Organizacija znanja, 24 (1–2), 2019, 1924005, https://doi.org/10.3359/oz1924005
Znanstveni članek / Scientific article

Merjenje kakovosti podatkov v bibliografskih in normativnih zapisih: študija primera izbranih podatkovnih elementov za fasetno omejevanje in izpis seznama zadetkov v COBISS+

Measuring the data quality of the bibliographic and authority records: case study of selected data elements for facet navigation and a displayed list of search results in COBISS+

Branka Badovinac[1]

IZVLEČEK: V prispevku predstavljamo teoretična in metodološka izhodišča za merjenje kakovosti podatkov v bibliografskih in normativnih zapisih, ki predstavlja enega od segmentov analize kakovosti podatkov. Na podlagi strokovne literature smo izpostavili načine oblikovanja mer, metode merjenja in analize rezultatov. Na podlagi profiliranja podatkov iz analize kakovosti zapisov dnevne produkcije v letih 2015–2018 smo zasnovali študijo primera, s katero smo merili kakovost podatkov v izbranih podatkovnih elementih, ki se uporabljajo pri fasetnem omejevanju in v izpisu seznama zadetkov v COBISS+. Prikazali smo primer oblikovanja mer in merjenja kakovosti podatkov, ki presega pristop štetja napak. Hkrati pa smo izpostavili nekatere podatkovne elemente, ki v katalogizacijski praksi (do zdaj) niso imeli večje veljave.

KLJUČNE BESEDE: kakovost podatkov, merjenje kakovosti podatkov, fasetna navigacija, OPAC, COBISS+

ABSTRACT: As one of the segments of data quality analysis, theoretical and methodological approaches towards measuring the quality of data in bibliographic and authority records are presented. On the basis of professional literature review, the paper focuses on metrics, methods of measurement and results analysis. Based on profiling the data set acquired by analyzing recommendations given to librarians within quality control of daily bibliographic records production in COBIB.SI in 2015–2018, the case study was conducted where the quality of data within selected data elements used in facet navigation and in the display of search results lists in COBISS+ was measured. This case study shows the possibility of metrics design, which could be more applicable if compared with the existing counting errors approach. At the same time the case study reveals a set of data elements that have had low value in cataloguing practices so far.

KEYWORDS: data quality, data quality measurement, facet navigation, OPAC, COBISS+

1 Uvod

Napake, povezane s podatki, ki jih uporabniki najdejo, imajo neposreden vpliv na uporabniško izkušnjo. Zato je kakovost zapisov in njihovih (meta)podatkov izrednega pomena, hkrati pa predstavlja teoretični in tehnični izziv. V katalogizaciji kakovost podatkov lahko razumemo v različnih kontekstih – z vidika uporabnika, enotne obdelave, katalogizatorja, racionalizacije delovnega procesa in tehnologije. V okviru aktivnosti spremljanja kakovosti zapisov dnevne

[1] Mag. Branka Badovinac, Institut informacijskih znanosti (IZUM), Maribor, Slovenija, branka.badovinac@izum.si.

produkcije smo h kakovosti pristopili z vidika enotne obdelave virov. Zanima nas, ali podatki ustrezajo standardom vnosa v format (referenčni viri). Vnos podatkov v knjižnične baze podatkov v primerjavi z drugimi sorodnimi ponudniki temelji na načelih natančne preslikave podatkov iz zelo različnih vrst virov, pri čemer so upoštevane tudi pomenske značilnosti podatkov, kar pogosto povečuje nivo kompleksnosti podatkov. Poleg težav s semantičnostjo se v COBIB.SI srečujemo z dinamičnostjo podatkov, saj se podatki v zapisih lahko spreminjajo.

Nekaj korakov k razumevanju kakovosti podatkov smo za potrebe aktivnosti spremljanja kakovosti dnevne produkcije že naredili. Tako smo raziskali, kako lahko pristopimo h kakovosti podatkov v katalogizaciji (Badovinac, 2017), in določili dimenzije kakovosti, s katerimi smo definirali, kakšne podatke (in posledično zapise) želimo v sistemu COBISS.SI. Po tej definiciji je podatek v bibliografskih in normativnih zapisih kakovosten, kadar je:

- naveden v edinstvenem zapisu (EDIN),
- semantično točen (SEMTOČ),
- točno prepisan (TOČPRE),
- strukturalno popoln (STRUPOP),
- vsebinsko popoln (VSEBPOP),
- strukturalno skladen (STRUSKLAD),
- vsebinsko skladen (VSEBSKLAD),
- oblikovno dosleden (OBLIDOS),
- predviden oziroma ni odvečen (ODVEČ),
- aktualen (AKT),
- (lahko) dodatno informativen (DODV) (Badovinac, 2018).

V tem prispevku bomo raziskali tudi teoretične in metodološke osnove merjenja kakovosti podatkov. Zato smo najprej predstavili spoznanja iz pregledane strokovne literature. Na podlagi profiliranja podatkov, pridobljenih v okviru aktivnosti spremljanja kakovosti dnevne produkcije, smo nato zasnovali študijo primera merjenja kakovosti na omejenem izboru podatkovnih elementov, ki se uporabljajo v COBISS+, in sicer za fasetno navigacijo in izpis seznama zadetkov.

2 Merjenje kakovosti podatkov v katalogizaciji

V strokovni literaturi zasledimo različne izraze za merjenje (angl. *measurement*), zato smo se odločili, da bomo od merjenja (preverjanja, prerez podatkov ipd.) ločili pojem ocenjevanja (angl. *assessment*, *auditing*), s katerim se v tej študiji nismo ukvarjali, saj gre za postopek (o)vrednotenja (evalvacije) rezultatov meritev na podlagi zahtev določenega podatkovnega modela.

Z merjenjem kakovosti želimo zmanjšati negotovost glede poznavanja stvari in pri odločitvah, kaj in kako izboljšati. To je tudi osrednja misel opredelitve merjenja na področju kakovosti podatkov.

Metrike oblikujemo zaradi:

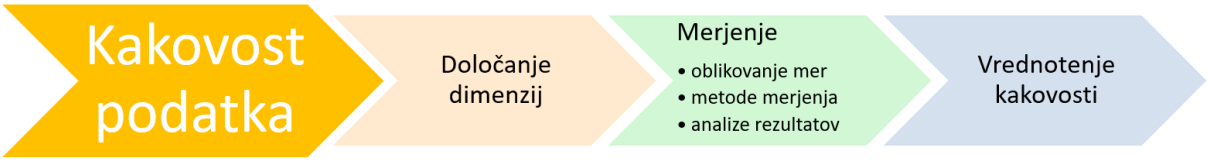
- informacije o želenih ciljih za ustvarjalce in upravljavce podatkov,
- razvoja standardov in
- vzpostavitve tehnike merjenja.

Pozorni moramo biti na to, da je merjenje:

- razumljivo in interpretativno (Merjenje ni le orodje za analizo, temveč je tudi orodje za komunikacijo. Če merimo nekaj izredno pomembnega na način, ki ga ljudje ne razumejo, je zelo verjetno, da merjenje ne bo učinkovito. Zato morajo biti mere že same po sebi jasno zastavljene; razvidno mora biti, kaj merimo.);
- ponovljivo (Instrumenti merjenja (enote, lestvice ipd.) in pogoji merjenja so zelo pomembni, saj naj bi omogočali konsistentne rezultate in razumevanje vzrokov, ki vplivajo nanje.);
- namensko (Razumeti moramo potrebe, vedeti moramo, zakaj merimo.) (Sebastian-Coleman, 2013).

Čeprav je merjenje v današnjih časih dokazovanja produktivnosti zaželeno, se moramo zavedati, da mora biti merjenje osmišljeno in osnovano na dobrih izhodiščih (Sebastian-Coleman, 2013; Loshin, 2011). Zaradi nejasnih stališč glede tega, kaj merimo ter ali merimo pravilno, na pravem mestu in v pravem času, ni težavno le merjenje, temveč lahko dobimo napačne rezultate ali jih napačno interpretiramo zaradi napačnih korelacij. Kadar recimo uporabimo vmesnik OPAC (Online Public Access Catalog), se v primeru nejasnih teoretičnih in metodoloških izhodišč pri merjenju kakovosti podatkov lahko zgodi, da namesto podatkov končni uporabniki v resnici ocenjujejo vmesnik.

Nekaj teoretičnih izhodišč lahko pripravimo že z definicijo kakovosti podatka, torej z naborom dimenzij (meril), ki opišejo, kakšne podatke želimo. Drugo pomembno izhodišče je, da merimo le tiste podatke oziroma podatkovne elemente, ki so v določenih kontekstih pomembni, izvedba merjenja kakovosti pa je racionalizirana (Kaiser, Klier in Heinrich, 2007). To zagotovo velja za področje katalogizacije, kjer imamo veliko število podatkovnih elementov, podatki pa se uporabljajo za različne namene in servise.



