze citatov in preko SPARQL točke še do vseh zunanjih prosto dostopnih triple store baz.
    - Native triple store baza kontrolnih podatkov, v katero bi uvozili prosto dostopne trojčke podatkov, ki niso dostopni v prosto dostopnih triple store bazah. Namenjeno iskanju kontrolnih podatkov za vnašanje v SICI bazo citatov.

Ker bo glavni poudarek na relacijah med publikacijami, smo se odločili uporabiti relacijske SQL baze. Izbrali bomo MySQL, katere glavna prednost je ta, da je nelastniška, zelo razširjena, dobro poznana med programerji, preizkušena in s tem posledično cenejša od komercialnih relacijskih baz. Hkrati tudi ne pričakujemo, da bo količina podatkov v naslednjih desetih letih tako velika, da bi bilo potrebno MySQL nadomestiti s kakšno komercialno relacijsko podatkovno bazo (npr. Oracle).

Osnovni podatkovni model SICI baz podatkov

Glavni namen Slovenskega indeksa citiranosti ne bo čim bolj natančno navajanje metapodatkov o publikacijah, temveč čim bolj natančno navajanje relacij med temi publikacijami.

Podatkovni model Slovenskega indeksa citiranosti bo zato temeljil na vzpostavljanju relacij med entitetami. Entitete tega podatkovnega modela so publikacije. Publikacija je besedilo v analogni ali digitalni obliki, ki lahko citira drugo publikacijo in/ali je lahko citirana s strani druge publikacije. Vsaka entiteta ima unikatno SICI identifikacijsko oznako.

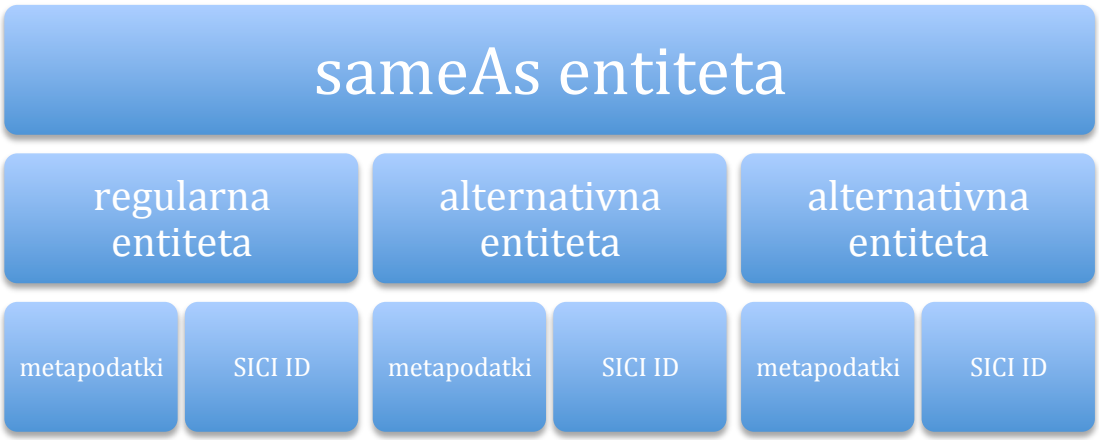
Podatki v SICI se delijo na štiri osnovne skupine:

- metapodatki o entitetah,
- metapodatki o relacijah sameAs med entitetami,
- metapodatki o relacijah citiranja med entitetami,
- metapodatki o hierarhičnih relacijah med entitetami.

Metapodatki o posamezni entiteti so v SICI bazo citatov načeloma lahko vneseni samo enkrat. V SICI vnesene entitete ni mogoče izbrisati.

Če so bili metapodatki o posamezni entiteti vneseni več kot enkrat in ima ta entiteta več kot eno unikatno SICI identifikacijsko oznako, se za eno od teh entit določi, da je regularna, in za ostale, da so alternativne. Regularna in alternativne entitete so med seboj povezane z relacijo sameAs (gl. Grafikon 7: Entiteta sameAs Slovenskega indeksa citiranosti.) Metapodatki in SICI ID regularne entitete so privzeti metapodatki sameAs entitete.

Grafikon 7: Entiteta sameAs Slovenskega indeksa citiranosti.



Metapodatke entitete se lahko popravlja. Vse verzije vpisov metapodatkov so trajno shranjene. Vsakič, ko shranimo metapodatke entitete, se shrani še časovna znamka in identifikacijska oznaka vnašalca. Zadnja shranjena verzija metapodatkov je privzeta verzija metapodatkov o entiteti.

Osnovno izhodišče pri vpisovanju metapodatkov entitete je čim bolj natančna identifikacija entitete in ne čim bolj razvejan opis entitete. Glavni namen SICI ni iskanje sorodnih entitet, temveč iskanje relacij med natančno identificiranimi entitetami. Zato metapodatkov o opisu vsebine, o deskriptorjih, ključnih besedah, UDK ipd. niso potrebni.

Metapodatki o relacijah sameAs med entitetami se delijo na dve osnovni skupini:

- Metapodatkov, ki jih ni mogoče spreminjati in izbrisati: relacija med SICI ID ene in SICI ID druge entitete, ki mora obstajata v okviru iste sameAs entitete. Med regularno in alternativno entiteto je lahko samo ena sameAs relacija.
- Metapodatki, ki jih ni mogoče izbrisati in jih je mogoče spreminjati: relacija je aktivna ali neaktivna. Avtomatsko se shrani še časovna znamka in indentifikacijska oznaka vnašalca, ki shrani te metapodatke.

