Izhodišča za proučevanje kakovosti podatkov v bibliografskih in normativnih zapisih: kakovost podatkov v kontekstu in raziskovalne usmeritve v katalogizaciji

Researching the quality of data of bibliographic and authority records: the data quality in the context and research directions in cataloging

Branka Badovinac

Oddano: 10. 3. 2017 – Sprejeto: 5. 5. 2017

1.02 Pregledni znanstveni članek
1.02 Review article
UDK 025.32

Izvleček

Namen: S prispevkom smo želeli opozoriti na nekatera izhodišča za proučevanje kakovosti podatkov v bibliografskih in normativnih zapisih. Specifično nas je zanimalo 1) razumevanje kakovosti podatkov v katalogizaciji, 2) teoretične osnove in pojmi o kakovosti podatkov ter 3) usmerjenost novejših raziskav o kakovosti podatkov s področja katalogizacije (vključujoč digitalne knjižnice).

Metodologija/pristop: Prva dela raziskave sta zasnovana eksplorativno, medtem ko smo se v študiji raziskovalnih usmeritev z metodo analize literature omejili na pomembnejše objave v letih 2003–2016.

Rezultati: Kakovost podatkov smo umestili v katalogizacijo tako, da smo poskusili opredeliti dejavnike, ki vplivajo na razumevanje kakovosti podatkov v katalogizaciji. V prispevku smo povzeli nekaj izhodišč z vidikov uporabnika, enotne obdelave, organizacije/racionalizacije delovnega procesa, katalogizatorjev ter tehnologije in programske opreme. Na podlagi drugega dela prispevka o teoretičnih osnovah o kakovosti podatkov smo nato raziskovalne usmeritve povzeli skozi prizmo opredeljevanja napak, dimenzij, uporabo mer in metod merjenja, vrednotenja in interpretacije ugotovitev, tehnike in aktivnosti pri odpravljanju napak. Rezultati so pokazali možnosti razširitve nabora dimenzij, ki opredeljujejo kakovost podatkov, ter nakazali problem

ekspertnega vrednotenja in smer razvoja bolj sofisticiranih avtomatiziranih postopkov merjenja in odpravljanja napak.

Omejitve raziskave: Zaradi obsežnosti raziskovalnega vprašanja in slabše sistematičnosti pri izboru virov so nekateri vidiki premalo problematizirani.

Izvirnost/uporabnost raziskave: Študija zapolnjuje nekatere teoretične vrzeli pri preučevanju kakovosti podatkov v katalogizaciji.

Ključne besede: *katalogizacija, kakovost podatkov, metapodatki, bibliografski zapisi, normativni zapisi*

Abstract

Purpose: The aim of the study is to explore starting points in researching the quality of data in bibliographic and authority records, specifically 1) how is the concept of data quality determined in cataloguing field, 2) what are basic theoretical concepts of data quality, and 3) what are research directions in data quality according to the contemporary professional literature in cataloguing (including digital libraries).

Methodology/approach: The first two parts are designed in exploratory research manner, the third part has been conducted by literature review of contemporary works in the field of cataloguing in the years 2003–2016.

Results: Five main factors were identified in comprehension of data quality in the field of cataloguing: users, contents and format standards, workflow rationalization, cataloguers, and cataloguing technology. From the theoretical basis the contemporary literature was disposed in the following viewpoints: determining the data problems, dimensions, metrics and evaluation, quality assessment and study interpretations, and techniques and activities of data enhancement. The results are showing trends in the extension of possible dimensions and development in automatic metric and evaluation. Moreover, the subjective evaluation from the expert point of view has been highlighted.

Research limitation: Because of the extensiveness of the research question and the arbitrariness in selecting the sources, some issues have not been explored in detail.

Originality/practical implications: The study is filling up some theoretical gaps in researching data quality.

Keywords: *cataloguing, data quality, metadata, bibliographic records, authority records*

1 Uvod

S prispevkom smo želeli opozoriti na nekatera izhodišča za proučevanje kakovosti podatkov v bibliografskih in normativnih zapisih, za katera ocenjujemo, da so v slovenskem prostoru premalo raziskana. Prispevek je razdeljen na tri dele. Razumevanje kakovosti je kontekstualno, zato smo pojem kakovosti podatkov

najprej umestili na področje katalogizacije in poskusili opredeliti dejavnike, ki vplivajo na razumevanje in s tem vrednotenje kakovosti podatkov. V drugem delu smo proučili teoretične osnove in pojme o kakovosti podatkov širše, torej povzeli smo spoznanja iz sorodnih in drugih raziskovalnih področij. Na podlagi teh spoznanj smo nato v tretjem delu izvedli raziskavo, kjer nas je zanimalo, katere so raziskovalne usmeritve v novejših študijah o kakovosti podatkov s področja katalogizacije (vključujoč tudi digitalne knjižnice). Prva dela oziroma poglavji v prispevku sta zasnovana eksplorativno, medtem ko smo se v študiji raziskovalnih usmeritev z metodo analize literature omejili na pomembnejše objave v letih 2003–2016.

Zanimajo nas **podatki v bibliografskem in normativnem zapisu**, zato smo se glede na uveljavljeno terminologijo[1] odločili, da so enota našega preučevanja podatki v najširšem smislu. Čeprav bi izraz metapodatek lahko obravnavali kot le en tip podatkov v zapisu (glej Aljumaili, Karim in Tretten, 2016; Bibliotekarski terminološki slovar, 2009), ga razumemo kot sinonim za našo enoto preučevanja. Metapodatek je navadno strukturiran, kodiran podatek, ki opiše značilnosti entitete z namenom identifikacije, odkrivanja, vrednotenja, upravljanja in ohranjanja. Poznamo tri tipe metapodatkov: opisni metapodatki (opišejo vsebino vira za identifikacijo, iskanje in poizvedovanje), strukturni metapodatki (opišejo arhitekturo in odnose različnih delov vira za navigacijo) in administrativni podatki (opišejo tehnične vidike vira za obdelavo in upravljanje) (prim. Zeng in Qin, 2008).[2]

[1] *Bibliografski podatek* je vsak podatek o delu, dokumentu, publikaciji, ki skupaj z drugimi podatki omogoča njegovo identifikacijo in/ali opis. *Bibliografski opis* so po strokovnih načelih izbrani in razporejeni bibliografski podatki; *popolni nivo bibliografskega opisa* je opis z vsemi mogočimi elementi, priporočen za nacionalne bibliografije (poznamo še skrajšani in minimalni bibliografski opis). *Bibliografski vpis* je osnovna enota knjižničnega kataloga, ki vsebuje bibliografski opis, razporejevalne, lokacijske in druge podatke; *kataložni vpis* je osnovna enota knjižničnega kataloga, ki vsebuje ali bibliografski vpis ali normativni vpis ali napotilni vpis. *Popolni kataložni vpis* obsega bibliografski opis, značnico, signaturo in razporejevalni ključ. *Bibliografski zapis* je računalniški zapis bibliografskega vpisa; celotni bibliografski zapis zajame vse bibliografske podatke v skladu s katalogizacijskimi pravili. *Normativni vpis* je v klasičnem katalogu kataložni vpis, ki se začenja z iskalnim elementom v obliki enotne značnice, lahko pa vsebuje še opombe, variantne in sorodne značnice. Normativni vpis postane v računalniškem okolju *normativni zapis* (Bibliotekarski terminološki slovar, 2009).

[2] Metapodatkovni opis je sicer sestavljen iz ene ali več metapodatkovnih izjav (elementov) o enem viru, metapodatkovni zapis pa je strukturirana prezentacija opisnih, administrativnih in strukturalnih informacij enote ali zbirke (prim. Zeng in Qin, 2008).

2 Umestitev kakovosti podatkov na področju katalogizacije

Pojem kakovost podatkov v katalogizaciji se je razvijal skupaj z oblikovanjem bibliografske kontrole in vzajemne katalogizacije. V ameriškem prostoru ga je mogoče zaznati že v 19. stoletju, sredi 20. stoletja se poudarja v kontekstu produktivnosti; v 70. in 80. letih prejšnjega stoletja pa je vzporedno s pojavom kooperativnih sistemov največ pozornosti dobil v okviru podatkovnih zbirk. Kasneje se pojem kakovost podatkov pojavi v vprašanju, kaj razume in želi uporabnik; z razvojem digitalnih knjižnic in novih generacij katalogov v zadnjih desetletjih pa se pogosto pojavi v kontekstu interoperabilnosti in migracij podatkov (prim. Graham, 1990; Snow, 2011; Schultz-Jones, Snow, Miksa in Hasenyager, 2012; Zeng in Qin, 2008; Moulasion Sandy in Dykas, 2016; Tani, Candela in Castelli, 2013). Med prvimi sta v slovenskem prostoru na kakovost podatkov oziroma »ekologijo vzajemnih in lokalnih baz« opozorila Urbajs in Šobot že leta 1991; zaradi zahtev uvajanja najrazličnejših aplikacij in vključevanju novih knjižnic v sistem COBISS.SI pa se je razvoj mehanizmov zagotavljanja kakovosti postopoma nadgrajeval (Seljak, 2006).

Večina strokovnjakov se strinja, da je *namen kakovosti zapisov v prvi vrsti zadovoljiti potrebe končnih uporabnikov, v skladu z načeli vzajemnosti, standardizacije in izmenjave podatkov v lokalnem in mednarodnem okolju*. Definicije poudarjajo potrebe končnega uporabnika. Najpogosteje je v rabi pragmatična opredelitev, da naj kakovost »ustreza namenu«, v katalogizaciji je to v prvi vrsti cilj knjižničnega kataloga (prim. Graham, 1990; Harmon, 1996; Thomas, 1996; Petek, 1998; Paiste, 2003; MacEwan in Young, 2004; Danskin, 2006; Hider in Tan, 2008; Stvilia in Gasser, 2008; Bade, 2008).

Pojem kakovost podatkov je vpet v opredelitve kakovostnega kataloga. Npr. po Myall in Chambers (2007) lahko kakovost kataloga opredelimo: 1) z vidika kakovosti posameznih kataložnih zapisov glede na zastavljeno funkcionalnost in dogovorjene standarde in 2) z vidika kakovosti celotne strukture in predstavitve kataloga ter 3) z vidika kakovosti katalogizacijske storitve kot dela celotne knjižnične dejavnosti (npr. pravočasnost vnosa zapisa). Z vidika ugotavljanja uspešnosti katalogizacijskega procesa lahko katalog ovrednotimo s kvantitativnimi (npr. število prirasta, čas obdelave) in kvalitativnimi kazalci (delež ustreznosti zapisov glede na standarde) (Massey, 2000). Po Petek (1998) pa kakovost kataloga lahko presojamo z vidika pravilnosti kataložnega vpisa in z vidika zadovoljstva uporabnikov (npr. uspešnostjo iskanja).

Na področju digitalnih knjižnic se kakovost podatkov umešča v splošne evalvacijske analize (prim. Tani idr., 2013), z vidika sistemov upravljanja znanja pa se kakovost podatkov lahko meri na nivoju vsebine informacij, vira informacij in kakovosti informacijskega sistema (prim. Aljumaili idr., 2016).

Na dojemanje kakovosti podatkov vpliva več dejavnikov, v nadaljevanju smo povzeli nekaj izhodišč z vidikov: *uporabnik kataloga in drugih storitev, enotna obdelava, organizacija/racionalizacija delovnega procesa, katalogizator, tehnologija in programska oprema*.

2.1 Uporabnik

Uporabnik je najpomembnejši dejavnik pri presoji kakovosti podatkov. Žal so študije uporabnikov o kakovosti podatkov redke, verjetno tudi zaradi njihove zahtevne izvedljivosti, posploševanje rezultatov pa je navadno omejeno na posamezne študije primera (manjše skupine uporabnikov). Študije uporabnikov so tesno vezane na preučevanje javnih računalniških katalogov (OPAC), kar povzroča metodološke in interpretativne nejasnosti, npr. ne vemo, kaj v resnici uporabnik vrednoti: kakovost podatkov ali funkcionalnost vmesnika, iskalnika, prikaz rezultatov itn. (prim. Ma, Mo in Luo, 2014). Ne nazadnje pa kakovosti podatkov ne moremo več pogojevati le v okviru kakovosti kataloga, saj knjižnični katalog ni več glavna vstopna točka iskanja virov in tudi ne edini produkt knjižnice.

V številnih študijah, ki so povezane s katalogom in pojmom kakovosti podatkov, razbiramo informacije o obsegu (naboru) in vrstah podatkov. Starejše študije so ugotavljale, da uporabniki uporabljajo zelo malo podatkov: najpogosteje so bili v rabi podatek o avtorju, naslovu in letu izida, manj podatki o ilustracijah, velikost dokumenta in knjižne zbirke, podobni rezultati pa so veljali tudi za knjižnično osebje. Uporabniki so najpogosteje iskali dejanski dokument ali pa dokumente na določeno temo; slednje je značilno tudi za tedanje računalniške kataloge (prim. Petek, 1998). Da je za uporabnike najpomembnejša identifikacija vira in da potrebujejo le malo podatkov, zapiše tudi Kavčič (2012), ki je pri presoji o kakovosti bibliografskih zapisov preverjala, ali lahko uporabniki najdejo določeni vir na podlagi v zapisu navedenih podatkov.