Slika 1: Posamezni segmenti analize kakovosti podatkov

Ko smo za posamezen podatkovni element določili, katere dimenzije so za kakovost podatka relevantne, sledi **oblikovanje mer** (angl. *metrics*), izbira **metode merjenja** in **analiza rezultatov** (slika 1). V okviru posamezne dimenzije lahko oblikujemo več različnih mer, ki nam podrobneje opišejo kriterije in predmet meritve. Tako Sebastian-Coleman (2013) ponudi trinivojski model razumevanja merjenja kakovosti podatkov, s čimer lahko lažje preidemo od abstraktnega h konkretnim izvedbam. Določimo lahko:

1) dimenzijo (angl. *dimension*), s katero odgovorimo, zakaj merimo,
2) tip meritve (angl. *measurement type*), s katerim določimo, kako bomo merili, in

3) specifično metriko kakovosti (angl. *specific data quality metric*), s katero določimo, kaj bomo merili (slika 2).

Dimenzija (zakaj merimo)	Hitrost dostave blage (časovnost)
Tip merjenja (kako merimo)	Primerjava med dejanskim časom dostave in predvidenim časom dostave
Specifične metrike kakovosti (kaj merimo)	Primerjava podatkov o dostavljenem času s časom dostave, navedenem v potrdilih za stranke.

Slika 2: Primer oblikovanje metrike za dimenzijo časovnost z vidika dostave blaga po modelu Sebastian-Coleman (2013)

Glede na kompleksnost podatkov v bibliografskih in normativnih zapisih je za področje katalogizacije treba izpostaviti tudi razumevanje zajema in vzorčenja podatkov. Metriko nekaterih podatkovnih elementov (npr. naslov) lahko izvedemo le z uporabo zunanjih virov, tj. t. i. validacijskih virov. Najbolj zaželeno je torej, da kakovost podatka preverjamo na osnovi dejanskega vira (Zeng in Qin, 2016). Žal je ta način, ki zajema ročno iskanje napak, preveč zamuden, sploh če želimo imeti dovolj obsežen vzorec. Drugi način sicer vključuje (pol)avtomatizirano iskanje slabih podatkov, vendar pa potrebujemo bolj kakovostno bazo podatkov, kjer se uporabljajo ista katalogizacijska pravila za vnos podatkov in programske kontrole. Tovrstnih kontrolnih baz v knjižničarskem okolju nimamo, razmeroma neuporabni so tudi založniški viri, ki običajno ne sledijo knjižničarskim standardom.

Ena izmed težav je razumevanje granulitete merjenja kakovosti podatkov. Zeng in Qin (2016) navajata tri možne nivoje merjenja: zbirko, zapis in podatkovni element. Te tri nivoje izpostavlja tudi Király (2015), ki pravi, da je merjenje na nivoju zapisa najbolj pogosto, saj s primerjavo med zapisi lahko dobimo tudi filter slabih zapisov. Težava tovrstnih raziskav na področju katalogizacije, pri katerih se preštevajo slabi podatki in podajajo izračuni povprečja napak na zapis, je, da s temi rezultati ne moremo pojasniti kakovosti zapisov, saj niti ne vemo, koliko je vseh podatkov v zapisu. Navkljub standardom je podajanje celotnega števila podatkov v zapisu svojevrsten izziv. Vnos podatka je namreč pogojen z več dejavniki, med katerimi so npr. tip in vrsta gradiva, lokalne katalogizacijske prakse, stopnja obveznosti podatka. Zapis je kakovosten, kadar vsebuje kakovostne podatke. Število možnih napak v zapisu je vsota vseh opredeljenih dimenzij, ki jih določimo pri posameznem podatkovnem elementu. Npr. podatek v podpolju 100b – *Oznaka za leto izida* mora biti semantično točen in vsebinsko skladen ter v določenih primerih strukturalno popoln, zato so v tem primeru možne tri napake.

Poleg nivoja lahko v meritvah določimo še druge spremenljivke. Za področje katalogizacije bi bili npr. zanimivi tudi čas kreiranja podatkov, ustanove in kreatorji, z vidika izboljšanja kakovosti dokumentacije pa tudi tip in vrsta gradiva ipd.

Po določitvi mer oz. kazalnikov se odločimo, katere **metode merjenja** bomo uporabili. Z merami preverimo odnos slabih podatkov (angl. *bad data*) do vseh relevantnih podatkov. Ta odnos se običajno preveri s kvantitativno metodo, kot je izračun deleža slabih podatkov, obstajajo pa tudi kvalitativni kazalci/kazalniki (angl. *indicators*) (npr. ankete, študije

uporabnikov). Pipino (2002) s sodelavci zagovarja kombinacijo oz. primerjavo subjektivnega in objektivnega ocenjevanja (merjenja) podatkov znotraj posamezne dimenzije, saj razumevanje razlik med obema ocenama omogoči bolj racionalizirano odpravo napak. Sicer pa Pipino (2002) v okviru kvantitativnega pristopa ponudi naslednje tri vrste metrik:

- metodo razmerja oz. deleža (ne)želenih podatkov glede na vse podatke,
- metodo agregatne funkcije oz. dovoljene najvišje in najnižje vrednosti (min in max), ki se uporabljata, kadar je v dimenzijo vključenih več spremenljivk, in
- metodo uteženega povprečja, pri čemer je vsakemu indikatorju (spremenljivki) določena utež glede na to, kako pomemben je za končno vrednost dimenzije, s čimer se izračuna povprečje. Utežene vrednosti so med 0 in 1, skupna vrednost je 1, tako se dobi normalizirana ocena. Posamezni indikator (spremenljivka) se izračunava z metriko enostavnega deleža.

Nekatere izračune lahko izvedemo avtomatizirano ali pa je podatke treba (pol)manualno zbirati in jih izračunavati. Slednje je zlasti v uporabi pri semantično pogojenih in nestrukturiranih podatkih, npr. semantična točnost podatka o avtorju kot točke dostopa ne zahteva le preverbe primarnega gradiva, temveč tudi oceno pravilnosti izbire normativnega zapisa, saj lahko obstaja več soimenjakov.

Verjetno tudi zato avtomatizirano merjenje, ki temelji na bolj ali manj kompleksnih algoritmih in statistično podprtih predvidevanjih kot obliki profiliranja podatkov/vzorcev, na področju katalogizacije redko zasledimo. Sicer pa, kot ugotavljata Ochoa in Duval (2009), je avtomatizirano podprto merjenje kakovosti podatkov tudi v primeru digitalnih zbirk le delno zanesljiva tehnika, ki je lahko uporabna le pri nekaterih merah. Ne glede na to so avtomatizirani kvantitativni pristopi k merjenju kakovosti na področju digitalnih knjižnic v trendu, zlasti zaradi potreb po vzpostavljanju standardov vnosa podatkov ter po bogatenju podatkov. Poleg tega predvidevajo metapodatkovne sheme digitalnih zbirk razmeroma malo različnih podatkovnih elementov, tako jih ima npr. poenostavljeni Dublin Core (DC) 15, Learning Object Metadata (LOM) pa 58. Metrika za dimenzijo *popolnost* pri DC predstavlja število vseh izpolnjenih polj v primerjavi s predvidenim številom polj v DC-standardu (Ochoa in Duval, 2009; Margaritopoulos et al., 2012).

Enega od primerov opisa mere z avtomatiziranim pristopom zbiranja/analize podatkov opiše Király (2015, 2019) na primeru Europeane. Njegovo izhodišče je, da z merjenjem strukturalnih elementov lahko napovemo kakovost metapodatkovnega zapisa. Osredotočil se je na naslednje tri vidike:

- merjenje značilnosti (tj. dimenzij), ki so neodvisne od sheme,
- zahteve najpomembnejših funkcij in
- nekateri že znani metapodatkovni problemi.

Királyjev konceptualni model sledi izhodiščem Brucea in Hillmanna (2004) ter Ochoe in Duvala (2009), na podlagi katerih izdela metrike za sedem dimenzij. Mera popolnosti recimo je delež med številom polj brez vrednosti in številom predvidenih polj glede na metapodatkovni standard, pri čemer lahko dodajamo uteži izračuna glede na stopnjo obveznosti podatka. S skupino za kakovost metapodatkov pri Europeani Király leta 2019 prvič predstavi izsledke dejanskih meritev, ki so se nanašale na napake, povezane z večjezičnostjo v podatkih. Avtorji

so k merjenju pristopili trinivojsko, od abstraktnega h konkretnemu, podobno kot to stori Sebastian-Coleman (2013) v prej omenjenem modelu. V okviru dimenzije *konsistentnost podatkov* npr. so za vidik različnosti označevanja jezikov uporabili mero *število različnih jezikovnih notacij* (Király et al., 2019).

Analizo rezultatov izvedemo skladno z metodološkimi določili podanih mer, spremenljivk in metod merjenja. Pri kvantitativni metodologiji lahko uporabimo vrsto statističnih izračunov in načine predstavitve rezultatov. Loshin (2011) izpostavlja pomen Shewhartovega diagrama, kjer rezultate opazujemo v okviru kontrolnih mej ter pri večjih odstopanjih poskušamo odkriti in odpraviti vzroke. Tu Loshin uporabi preneseno Paretovo načelo, ki pravi, da za mnoge pojave velja, da 20 % vzrokov povzroči 80 % posledic.