Metapodatki o relacijah citiranja med entitetami se delijo na dve osnovni skupini:

- Metapodatkov, ki jih ni mogoče spreminjati in izbrisati: relacija med SICI ID ene in SICI ID druge entitete, ki ne smeta biti znotraj iste same As

entitet. Med entiteto, ki citira in entiteto, ki je citirana, je lahko najmanj ena in največ neskončno relacij citiranja.

- Metapodatki, ki jih ni mogoče izbrisati in jih je mogoče spreminjati:
    - Relacija je aktivna ali neaktivna (min. 1, max. 1). Avtomatsko se shrani še časovna znamka in identifikacijska oznaka vnašalca, ki shrani te metapodatke.
    - Številka citiranih strani (min. 0, max. nesk.) (char). Avtomatsko se shrani še časovna znamka in identifikacijska oznaka vnašalca, ki shrani te metapodatke.

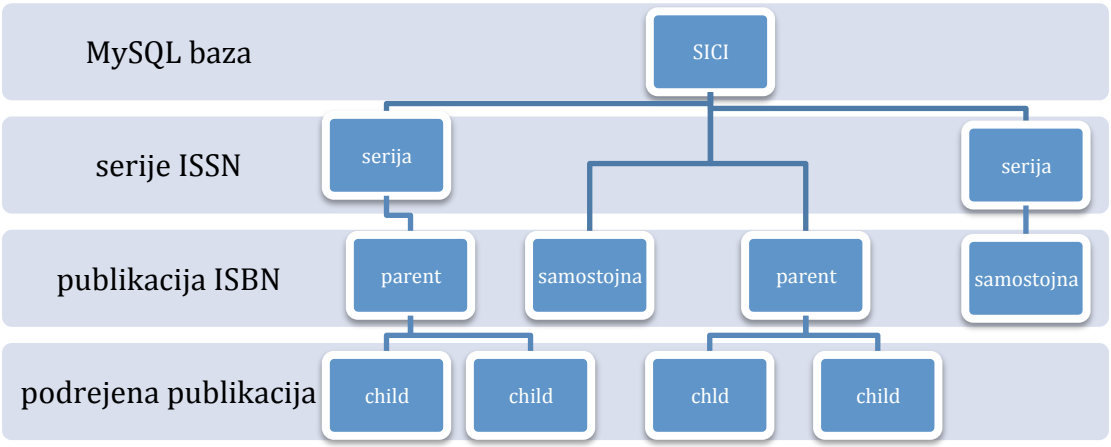
Metapodatki o hierarhičnih relacijah med entitetami se delijo na dve osnovni skupini:

- Metapodatki, ki jih ni mogoče spreminjati in izbrisati: relacija med SICI ID ene in SICI ID druge entitete, ki ne smeta biti znotraj iste sameAs entitete. Med nadrejeno in podrejeno entiteto je lahko samo ena relacija.
- Metapodatki, ki jih ni mogoče izbrisati in jih je mogoče spreminjati: relacija je aktivna ali neaktivna. Avtomatsko se shrani še časovna znamka in identifikacijska oznaka vnašalca, ki shrani te metapodatke.

Hierarhične relacije med entitetami temeljijo na sledečem razlikovanju med skupinami publikacij (prim. Grafikon 8: Hierarhična razdelitev entitet na skupine publikacij v sistemu SICI):

- Samostojne publikacije (monografije, baze raziskovalnih podatkov, ...).
- Podrejene child publikacije (članki v revijah, članki v časopisih, članki v zbornikih, samostojna poglavja v monografijah, gesla v enciklopedijah, ...). V terminologiji SICI so to child publikacije. Vsaka child publikacija je nujno del parent publikacije.
- Parent publikacije (posamezne številke revij in časopisov z identificiranimi posameznimi članki, zbornik, monografija s samostojnimi poglavji, posamezen zvezek enciklopedije z identificiranimi posameznimi gesli, ...).
- Serijske publikacije (revija, časopis, zbirke monografij ali zbornikov). Serijske publikacije imajo praviloma samostojno identifikacijsko oznako ISSN. Pri starejših publikacijah, ki nimajo identifikacijskih bibliografskih oznak, upoštevamo analogijo glede na sorodne sodobne publikacije.

Grafikon 8: Hierarhična razdelitev entitet na skupine publikacij v sistemu SICI



Citate se vnaša zgolj iz:

- samostojnih publikacij in/ali
- child publikacij.

Citirane so lahko:

- samostojne publikacije in
- child publikacije ali parent publikacije.

Glavni cilj takšnega hierarhičnega razmerja med entitetami je čim bolj avtomatična določitev vrste publikacije tudi v primerih, ko ta razmerja iz podatkov o citiranem delu sprva niso povsem jasno razvidna. Obenem bo na ta način čim bolj optimalno mogoče razrešiti tudi primere napačnega citiranja. To po posledično imelo velik pomen pri realnem merjenju znanstvene uspešnosti. V primerih, ko je bila namesto child publikacije citirana parent publikacija, se lahko podatke o child publikaciji izlušči iz parent publikacije glede na podatke o citiranih straneh.[5]

[5] Ena publikacija lahko npr. (pravilno) citira samostojno poglavje v monografiji, medtem kot druga publikacija citira samo monografijo s samostojnimi poglavji, brez navedbe ustreznega poglavja. Glede na podatke iz Zgodovinarskega indeksa citiranosti je takšna praksa citiranja relativno zelo pogosta v humanistiki. V primeru, da bi pri indeksu citiranosti upoštevali samo zbornik in ne samostojnega poglavja, bi bile točke znanstvene uspešnosti avtomatično pripisane uredniku zbornika in ne dejanskemu avtorju posredno citiranega poglavja tega zbornika.