Ne nazadnje Kavčič in Velkavrh (2009) argumentirata tudi, da je knjižničar med drugim uporabnik. Odmevna OCLC-jeva študija iz leta 2009 pa kaže, da obstajajo razmeroma velike razlike med potrebami končnih uporabnikov in knjižničarji kot uporabniki (Calhoun, Cantrell, Gallagher in Hawk, 2009). Ugotovili so, da končni uporabniki spletnih (online) katalogov pričakujejo več kot gole podatke, potrebujejo dodane vrednosti podatkov (povzetke, kazalo vsebin) in pomoč pri iskanju informacij z različnimi orodji (npr. omejevanje) ter sorodne predmetne oznake. Ugotovitev, da uporabniki od knjižničnih katalogov pričakujejo več podatkov, kot jih navaja ISBD; medtem ko so nekateri elementi verjetno odveč, potrjujeta tudi Švab in Žumer (2016). Petrucciani (2015) pa opozori, da četudi so podatki točni, težave izhajajo iz razlik med »katalogizacijskim jezikom« in obliko podatkov, ki

bi jih uporabniki razumeli. »Jezik katalogizacije« je namreč abstrakten, medtem ko je jezik končnih uporabnikov specifičen.

2.2 Enotna obdelava

Enotna obdelava je izrazito vpeta v mednarodne standarde za zajem vsebine, strukture in vrednosti[3] podatkov in je v strokovni literaturi deležna največ pozornosti. V zadnjih dveh desetletjih so standardi v postopkih posodabljanja. Novo izhodišče v katalogizaciji je Iflina družina konceptualnih modelov FZBZ,[4] ki podajajo osnovo, kako informacijo opišemo in organiziramo oziroma jo identificiramo in določimo odnose. Zanimivo je, da niti model Funkcionalne zahteve za bibliografske zapise – FZBZ ne temelji na empiričnih raziskavah, nekaj kasnejših študij pa je potrdilo intuitivnost modela (prim. Pisanski in Žumer, 2009; Zhang in Salaba, 2012; Cossham, 2013).

Ne glede na pomanjkljivosti modela (prim. Petrucciani, 2015) so na teh osnovah oblikovana nova katalogizacijska načela (Izjava, 2009; Statement, 2016). Ta načela se nanašajo na izdelavo katalogizacijskega pravilnika, cilje in funkcije kataloga, zahteve iskanja in najdenja. Poleg tega še določajo, da morajo pravilniki upoštevati entitete, atribute in odnose v Iflinih konceptualnih modelih (da mora bibliografski opis, tj. opisni del bibliografskega zapisa, temeljiti na Iflinem standardu ISBD), način izbire točk dostopa[5] v bibliografskih podatkih (zapisih)[6] ter obliko in izbiro točk dostopa v normativnih podatkih (zapisih). V izdajo načel iz leta 2016 so vključene še značilnosti novih kategorij uporabnikov in spremenjenih uporabniških vedenj, novih orodij za odkrivanje informacij, interoperabilnost in dostopnost podatkov.

Pri izdelavi katalogizacijskega pravilnika je najpomembnejše načelo uporabnik – ta določa tudi, kakšno besedišče in katere elemente opisa uporabiti; slednji morajo biti v bibliografskem smislu pomembni. Podatki morajo biti, po načelih Izjave (2009), točni, prepisani tako, kot so na viru, navedeni ter oblikovani in strukturirani na podlagi standardov, kar pomeni njihovo doslednost. Opisi za vse vrste gradiva naj temeljijo na enotnih pravilih. Med bolj poudarjenimi načeli

[3] Angl. izraz *data value* na tem mestu ne pomeni vrednotenje podatka, temveč gre za dejanski zapis podatka (npr. številčna vrednost).

[4] Iflina družina konceptualnih modelov: http://www.ifla.org/node/2016.

[5] Točka dostopa je ime, pojem, kodiran podatek itd., s pomočjo katerega iščemo bibliografske in normativne podatke in jih identificiramo (prim. Izjava, 2009).

[6] V dopolnjeni izdaji iz leta 2016 je namesto bibliografskega/normativnega zapisa uporabljen termin bibliografski/normativni podatki; zapis je namreč le en način agregacije in predstavitve podatkov.

je tudi interoperabilnost, ki omogoča izmenjavo podatkov znotraj in zunaj knjižničarske skupnosti. V poglavju o iskanju in najdenju Izjava določa minimalni nivo podatkov, tj. seznam bistvenih točk dostopa bibliografskih in normativnih podatkov.

Kot določa Izjava, je ISBD osrednji standard za opisno katalogizacijo vira. Standard je bil zasnovan v listkovnem obdobju kataloga z namenom, da se ne glede na tehnološki razvoj in kulturne ter jezikovne razlike omogoči mednarodna izmenjava podatkov (skladno z Iflinim programom UBC). Standard določa nabor, strukturo (vrstni red) in predstavitev (prepis) podatkov. V kontekstu FZBZ-ja se ISBD ukvarja s praktično aplikacijo opisa za identifikacijo pojavne oblike. Nova združena izdaja naj bi prinesla korenite spremembe standarda glede na tehnološke možnosti in zahteve uporabnikov. Npr. nova izdaja se odziva na potrebe granulacije informacij, medtem ko interpunkcija vztraja – ta predstavlja sintakso oziroma določa kontekst med elementi v opisu kot celoti. Med argumenti nove izdaje je tudi večja kakovost podatkov, saj standard omogoča konsistentnost podatkov iz vseh vrst objavljenih virov na različnih nivojih nabora podatkov za različne uporabnike (ustanove) ter stopnjo obveznosti posameznih elementov (Escolano Rodriguez, 2014).

Na izvedbeni ravni so določila ISBD-ja vključena v posameznem katalogizacijskem pravilniku, ki odraža kulturne posebnosti določene skupnosti. Zaradi ekonomičnosti se številne države odločajo za (delni) prevzem novega ameriškega pravilnika Resource Description and Access (v nadaljevanju RDA), ki ima težnjo postati mednarodni standard. RDA vsebuje pravila, katere informacije potrebujemo za identifikacijo vira in njegovih odnosov z drugimi viri. Obeta se, da bo RDA zagotovil večjo kakovost podatkov, saj se namesto na zapis osredotoča prav na podatek. Zapis tako ni več statičen, temveč dinamičen, omogoča pa tudi uvedbo mreže povezav med podatki. Pravilnik RDA daje večji poudarek zapisovanju oziroma zajemanju podatkov, medtem ko je predstavitev podatkov odvisna od posameznih tehnoloških zmožnosti in potreb. Natančnost in verodostojnost popisa podatkov izhajata iz metodološkega načela: »vzemi, kar vidiš« (angl. »take what you see«) in »sprejmi, kar dobiš« (angl. »accept what you get«) (Bianchini in Guerrini, 2016). V sodobnem pravilniku značnica nima več osrednjega pomena,[7] v elektronskem okolju so vsi elementi vira najdljivi – vir ima namreč več točk dostopa, nekatere med njimi so kontrolirane. Slednje so bistveni korak k večji konsistentnosti podatkov (prim. Seljak idr., 2004; Zalokar, 2006), njihov nabor pa se še širi z različnimi projekti, npr. normativne kontrole založnikov (Connaway in Dickey, 2011).

[7] Koncept značnice je lahko uporaben le še npr. v izpisih citatov, literature ipd.

Podatke o viru v računalniškem okolju vpisujemo v paket, t. i. komunikacijski format, ta format je sedaj v katalogizaciji prevladujoča družina formatov MARC, ki je izpeljan iz standarda ISO 2709 in je bil zasnovan za shranjevanje in izmenjavo podatkov ter za izpis listka. V primerjavi s številom podatkov ISBD se je nabor podatkov, ki jih katalogizator navaja, s formatom MARC razširil. Format je sestavljen iz treh elementov: struktura zapisa (sistemska polja), označevalci (polja, podpolja, indikatorji) in vsebina (to določajo pravilniki za opisno in vsebinsko katalogizacijo). Zajema torej opisne podatke, administrativne podatke o zapisu, identifikacijske podatke (npr. DOI), relacijske podatke (povezave z drugimi viri), podatke o pravicah in tehnične podatke (npr. podatek o programu za uporabo vira).

MARC je zelo strukturiran in semantično bogat z metapodatki, a je v primerjavi z drugimi metapodatkovni standardi neracionalen tako časovno kot finančno. Po eni strani ima preobsežno strukturo elementov, po drugi strani pa je omejen tudi za opise nekaterih tipov virov ter zastarel glede tehnoloških zmožnosti, ki jim niti pravilniki niti razvoj MARC-a niso sledili (prim. On the record, 2008; Leckie, Given in Campbell, 2009). Leta 2011 je Kongresna knjižnica predlagala, da se ugotovijo vidiki trenutne metapodatkovne sheme (kateri naj se obdržijo in kateri dodajo), preučijo možnosti semantičnega spleta in povezovanja podatkov, pospešijo možnosti ponovne uporabe metapodatkov, omogočijo možnosti krmarjenja med različnimi entitetami (FZBZ), preučijo možnosti prezentacije metapodatkov, ki presegajo sedanje sisteme MARC-a, opredelijo načini prehoda v strokovni skupnosti in načrtujejo prenosi metapodatkov v nove bibliografske sisteme glede na sedanjo tehnično infrastrukturo (Transforming, 2011). Tehnologija novih orodij v katalogizaciji naj bila prostodostopna, temeljila naj bi na formatu, ki je fleksibilen in nadgradljiv ter prilagojen potrebam mobilnega sveta.

Nastala je iniciativa BIBFRAME,[8] okvirni model, ki bo omogočil enostavno implementacijo pravilnika RDA in druge sorodne koncepte. Sedanja različica sheme BIBFRAME 2.0 ima v primerjavi z RDA-jem širši nabor podatkov z namenom, da omogoči opis čim več tipov virov, podpre nove uporabniške vmesnike in zadosti potrebam nadgradnje, granularnosti in kompleksnega povezovanja virov.

Poleg trdnih konceptualnih izhodišč pa se v stroki poudarja tudi potreba po kvalitetni in ažurni dokumentaciji enotne obdelave (prim. Seljak, 2000; Pesjak in Petek, 2010); nekaj raziskav kaže, da so napake v zapisih med drugim posledica rabe katalogizacijskih pravil in priročnikov, ki so po mnenju katalogizatorjev

[8] BIBFRAME: https://www.loc.gov/bibframe.

preveč nejasna, podrobna, nepovezana, hitro spremenljiva (Romero in Romero, 1992).

2.3 Organizacija/racionalizacija delovnega procesa

Bade (2008) poudarja, da je stroka preveč obremenjena z idejo o »popolnem bibliografskem zapisu« (angl. perfect bibliographic record), s čimer se v bistvu izogiba debati o kakovosti v katalogizaciji. Ker ne vemo, kaj naj bi popolni bibliografski zapis predstavljal, Bade predlaga pragmatični pristop, ki bo pomagal pri določitvi, kateri elementi so pomembni za uporabnika in koliko teh elementov si ustanova lahko privošči.

Posodabljanje procesa katalogizacije naj bo podprto z ekonomičnimi rešitvami vzajemnosti in povezovanja, saj knjižničarji sami ne bodo uspeli uresničevati ciljev informacijsko bogatega knjižničnega kataloga (prim. Calhoun idr., 2009; Krstulović, 2006). ALA je leta 2010 pozvala strokovno javnost k preučevanju stroškov in vrednosti bibliografske kontrole, v literaturi lahko zasledimo nekaj študij o dejanskih stroških katalogizacije (npr. Kont, 2015), vendar matrika stroškov in koristi (angl. cost-benefit) še ni povsem razvita. Velik problem je vrednotenje podatkov in zapisov ipd. (Stalberg in Cronin, 2011). Kavčič (2012) na primer predlaga določitev nivoja zapisov ter zmanjšanje nabora podatkov na srednjem nivoju bibliografskega opisa.

Katalogizacija predstavlja sorazmerno velik strošek, zato obstajajo različni pogledi med skupino zagovornikov, za katere je katalogizacija oblika intelektualnega dela, in predstavniki, ki katalogizacijo obravnavajo kot tehnično, rutinsko opravilo. Prvi se zavzemajo za čim natančnejši kataložni zapis, lokalno katalogizacijo glede na potrebe dejanskega uporabnika in podrobna katalogizacijska pravila ter za strokovni in usposobljen kader, ki lahko zagotovi intelektualno osnovo pravilnosti zapisa. Medtem drugi očitajo prvim časovno neučinkovitost, zlasti za tisto gradivo, ki je manj v uporabi, in preširok nabor bibliografskih podatkov, ki jih povprečni uporabniki ne potrebujejo. Sami pa podpirajo zaposlovanje tehničnih delavcev in zunanje izvajanje storitev (angl. outsourcing) ter vse večjo stopnjo avtomatizacije, tj. uvažanje paketov zapisov različnih ponudnikov (prim. Graham, 1990; Thomas, 1996; Massey, 2000; Paiste, 2003).

Primer tovrstnega spora je uvedba vzajemnih bibliografskih sistemov v 70. letih 20. stoletja, ko se je pojavil strah pred nekakovostnimi bibliografskimi zbirkami in deprofesionalizacijo. V resnici so bile kritike o nekakovostnih zapisih pretirane (Intner, 1989); Cook in Payn (1991 cv: Massey, 2000) pa sta dokazala, da so spletni katalogi popolnejši in imajo manj napak v primerjavi z listkovnimi katalogi.