3 Pomen podatkov v tretji generaciji knjižničnih katalogov: fasete in prikaz rezultatov iskanja v COBISS+

Fasete so bile med najbolj pričakovanimi in obetavnimi značilnostmi tretje generacije knjižničnih katalogov, saj bi omogočale nove načine navigacije in omejevanja rezultatov iskanja. Kot postkoordinirana tehnika fasete omogočajo zmanjševanje obsežnih rezultatov iskanja na manjše, bolj obvladljive skupke, uporabnika kataloga pa ta tehnika nikoli ne vodi v poizvedbo brez zadetkov, saj so fasete oziroma njihova vsebina vidne le, če so dejansko povezane z virom v rezultatih iskanja. Raziskave so pokazale, da imajo fasete tudi pasivno vlogo, zlasti pri evalvaciji virov, saj že vsebina fasete uporabniku ponudi določeno analizo podatkov o značilnostih najdenih virov (Hall, 2016).

Fasete se uporabljajo v vseh fazah iskalnega procesa, zlasti pri odprtih (splošnih) poizvedbah (Niu, Fan in Zhang, 2019). Študije zadovoljstva uporabnikov kažejo, da so uporabniki s fasetami zelo zadovoljni ter da so fasete lahko razumljive, iskanje pa hitrejše in uspešnejše (Salaba in Zhang, 2009; Hall, 2016). Pri implementaciji faset se je izkazalo, da morajo oblikovalci paziti predvsem na jasno poimenovanje in taksonomijo faset (Gallaway in Hines, 2012; Niu in Hemminger, 2015; Hall, 2016).

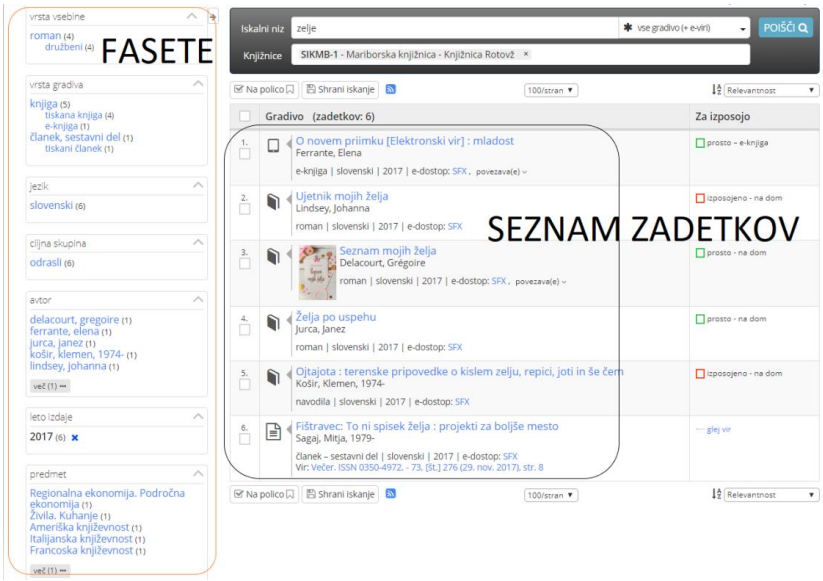
Prikazovanje podatkov v fasetah je vplivalo tudi na katalogizacijsko prakso. Z uvedbo te funkcionalnosti so nekateri podatki, ki v preteklih obdobjih niso imeli večje veljave, postali izredno koristni. Tu gre predvsem za nabor podatkov, ki presega opis vira po standardu ISBD. Ko so ti podatki postali vidni, so postale vidne tudi njihove pomanjkljivosti, ki so nastale bodisi zaradi slabo zasnovanih katalogizacijskih pravil (in prakse) bodisi zaradi slabo strukturirane podatkovne sheme (formata) ali pa dejanskih napak (Wynne in Hanscom, 2011; Schultz-Jones et al., 2012; Hall, 2016).

COBISS+ je ime za OPAC v sistemu COBISS. Uveden je bil leta 2017, njegove fasete ponujajo klasifikacijo zapisov po izvedenem iskanju po:

- vrsti vsebine (bibliografski podatki),
- vrsti gradiva (bibliografski podatki),
- jeziku (bibliografski podatki),
- ciljni skupini bralcev (bibliografski podatki),
- avtorju (bibliografski podatki v odnosu do normativne baze podatkov),
- letu izdaje (bibliografski podatki),

- predmetu oz. vsebini (bibliografski podatki) ter
- zalogi v knjižnicah – pri iskanju po več katalogih knjižnic hkrati (podatki iz zaloge) ali
- oddelku – pri iskanju po katalogu knjižnice z oddelki (podatki iz zaloge).

V pogledu po izvedbi iskanja so uporabniku vidni še podatki, ki so vključeni pri izpisu seznama zadetkov, torej rezultatov iskanja. Ta segment knjižničnega kataloga v strokovni literaturi ni posebej poudarjen, zasledimo pa nasvete o tem, da naj ti podatki uporabniku že takoj omogočijo informacijo o dostopu do celotnega besedila. Tudi v COBISS+ se za posamezen rezultat prikaže kratek nabor podatkov, ki omogoči hitro identifikacijo in dostop do vira (slika 3).



Slika 3: Zaslonska slika s primeri podatkov v izpisu zadetkov in fasetah v COBISS+ (z dne 21. 3. 2019)

4 Profiliranje podatkovnih elementov v fasetah in izpisu zadetkov COBISS+

Osnovno profiliranje kakovosti podatkov za izbor podatkovnih elementov temelji na analizah rezultatov pregleda zapisov dnevne produkcije, ki je le del modela zagotavljanja kakovosti v COBISS.SI in poteka od junija 2015 z metodo vzorčenja 10 % zajetih bibliografskih zapisov, kreiranih na določen dan, s pripadajočimi normativnimi zapisi.

Zapise zajamemo v COBIB.SI z iskalno zahtevo, pri čemer izločimo: zbirne zapise (dt=c), zapise, označene za brisanje (rs=d), predhodne nepopolne kataložne zapise oz. CIP-zapise (rs=p), prve vnose zapisa (rs=i), zapise, ki so bili vpisani s konverzijami lokalnih baz (cr=*old), programsko kreirane zapise, ki so preneseni iz baz Springer (cr=ctk springer) in Ebrary (cr=uplsi*) in drugih virov, npr. iz baze ISSN v bazo ELINKS ipd. (cr=knt izum_), ter zapise, ki so jih kreirali ali redigirali katalogizatorji iz Narodne in univerzitetne knjižnice (NUK) (cr=nuk*, re=nuk*) (Dornik et al., 2017).

Poglavitni cilj aktivnosti je takojšnja odprava morebitnih napak v novih zapisih (v sodelovanju s kreatorjem zapisa), ugotovitve analiz pa se uporabijo tudi na področju izobraževanja in usposabljanja za delo v COBISS.SI, pri oblikovanju programskih kontrol, izboljšavah dokumentacije ipd. Pri interpretaciji rezultatov, ki jih bomo predstavili, je treba razumeti, da:

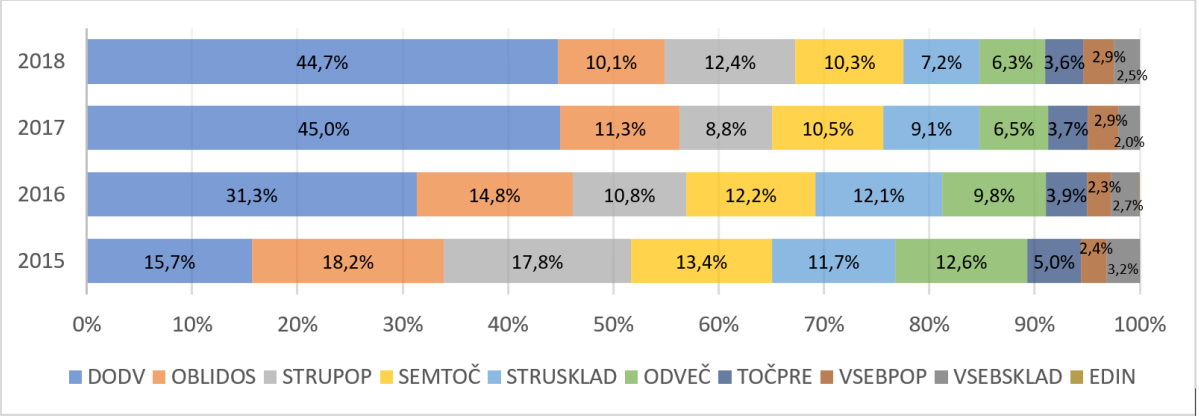
- gre za majhen vzorec (letno okoli 2.500 zapisov),
- gre za pregled brez primarnega vira,
- gre za pregled z metodo ekspertnega mnenja,
- so pri nekaterih podatkovnih elementih vključene programske kontrole, s katerimi se izognemo morebitnim napakam ali pa na pomanjkljivosti le opozorijo (prim. priročnik COBISS3/Katalogizacija, 2019).