Četudi se je potreba po zagotavljanju in kontroli kakovosti v decentraliziranih sistemih večala, so se zaradi ekonomske učinkovitosti ukinjali katalogizacijski oddelki in s tem notranja kontrola kakovosti (Hafter, 1986).[9]

Največji problem v katalogizaciji je podvajanje dela. Zunanji izvajalci po mnenju prve skupine ogrožajo status in avtonomijo katalogizacije, kar se kaže v zmanjšani potrebi po usposobljenih katalogizatorjih, ki so s svojim ekspertnim znanjem nepogrešljivi pri najrazličnejših razvojnih projektih. Razvojniki digitalnih knjižnic so recimo sprva razglašali, da je katalogizacija zastarelo opravilo in da so nabori bibliografskih podatkov preobsežni. Izkazalo se je, da tudi obsežne zbirke s celotnimi besedili potrebujejo kakovostne metapodatke, ustvarjene z intelektualnim delom; številni projekti zaradi čim hitrejše izgradnje digitalnih knjižnic ne predvidevajo ali si ne morejo privoščiti kontrole kakovosti, npr. že pri dodajanju metapodatkovnih zapisov – postopek, ki je v katalogizacijski praksi sicer običajen. Poleg tega pa nove zahteve v katalogizaciji kažejo potrebo po več podatkih s semantičnimi povezavami, ki zahtevajo več intelektualnega dela. Uravnoteženost med obema poloma zatorej izhaja iz možnosti avtomatizacije katalogiziranja z določenim, a neobhodnim intelektualnim vložkom (prim. Thomas, 1996; Zeng in Qin, 2008; Cox in Myers, 2010; Tani idr., 2013; Diao in Hernadez, 2014; Moulaison Sandy in Dykas, 2016).

2.4 Katalogizator

Katalogizator ima po Krstuloviću (2006) najodgovornejšo nalogo. Zagotoviti mora, da se »struktura podatkov, ki jo opredeljujejo katalogizacijska pravila, ustrezno razvrsti (pravilno preslika) v elemente podatkovnega formata bibliografskega sistema«. Avtor zato zaključi, da je podatkovni niz težko preverjati izključno z algoritmi in da je ne glede na izbor katalogizacijskih pravil pri zagotavljanju kakovosti bibliografskih podatkov odločujoč subjektivni dejavnik, tj. katalogizator.

Tudi Dimec (1994 cv: Likar, 2003) je ugotavljala, da na kakovost zlasti vpliva usposobljenost katalogizatorjev, saj so težave pri vnosu podatkov v COMARC večinoma odraz preslabega poznavanja strokovne obdelave gradiva in nezadostne uporabe priročnikov, pa tudi, da na kakovost zapisov vpliva pomanjkljiva priprava in podcenjevanje katalogizacijskega dela. Podobno sta tudi Pesjak in Petek

[9] Tako so v ameriškem prostoru oblikovali program za vzajemno katalogizacijo (angl. PCC, http://www.loc.gov/aba/pcc/), ki je razvil standardni zapis (angl. standard record). Ta naj bi predstavljal zanesljiv, točen in veljaven zapis, katerega osnovni namen je zmanjšati stroške obdelave in povečati kakovost.

leta 2010 preverjala, ali uporaba priročnikov lahko vpliva na kakovost bibliografskih zapisov. Ugotovila sta, da manjšina, ki ne uporablja priročnikov, meni, da v prvi vrsti priročniki ne ponujajo ustreznih rešitev, priročnikov pa ne uporabljajo tudi zaradi pomanjkanja časa ter zadostnega poznavanja pravil – raba katalogizacijskih priročnikov namreč z izkušnjami pada. Zanimivo je, da več kot polovica respondentov želi več kontrole, saj bi se kakovost zapisov zvišala. Sicer pa si, kot kaže OCLC-jeva študija, knjižničarji najbolj želijo odpraviti podvojene zapise in tipkarske napake, nato pa dopolniti kratke zapise in dodati druge podatke: kazala vsebin, povzetke in slike ovitkov. Slednje pa ne velja za katalogizatorje, ki želijo zlasti orodje za popravljanje zapisov (Calhoun idr., 2009).

Tudi Schultz-Jones in sodelavci (2012) poročajo o rezultatih treh študij, kjer so poudarjeni katalogizatorji. V prvi razkrivajo, da katalogizatorji v splošnih knjižnicah priročnike uporabljajo redko ali občasno, ker jih ne poznajo in ker jim primanjkuje finančnih sredstev in osebja. Zelo veliko zaupanja imajo v zapise različnih dobaviteljev. Druga študija med profesionalnimi in pomožnimi katalogizatorji v visokošolskih knjižnicah razkriva, da kakovost v katalogizaciji najpogosteje zanje predstavlja 1) tehnično podrobnost bibliografskega zapisa (točnost in popolnost podatkov), 2) vpliv katalogizacije na uporabnike (npr. najdljivost, dostopnost idr.), 3) usklajenost s katalogizacijskimi standardi ter (4) katalogizacijski postopek (npr. delitev dela). Med najpomembnejšimi podatki v formatu MARC so polja točk dostopa (z izjemo leto izdaje): naslov, avtor, predmetne oznake. Tretja študija pa se je izvajala med šolskimi knjižničarji in je med drugim razkrila, da je katalogizacija drugotnega pomena, da obstaja veliko zaupanje v zapise dobaviteljev in da jih večina še ni slišala za uvedbo novega pravilnika RDA ter da se o tem tudi ni pogovarjala z dobavitelji.

Zanimiva je tudi študija napak na manjšem vzorcu študentov, ki niso imeli praktičnih izkušenj s katalogizacijo. Po kratkem usposabljanju z AACR2 so kreirali zapise za glasbene tiske. Analiza napak je pokazala, da so v bistvu katalogizirali po načinu RDA, čeprav se z njim niso srečali, kar kaže na enostavnost in intuitivnost novega pravilnika ter obete za lažje izobraževanje novih katalogizatorjev (Harden, 2012). Kavčič (2012) pa poudarja, da posodobitev pravil ni edini pogoj za odpravo neupoštevanja pravil in površnost katalogizatorjev, ampak so nujna tudi dodatna izobraževanja. Po njenem mnenju ima dober katalogizator osnovno znanje iz katalogizacije in povprečno splošno razgledanost ter je zelo natančen pri prepisu podatkov. Podobno tudi Romero in Romero (1992) opozarjata na pomanjkanje specifičnega znanja (npr. znanje tujih jezikov), omejenosti znanja (npr. interdisciplinarno gradivo) in nepoznavanje katalogizacijskih pravil pri katalogizatorjih.

2.5 Tehnologija in programska oprema

Tehnologija in programska oprema v procesu katalogizacije omogočata in določata izvedljivost modelov in učinkovitost katalogizacijskih postopkov. Vprašanje odnosa med kakovostjo zapisa in vmesnika za katalogizacijo sta poudarili tudi Likar in Žumer (2004), ko sta preverili, kakšen odnos imajo katalogizatorji do segmenta COBISS2/Katalogizacija. Študija je pokazala, da so bili katalogizatorji zadovoljni z zanesljivostjo, hitrostjo in kakovostjo segmenta, razmeroma tudi z enostavnostjo in zaslonskim prikazom.

V skladu z razvojem formatov se razvijajo tudi nova katalogizacijska orodja, npr. orodje BIBFRAME EDITOR za katalogizacijo v formatu BIBFRAME je usklajeno s standardom povezanih podatkov. Tehnološki izzivi tovrstnega orodja so shranjevanje podatkov, oblikovanje katalogov s tehnologijo, ki bo razumela podatke formata BIBFRAME, in usklajevanje besednjaka z drugimi shemami (prim. Yang in Li, 2015).

Veliko kakovostnih podatkov zahtevajo nove generacije katalogov in migracije v nove modele t. i. FZBZ-izacija. Pri njihovem uvajanju so se pokazale najrazličnejše težave s podatki, npr. zaradi nekonsistentnosti v kodiranju, določanja točk dostopa, nepravilnih predmetnih oznak, pomanjkanja klasifikacije e-virov, integracije kontroliranih in nekontroliranih geslovnikov, integracije metapodatkov v shemo MARC itn. (prim. npr. Wayne in Hanscom, 2011 cv: Schultz-Jones idr., 2012; Mitchell in McCallum, 2012).

Knjižnični metapodatki so bili vedno kompleksen pojem, bodisi v tehnološkem bodisi v vsebinskem smislu. Čeprav posamezne podatke v zapisu lahko upravljamo ločeno, je bil bibliografski zapis vedno uporabljen kot smiselna celota o pojavni obliki. Zapisi so vedno del podatkovne baze in s težavo sodelujejo v svetu povezanih podatkov (angl. linked data). Zaradi zaprtosti sistema uporabniki najraje začnejo iskanje pri drugih ponudnikih. Podatki naj bodo strukturirani s povezavami na druge podatke zato, da bodo lažje odkriti, vloga kataloga je v tem okviru povezava med virom (znanjem) na spletu in knjižnico. Ko so bibliografski metapodatki dostopni/objavljeni na spletu, se razširi tudi funkcija bibliografske kontrole; ta bo sodelovalna, decentralizirana, internacionalizirana in zasnovana na spletu. Coyle (2010) ugotavlja, da težave z interpretacijo in rabo podatkov za orodja semantičnega spleta izhajajo iz tega, da FZBZ in RDA nista povsem razvita v skladu z zahtevami standardov povezanih podatkov, čeprav vsebujeta koncepte semantičnega spleta.

3 Kratek uvod v kakovost podatkov

Kakovost podatkov je ključnega pomena za zanesljivost in učinkovitost ter uspešnost ustanov in informacijskih sistemov. Zaradi nekakovostnih podatkov nastanejo neposredni in posredni stroški (npr. stroški odprave napak, izguba ugleda itn.). Kakovost podatkov je vpeta v zakonodajne okvire, recimo v direktivo 2003/98/CE Evropske unije (z dopolnitvijo 2013/37/EU), ki govori o ponovni uporabi javnih podatkov;[10] prav tako je predmet različnih mednarodnih standardov (npr. ISO 25012, ISO 8000, ISO 9000).

Zato ni presenetljivo, da obstaja precejšnje število opredelitev kakovosti podatkov. Ena izmed najosnovnejših je t. i. naravna oziroma inherentna definicija, ki opisuje stopnjo, s katero podatek najboljše oziroma čim bolje odraža realni svet, medtem ko pragmatična ali stvarna kakovost podatkov opisuje stopnjo uporabnosti podatkov za neki namen. Problematika kakovosti podatkov narašča z različnostjo podatkov (npr. strukturirani, nestrukturirani itn.), kategorijami podatkov (npr. glavni, začasni, zgodovinski) in tipi informacijskih sistemov (npr. monolitni, distribuirani, kooperativni itn.). Nekateri strokovnjaki opozarjajo tudi na razliko med kakovostjo podatkov in kakovostjo informacij. Kakovost informacij je načeloma širši pojem od kakovosti podatkov, saj je odvisna od konteksta uporabe in vsebine (Stvilia, Gasser, Twidale in Smith, 2007) oziroma namena in znanja avtorja, intertekstualnosti, družbenih in kulturnih norm o vsebini vira ter aktivnosti oziroma interesov bralca (Mai, 2013).

Definicija kakovosti podatkov je odvisna zlasti od vključenih dimenzij, to so značilnosti podatka, ki jih lahko merimo in ovrednotimo glede na zastavljene standarde. Dimenzije se lahko nanašajo na vrednost podatka (angl. data value) ali na shemo (npr. shema podatkovne baze); vse dimenzije so opisane na kvalitativni način. Poimenovanj, opredelitev dimenzij in njihove kategorizacije je v literaturi veliko, najpogostejše so naslednje (prim. Batini in Scannapiecco, 2016):

- Točnost (angl. accuracy): ali podatek pravilno predstavlja realni svet ali dogodek? V grobem jo delimo v dve različni obliki: sintaktična točnost (npr. tipkarska napaka) in semantična točnost (npr. napačno avtorstvo).
- Popolnost (angl. completness): ali podatek vsebuje vse dele podatka, ki predstavljajo en subjekt ali dogodek? Do katere stopnje vneseni podatek opisuje realni svet?
- Konsistentnost (oziroma doslednost) (angl. consistency): koliko so upoštevana semantična pravila?

[10] Direktiva določa, da morajo biti dokumenti skupaj s svojimi metapodatki splošno dostopni, in sicer v obliki – formatu, ki ni odvisen od specifičnega programja, pojem ponovne uporabe pa je zagotovilo, da bodo podatki kakovostni, točni in aktualni.

– Skupina časovnih dimenzij, ki vključujejo spremembe in posodobitve podatkov skozi čas:
  ▸ Aktualnost (angl. currency): kako hitro so podatki posodobljeni, npr. datum zadnje posodobitve?
  ▸ Pravočasnost (angl. timeliness): kako aktualen je podatek za neki cilj, nalogo ipd.?
  ▸ Nestanovitnost (angl. volatility): kako pogosto se podatki spreminjajo skozi čas?

V uporabi pa so še naslednje dimenzije: edinstvenost (angl. uniqueness) (npr. delež dvojnikov), ustrezna količina podatkov (angl. appropriate amount of data), dostopnost oziroma razpoložljivost (angl. accessibility), verodostojnost (angl. credibility), interpretativnost (angl. interpretability), uporabnost (angl. usability), jedrnatost (angl. conciseness), izčrpnost (angl. comprehensiveness), pravilnost (angl. correctness) in ustreznost oziroma relevantnost (angl. relevancy). Za shemo lahko poudarimo še dimenzijo minimalnosti (angl. minimality), tj. vsak element v shemi je predviden le enkrat (npr. odvečni atribut za entiteto). Dodamo pa lahko tudi dimenzije, ki so bliže pojmu kakovosti informacij, npr. stopnja nevtralnosti, kredibilnosti, objektivnosti (prim. Mai, 2013).

Za primer kategorizacije dimenzij najprej opozorimo na znano Wangovo in Strongovo (1996) opredelitev. Raziskovalca sta na podlagi študije kupcev opredelila 15 dimenzij in jih razvrstila v štiri kategorije (Preglednica 1).

Preglednica 1: *Primer kategorizacije dimenzij kakovosti podatkov (po Wang in Strong (1996))*

Kategorija kakovosti podatkov	**Dimenzije**
Notranja, neločljiva kakovost podatkov (angl. intrinsic) (zajeti kakovost podatka takšnega, kot je)	Točnost (natančnost), objektivnost (celostni in nepristranski podatki), verodostojnost, ugled (podatki so zanesljivi glede na njihov vir in vsebino)
Kontekstualna kakovost podatkov (upošteva kontekst uporabe podatkov)	Dodana vrednost, relevantnost, pravočasnost, popolnost, ustreznost količine podatkov
Predstavitvena kakovost podatkov (angl. representational) (oblika podatkov)	Interpretativnost, razumljivost, doslednost, natančna predstavitev
Dostopnost	Dostopnost, varnost

Primer naslednje kategorizacije dimenzij pa povzemamo po Redman, Fox in Levitin (2009), ki do dimenzij pristopajo iz opredelitve podatka. Ta je v določenem *konceptualnem modelu* (oziroma podatkovnem modelu) opredeljen z *vrednostjo*, ki je izbrana iz *domene atributov* za določeno *entiteto*. Skupino entitet, ki si v konceptualnem modelu delijo nabor atributov, imenujemo *tip entitete* (oziroma

razred, kategorija, set). Podatek pa zapišemo v skladu s pravili za *reprezentacijo podatka* s pomočjo *formata* za zapisovanje vrednosti podatka. Iz tega sledi, da kakovost podatkov določajo značilnosti (dimenzije), ki jih lahko kategoriziramo v tri skupine: 1) *kakovost konceptualnega modela (dimenzije sheme)*, 2) *kakovost vrednosti* in 3) *kakovost reprezentacije podatka v formatu* (Preglednica 2).

Preglednica 2: *Kategorizacija dimenzij po Redman, Fox in Levitin (2009)*

1. *Kakovost konceptualnega modela (dimenzije sheme)* se nanaša na: – vsebino (značilnosti dejstev, ki jih predstavlja podatek) z dimenzijami: ustreznost, natančnost definicij (nedvoumne opredelitve sestavnih komponent v modelu), pridobivanje podatkov (za entitete in atribute) (npr. pravna podlaga za zbiranje podatkov); – obseg modela z dimenzijami: izčrpnost in poglavitnost (angl. essential) (obširen tako, da zadovolji vse potrebe uporabnikov in izključi nepotrebne informacije); – nivo specifičnosti z dimenzijami: granularnost atributov (npr. bolj podrobni podatki omogočajo več možnosti za uspešno iskanje in popravljanje napak, vendar jih je težje in dražje pridobiti in vzdrževati) in natančnost domene (nanaša se na izbiro mere/klasifikacije izbranega elementa); – kompozicijo (strukturiranje in grupiranje dejstev, ki jih predstavljajo podatki) z dimenzijami: naravnost (npr. vsak podatek naj ustreza svoji skupini atributov in entitet), identifikacijska oznaka (unikatna oznaka, primarni ključ) in homogenost tipov entitet (vsak atribut se lahko uporabi pri vseh entitetah istega tipa); – konsistentnost z dimenzijami: semantična doslednost (med definicijami sorodnih komponent v modelu) in strukturna doslednost (ko kombiniramo komponente modela); – odziv na spremembe z dimenzijami: možnost dodajanja/brisanja entitet, spreminjanja vrednosti in domen ter uvedba novega atributa ali tipa entitete.
2. *Kakovost vrednosti* (angl. data value) zajema dimenzije: točnost, aktualnost, popolnost (in edinstvenost) ter konsistentnost (in celovitost).
3. *Kakovost reprezentacije podatka v formatu* zajema dimenzije: primernost, nedvoumnost, univerzalnost (razumljivost podatka za večino uporabnikov), natančnost, fleksibilnost formata, sposobnost predstavitve praznih vrednosti (angl. null values) in učinkovita uporaba formata (angl. recording medium).

Dimenzije niso neodvisne; med njimi obstajajo korelacije, ki so določene s specifiko proučevanega področja, npr. med konsistentnostjo in popolnostjo je razmerje pogosto obratno sorazmerno. Odločitev, katere dimenzije so pomembnejše za določeno opredelitev kakovosti, lahko izhaja iz teoretskega, intuitivnega ali raziskovalnega pristopa (prim. Wang in Strong, 1996). Dimenzije ne dajejo kvantitativnih mer (angl. metrics); te so ločene lastnosti dimenzij. Za vsako mero obstaja več metod merjenja, glede na to, kje je merjenje izvedeno, kateri podatki so vključeni, s katerim merilnim instrumentom ter po kateri lestvici so rezultati predstavljeni (Batini in Scannapiecco, 2016).

Dimenzije uporabljamo v različnih vlogah, v različnih tehnikah in modelih. Modele v podatkovnih bazah uporabljamo za opis podatka in podatkovne sheme (npr. model entitet in povezav); v informacijskih sistemih z modeli opišemo

poslovne procese organizacije. Tehnike so algoritmi, procedure in postopki, s katerimi rešujemo specifični problem kakovosti podatkov v okviru določene aktivnosti (npr. ročni popravki nabora zapisov). Aktivnosti so lahko osvežitve oziroma dopolnitve z novimi podatki, standardizacija (oziroma normalizacija), identifikacija objektov[11] in odprava dvojnikov, integracija podatkov, določanje zanesljivosti podatkov, lokalizacija oziroma odkrivanje ter odpravljanje napak in optimizacija stroškov (npr. razmerje med ceno in kakovostjo uvoženih podatkov) (prim. Batini in Scannapiecco, 2016).

Za zagotavljanje kakovosti podatkov (angl. data quality assurance) potrebujemo splošni okvir, (metodologijo, model), ki glede na potrebe organizacij ali informacijskih sistemov daje smernice za:
- **ugotovitev** stanja kakovosti podatkov (npr. kontekstualne informacije o podatkih – taksonomija nepravilnih podatkov (prim. Kim idr., 2003), študije uporabnikov itn.),
- najbolj učinkovito in stroškovno upravičeno **merjenje**[12] kakovosti podatkov (glede na določene dimenzije) ter
- **izboljšanje** kakovosti podatkov oziroma uvedbo kontrole preprečevanja težav (prim. Vetro idr. 2016).

Na voljo so številni okviri, ki so navadno vpeti v celovito zagotavljanje kakovosti organizacije, sistema, najbolj znana sta npr. Epplerjev *Information Quality Measurement* (IQM) in Englishev *Total Information Quality Management* (TIQM). Sicer pa jih lahko kategoriziramo v štiri splošne skupine: 1) celovite metodologije (okviri), ki zagotavljajo podporo za ugotavljanje/merjenje in izboljšavo ter poudarjajo tehnične in ekonomske vidike; 2) revizijske metodologije, ki se osredotočajo na ugotavljanje in merjenje z manjšim poudarkom na izboljšavah; 3) operativne metodologije, ki so osredotočene na tehnične vidike ugotavljanja/merjenja in izboljšav, brez ekonomskih vidikov; 4) ekonomske metodologije, ki so osredotočene na stroške evalvacije (prim. Batini in Scannapiecco, 2016).

Številni avtorji opozarjajo, da imajo napake, ki jih uporabniki najdejo, neposreden vpliv na zadovoljstvo uporabnikov, ki je končno merilo kakovosti podatkov. Prav tako poudarjajo, da je iskanje in popravljanje napak občutno dražje v primerjavi s kontrolo vnosa podatkov. Zato Redman, Fox in Levitin (2009) priporočajo, da naj menedžment kakovosti zajema hkrati tri pristope: s prvim pristopom se osredotočimo na nadzor in predelavo končnih produktov, z drugim se usmerimo

[11] Angl. izrazi: data matching, record linkage, data linkage, entity resolution, object identification, field matching. S to aktivnostjo identificiramo, uskladimo in tudi lahko združimo zapise, ki se nanašajo na isto entiteto ene ali več podatkovnih baz.

[12] V uporabi so tudi izrazi evalvacija, ocenjevanje, vrednotenje, preverjanje, revizija ipd.

na zmanjšanje vzrokov napak v procesu izdelave in dostave končnih produktov, medtem ko se pri tretjem pristopu osredotočimo na celostno oblikovanje procesov, s katerimi onemogočimo pojavljanje napak (npr. poenostavljanje postopkov, implementacija novih tehnologij ipd.). Sicer pa so problemi, zaradi katerih nastajajo napake, večplastni. Galway in Hank (2011) jih npr. opredelita na treh nivojih:

- *Operativni:* podatki so netočni, nepopolni ali manjkajoči zaradi težav pri zajemanju ali prenosu podatkov (npr. odvečno ponavljanje vnosa podatkov, težave zaradi metode zajemanja podatkov, kontrole vnosa, kodirnih shem itn.).
- *Konceptualni:* podatki so netočni, nepopolni ali manjkajoči zaradi slabo definiranih podatkov oziroma podatkov, ki niso namenjeni uporabi (npr. slabo opredeljeni nabor elementov, slabo opredeljene kodirne sheme itn.).
- *Organizacijski*: nenehni operativni in konceptualni problemi zaradi različnih težav med tistimi, ki zbirajo podatke, in tistimi, ki podatke uporabljajo.

4 Raziskovalne usmeritve v študijah kakovosti (meta)podatkov na področju katalogizacije (vključujoč digitalne knjižnice)

4.1 Namen in metodološka zasnova

Z analizo novejše literature smo želeli ugotoviti, katere so raziskovalne usmeritve na temo kakovosti (meta)podatkov na področju katalogizacije, vključujoč digitalne knjižnice in repozitorije. Zanimalo nas je torej, kakšni so načini raziskovanja in obeti z vidika raziskovalnih in metodoloških pristopov ter spoznanj. Tuje znanstvene vire smo iskali v podatkovnih zbirkah (npr. LISTA, ESBCO, LISA, Web of Science, Scopus), omejili smo se na vire s področja katalogizacije in knjižničnih katalogov ter digitalnih knjižnic v obdobju zadnjega desetletja. V iskanju smo uporabili kombinacije ključnih besed (v angl.): *quality, data, bibliographic, records, metadata, information, auditing, assessing, evaluation, database, control* ipd. Nabor gradiva smo dopolnili z metodo sledenja referenc navedenih v prispevkih in z viri s slovenskega prostora. Vire smo nato v končni vzorec uvrstili glede na njihovo odmevnost in presojo o izvirnosti, uporabnosti ipd. v primerjavi s celotnim pregledanim korpusom gradiva. V končni vzorec smo uvrstili 27 virov, objavljenih v obdobju 2003–2016.

4.2 Ugotovitve

Preglednica v Prilogi 1 navaja vire v kronološkem redu z informacijami o namenu študije, metodološkem pristopu, navedenimi (oziroma identificiranimi) dimenzijami in ugotovitvami raziskave. V večini primerov gre za aplikativne raziskave, vključeni pa so tudi nekateri zanimivi teoretični prispevki. V prikazu ugotovitev

vire ne primerjamo med seboj in jih ne ovrednotimo, temveč jih smiselno povzemamo po problemskih sklopih, ki smo jih oblikovali glede na zgoraj predstavljeno poglavje o splošnem teoretičnem uvodu o kakovosti podatkov. Ti sklopi so: opredelitev napak, dimenzije, mere, metode merjenja in evalvacije, vrednotenje in interpretacija rezultatov, tehnike in aktivnosti za odpravljanje napak.

4.2.1 Opredelitev napak

Ugotavljanje vrste napak in njihovo štetje je precej uveljavljen pristop, značilen zlasti za starejše študije (prim. Romero in Romero, 1992; Massey, 2000; Shin 2003). V slovenskem prostoru ga zasledimo tudi pri Crnčić (2010), ki je v vzorec zajela 50 zapisov za monografske publikacije ter z njihovo primerjavo s katalogizacijskimi priročniki ugotovila, da je 72 % zapisov imelo skupaj 63 različnih napak, največ v območju založništva in distribucije. Enci (2011) pa je na podlagi vzorca 50 zapisov za serijske publikacije ugotovila, da je 94 % zapisov vsebovalo skupaj 128 napak, opredeljenih glede na priporočila ISBD(CR). Največ napak je bilo v območju založništva in fizičnega opisa ter območju opomb. Te napake po mnenju Encijeve nimajo vpliva na najdenje publikacije, saj večina teh podatkov po ISBD(CR)-ju ni obveznih.

Več pozornosti opredelitvi kakovosti podatkov se v zadnjem času posveča na področju digitalnih knjižnic, npr. Yaser (2011) je opredelil pet problemov: netočna vrednost, nepravilen element, manjkajoči podatek (informacija), izgubljena informacija (npr. zaradi konverzij), nedosledna vrednost. Wisser (2014) je na primeru agregiranih podatkov za imena korporacij in osebnih imen razkrila 30 različnih vrst napak, med drugim: napačno kodiranje, napačna interpunkcija in tipkarske napake, prisotnost kvalifikatorjev, krajšave. Najpogostejši napaki sta bili nekonsistentna vsebina (oblikovanje podatka) in nekonsistentni format (osebno ime/korporacija).

Parkova (2006) pa je izhajala iz formata; z raziskavo je želela ugotoviti katera poimenovanja polj v shemi Dublin Core so težko razumljiva, kar povzroča netočnost podatkov in s tem neinteroperabilnost med različnimi zbirkami. Z analizo zapisov treh različnih zbirk je ugotavljala, v katerih poljih je največ napak, ki so rezultat neenotnega razumevanje vsebine oziroma namena posameznega metapodatkovnega elementa. Ugotovila je, da je največ težav pri razumevanju pri naslednjih metapodatkovnih elementih: format, tip, opis, vir in relacije. Avtorica je zato predlagala, da bi pomene in razumevanje metapodatkovnih elementov poenotili.