Raven zanesljivosti priporočil zvišujemo s strokovnimi posveti sodelavcev in odzivi katalogizatorjev.

V letih 2015–2018 smo iz 156 zajemov (skupaj 85.387 zapisov) vzorčili 8.732 bibliografskih zapisov s pripadajočimi normativnimi zapisi. V vzorcu so prevladovali monografski tiskani viri (44 %), 10 % vseh zapisov v vzorcu pa predstavljajo elektronski viri, med katerimi so prevladovali sestavni deli. Delež zapisov za izvedena dela je v vzorcu 8-odstotni.

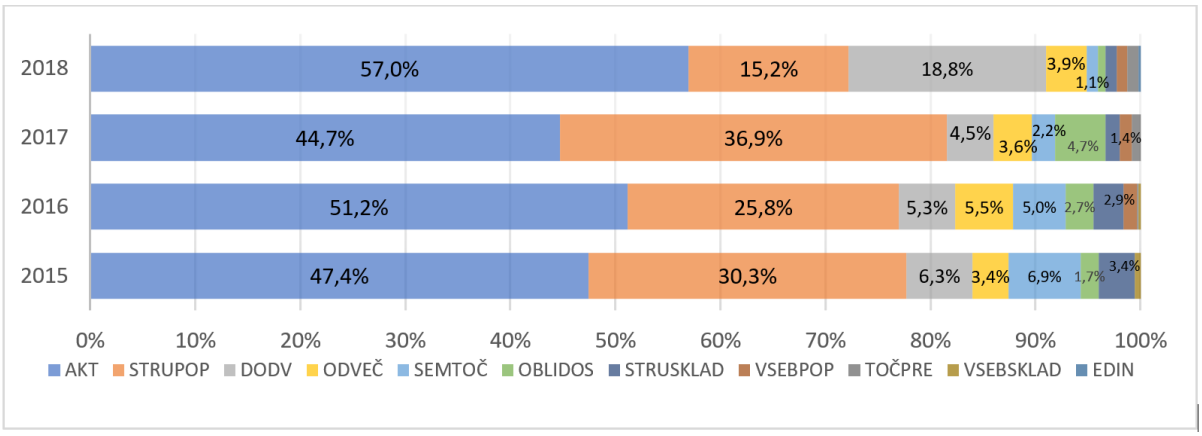
Zapise v vzorcu je kreiralo 540 različnih katalogizatorjev za 238 različnih ustanov. Ugotovili smo, da je bilo brez napak nekaj več kot 45 % pregledanih zapisov, v 28 % primerov smo zapisali priporočila z manjšo pomanjkljivostjo, v preostalih 24 % smo zasledili večje pomanjkljivosti, kot je to določeno s *Kriteriji za ocenjevanje bibliografskih in normativnih zapisov v COBISS.SI* (2009), ali pa vključujejo napake, povezane s formatom (zlasti blok 1XX). Na pomanjkljivosti smo opozorili s 1.473 elektronskimi sporočili. Raven odzivnosti katalogizatorjev, ki jo od leta 2016 preverimo letno, in sicer s ponovnim pregledom omejenega nabora zapisov, je v povprečju 80-odsotna. V večini primerov so popravki glede na priporočila ustrezni (Badovinac, 2019).

Glede na dimenzije kakovosti smo v 4.001 bibliografskem zapisu z majhno ali večjo pomanjkljivostjo analizirali 7.405 slabih podatkov, tj. podatkov po posameznih podatkovnih elementih oziroma določilih sheme dimenzije kakovosti. Ugotovili smo, da so v obdobju 2015–2018 najbolj pogosto manjkali priporočljivi podatki (DODV; 37,04 %) in drugi podatki (STRUPOP; 11,44 %). Beležili smo težave z oblikovanjem (OBLIDOS; 12,92 %) in točnostjo (SEMTOČ; 11,34 %) podatkov. Sledile so napake zaradi vpisa podatkov v napačno (pod)polje ali masko ipd. (STRUSKLAD; 9,38 %), nekaj podatkov je bilo odvečnih (ODVEČ; 8,24 %), drugi so bili pomanjkljivi (VSEBPOP; 2,66 %) ali pa se niso ujemali (VSEBSKLAD; 2,51 %). Zatipkanih podatkov je bilo nekaj manj kot štiri odstotke (TOČPRE; 3,89 %), dvojnikov zapisov (EDIN) pa 0,01 %. Primerjava med posameznimi leti kaže le manjša odstopanja med dimenzijami, se je pa število priporočil za vnos neobveznih podatkov (DODV) z leti zvišalo (slika 4).



Slika 4: Delež slabih podatkov v bibliografskih zapisih po dimenzijah in letih (jun 2015–2018) (n = 7.405)

Pri normativnih zapisih, ki so pripadali zajetim bibliografskim zapisom, smo v istem obdobju zasledili 1.659 slabih podatkov glede na posamezne podatkovne elemente. Zanje je bilo značilno, da smo poleg priporočanja vnosa dodatnih podatkov (STRUPOP, 25 %, in DODV, 9 %) priporočali zlasti ažuriranje celotnih zapisov na sploh (AKT, 51 %). Glede na podatke tudi tu ni večjih odstopanj med leti, v zadnjem letu je bilo nekaj več priporočil za vnos neobveznih podatkov (slika 5).



Slika 5: Delež slabih podatkov v normativnih zapisih po dimenzijah in letih (junij 2015–2018) (n = 1.659)

Fasetno omejevanje in izpis seznama zadetkov v COBISS+ uporablja 90 različnih podatkovnih elementov, kar vključuje tudi 11 podatkovnih elementov iz normativnih zapisov (baza CONOR). Nekateri podatkovni elementi so prisotni le v fasetah (npr. 100c – *Leto izida 1*, 100d – *Leto izida 2*) ali samo v izpisu zadetkov (npr. 210d – *Leto izida, distribucije itd.*), drugi pa so prisotni v obeh funkcionalnostih kataloga (npr. 001t – *Tipologija dokumentov/del*). Trije podatkovni elementi se obravnavajo na nivoju polja, pri poljih 140 – *Antikvarno gradivo – splošno* in 141 – *Antikvarno gradivo – značilnosti izvoda* se preverja prisotnost polj v zapisu, pri 464 – *Matična enota (monografska publikacija)* pa so polja in podpolja vgrajena. Pod določenimi katalogizacijskimi pogoji je obveznih 64 podatkovnih elementov (npr. obveznost za določene vrste gradiva, obveznost vnosa, če podatek obstaja ipd.). Podpolja so lahko ponovljiva znotraj polja, na nivoju polja ali v obeh primerih. Ne nazadnje pa je pri razumevanju nabora podatkov treba upoštevati tudi to, da so nekatera podpolja prisotna le pri posameznih

vrstah gradiva in v maskah vnosa. Nekateri podatkovni elementi so vključeni tudi v programsko preverjanje napak ob shranjevanju in/ali vnosu podatka (COBISS3/Katalogizacija, 2019). Glede na dimenzije kakovosti smo ocenili, da je v vseh podatkovnih elementih, ki se uporabljajo za fasetno navigacijo in izpis rezultatov iskanja, možnih 488 napak.

V okviru analize priporočil spremljanja dnevne produkcije v letih 2015–2018 smo zabeležili 3.556 slabih podatkov v 54 različnih podatkovnih elementih, ki jih uporabljamo v fasetah in izpisu seznama zadetkov, 40 napak se je zabeležilo na nivoju celotnega zapisa (en zapis je bil dvojnik). Največ pomanjkljivosti smo pripisali monografskim tiskanim in elektronskim virom (61 %), 80 % vseh pomanjkljivosti pa smo zabeležili v 12 različnih podpoljih, najbolj pogoste so v podpoljih 100e – *Koda za namembnost*, 100b – *Oznaka za leto izida*, 200a – *Stvarni naslov* in 200e – *Dodatek k naslovu*.

V okviru podatkovnih elementov v fasetah COBISS+ smo analizirali 2.213 priporočil (brez splošnih opomb na celotni zapis) in ugotovili, da smo največkrat priporočali vnos podatka za faseto *Ciljna skupina* (DODV, 64 %), sledile so napake, vezane na faseto *Leto izdaje* (16 %), v okviru katere smo zasledili več napak, povezanih s točnostjo podatka (SEMTOČ, 8 %) (tabela 1).