Dimenzije. Izbira dimenzij oziroma kazalnikov je po mnenju Zeng in Qin (2008) navadno stvar intuitivnega razumevanja problema, izkustev in strokovne

literature, v kateri sicer ni soglasja. Pri aplikativnih raziskavah sta najpogosteje uporabljeni dve dimenziji: točnost in popolnost, medtem ko jih v teoretičnih prispevkih predvidevajo več.[13] Bruce in Hilman (2004) sta opredelila sedem dimenzij: popolnost, izvor, točnost, skladnost s pričakovanim, logična konsistentnost in skladnost, pravočasnost in dostopnost. Avtorja razlikujeta tri nivoje kakovosti metapodatkov: semantično strukturo (format), sintaktično strukturo (shema) in dejanske podatke. Stvilia s sodelavci (Stvilia idr., 2007; Stvilia in Gasser, 2008) je pri oblikovanju splošnega modela kakovosti informacij opredelil kar 22 dimenzij. Te so nato kategorizirali v tri večje skupine, ki se nanašajo na: 1) dogovorjene standarde (skupina notranjih dimenzij), 2) odnos med informacijo in njeno uporabo (skupina relacijskih/kontekstualnih dimenzij) ter 3) merodajnost danega podatka/informacije (skupina uglednost).

Zanimiv je tudi pristop Moulaisonove (2015), pri kateri med drugim zasledimo časovne dimenzije. Avtorica je preučila, kako se normativnim zapisom za osebe po RDA-ju dodajajo atributi skozi čas, natančneje v enem letu. Glede na popolnost je študija pokazala nizko kakovost zapisov. Število metapodatkov, povezanih z osebo, je malo in zelo počasi naraščajo. Najpogostejši atribut je letnica, sledijo jezik, država in spol entitete.

Taniguchi (2005, 2007) je predlagal sistem, v katerem se pri katalogizaciji poleg podatkov v bibliografskem zapisu navedejo tudi dokazi (npr. pravila), na osnovi katerih je bil določen podatek v bibliografskem zapisu izbran in oblikovan. Sistem navajanja dokazov omogoča razumevanje, zakaj je določena vrednost podatka v bibliografskem zapisu navedena. Na takšen način bi se povečali izraznost (angl. expressivity) in zanesljivost podatkov ter s tem interoperabilnost in življenjska doba bibliografskih zapisov ter deskriptivnih metapodatkov. Shranjevanje dokazov je uporabno tudi za razumevanje posameznih podatkov ter morebitno zasnovo za avtomatizacijo katalogizacije virov nasploh. V postopku testiranja modela je Taniguchi predvidel tudi način (pol)avtomatizacije beleženja dokazov že v postopku kreiranja in redigiranja zapisov.

4.2.2 Mere

Za merjenje kakovosti zapisa in bibliografskih podatkov sta najpogosteje v rabi dve meri: 1) količina podatkov v zapisu in 2) delež napak (vključujoč tudi izpuščanje) na zapis v primerjavi s popolnim zapisom, tj. zapisom, kot ga določa neka ustanova glede na dogovorjene standarde (Hider in Tan, 2008). Najpogostejši je

[13] Za primere starejših študij glej npr. Thomas (1996) in Tenopir (1990).

prvi način, to so obsežne študije štetja polj/podpolj oziroma ugotavljanja rabe označevalcev v formatu. Petek (2012) je npr. v okviru pomembnosti enotnega naslova kot podatka opravila študijo o uporabi polj 300 in 500 v COBIB-u in CROLIST-u ter ugotovila, da večina zapisov nima želenega podatka, uporaba polj pa je nedosledna. Lundy (2006) je predstavil analizo primerjave uporabe polj v zapisih MARC21 v dveh katalogih glede na navodila PCC za antikvarno gradivo in redke knjige. Ugotovil je, da je standard deloma uporaben, vendar v praksi katalogizatorji za to vrsto gradiva potrebujejo večji nabor polj, zato je treba standard dopolniti.

Eklund (2009) je s sodelavci izvedel obsežno študijo o rabi več kot 2000 označevalcev v zapisih MARC. Analiza 56 milijonov zapisov v WorldCatu, ki so bili razdeljeni na 10 setov glede na vrsto gradiva oziroma 20 podsetov glede na popolnost zapisa (nivo), je pokazala, da se povprečno uporablja zelo omejen nabor različnih polj in podpolj, ki so katalogizatorjem na voljo.[14] Rezultate so nato primerjali z navodili Kongresne knjižnice za minimalni nivo obdelave in standardnimi zapisi PCC BIBCO. Ugotovljene razlike med polji kažejo, da navodila niso bila pripravljena na osnovi empiričnih raziskav, temveč v skladu z interesi zmanjševanja stroškov in standardizacije. V okviru projekta se je kasneje pojavilo tudi vprašanje, kako ti zapisi podpirajo uporabniška opravila po FZBZ-ju; za monografije so npr. ugotovili, da le 16 % uporabljenih polj podpira opravilo najdi, 12 % uporabljenih polj podpira opravilo identifikacije, 16,3 % podpira opravilo izbire in 11,9 % podpira opravilo pridobi. Sicer pa te ugotovitve ne odgovarjajo na vprašanje, ali so ti rezultati zadovoljivi in ali prisotnost polj v resnici kaže na njihovo rabo pri poizvedovanju informacij (prim. Miksa, 2007).

Kasneje so pri OCLC izvedli podobno študijo na primeru 145 milijonov zapisov. Poleg končnih uporabnikov, knjižničarjev (in katalogizatorjev) so med uporabnike metapodatkov vključili tudi strojne aplikacije, ki se uporabljajo za skupno iskanje in indeksiranje agregiranih podatkov vzajemnih podatkovnih zbirk. Ugotovili so, da se v zapisih od 200 zajetih polj uporablja le manjši nabor polj (npr. najpogosteje polja naslova in odgovornosti, založništva in fizičnega opisa), prav tako se le manjši nabor indeksira v večini vmesnikov knjižničnih sistemov. Razkrili sta se tudi nekonsistentnost uporabe kod in nezanesljivost oznake o popolnosti zapisa. Avtorji so med drugim opozorili na razlike med samimi strojnimi aplikacijami, ki najpogosteje uporabljajo poenostavljene algoritme. Sicer pa strojne aplikacije težko interpretirajo opombe, čeprav so ta polja v zapisih relativno pogosta. Ne nazadnje so tudi ugotovili, da analize dnevnikov (angl. log analysis) ne zagotavljajo

[14] Oziroma se, kot ugotovijo Moen idr. (2006), v zapisih za monografije, kreirane v LC, le 15 % polj pojavi v več kot 50 % zapisov in le 167 različnih polj se pojavi vsaj v enem zapisu.

zadostnih informacij o vedenju uporabnikov tj. pogostnost uporabe polj v odnosu do uresničitev uporabniških potreb (Smith-Yoshimura idr., 2010).

Po Zeng in Qin (2008) lahko merimo kakovost na nivoju zbirke, zapisa in elementa. Čeprav je običajna enota merjenja metapodatkovni zapis v izbranem vzorcu podatkovne zbirke, v zadnjem času postaja zanimiv tudi posamezni element v zapisu, tj. niz podatkov o posameznem viru (izjava). Predlog mer za splošni model kakovosti informacij podajo tudi Stvilia idr. (2007). Na primer problem manjkajočih elementov glede na priporočljivi nabor sovpada z dimenzijo relacijske popolnosti, ki se jo lahko izračuna na podlagi formule FZBZ in identificira avtomatizirano. Medtem identifikacija elementov, ki vsebujejo napačne vrednosti glede na dani standard (dimenzija relacijske semantične konsistentnosti), zahteva polavtomatizirano štetje tovrstnih napak.

4.2.3 Metode merjenja in evalvacije

Najpogostejša metoda identifikacije/analize napak je ročna, na osnovi ekspertnega mnenja, ki ga lahko deloma dopolnimo z avtomatizirano identifikacijo napak na osnovi danih algoritmov. Pri konverzijah se lahko izvajajo primerjave med zapisi, sicer pa je po Zeng in Qin (2008) primerjava zapisov in primarnega gradiva nujna, čeprav so v praksi redke. Npr. Likarjeva (2003) je z metodologijo CAT-ASSES (prim. Chapman in Massey, 2002[15]) ugotavljala kakovost obdelave slovenskih spletnih (online) serijskih publikacij in tistih tiskanih serijskih publikacij, ki imajo vzporedno izdajo v spletni (online) obliki. Ekspertna študija ročne primerjave zapisov s knjižničnim gradivom je zajela celotni zapis s posebnim poudarkom na elementih, ki omogočajo najdljivost, identifikacijo in dostopnost. Likar (2003) zaključuje, da je bila le slaba tretjina zapisov zadovoljivih, največ pomanjkljivosti je bilo v poljih za vsebinsko obdelavo ter v poljih za elektronsko lokacijo in dostop (polje 856) ter pri določanju indikatorjev v tem polju. Pri publikacijah z vzporedno spletno izdajo pa je bilo največ napak pri polju 856 in njegovem indikatorju ter v polju za vnos ISSN-ja.

Ochoa in Duval (2009) sta izhajala iz ugotovitve, da ročna evalvacija vzorca ni uporabna metoda za hitro naraščajoče digitalne zbirke, zato sta poskusila zasnovati avtomatizirano evalvacijo. Po njunem mnenju mora ta imeti dve značilnosti:

[15] Chapman in Massey (2002), ki sta zasnovala metodo in orodje za evalvacijo katalogov (CAT-ASSESS), sta predlagala dva simultana ročna testa: »katalog-zbirka« in »zbirka-katalog«. Pri prvem se na osnovi naključnega vzorčenja in kodiranja napak po določeni shemi preveri pravilnost zapisa z dejanskim gradivom, pri drugem pa lahko poleg napak najdemo tudi primere nekatalogiziranega gradiva in dvojnikov, s čimer ugotavljamo popolnost kataloga.

nadgradljivost, tj. avtomatski izračun za vsak nov vneseni metapodatkovni primerek (angl. scalability), in uporabnost mere (angl. meaningfulness). Za avtomatizirano zagotavljanje kakovosti sta avtorja izbrala sedem dimenzij, ki sta jih predlagala Bruce in Hilmann (2004), in jim dodelila mere. Npr. mera za popolnost se lahko izračuna s prisotnostjo elementa v zapisu, izračun je podprt tudi s pomembnostjo elementa (pomembnost pa lahko razberemo recimo analiz dnevnikov iskanja (log datotek)). Pri primerjavi avtomatiziranega izračuna in rezultatov skupine pregledovalcev zapisov se je izkazalo, da rezultati kakovosti podatkov ne sovpadajo, neekspertno ocenjevanje zapisov pa je nezanesljiva metoda. Pri primerjavi visoko in nizko kakovostne zbirke se je izkazalo, da je avtomatizacija merjenja kakovosti zadovoljiva. Medtem je primerjava med ekspertnim in avtomatiziranim določanjem najslabše kakovosti metapodatka glede posameznih dimenzij v zapisu pokazala, da je kljub razlikam avtomatiziran postopek merjenja nekaterih dimenzij uporaben. S študijo so ugotovili zelo nizek konsenz ekspertne (ročne) evalvacije, avtomatizacija je uporabna le pri nekaterih merah (npr. stopnja izpolnjevanja zahtev skupnosti za določeno uporabniško nalogo), druge pa, četudi ne sovpadajo z rezultati ročne evalvacije, so zaradi svoje učinkovitosti pogojno uporabne pri zagotavljanju kakovosti podatkov.

Metodologije so lahko vpete tudi v splošni evalvacijski model, npr. na področju digitalnih knjižnic, kjer se dopolnjujeta metoda zbiranja mnenj uporabnikov (npr. intervjuji) in metoda zbiranja podatkov o uporabi (npr. analize log datotek) (prim. Zeng in Qin, 2008).

4.2.4 Vrednotenje in interpretacija rezultatov

Zeng in Qin (2008) ločita dva nivoja vrednotenja: makrovrednotenje izvedemo z analizo na nivoju zbirke, mikrovrednotenje pa nivoju elementov. Po Krstulović (2006) morajo bibliografski podatki ustrezati semantični opredelitvi (pod)polja, njihovo pravilnost in ustreznost pa dokončno potrdi kontekst zapisa kot celote. Na semantični ravni namreč ocenjujemo le natančnost vnosa podatkov, medtem ko na kontekstualni ravni raziščemo, ali so izpolnjena vsa za identifikacijo publikacije nujna polja in ali pravilni podatki v posameznih (pod)poljih niso v medsebojnem nasprotju ali navzkrižju. Torej je podatek na ravni (pod)polja lahko pravilen, a je napačen v kontekstu bibliografskega zapisa kot celote.

Kavčič (2012) je z manjšo študijo poskusila pokazati, da so slabe ocene kakovosti zapisov v postopku preverjanja naključnih zapisov pretirane glede na potrebe splošnega uporabnika, saj so viri najdljivi kljub omejenim obsegu podatkov. Ocene izhajajo iz strokovnih zahtev, zato predlaga spremembo kriterijev ocenjevanja. Obenem pa ugotavlja, da katalogizatorji pomanjkljivosti ocenjenih zapisov

pogosto ne odpravijo zaradi nestrinjanja z redaktorjem, zaradi nepoznavanja pravil ali pa pomanjkanja interesa.

Z metodološkega vidika preučevanja kakovosti podatkov je štetje napak nesmiselno, če vrednotenje ni povezano z uporabniškimi potrebami. Po MacEwan in Young (2004) to lahko storimo s pomočjo uporabniških opravil FZBZ-ja. Za britansko nacionalno knjižnico so razvili matriko pomembnosti posameznih atributov pojavne oblike, s katero so omogočili izračun odstotka kakovosti opisne in vsebinske obdelave ter splošne kakovosti posameznega zapisa, kar so preverili tudi na osnovi manjšega vzorca zapisov v MARC21, ki so jih primerjali s predlogo – virom.

Kritična do vrednotenja na osnovi ekspertnega mnenja sta zlasti Hider in Tan (2008); različni knjižničarji bi namreč pod okriljem enakih standardov različno katalogizirali isto gradivo, prav tako različni katalogizatorji različno ocenjujejo kakovost zapisov. Primerjalne analize med študijami so zaradi različnosti vzorčenja, metode analize in taksonomije napak in interpretacije precej otežene. Nekateri ugotavljajo nizko kakovost zapisov, spet drugi zagovarjajo, da so napake le kozmetične (npr. El-Sherbini, 2010).

Hider in Tan (2008) se zato zavzemata za katalogizacijo, ki bo podprta z dokazi (angl. evidence-based cataloging), sama pa sta zasnovala študijo, ki je sicer še vedno temeljila na napakah kot osnovni enoti, a so bile te ovrednotene glede na dejansko uporabo kataloga – identifikacijo in izbiro virov med uporabniki singapurskih splošnih knjižnic. Z intervjuji in anketami ter raziskovalno metodo glasnega razmišljanja sta ugotavljala, kako uporabniki v javnem računalniškem katalogu ovrednotijo bibliografske elemente v različnih mogočih scenarijih. Rezultati so pokazali, da so pri identifikaciji najbolj uporabni podatki o naslovu in avtorju ter povzetki in slika ovoja, pri izbiri pa je bolj kot slika ovoja pomembna vsebina. Uporabnike so povprašali tudi o učinku štirih tipov napak – izpuščanje, razlikovanje, tipkarske napake in napačna raba ločil – na identifikacijo in izbiro virov. Rezultati so pokazali, da ni velikih odstopanj med pomembnostjo tipov napak, najmanjšo težo imajo napake, povezane z rabo ločil. Na podlagi rezultatov sta avtorja nato oblikovala preglednico ovrednotenih napak za najbolj uporabljena polja in podpolja formata MARC21. Razen napak z ločili in »oblikovnih« napak imajo vse druge napake vpliv na poizvedovanje, zato sta jih razdelila v dve skupini: bistvene napake (vključujoč zatipkanost) in oblikovne napake (npr. napačna raba ločil, z nekaterimi izjemami, npr. navedbo URL-ja). Vrednotenje posameznega polja sta izvedla z lestvico pomembnosti od 0 do 9 (višja stopnja pomeni večjo pomembnost) in skupaj z naborom polj in podpolj predstavlja priporočilo pri katalogizaciji gradiva.

4.2.5 Tehnike in aktivnosti izboljšanja kakovosti podatkov

Tani, Canela in Castelli (2013) so na področju digitalnih knjižnic poskušali izluščiti naslednje pristope za odpravo problemov:
- uveljavitev skupne obdelave (standardi in navodila);
- evalvacija metapodatkov s podpornimi orodji: analitično usmerjeni pristopi opredelitve problemov (npr. avtomatizirana validacija – programske kontrole, črkovalniki) in pristop, usmerjen v uporabnike (odziv končnih uporabnikov in administratorjev, npr. orodja za sporočanje napak, programsko preverjanje zapisov);
- polavtomatizirano generiranje metapodatkov (ekstrakcija in zbiranje metapodatkov neposredno iz virov);
- odprava, izboljšanje in dopolnitev metapodatkov (projekti normativne kontrole, avtomatizirano popravljanje napak).

V strokovni literaturi zasledimo veliko poročil o posameznih aktivnostih, kot je projekt *Izboljšanja kakovosti v COBIB.SI*, ki je bil izveden leta 2005 in 2006. V tem projektu so se osredotočili na probleme pri zapisih za posamezne vrste publikacij, na probleme pri zapisih določenega statusa, pa tudi na probleme v posameznih delih zapisov (bibliografski del, podatki o zalogi), in z njim odpravili vrsto težav, kot je poenotenje polj za opombe, kontrola neverificiranih ISSN, kontrola vnosa v polje 115b itn. (Rogina, 2007).

Najpogostejše aktivnosti so uvajanje programske kontrole podatkov, avtomatizacija vrednotenja posameznih vidikov ter zaznavanje tipkarskih in podobnih napak, uvajanje normativne kontrole in kontrole zaznavanja duplikatov. Po Zen in Qin (2008) k večji kakovosti prispevajo tudi t. i. vnaprej pripravljene predloge, ki imajo določene tudi stopnje obveznosti posameznih metapodatkovnih elementov. Likarjeva (2003) je npr. predlagala izdelavo smernic, priročnikov, navodil in vzorčnih zapisov. Van Kleeck in sodelavci (2016) pa na študiji primera predstavljajo uvedbo mednarodnih standardov za katalogizacijo in načine pogajanja z dobavitelji podatkov. Pri kontroli so se usmerili na urejanje zapisov, postopke »čiščenja« paketnih zapisov, identifikacijo potencialnih napak in analizo trendov. Uporabili so tudi orodje CatQC, ki poroča o potencialno manj kakovostnih uvoženih zapisih, uvedli pa so tudi možnost, da na napake opozorijo uporabniki.

5 Razprava

Čeprav so zaradi obsežnosti izbrane raziskovalne teme in slabše sistematičnosti pri izboru virov nekateri vidiki v tem prispevku ostali neproblematizirani,

ocenjujemo, da smo z našo pregledno raziskavo poudarili pomembna izhodišča za proučevanje kakovosti podatkov v bibliografskih in normativnih zapisih.

Najprej smo izhajali iz predpostavke, da je razumevanje kakovosti podatkov kontekstualno. Na podlagi strokovnega diskurza smo poudarili pet dejavnikov, ki lahko vplivajo na dojemanje in s tem na vrednotenje kakovosti. V prispevku smo povzeli nekaj izhodišč z vidika: končnih uporabnikov, enotne obdelave, katalogizatorja, organizacije/racionalizacije delovnega procesa in tehnologije in programske opreme.

Ugotovitve kažejo, da je kakovost podatkov v katalogizaciji pogosto poudarjen pojem, posebej z mislijo na končne uporabnike. Opredelitve kakovosti so navadno pragmatične, najpogosteje v odnosu s knjižničnim katalogom, čeprav ta ni več primarna izhodiščna uporabniška točka niti edina storitev, pri kateri knjižnice in servisi uporabljajo podatke bibliografskih in normativnih zapisov. Razumevanje kakovosti podatkov z vidika uporabnika kot najpomembnejšega presojevalca kakovosti je žal redko podprto z empiričnimi raziskavami, verjetno tudi zaradi njihove težje izvedljivosti. Iz večine (ne)posrednih raziskav se lahko razbereta predvsem nabor in pomembnost podatkov. Zato pa je lažje kot končne uporabnike v metodološkem smislu preučevati katalogizatorje, ki se v literaturi tudi kažejo kot poglavitni dejavnik kakovosti podatkov. Te raziskave se usmerjajo na pogostnost uporabo priročnikov, usposabljanje in znanje katalogizatorja pri preslikavi podatkov v format; poleg tega katalogizator lahko nastopi tudi v vlogi presojevalca kakovosti.

Pomanjkanje študij končnih uporabnikov zasledimo tudi pri razvoju novih modelov in enotnih dogovorov o obdelavi, čeprav se ti v praksi že udejanjajo. Med poglavitnimi argumenti novih dokumentov o enotni obdelavi je prav večanje kakovosti podatkov, četudi nekatera pričakovanja strokovne javnosti, ki bi lahko vplivala na podatke, niso uresničena (npr. odprava interpunkcije v ISBD-ju). Zanimivo je tudi, da so osrednja tema zanimanja predvsem podatki bibliografskega opisa, čeprav se je nabor podatkov v elektronskem okolju s formatom MARC že pred desetletji razširil (npr. podatki o zapisu, podatki o elementih bibliografskega zapisa itn.). Vmes se je zgodila tudi dekonstrukcija bibliografskega zapisa – ta ne temelji več na statičnih podatkih, temveč se nenehno dopolnjuje in osvežuje. Težave z zastarelostjo formata MARC pa so vodile v oblikovanje novega, še razvijajočega se fleksibilnega formata BIBFRAME. Sicer pa, kot kažejo študije, je dobro urejena in dostopna dokumentacija o enotni obdelavi pomemben dejavnik pri kakovostni katalogizaciji.

Opazili smo tudi potrebo po vnosu vse več podatkov, kot jih je mogoče razbrati iz vidika uporabnika in enotne obdelave. Nekateri podatki oziroma dodane vrednosti zapisa so bile v dosedanjih standardih za opis vira že predvidene (npr.

kazalo vsebin, povzetki, relacije), a zanje ni bilo dovolj interesa. Realizacija teh potreb je namreč odvisna tudi organizacije delovnega procesa katalogizacije, ki predstavlja enega največjih izdatkov knjižnic. K zmanjševanju stroškov se je pristopalo z odpravljanjem podvajanja del in zmanjševanjem števila strokovnih delavcev, kar je vodilo tudi v deprofesionalizacijo v katalogizaciji. Racionalizacija v obliki poenostavljanja katalogizacije in nabora podatkov je vplivala na kakovost podatkov, npr. večina zunanjih dobaviteljev podatkov ne dosega katalogizacijskih standardov za opis vira. V tem oziru je umestna Grahamova teza (1990), da sta produktivnost in kakovost v katalogizaciji obratno sorazmerni. Ne nazadnje kakovostne podatke potrebujejo tudi strojne aplikacije, ki predstavljajo novi tip uporabnika. Zanimivo je torej, da tehnologija v tem primeru predstavlja porabnika podatkov, ne pa dejavnik, ki omogoča pogoje za boljšo kakovost podatkov (npr. programska oprema za katalogizacijo).

V drugem delu prispevka smo se v teoretičnem uvodu seznanili z osnovnimi pojmi in problematiko preučevanja kakovosti podatkov nasploh. Poudarjen je bil pomen dimenzij pri opredelitvi kakovosti podatkov in njihove kategorizacije. Med drugim smo nakazali strukturo splošnega okvira zagotavljanja kakovosti podatkov. Poudarili smo tudi, da imajo napake, ki jih uporabniki najdejo, neposreden vpliv na uporabniško izkušnjo.

Na podlagi teoretičnega pregleda o kakovosti podatkov smo v tretjem delu prispevka prikazali izsledke analize literature o raziskovalni usmeritvah v obdobju 2003–2016, tako da smo oblikovali šest tematskih sklopov: opredelitev napak, dimenzije, mere, metode merjenja in evalvacije, vrednotenje in interpretacija rezultatov, tehnike in aktivnosti za odpravljanje napak. Analiza literature je pokazala, da precej pozornosti še vedno vzbuja pristop iskanja oziroma opredeljevanja napak s poskusi njihovega vrednotenja. Predlagani so tudi bolj sofisticirani in obsežnejši nabori dimenzij (npr. časovne dimenzije), čeprav sta največkrat v rabi dve: popolnost in točnost, ki se ju meri z ročnim štetjem na osnovi določene kategorizacije napak. Osrednja enota preučevanja so bibliografski in normativni zapisi, pokazal pa se je tudi trend granulacije – usmeritev k posameznem podatku – elementu zapisa. Strokovnjaki opozarjajo, da se kakovost največkrat vrednoti na osnovi (subjektivnega) ekspertnega mnenja; metode analize potekajo ročno – pregledi zapisov/podatkov glede na veljavno dokumentacijo enotne obdelave, in to brez primarnega gradiva, čeprav je slednje zaželeno. In čeprav naj bi standardi za kakovost postajali bolj eksplicitni in objektivni, primerjave med raziskavami niso zaželene, saj se uporablja različno vzorčenje, tipologijo napak in metodologijo vrednotenja.

Veliko zanimanja je tudi za avtomatizacijo merjenja kakovosti in evalvacijo ter za aktivnosti pri izboljšanju kakovosti podatkov. Prav tako so v porastu obsežne

študije, ki proučujejo pogostnost rabe posameznih označevalcev/polj, kot osnovo za razvoj navodil in novih modelov enotne obdelave; poudarjajo pa se potrebe po vrednotenju na osnovi uporabniških opravil. Ne nazadnje strokovnjaki poročajo tudi o težavah s kakovostjo podatkov pri postopkih interoperabilnosti in prenosom podatkov v druge modele.

6 Zaključki

Pregled strokovne literature o kakovosti podatkov nakazuje možnosti nadaljnjega proučevanja tega področja, saj je kakovost podatkov pomemben aspekt najrazličnejših informacijskih sistemov, posebej tako kompleksnih, kot so vzajemni bibliografski sistemi. Z našo pregledno raziskavo smo najprej želeli pokazati, da kakovost podatkov lahko razumemo z različnih vidikov. Na primer, v kontekstu uporabnika in enotne obdelave zasledimo miselnost, da bi se z razširitvijo oziroma reorganizacijo nabora elementov povečala kakovost podatkov bibliografskega in normativnega zapisa. Bibliografski in normativni zapisi z dodano vrednostjo bi tako bolje zadovoljili zahteve naprednih knjižničnih katalogov in storitev. Velik poudarek pa je tudi na potrebah strojnih aplikacij kot novega tipa uporabnika kakovostnih podatkov. Poleg normativne kontrole še vedno obstaja velika težnja po odpravljanju podvajanja dela in vključevanju drugih partnerjev, s katerimi bi bibliografske zapise lahko še bolj obogatili.