Tabela 1: Število slabih podatkov v podatkovnih elementih v fasetah COBISS+ po dimenzijah (junij 2015–2018; n = 2.213)

Fasete	Podatkovni element – fasete	DODV	SEMTOČ	VSEBSKLAD	STRUSKLAD	STRUPOP	ODVEČ	TOČPRE	Skupaj
CILJNA SKUPINA	Skupaj	1.421	14	1	0	0	0	0	1.436
	100e	1.421	14	1	0	0	0	0	1.436
LETO IZDAJE	Skupaj	167	67	109	2	9	13	0	367
	100b	167	48	68	0	2	0	0	285
	100c	0	13	37	1	1	0	0	52
	100d	0	6	4	1	6	13	0	30
AVTOR	Skupaj	51	23	20	42	33	23	2	194
	702	17	4	3	15	18	15	1	73
	701	34	5	6	12	6	3	0	66
	700	0	1	11	14	7	5	1	39
	7003	0	9	0	0	0	0	0	9
	7013	0	2	0	0	1	0	0	3
	70X	0	0	0	1	1	0	0	2
	7023	0	2	0	0	0	0	0	2
VRSTA VSEBINE	Skupaj	60	46	1	1	23	5	0	136
	105b	33	4	1	0	5	2	0	45
	135a	0	17	0	0	9	0	0	26
	135b	0	9	0	0	7	0	0	16
	001t	0	15	0	1	0	0	0	16
	105f	11	1	0	0	0	1	0	13
	105g	12	0	0	0	0	0	0	12

Fasete	Podatkovni element – fasete	DODV	SEMTOČ	VSEBSKLAD	STRUSKLAD	STRUPOP	ODVEČ	TOČPRE	Skupaj
	125c	2	0	0	0	1	0	0	3
	116	0	0	0	0	0	1	0	1
	128	1	0	0	0	0	0	0	1
	115	0	0	0	0	0	1	0	1
	126	0	0	0	0	1	0	0	1
	115a	0	1	0	0	0	0	0	1
VRSTA GRADIVA	Skupaj	0	17	0	30	0	0	0	47
	001c	0	7	0	18	0	0	0	25
	001b	0	10	0	12	0	0	0	22
PREDMET	Skupaj	0	26	0	0	0	0	0	26
	675c	0	26	0	0	0	0	0	26
JEZIK	Skupaj	0	2	1	4	0	0	0	7
	101a	0	2	1	4	0	0	0	7
Skupaj		1.700	194	132	79	65	41	2	2.213

Legenda: DODV – dodatna vrednost; SEMTOČ – semantična točnost; VSEBSKLAD – vsebinska skladnost; STRUSKLAD – strukturalna skladnost; STRUPOP – strukturalna popolnost; ODVEČ – odvečnost; TOČPRE – točnost prepisa; 001b – Vrsta zapisa; 001c – Bibliografski nivo; 001t – Tipologija dokumentov/del; 100b – Oznaka za leto izida; 100c – Leto izida 1; 100d – Leto izida 2; 100e – Koda za namembnost; 101a – Jezik besedila; 105b – Kode za vrsto vsebine; 105f – Koda za literarno vrsto; 105g – Koda za biografijo; 115 – Projicirno gradivo, videoposnetki in filmi; 115a – Vrsta gradiva; 116 – Slikovno gradivo; 125c – Oznaka za govorjeno besedilo; 126 – Zvočni posnetki – fizični opis; 128 – Glasbene izvedbe in partiture; 135a – Vrsta elektronskega vira; 135b – Fizična oblika; 675c – Vrstilec za iskanje; 700 – Osebno ime – primarna odgovornost; 7003 – Številka normativnega zapisa; 701 – Osebno ime – alternativna odgovornost; 7013 – Številka normativnega zapisa; 702 – Osebno ime – sekundarna odgovornost; 7023 – Številka normativnega zapisa; 70X – blok 7

Pri podatkovnih elementih, ki niso v fasetah, ampak se uporabljajo le pri izpisu zadetkov, smo od vseh 1.303 slabih podatkov beležili največ težav z oblikovanjem podatka (OBLIDOS, 25 %) in strukturno skladnostjo (STRUSKLAD, 20 %). Rezultati kažejo tudi razmeroma visok delež napak, povezanih s točnostjo (SEMTOČ, 13 %) in zatipkanostjo (TOČPRE, 9 %) podatkov, pri čemer bi posebej izpostavili podpolji 200a – *Stvarni naslov* in 856u – *Enotna lokacija vira (URL).* V skupini *strukturalna popolnost* (STRUPOP) pa bi posebej izpostavili manjkajoče letnice izida vira v podpolju 210d – *Leto izida, distribucije itd*., s katerim je sicer povezanih tudi več drugih vrst napak – kar 11 % v primerjavi z vsemi podatkovnimi elementi (tabela 2).

Tabela 1: Število slabih podatkov v podatkovnih elementih v izpisu zadetkov COBISS+ po dimenzijah (junij 2015–2018, n = 1.303)

Podatkovni element – izpis zadetkov	OBLIDOS	STRUSKLAD	SEMTOČ	TOČPRE	STRUPOP	VSEBSKLAD	VSEBPOP	DODV	ODVEČ	Skupaj
200a	65	82	24	55	0	0	6	11	2	245
200e	78	64	13	28	1	0	31	6	3	224

Podatkovni element – izpis zadetkov	OBLIDOS	STRUSKLAD	SEMTOČ	TOČPRE	STRUPOP	VSEBSKLAD	VSEBPOP	DODV	ODVEČ	Skupaj
210d	24	3	9	1	15	90	1	1	0	144
215a	67	7	3	19	10	1	11	6	5	129
856u	5	0	62	3	27	0	1	28	1	127
200	0	57	0	0	0	0	0	0	0	57
215h	26	3	5	2	0	0	15	0	4	55
200b	3	7	10	0	29	0	0	0	5	54
710	0	15	2	0	8	3	0	0	14	42
215k	10	1	5	3	0	4	4	1	9	37
200d	9	12	1	6	5	0	2	0	1	36
200ind1	0	0	32	0	0	0	0	0	0	32
215i	16	2	1	2	0	0	8	2	1	32
710a	12	1	10	3	0	0	1	0	0	27
200i	8	9	0	1	2	0	0	0	1	21
017a	0	0	3	3	1	0	0	7	0	14
710c	3	0	0	0	0	0	0	1	1	5
710b	0	0	0	0	2	0	1	0	1	4
710d	0	0	0	1	3	0	0	0	0	4
711	0	3	0	0	0	0	0	0	1	4
711a	2	0	1	0	0	0	0	0	0	3
7XX	0	1	0	0	1	0	0	0	0	2
011a	0	0	1	0	0	1	0	0	0	2
011e	0	1	0	0	0	0	0	0	0	1
011s	0	0	0	0	0	0	0	0	1	1
711b	0	1	0	0	0	0	0	0	0	1
Skupaj	328	269	182	127	104	99	81	63	50	1.303

Legenda: OBLIDOS – oblikovna doslednost; STRUSKLAD – strukturalna skladnost; SEMTOČ – semantična točnost; TOČPRE – točnost prepisa; STRUPOP – strukturalna popolnost; VSEBSKLAD – vsebinska skladnost; VSEBPOP – vsebinska popolnost; DODV – dodatna vrednost; ODVEČ – odvečnost; 011a – ISSN pri članku; 011e – Veljavni ISSN; 011s – ISSN pri članku v seriji s podserijo ali v prilogi; 017a – Identifikator; 200 – Naslov in navedba odgovornosti; 200a – Stvarni naslov; 200b – Splošna oznaka gradiva; 200d – Vzporedni stvarni naslov; 200e – Dodatek k naslovu; 200i – Naslov podrejenega dela; 200ind1 – Pomembnost naslova; 210d – Leto izida, distribucije itd.; 215a – Posebna oznaka gradiva in obseg; 215h – Številčenje – prvi nivo; 215i – Številčenje – drugi nivo; 215k – Kronologija; 710 – Ime korporacije – primarna odgovornost; 710a – Začetni element; 710b – Podrazdelek; 710c – Dodatek k imenu ali kvalifikator; 710d – Zaporedna številka sestanka; 711 – Ime korporacije – alternativna odgovornost; 711a – Začetni element; 711b – Podrazdelek; 7XX – blok 7; 856u – Enotna lokacija vira (URL)

5 Merjenje kakovosti podatkov na primeru podatkovnih elementov za fasetno omejevanje in izpis zadetkov v COBISS+

5.1 Metodološka izhodišča

Na podlagi profiliranja podatkovnih elementov smo za namen prikaza možnosti oblikovanja mer in analize rezultatov izbrali tiste ključne podatkovne elemente, ki ne zahtevajo večjega ročnega preverjanja in priprave podatkov. V okviru štirih dimenzij smo izmerili kakovost podatkov iz sedmih različnih podatkovnih elementov, ki se uporabljajo za fasetno navigacijo in v prikazu seznama zadetkov.

Z izborom podatkovnih elementov smo želeli preveriti:

1) usklajenost podatkov med fasetami in seznamom zadetkov (Za ta namen smo preverili podatek o letnici izida. Posebej smo preverili dve kombinaciji vsebinske skladnosti podpolj 100c – Leto izida 1, 100d – Leto izida 2 in 210d – Leto izida, distribucije itd. ter izpolnjenost oziroma strukturalno popolnost podpolja 210d – Leto izida, distribucije itd.);
2) semantično točnost podatka (Preverili smo jo le za podpolje 200b – Splošna oznaka gradiva, ki v seznamu zadetkov neposredno sledi podatkovnemu elementu stvarnega (in podrejenega) naslova. Vnos tega podatkovnega elementa je sicer neobvezen, vendar je skozi katalogizacijsko prakso ta element postal ustaljen podatek – razen za vire, kjer prevladuje tiskano besedilo.);
3) stopnjo izpolnjenosti podatkovnih elementov v fasetah, katerih vnos je po referenčnih virih le priporočljiv (V okviru dimenzije dodane vrednosti smo preverili vnos kode za namembnost (podpolje 100e), ki se uporablja v faseti Ciljna skupina, ter podatka o vrsti vsebine (podpolje 105b) in o literarni vrsti (podpolje 105f), ki se uporabljata v faseti Vrsta vsebine.).