Pomembno izhodišče je tudi poznavanje raziskovalnih in metodoloških pristopov ter izsledkov s področja proučevanja kakovosti podatkov. Kakovost podatkov je treba natančno opredeliti z dobro definiranimi dimenzijami, merami, metodami merjenja in vrednotenja. Analiza literature raziskovalnih usmeritev s področja katalogizacije (vključujoč digitalne knjižnice) je pokazala na možnosti razširitve dimenzij, ki opredeljujejo kakovost podatkov, nakazal se je tudi problem ekspertnega vrednotenja in trend k razvijanju bolj sofisticiranih avtomatiziranih postopkov merjenja in odpravljanja napak, čeprav je vnos podatkov še vedno ključna faza v modelu zagotavljanja kakovosti. V tem oziru je katalogizator s svojim znanjem nedvomno najpomembnejši dejavnik.

V nadaljnjih študijah o kakovosti podatkov v bibliografskih in normativnih zapisih bi bilo smiselno raziskati tudi možnosti za enotnejše razumevanje in vrednotenje kakovosti, npr. z oblikovanjem modela kakovosti podatkov bibliografskih in normativnih zapisov. V skladu s tem bi bilo treba določiti taksonomijo napak na osnovi FZBZ-ja in mapiranje dimenzij, določiti pa bi bilo treba tudi mere in mogoče tehnike merjenja kakovosti podatkov v okviru posamezne dimenzije.

Navedeni viri

Aljumaili, M., Karim, R. in Tretten, P. (2016). Metadata-based data quality assessment. *Vine*, *46*(2), 232–250.

Bade, D. (2008). The perfect bibliographic record: platonic ideal, rhetorical strategy or nonsense?. *Cataloging and classification quarterly*, *46*(1), 109–133.

Batini, C. in Scannapieca, M. (2016). *Data and information quality: dimensions, principles and techniques*. Berlin: Springer.

Bianchini, C. in Guerrini, M. (2016). RDA: a content standard to ensure the quality of data. *JLIS*, *7*(2). 83–98

Bibliotekarski terminološki slovar. (2009). Ljubljana: Zveza bibliotekarskih društev Slovenije.

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* (str. 238–256). Chicago: American Library Association. Pridobljeno 12. 3. 2012 s spletne strani: http://www.ecommons.cornell.edu/handle/1813/7895

Calhoun, K., Cantrell, J., Gallagher, P. in Hawk, J. (2009). *Online catalogs: what users and librarians want: an OCLC report*. Dublin, Oh.: OCLC. Pridobljeno 12. 12. 2016 s spletne strani: http://www.oclc.org/us/en/reports/onlinecatalogs/fullreport.pdf

Chapman, A. in Massey, O. (2002). A catalogue quality audit tool. *Library and information research news*, *26*(82), 26–37.

Colyle, K. (2010). Library data in the web world. *Library technology reports*, *46*(2), 5–11.

Connaway, L. S. in Dickey, T. J. (2011). Publisher names in bibliographic data: an experimental authority file and a prototype application. *Library resources and technical services*, *55*(4), 182–194.

Cossham, A. F. (2013). Bibliographic records in an online environment. *Information research*, *18*(3), paper C42.

Cox, E. J. in Myers, A. K. D. (2010). What is a professional cataloger?: perception differences between professionals and paraprofessionals. *Library resources and technical services*, *54*(4), 212–226.

Crnčić, M. (2010). *Kakovost bibliografskih zapisov v vzajemnem katalogu COBIB*. Diplomsko delo. Ljubljana: Filozofska fakulteta.

Danskin, A. (2006). What difference does it make? Measuring the quality of cataloguing and the catalogue**.** *Catalogue and index*, (154), 9–12.

Diao, J. in Hernadez, M. A. (2014). Transferring cataloging legacies into descriptive metadata creation in digital projects: catalogers' perspective. *Journal of library metadata*, *14*(2), 130–145.

Eklund, A. P., Miksa, S. D., Moen, W. E., Snyder, G. in Polyakov, S. (2009). Comparison of MARC content designation utilization in OCLC WorldCat records with national, core, and minimal level record standards. *Journal of library metadata*, *9*(1–2), 36–64. doi: 10.1080/19386380903095073

El-Sherbini, M. (2010). Program for cooperative cataloging: BIBCO records: analysis of quality. *Cataloging and classification quarterly*, *48*(2–3), 221–236.

Enci, T. (2011). *Kakovost bibliografskih zapisov za serijske publikacije v vzajemnem katalogu COBIB.SI*. Diplomsko delo. Ljubljana: Filozofska fakulteta.

Escolano Rodriguez, E. (2014). Consolidated edition of ISBD, International Standard Bibliographic Description: a standard to trust, a quality brand. *Cataloging and classification quarterly*, *52*(8), 835–854.

Galway, L. A. in Hank, C. H. (2011). Classifying data quality problems. *IDQ newsletter*, *7*(4), 3 str.

Graham, P. S. (1990). Quality in cataloguing: making distinctions. *Journal of academic librarianship*, *16*(4), 213–218.

Hafter, R. (1986). *Academic librarians and cataloging networks: visibility, quality control, and professional status*. New York: Greenwood Press.

Harden, J. (2012). Inadvertent RDA: new catalogers' errors in AARC2. *Journal of library metadata*, *12*(2–3), 264–278.

Harmon, J. C. (1996). The death of quality cataloging: does it make a difference for library users?. *Journal of academic librarianship*, *22*(4), 306–307.

Hider, P. in Tan, K. (2008). Constructing record quality measures based on catalog use. *Cataloging and classification quarterly*, *46*(4), 338–361.

Intner, S. S. (1989). Much ado about nothing: OCLC and RLIN cataloguing quality. *Library journal*, *114*(2), 38–40.

Izjava o mednarodnih katalogizacijskih načelih. (2009). Ljubljana: Narodna in univerzitetna knjižnica. Pridobljeno 12. 12. 2016 s spletne strani: http://old.nuk.uni-lj.si/infobib/images/stories/Dokumenti/Izjava_o_mednarodnih_katalogizacijskih_nacelih.pdf

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

Kavčič, I. in Velkavrh, V. (2009). Kakovost knjižničnih katalogov (bibliografskih baz podatkov). *Knjižničarske novice*, *19*(11), 8–10.

Kim, W., Choi, B. J., Hong, E. K., Kim, S. K. in Lee, D. (2003). A taxonomy of dirty data. *Data mining and knowledge discovery*, *7*(1), 81–99.

Kont, K. (2015). How much does it cost to catalog a document?: a case study in Estonian university libraries. *Cataloging and classification quarterly*, *53*(7), 825–850.

Krstulović, Z. (2006). Katalogizacijska pravila in kakovost bibliografskih podatkov. *Organizacija znanja*, *11*(4), 215–218.

Leckie, J. G., Given, L. in Campbell, G. (2009). Technologies of social regulation: an examination of library OPACs and web portals. V G. J. Leckie in J. E. Buschman (ur.), *Information technology in librarianship: new critical approaches* (str. 221–259). Wesport: Libraries Unlimited.

Likar, T. (2003). Enotna obdelava knjižničnega gradiva: pogoj za izmenjavo in skupno uporabo bibliografskih zapisov. *Knjižnica*, *47*(1–2), 7–34.

Likar, T. in Žumer, M. (2004). Mnenja katalogizatorjev o modulu za katalogizacijo v sistemu COBISS. *Knjižnica*, *48*(1–2), 83–122.

Lundy, M. W. (2006). Evidence of application of DCRB Core Standard in WorldCat and RLIN. *Library resources and technical services*, *50*(1), 42–57.

Ma, F., Mo, Z. in Luo, Y. (2014). Empirical research on a model to measure end-user satisfaction with the quality of database search results. *Journal of academic librarianship*, *40*(2), 194–201.

MacEwan, A. in Young, T. (2004). Quality vs. quantity: developing a systematic approach to a perennial problem. *Catalogue and index*, (152), 1–7.

Mai, J. E. (2013). The quality and qualities of information. *JASIST*, *64*(4), 675–688.

Massey, O. (2000). *Auditing catalogue quality by random sampling*. Master's dissertation. Loughborough: Centre for information management. Pridobljeno 12. 2. 2016 s spletne strani: http://users.ox.ac.uk/~bodl0842/dissertation/index.html

Miksa, S. D. (2007). Functional requirements for bibliographic records: understanding support of FRBR's four user tasks in MARC-encoded bibliographic records. *Bulletin of the Association for the Information Science and Technology*, *33*(6), 24–26.

Mitchell, E. in McCallum, C. (2012). Old data, new scheme: an exploration of metadata migration using expert-guided computational techniques. *Proceedings of the Association for Information Science and Technology*, *49*(1), 1–10.

Moen, W. E., Miksa, S. D., Eklund, A., Polyakov, S. in Snyder, G. (2006). Learning from artifacts: metadata utilization analysis. V *JCDL '06: proceedings of the 6th ACM/IEEE-CS joint conference on digital libraries* (2 str.). New York: Association for Computing Machinery. Pridobljeno 12. 12. 2016 s spletne strani: http://dl.acm.org/citation.cfm?id=1141813

Moulaison Sandy, H. in Dykas, F. (2016). High-quality metadata and repository staffing: perceptions of United States-Based OpenDOAR participants. *Cataloging and classification quarterly*, *54*(2), 101–116.

Moulaison, H. L. (2015). The expansion of the personal name authority record under Resource Description and Acess: current status and quality considerations. *IFLA journal*, *41*(1), 13–24.

Myall, M. in Chambers, S. (2007). Copy cataloging for print and video monographs in two academic libraries: a case study of editing required for accuracy and completeness. *Cataloging and classification quarterly*, *44*(3–4), 233–257.

Ochoa, X. in Duval, E. (2009). Automatic evaluation of metadata quality in digital repositories. *International journal on digital libraries*, *10*(2–3), 67–91.

On the record: report of the Library of Congress Working Group on the Future of Bibliographic Control. (2008). Washington, D.C.: Library of Congress. Pridobljeno 12. 4. 2012 s spletne strani: http://www.loc.gov/bibliographic-future/news/lcwg-ontherecord-jan08-final.pdf

Paiste, M. S. (2003). Defining and achieving quality in cataloging in academic libraries: a literature review. *Library collections, acquisitions and technical services*, *27*(3), 327–338.

Park, J. (2006). Semantic interoperability and metadata quality: an analysis of metadata item records of digital image collections. *Knowledge organization*, *33*(1), 20–34.

Pesjak, D. in Petek, M. (2010). Kakovost bibliografskih zapisov v COBIB in uporaba katalogizacijskih priročnikov. *Knjižnica*, *54*(3), 15–33.

Petek, M. (1998). Vrednotenje knjižničnih katalogov s stališča uporabnikov. *Knjižnica*, *42*(4), 127–147.

Petek, M. (2012). Enotni naslov v teoriji in v slovensko-hrvaški katalogizacijski praksi. *Knjižnica*, *56*(1–2), 127–148.

Petrucciani, A. (2015). Quality of library catalogs and value of (good) catalogs. *Cataloging and classification quarterly*, *53*(3–4), 303–313.

Pisanski, J. in Žumer, M. (2009). Funkcionalne zahteve za bibliografske zapise (FZBZ): analiza uporabnosti konceptualnega modela bibliografskega sveta. *Knjižnica*, *53*(1–2), str. 61–67.

Redman, T. C., Fox, C. in Levitin, A. (2009). Data and data quality. V *Encyclopedia of library and information sciences* (str. 1420–1431). New York: Taylor and Francis.

Rogina, A. (2007). Projekt Izboljšanje kakovosti zapisov v COBIB.SI. *Organizacija znanja*, *12*(2), 58–67.

Romero, L. in Romero, N. (1992). Original cataloging in a decentralized environment: an identification and explanation of errors. *Cataloging and classification quarterly*, *15*(4), 47–65.

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

Seljak, M. (2000). Poti do konsistentnih katalogizacijskih pravil. *Organizacija znanja*, *5*(4).

Seljak, M. (2006). Izvajanje določil pravilnika o izdaji dovoljenja za vzajemno katalogizacijo v sistemu COBISS.SI. *Organizacija znanja*, *11*(4), 208–214.

Seljak, M., Brešar, T., Curk, L., Zalokar, M., Tominac, A., Mazić, G., … Urbajs, A. (2004). Vzpostavitev normativne kontrole v knjižničnem informacijskem sistemu COBISS.SI, Slovenija. *Organizacija znanja*, *9*(2), 37–46.

Shin, H. (2003). Quality of Korean cataloging records in shared databases. *Cataloging and classification quarterly*, *36*(1), 55–90.

Smith-Yoshimura, K., Argus, C., Dickey, T. J., Naun, C. C., Rowlison de Ortiz, L. in Taylor, H. (2010). Implication of MARC tag usage on library metadata practices. Ohio: OCLC. Pridobljeno 12. 12. 2016 s spletne strani: http://www.oclc.org/content/dam/research/publications/library/2010/2010-06.pdf

Snow, K. (2011*). A study of the perception of cataloging quality among catalogers in academic libraries*. Dissertation. Denton, Texas: University of North Texas Libraries. Pridobljeno 12. 12. 2016 s spletne strani: https://digital.library.unt.edu/ark:/67531/metadc103394/m2/1/high_res_d/dissertation.pdf

Stalberg, E. in Cronin, C. (2011). Assessing the cost and value of bibliographic control. *Library resources and technical services*, *55*(3), 124–137.

Statement of international cataloging principles (ICP). (2016). Haag: IFLA. Pridobljeno 23. 12. 2016 s spletne strani: http://www.ifla.org/publications/node/11015

Stvilia, B. in Gasser, L. (2008). Value-based metadata quality assessment. *Library and information science research*, *30*(1), 67–74.

Stvilia, B., Gasser, L., Twidale, M. B. in Smith, L. C. (2007). A framework for information quality assessment. *Journal of the American Society for Information Science and Technology*, *58*(12), 1720–1733.