5.2 Vzorčenje in analiza

V okviru posameznega podatkovnega elementa smo glede na izbrano dimenzijo določili mero in oblikovali vzorčenje za izbrano obdobje kreiranja skupine bibliografskih zapisov. Vzorčenje smo 26. 8. 2019 izvedli na nivoju vzajemne baze COBIB.SI, razen vzorčenja usklajenosti podatkov v 100c – *Leto izida 1*, 100d – *Leto izida 2* in 210d – *Leto izida, distribucije itd*., ki je bilo izvedeno 3. 4. 2019.

Pri vzorčenju smo uporabili osnovni iskalni niz, ki ga uporabljamo za spremljanje kakovosti dnevne produkcije in ki smo mu dodali potrebne omejitve glede na dana izhodišča. Pri nekaterih vzorcih smo podatke še ročno obdelali in vzorec naknadno uskladili z izhodišči posamezne mere (tabela 3).

Neusklajenost podatkov v podpoljih 100c – *Leto izida* 1 in 210d – *Leto izida, distribucije itd*. smo preverili z izračunom deleža bibliografskih zapisov z neujemajočimi podatki glede na število vseh bibliografskih zapisov s podatki v podatkovnih elementih 100c – *Leto izida 1* in 210d – *Leto izida, distribucije itd*. Zajeli smo zapise, ki so bili kreirani v obdobju od 1. 1. 2019 do 4. 2. 2019, pravilnost podatkov v zapisih smo preverjali ročno. Postopek smo za isto obdobje ponovili tudi pri zajemu podatkov, ko smo preverjali usklajenost med podpolji 100c – *Leto izida 1*, 100d – *Leto izida 2* in 210d – *Leto izida, distribucije itd*. (tabela 3).

Število bibliografskih zapisov z manjkajočim podatkom v podpolju 210d – *Leto izida, distribucije itd*. glede na število vseh relevantnih bibliografskih zapisov v letu 2018 smo dobili z zajemom, ki ni vključeval zapisov z masko za monografske publikacije (000a=*001*), kjer je podpolje 210d – *Leto izida, distribucije itd*. obvezno, ter za sestavne dele (dt=a) in izvedena dela (dt=d), kjer je podatek v 210d – *Leto izida, distribucije itd*. odvečen. Skeniranja ni mogoče začeti z operatorjem &NOT, zato smo najprej navedli kriterij (001a=*), ki ustreza vsakemu zapisu.

Zajem bibliografskih zapisov z napačnim podatkom v podpolju 200b – *Splošna oznaka gradiva* se je omejil na zapise elektronskih virov, ki v podpolju 200b – *Splošna oznaka gradiva* niso imeli navedenega začetka besede elektronski (»elektr*«) in so bili kreirani v letu 2018.

Pri zajemu bibliografskih zapisov, kreiranih v letu 2018 in z manjkajočim podatkom v podpolju 105f – *Koda za literarno vrsto*, smo izključili zapise za sestavne dele (dt=a) in izvedena dela (dt=d) ter zapise z navedeno tipologijo (001t – *Tipologija dokumentov/del*), tako da smo se omejili na zapise za knjižno gradivo (/bma), ki imajo v podpolju 675c – *Vrstilec za iskanje* začetek UDK-vrstilca 821*.

Pri podatkovnem elementu 105b – *Kode za vrsto vsebine* smo iz zajema zapisov, kreiranih v letu 2018, izključili zapise za sestavne dele (dt=a) in izvedena dela (dt=d) ter zapise, ki so vsebovali podatke v podpoljih 105f – *Koda za literarno vrsto*, 001t – *Tipologija dokumentov/del* in 105g – *Koda za biografijo*. Pri zajemu zapisov z manjkajočimi podpolji 100e – *Koda za namembnost* smo se omejili na zapise, kreirane v letu 2018, in na zapise knjižnega gradiva (2018*/bma). Izključili smo tudi zapise za izvedena dela (dt=d). Z zajemom sken1 smo dobili število zapisov, ki vsebujejo podpolje 105b – *Kode za vrsto vsebine*; dobljeno število smo odšteli od števila vseh relevantnih zapisov (tabela 3).

Vse mere so oblikovane na osnovi metrike *metode razmerja* (Pipino et al., 2002), vsi deleži so izračunani za posamezno vzorčenje: razen prvih dveh so med seboj neprimerljivi, saj zajemajo različna obdobja in različne predpostavke, ki hkrati veljajo tudi za omejitve posploševanja rezultatov.

Tabela 3: Mere in vzorčenje glede na posamezno dimenzijo izbranih podatkovnih elementov

Podatkovni element	Uporaba	Dimenzija	Mera	Vzorčenje
100c/210d	faseta/ izpis zadetkov	VSEBSKLAD	število bibliografskih zapisov z neujemajočimi podatki glede na število vseh bibliografskih zapisov s podatki v podatkovnih elementih 100c in 210d	→zajem: dm=20190101:20190402 not (dt=c or rs=d or rs=p or cr=*old or cr=ctk springer* or cr=uplsi* or cr=knt izum_ or cr=nuk* or re=nuk* or rs=i or dt=d or dt=a) sken-->210d=* & 100c=* &not 100d=* → ročna preverba podatkov
100c/100d/210d	faseta/ izpis zadetkov	VSEBSKLAD	število bibliografskih zapisov z neujemajočimi podatki glede na število vseh bibliografskih zapisov s podatki v podatkovnih elementih 100c, 100d in 210d	→zajem: dm=20190101:20190402 not (dt=c or rs=d or rs=p or cr=*old or cr=ctk springer* or cr=uplsi* or cr=knt izum_ or cr=nuk* or re=nuk* or rs=i or dt=d or dt=a) sken--> 210d=* & 100c=* &

Podatkovni element	Uporaba	Dimenzija	Mera	Vzorčenje
				100d=* → ročna preverba podatkov
210d	izpis zadetkov	STRUPOP	število bibliografskih zapisov z manjkajočim podatkom glede na število vseh relevantnih bibliografskih zapisov	→ zajem: dm=2018* not (dt=c or rs=d or rs=p or cr=*old or cr=ctk springer* or cr=uplsi* or cr=knt izum_ or cr=nuk* or re=nuk* or rs=i or dt=d or dt=a) sken 1-->001a=* &not 000a=*001* &not 210d=* sken 2-->001a=* &not 000a=*001*
200b	izpis zadetkov	SEMTOČ	število bibliografskih zapisov z napačnim podatkom glede na število vseh relevantnih bibliografskih zapisov	→ zajem: dm=2018* not (dt=c or rs=d or rs=p or cr=*old or cr=ctk springer* or cr=uplsi* or cr=knt izum_ or cr=nuk* or re=nuk* or rs=i or dt=d) sken 1-->001b=l & 200b=* &not 200b=elektr* sken 2-->001b=l & 200b=*
105f	faseta	DODV	število bibliografskih zapisov z manjkajočim podatkom glede na število vseh relevantnih bibliografskih zapisov	→ zajem: dm=2018*/bma not (dt=c or rs=d or rs=p or cr=*old or cr=ctk springer* or cr=uplsi* or cr=knt izum_ or cr=nuk* or re=nuk* or rs=i or dt=d or dt=a) sken 1-->675c=821* &not 001t=* &not 105f=* sken 2-->675c=821* &not 001t=*
105b	faseta	DODV	število bibliografskih zapisov z manjkajočim podatkom glede na število vseh relevantnih bibliografskih zapisov	→ zajem: dm=2018*/bma not (dt=c or rs=d or rs=p or cr=*old or cr=ctk springer* or cr=uplsi* or cr=knt izum_ or cr=nuk* or re=nuk* or rs=i or dt=d or dt=a) sken 1-->001a=* & 105b=* &not 105f=* &not 001t=* &not 105g=* sken 2-->001a=* &not 105f=* &not 001t=* &not 105g=*
100e	faseta	DODV	število bibliografskih zapisov z manjkajočim podatkom glede na število vseh relevantnih bibliografskih zapisov	→ zajem: dm=2018* not (dt=c or rs=d or rs=p or cr=*old or cr=ctk springer* or cr=uplsi* or cr=knt izum_ or cr=nuk* or re=nuk* or rs=i or dt=d) sken 1-->001a=* &not 100e=*

Legenda: 100c – Leto izida 1; 100d – Leto izida 2; 210d – Leto izida; distribucije itd.; 200b – Splošna oznaka gradiva; 105b – Kode za vrsto vsebine; 105f – Koda za literarno vrsto; 100e – Koda za namembnost, VSEBSKLAD – Vsebinska skladnost; STRUPOP – Strukturalna popolnost; SEMTOČ – Semantična točnost; DODV – Dodana vrednost

5.3 Rezultati

Rezultati kažejo, da je vsebinska skladnost (VSEBSKLAD) podatkov v bibliografskih zapisih, ki so bili kreirani med 1. 1. 2019 in 2. 4. 2019 ter imajo le podpolji 100c – *Leto izida 1* in 210d – *Leto izida, distribucije itd*., večja kot v zapisih, ki so bili kreirani v istem obdobju in kjer so

navedeni trije podatkovni elementi (100c – *Leto izida 1*, 100d – *Leto izida 2* in 210d – *Leto izida, distribucije itd*.). Glede na vzorčenje strukturalne popolnosti (STRUPOP) podatkovnega elementa 210d – *Leto izida, distribucije itd*., s katerim smo zajeli zapise, kreirane v letu 2018, smo zabeležili nekaj več kot dva odstotka zapisov, pri katerih je manjkal obvezen podatek v podpolju 210d – *Leto izida, distribucije itd*. (tabela 4).

Tabela 2: Vsebinska skladnost podatkov (VSEBSKLAD) v 100c – Leto izida 1, 100d – Leto izida 2 in 210d – Leto izida, distribucije itd. in strukturalna popolnost (STRUPOP) v 210d – Leto izida, distribucije itd.

Podatkovni element	Uporaba	Dimenzija	Število vseh relevantnih bibliografskih zapisov	Število bibliografskih zapisov s slabimi podatki	Delež zapisov s slabimi podatki (v %)
100c/210d	faseta/izpis zadetkov	VSEBSKLAD	18.331	195	1,1
100c/100d/210d	faseta/izpis zadetkov	VSEBSKLAD	1.147	103	9,0
210d	izpis zadetkov	STRUPOP	12.042	306	2,5

Pri podpolju 200b – *Splošna oznaka gradiva* smo preverili semantično točnost (SEMTOČ) le za elektronske vire, ki so bili kreirani v letu 2018, in ugotovili, da je stopnja točnosti podatka za izbrani vzorec visoka, saj je delež zapisov s slabimi podatki le 0,4 % (tabela 5).

Tabela 3: Semantična točnost podatkovnega elementa 200b – Splošna oznaka gradiva v letu 2018

Podatkovni element	Uporaba	Dimenzija	Število vseh relevantnih bibliografskih zapisov	Število bibliografskih zapisov s slabimi podatki	Delež zapisov s slabimi podatki (v %)
200b	izpis zadetkov	SEMTOČ	16.970	70	0,4

Za skupino priporočljivih podatkov (DODV) je odstotek manjkajočih podatkov razumljivo višji (tabela 6), saj je njihov vnos neobvezen, čeprav so v sedanji katalogizacijski praksi prepoznani kot koristni. Kot kažejo rezultati za vzorec zapisov, kreiranih v letu 2018, je pri katalogizatorjih večje zavedanje pri navajanju kode za literarno vrsto (podpolje 105f), medtem ko so kode za vrsto vsebine (podpolje 105b) nekoliko manj v uporabi, verjetno zaradi kodiranja v drugih podatkovnih elementih, morebiti pa tudi zaradi omejenih možnosti nabora kod za opis vsebine (forme) publikacije in njenih pomembnih delov.

Informacija o namembnosti, tj. podatek v podpolju 100e – *Koda za namembnost*, je pri lokalnih katalogih, ki so usmerjeni le na eno izmed ponujenih populacij bralcev (odrasli ali otroci), morebiti res odveč, vendar predstavlja na nivoju vzajemnega kataloga, ki je običajno tudi prva vstopna točka uporabnikov, eno izmed zelo pomembnih možnosti omejevanja bibliografskih zapisov. Zato si na tem mestu želimo, da bi tudi preostala tretjina zapisov, kot to kaže analiza zapisov, kreiranih v letu 2018, vsebovala informacijo o ciljni skupini bralcev (tabela 6).

Tabela 4: Dodana vrednost podatkovnih elementov 105f – Koda za literarno vrsto, 105b – Koda za vrsto vsebine in 100e – Koda za namembnost v letu 2018

Podatkovni element	Uporaba	Dimenzija	Število vseh relevantnih bibliografskih zapisov	Število bibliografskih zapisov s slabimi podatki	Delež zapisov s slabimi podatki (v %)
105f	faseta	DODV	10.345	1.847	17,9
105b	faseta	DODV	29.738	17.238	58
100e	faseta	DODV	131.807	44.114	33,5

6 Razprava

Kakovost podatkov je izredno kompleksna, zato je razumevanje merjenja kakovosti podatkov ključna, saj je rezultate treba umestiti v prostor in čas izvedbe ter način vzorčenja in analize. Na osnovi pregleda splošne strokovne literature smo ugotovili, da je oblikovanje mer specifično glede na zastavljena izhodišča zagotavljanja kakovosti in modeliranja nabora dimenzij. Na voljo imamo različne metode zbiranja in analize rezultatov.

Čeprav se o kakovosti podatkov na področju katalogizacije veliko razpravlja, je presenetljivo, da v strokovni literaturi ni zaslediti raziskav o metodoloških vprašanjih merjenja kakovosti podatkov, ki bi presegale štetje napak. Razmeroma malo je tudi poskusov meritev, ki so sicer bolj značilne za področje digitalnih knjižnic, kjer se nagibajo predvsem h kvantitativnim merjenjem z avtomatiziranimi tehnikami.

V primerjavi s podatkovnimi shemami digitalnih knjižnic so knjižničarske podatkovne sheme (formati) veliko bolj strukturirane, saj vsebujejo veliko število metapodatkov (podatkovnih elementov), vezanih na specifične standarde vsebine. Posebej to velja za kooperativne sisteme, kot je COBISS, kjer se podatki uporabljajo za različne skupine uporabnikov in aplikacij. Zato smo se za potrebe tega prispevka v skladu z napotki strokovne literature omejili le na dve funkcionalnosti online kataloga COBISS+.

Študija primera je imela tako dva namena: ponuditi način merjenja podatkov na osnovi oblikovanja mere v izbrani dimenziji ter preveriti kakovost podatkov v fasetnem omejevanju in izpisu rezultatov iskanja v COBISS+, ki sta pomembna za uporabniško navigacijo in identifikacijo virov. Slednje je zahtevalo še predhodno profiliranje podatkov in ugotavljanje, kateri podatkovni elementi bi bili primerni za merjenje. Pri tem se je pokazala težava zlasti pri iskanju ustreznih načinov preverbe brez primarnega vira ali validacijskega vira. Zato smo za študijo primera izbrali tiste podatkovne elemente, ki za dane mere potrebujejo manj ročne priprave in preverbe podatkov, a so še vedno dovolj pomembni za njihovo izpostavitev v kontekstu zahtev nove generacije knjižničnih katalogov. Pri razumevanju rezultatov je treba vedeti tudi, da so iz vzorčenja izločeni zapisi, ki so jih kreirali ali redigirali katalogizatorji iz NUK-a.

V prvo skupino smo uvrstili podatkovne elemente, ki vsebujejo podatek o letnici izida vira, s katero smo preverjali usklajenost podatkov med fasetami in izpisi rezultatov, saj je ta pomembna za identifikacijo in izbor vira. Z vidika vsebinske skladnosti smo oblikovali dve meri:

1) število bibliografskih zapisov z neujemajočimi podatki glede na število vseh bibliografskih zapisov s podatki v podatkovnih elementih 100c – *Leto izida 1* in 210d – *Leto izida, distribucije itd*. in
2) število bibliografskih zapisov z neujemajočimi podatki glede na število vseh bibliografskih zapisov s podatki v podatkovnih elementih 100c – *Leto izida 1*, 100d – *Leto izida 2* in 210d – *Leto izida, distribucije itd.*

Rezultati kažejo, da je več neusklajenosti podatkov pri zapisih o virih, kjer so preslikave podatkov o letnici izida bolj kompleksne; predvidevamo, da temu botruje pomanjkanje znanja o uporabi posameznih podatkovnih elementov.

Z dimenzijo *strukturalna popolnost* smo preverili število bibliografskih zapisov z manjkajočim podatkom glede na število vseh relevantnih bibliografskih zapisov in ugotovili, da je teh zapisov razmeroma malo.

Semantično točnost smo preverili le na primeru podatkovnega elementa 200b – *Splošna oznaka gradiva* s preverbo števila bibliografskih zapisov z napačnim podatkom glede na število vseh relevantnih bibliografskih zapisov. Ugotovili smo, da je število zapisov z netočnim podatkom zanemarljivo. Predvidevamo, da je to tudi posledica že izvedenih aktivnosti za zagotavljanje kakovosti COBIB.SI v preteklosti ter uvedbe vnosa podatkov z uporabo šifranta za lažji vnos.

S tretjo skupino podatkovnih elementov smo želeli izpostaviti zlasti trend razširitve nabora podatkov, ki v pretekli katalogizacijski praksi niso imeli večje veljave oziroma je bil njihov vnos le priporočljiv. Rezultati naše raziskave kažejo, da novejšim zapisom v COBIB.SI manjkajo podatki, ki prinašajo dodano vrednost v okviru funkcionalnosti COBISS+.

Iz nekaterih razprav v slovenskem okolju je mogoče razbrati ugotovitve, da uporabniki primarno uporabljajo le dva podatka: naslov in avtorja (Mrđenović, 2018; Kavčič, 2012). Tudi študija o obstoječi rabi knjižničnih katalogov, kot je npr. OCLC-jeva raziskava dnevniških datotek WorldCat, je pokazala, da se katalog uporablja predvsem za iskanje že znanih oz. drugje identificiranih virov (Wakeling et al., 2017). Podobna ugotovitev velja tudi za digitalne knjižnice (Niu, Fan in Zhang, 2019). Glede na mednarodna katalogizacijska načela pa to zagotovo ni dovolj, saj bi katalog moral postati vstopna točka tudi za raziskovanje (Izjava o mednarodnih katalogizacijskih načelih, 2017). Poleg tega se moramo zavedati, da so uporabniki vsaj z vidika vzajemnega kataloga v COBISS+ izredno raznolika skupina. Z uporabo fasetnega omejevanja se izboljša možnost navigacije pri raziskovanju ustreznega gradiva.

7 Zaključek

S teoretičnimi in metodološkimi izhodišči o kakovosti podatkov želimo osvetliti vidike in možnosti proučevanja kakovosti podatkov na področju katalogizacije. S tem prispevkom ponujamo v razmislek idejo, da je treba preseči obstoječi pristop, ki zajema preštevanje napak v zapisih. K merjenju in s tem k razumevanju kakovosti je treba pristopiti z bolj oblikovanimi mehanizmi in nivoji merjenja podatkov, ki smiselno izpostavljajo težave glede kakovosti podatkov.

Poleg tega smo s študijo primera merjenja kakovosti podatkov izpostavili specifični nabor podatkovnih elementov, ki izstopajo v rezultatih spremljanja kakovosti dnevne produkcije, saj

se z razvojem knjižničnega kataloga v sistemu COBISS.SI pojavi potreba po vnosu podatkov, ki v preteklosti niso imeli večje veljave. Z objavo podatkov v fasetah se kaže potreba po vnosu ter usklajenosti podatkov in semantični točnosti, ki podatkovni bazi COBIB.SI zagotavlja večjo zanesljivost in s tem višji ugled.

Glede na zastavljene korake je v prihodnje treba raziskati še vidik vrednotenja kakovosti podatkov, ki pa posega v razprave o aktualnosti katalogizacijskih pravilnikov in formatov ter novih podatkovnih modelov.

Reference

Badovinac, B., 2017. Izhodišča za proučevanje kakovosti podatkov v bibliografskih in normativnih zapisih: kakovost podatkov v kontekstu in raziskovalne usmeritve v katalogizaciji. *Knjižnica*, 61 (1–2), str. 119–154. Dostopno na: http://www.dlib.si/details/URN:NBN:SI:doc-QEBXUT6A [16. 4. 2019].

Badovinac, B., 2018. Nabor dimenzij za opredelitev kakovosti podatkov v bibliografskih in normativnih zapisih. *Organizacija znanja*, 23(1–2), str. 2–10. Dostopno na: http://dx.doi.org/10.3359/oz1812002 [16. 4. 2019].

Badovinac, B., 2019. »Pikice in vejice« pod drobnogledom: spremljanje kakovosti zapisov v letu 2018. *Blog COBISS*, 27. 5. 2019, http://blog.cobiss.si/2019/05/27/pikice-in-vejice-pod-drobnogledom/ [16. 6. 2019].

Bruce, T. R. in Hillman, D. I., 2004. The continuum of metadata quality: defining, expressing, exploiting. V: D. Hillmann in E. Westbrooks ur. *Metadata in practice*. Chicago: American Library Association. Str. 238−256. Dostopno na: http://www.ecommons.cornell.edu/handle/1813/7895 [14. 6. 2018].

Dornik, E., Badovinac, B., Kos, J. in Farkaš, B., 2017. Sistem zagotavljanja kakovosti COBIB.SI: izbrane aktivnosti za leto 2016. *Knjižnica*, 61(1–2), str. 191–205. Dostopno na: http://www.dlib.si/details/URN:NBN:SI:doc-6G3T8BQO [16. 6. 2019].

COBISS3/Katalogizacija, 2019. Maribor: Institut informacijskih znanosti. Dostopno na: https://izobrazevanje.izum.si/EntryFormDesktopDefault.aspx?tabid=38&type=manual&manual=1_COBISS3_Katalogizacija_svn [16. 6. 2019].

Hall, C. E., 2016. *Facets in library catalogs: the beliefs, behaviors, policies and practices that guide implementation*. Philadelphia: Faculty of Drexel University. Dostopno na: https://idea.library.drexel.edu/islandora/object/idea%3A7078 [16. 6. 2019].

Gallaway, T. in Hines, M., 2012. Competitive usability and the catalogue: a process for justification and selection of a next-generation catalogue or web-scale discovery system. *Library Trends*, 61(1), str. 173–185.

Izjava o mednarodnih katalogizacijskih načelih (ICP), izdaja 2016 z manjšimi popravki, 2017. *Knjižnica*, 61(1–2), str. 261–278.

Kaiser, M., Klier, M. in Heinrich, B., 2007. How to measure data quality? A metric-based approach. V: *ICIS 2007 Proceedings*. Atlanta: Association for Information Systems. Str. 108. Dostopno na: https://aisel.aisnet.org/icis2007/108 [16. 6. 2019].

Kavčič, I., 2012. Kakovost zapisov v vzajemni bibliografsko-kataložni bazi podatkov COBIB.SI. *Knjižničarske novice*, 22(6), str. 1–19.

Király, P., 2015. *A metadata quality assurance framework*. Göttingen: Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen. Dostopno na: http://pkiraly.github.io/metadata-quality-project-plan.pdf [16. 6. 2019].

Király, P., Stiller, J., Charles, V., Bailer, W. in Freire, N. 2019. Evaluating data quality in Europeana: metrics for multilinguality. V: E. Garoufallou, F. Sartori, R. Siatri in M. Zervas, ur. *Metadata and semantic research. MTSR 2018.* Cham: Springer. Str. 199–211. Dostopno na: https://link.springer.com/chapter/10.1007%2F978-3-030-14401-2_19 [16. 6. 2019].

Loshin, D., 2011. *The practitioner's guide to data quality improvement*. Amsterdam: Elsevier.

Margaritopoulos, M., Margaritopoulos, T., Mavridis, I. in Manitsaris, A. 2012. Quantifying and measuring metadata completeness. *JASIST* 63(4), str. 724–737. Dostopno na: https://doi.org/10.1002/asi.21706 [16. 6. 2019].

Mrđenović, B. 2018. *Katalogizacijske napake v COBIB-u z vidika katalogizatorjev in uporabnikov: magistrsko delo*. Ljubljana: B. Mrđenović.

Niu, X. in Hemminger, B., 2015. Analyzing the interaction patterns in a faceted search interface. *JASIST*, 66(5), str. 1030–1047.

Niu, X., Fan, X. in Zhang, T., 2019. Understanding facet search from data science and human factor perspectives. *ACM Transactions on Information Systems*, 37(2), 14. Dostopno na: https://dl.acm.org/citation.cfm?id=3284101 [16. 6. 2019].

Ochoa, X. in Duval, E., 2009. Automatic evaluation of metadata quality in digital repositories. *International Journal on Digital Libraries*, 10(2−3), str. 67−91.

Pipino, P., Lee, Y. W. in Wang, R. Y., 2002. Data quality assessment. *Communication of the ACM*, 45 (4), str. 211–218.

Salaba, A. in Zhang, Y., 2009. User perspectives on NextGen catalog features. *Proceedings of the American Society of Information Science and Technology*, 46(1), 1–4. Dostopno na: https://doi.org/10.1002/meet.2009.1450460372 [16. 6. 2019].

Schultz-Jones, B., Snow, K., Miksa, S. in Hasenyager Jr., R. L., 2012. Historical and current implications of cataloguing quality for next-generation catalogues. *Library Trends*, 61(1), str. 49–82.

Sebastian-Coleman, L., 2013. *Measuring data quality for ongoing improvement: a data quality assessment framework*. Amsterdam: Elsevier.

Wakeling, S., Clough, P., Connaway, L. S., Sen, B. in Tomas, D., 2017. Users and uses of a global union catalog: a mixed methods study of Worldcat.org. *JASIST*, 68(9), str. 2166–2181.

Wynne, S. in Hanscom, M., 2011. The effects of next-generation catalogs on catalogers and cataloging functions in academic libraries. *Cataloging & Classification Quarterly*, 49(3), str. 179–207.

Zeng, M. L. in Qin, J., 2016. *Metadata.* Chicago: Neal-Schuman.