Švab, K. in Žumer, M. (2016). Izbira leposlovja – še vedno izziv za knjižnice?: kriteriji, ki so pomembni za uporabnike. *Knjižnica*, *60*(2–3), 127–149.

Tani, A., Candela, L. in Castelli, D. (2013). Dealing with medatada quality: the legacy of digital library efforts. *Information processing and managament*, *49*(6), 1194–1205.

Taniguchi, S. (2005). Recording evidence in bibliographic records and descriptive metadata. *Journal of the American Society for Information Science and Technology*, *56*(8), 872–882.

Taniguchi, S. (2007). A system supporting evidence recording in bibliographic records. Part II: What is valuable evidence for catalogers?. *Journal of the American Society for Information Science and Technology*, *58*(6), 823–841.

Tenopir, C. (1990). Database quality revisited. *Library journal*, *115*(16), 64–67.

Thomas, S. E. (1996). Quality in bibliographic control. *Library trends*, *44*(3), 491–505.

Transforming our bibliographic framework: a statement from the Library of Congress. (2011, 13. maj). Washington, D.C.: Library of Congress. Pridobljeno 12. 4. 2012 s spletne strani: http://www.loc.gov/marc/transition/news/framework-051311.html

Urbajs, A. in Šobot, P. (1991). Vidiki kvalitete vzajemne in lokalnih baz v sistemu vzajemne katalogizacije. V T. M. Šercar (ur.), *Tretiranje znanstvenih in strokovnih publikacij in polpublikacij v online dostopnih bazah podatkov za znanost in tehnologijo, 14. posvetovanje o znanstvenih in strokovnih publikacijah in polpublikacijah, Maribor, 16.–18. 12. 1991* (str. 269–273). Maribor: Univerza, Institut informacijskih znanosti.

Van Kleeck, D., Langford, G., Lundgren, J., Nakano, H., O'Dell, A. J. in Shelton, T. (2016). Managing bibliographic data quality in a consortial academic library: a case study. *Cataloging and classification quarterly*, *54*(7), 452–467.

Vetrò, A., Canova, L., Torchiano, M., Orozco Minotas, C., Iemma, R. in Morando, F. (2016). Open data quality measurement framework: definition and application to open government data. *Government information quarterly*, *33*(2), 325–337.

Wang, R. Y. in Strong, D. M. (1996). Beyond accuracy: what data quality means to data consumers. *Journal of management information systems*, *12*(4), 5–33.

Wisser, K. M. (2014). The errors of our ways: using metadata quality research to understand common error patterns in the application of name headings. V S. Closs idr. (ur.), *Metadata and semantics research. MTSR 2014* (str. 83–94). Cham: Springer.

Yang, S. in Li, L. (2015). Emerging technologies for librarians: a practical approach to innovation. Amstardam: Chandos.

Yasser, C. M. (2011). An analysis of problems in metadata records. *Journal of library metadata*, *11*(2), 51–62.

Zalokar, M. (2006). Razvoj splošnega geslovnika COBISS.SI. *Organizacija znanja*, *11*(4), 224–229.

Zeng, M. in Qin, J. (2008). *Metadata*. London: Facet.

Zhang, Y. in Salaba, A. (2012). What do users tell us about FRBF-based catalogs?. *Cataloging and classification quarterly*, *50*(5–7), 705–723.

mag. Branka Badovinac

Institut informacijskih znanosti Maribor, Prešernova ulica 17, 2000 Maribor
e-pošta: branka.badovinac@izum.si

Priloga 1: Preglednica virov po kronološkem redu

Vir in namen študije	Metodološki pristop	Dimenzije	Ugotovitve
Likar (2003) Preverjanje kakovosti enotne obdelave slov. serijskih publikacij z online dostopom	Primerjava zapisov s primarnim gradivom po Chapman in Masey (2002) Ekspertna analiza	Točnost, popolnost Podatek je naveden v napačnem polju Nedosledna uporaba slov. jezika Tipkarske napake	Tretjina ustreznih zapisov Predlog ukrepov za izboljšanje kakovosti
Bruce in Hilman (2004) Splošni opis dimenzij kakovosti metapodatkov	Teoretični pristop	Točnost, izvor, točnost, skladnost s pričakovanim, logična konsistentnost in skladnost, pravočasnost in dostopnost	Opis dimenzij brez metrike, namenjeno za ekspertne analize
MacEwan in Young (2004) Razvoj metodologije točkovanja kakovosti zapisov za merjenje učinkovitosti katalogizacije	Primerjava zapisov z virom na podlagi matrike posameznih atributov po pomembnosti FRBR-jevih uporabniških opravil	Točnost, popolnost	Model FRBR je uporaben za vrednotenje kakovosti zapisov
Krstulović (2006) Vpliv katalogizacijskih pravil na kakovost v COBISS-u	Teoretični pristop	*(dimenzije niso navedene oz. jih ni mogoče določiti)*	Največji vpliv na kakovost podatkov ima katalogizator (subjektivni dejavnik)
Lundy (2006) Vzroki neuporabe standarda PCC za antikavarno gradivo in redke knjige (DCRB) na osnovi primerjava polj v zapisih MARC21 in v standardu v dveh katalogih	Primerjalna študija, statistična analiza	Popolnost	Standard je omejen, praktičen le za gradivo z manj podatki, sicer je nujna razširitev (npr. opombe)
Park (2006) Semantična inteoperabilnost v konverziji treh digitalnih zbirk slik v shemo Dublin Core	Kvalitativna in kvantitativna analiza metapodatkovnih zapisov	Točnost, popolnost	Problem netočnih elementov – podatek je naveden v napačnem polju Potreba po poenotenju pomenov in razumevanja rabe posameznih konceptov/elementov v metapodatkovni shemi ter navodila katalogizacije metapodatkov
Taniguchi (2005, 2007) Zasnova ekonomičnega sistema navajanja dokazov, ki kažejo, zakaj in kako so vrednosti podatka izbrani	Teoretični pristop modeliranja in testiranje uporabnosti	Izraznost (ang. expressivity), zanesljivost	Sistem je funkcionalen in uporaben
Myall in Chambers (2007) Dopolnjevanje bibliografskih zapisov OCLC za monografije in video gradivo dveh ustanov	Študija primera s primerjav izhodiščnih in končnih bibliografskih zapisov (kakovostni zapis = nič popravkov)	Točnost, popolnost	Zapisi za video gradivo potrebujejo več popravkov Za video gradivo potrebni nacionalni kooperativni program (npr. PCC)

Vir in namen študije	Metodološki pristop	Dimenzije	Ugotovitve
Rogina (2007) Predstavitev projekta Izboljšanje kakovosti zapisov v COBIB.SI	Avtomatizirana odprava napak/ konverzija podatkov	*(dimenzije niso navedene oz. jih ni mogoče določiti)*	Konverzija podatkov v več kot 400.000 zapisih Dopolnitev programske opreme Sinhronizacija normativnih zapisov
Stvilia idr. (2007) Opredelitev splošnega modela kakovosti informacij	Teoretični pristop z analizo literature: izhodišče iz vzrokov spremembe v kakovosti	22 dimenzij v 3 skupine	Tipologija problemov v kakovosti informacij Taksonomija dimenzij
Stvilia in Gasser (2008) Oblikovanje metode in modela vrednotenja kakovosti na podlagi pomembnosti metadapotkov	Analiza zbirke z analitičnim in empiričnim pristopom ugotoviti osnovni nivo kakovosti metadapotkov	Popolnost	Različna pomembnost istih metapodatkov pri ponudnikih in uporabnikih
Hider in Tan (2008) Načini vrednotenja kakovosti zapisa: zanesljivost ekspertnega mnenja, vrste napak in njihova pomembnost z vidika končnih uporabnikov	Anketa, intervju, metoda glasnega razmišljanja	Točnost, popolnost	Nezanesljivost ekspertnega mnenja Ni večjih razlik med pomembnostjo napak
Zeng in Qin (2008) Merjenje in izboljšanje kakovosti v metapodatkovnih projektih	Pregledni teoretični pristop	Popolnost, pravilnost, doslednost, deduplikacija (edinstvenost)	Povzetek o merjenju/vrednotenju, merah (kazalcih), evalvacijski metodologiji in aktivnostih izboljšav
Ochoa in Duval (2009) Dopolnjevanje modela kakovosti metapodatkov z merami za avtomatizirani postopek evalvacije (tj. avtomatizirano zagotavljanje kakovosti)	Testiranje s statistično analizo	Dimenzije po Bruce in Hilman (2004)	Enostavna implementacija mer; standardizirane v stroki; namenjene za analizo tekstovnih in numeričnih znakov; namenjene za relativno stabilno metapodatkovo shemo; normalizacija mere ni vedno mogoča; nujna kombinacija mer Nekaterih mer ni mogoče še avtomatizirati
Eklund idr. (2009) Primerjava najpogosteje uporabljenih polj in podpolj v WorldCat z elementi v priporočilih LC nivojih in programi PCC BIBCO, CONSER	Statistična analiza zapisov MARC21 in primerjalna analiza	Popolnost	Raba majhnega nabora polj/podpolj

Vir in namen študije	Metodološki pristop	Dimenzije	Ugotovitve
El-Sherbini (2010) Analiza kakovosti PCC zapisov	Primerjava zapisov po izboljšavi z vidika zunanjega ponudnika storitev Tipologija napak glede na možnost poizvedovanja: manjše, večje	*(dimenzije niso navedene oz. jih ni mogoče določiti)*	Kljub velikem številu popravkov večina napak ne vpliva na poizvedovanje virov
Smith-Yoshimura idr. (2010) Raziskava prakse uporabe metadatkov v WorldCat zapisih, tudi glede na značilnosti strojnih aplikacij	Kombinacija metod, analiza 145 milijonov zapisov	Popolnost	Raba majhnega nabora polj Nekonsistetnost Nepopolnost
Crnčić (2010) Ugotavljanje kakovosti bibliografskih zapisov v slovenskem vzajemnem katalogu COBIB	Ekspertna analiza 50 zapisov za monografije glede na pravilnik in format	*(dimenzije niso navedene oz. jih ni mogoče določiti)*	Največ napak v območju opomb in v območju založništva in distribucije, glede na priročnik COMARC pa največ v bloku šifriranih podatkov in bloku podatkov o odgovornost Večja težava v katalogizacijskem pravilniku kot pa v formatu
Enci (2011) Ugotavljanje kakovosti bibliografskih zapisov za serijske publikacije v vzajemnem katalogu COBIB.SI	Ekspertna analiza 50 zapisov s standardom ISBD(CR) in format	*(dimenzije niso navedene oz. jih ni mogoče določiti)*	Največ napak v območju fizičnega opisa, glede na format pa v bloku šifriranih podatkov Vzrok slabe kakovosti je neuporaba priročnikov
Yaser (2011) Opredelitev problemov pri kakovosti metapodatkov	Analiza literature	*(dimenzije niso navedene oz. jih ni mogoče določiti)*	Pet problemov: netočna vrednost, nepravilni element, manjkajoči podatek (informacija), izgubljena informacija, nedosledna uporaba vrednosti (value representation)
Mitchel in McCallum (2012) Evalvacija kakovosti metapodatkov po migraciji v konceptualni model FRBR z uporabo ekspertno-avtomatiziranih tehnik	Primerjava kakovosti podatkov med ekspertno in avtomatizirano migracijo podatkov	*(dimenzije niso navedene oz. jih ni mogoče določiti)*	Hibridni pristop se je izkazal za učinkovito metodo migracije podatkov

Vir in namen študije	Metodološki pristop	Dimenzije	Ugotovitve
Kavčič (2012) Glede na uporabnika in katalogizacijski pravilnik ugotavljanje dejanske kakovosti ocenjenih zapisov v postopku preverjanja 50 naključnih zapisov Odpravljanje napak po rezultatih preverjanja 50 naključnih zapisov	Anketa uporabnikov (2/3 knjižničarji) o pomembnosti podatkov pri iskanju in identifikaciji vira ter primerjava z napakami v ocenjenih zapisih Ekspertna analiaza za primerjavo zapisov	*(dimenzije niso navedene oz. jih ni mogoče določiti)*	Večina zapisov je glede uporabnika ustrezna, tudi tisti, ki so označeni kot večja napaka Potreba po novih kriterijih za ocenjevanje zapisov Manj kot polovica katalogizatorjev zapisov ni popravila
Petek (2012) Uporaba enotnega naslova kot zbirna funkcija kataloga COBIB in CROLIST	Primerjalna študija o uporabi polj 300 in 500 za prevode avtorskih in anonimnih del	Popolnost	Nedosledna in pomanjkljiva uporaba polj 300 in 500 Manj podatkov v CROLIST-u, a tudi v COBIB-u premalo zavedanja o uporabnosti polja 500
Tani, Candela in Castelli (2013) Pregled raziskav o kakovosti podatkov v digitalnih knjižnicah, s poudarkom na splošnih okvirih	Analiza literature	*(dimenzije niso navedene oz. jih ni mogoče določiti*)	Zanimanje za kakovost podatkov in oblikovanje splošnih okvirov narašča Potreba po enotnem splošnem okviru, avtomatizaciji vnosa podatkov, razvoju orodja za preverjanje kakovosti in veljavnosti podatkov
Wisser (2014) Opredelitev napak pri imenih korporacij in osebnih imen iz agregiranih podatkov v zapisih za arhive EAC-CPF	Analiza zapisov za isto entiteto	Točnost, popolnost, konsistetnost	Identifikacija 30 različnih napak
Moulaison (2015) Preučevanje dodajanja atributov v normativnih zapisov po RDA-ju	Longititudinalna študija primera	Popolnost	Normativni zapisi se redko dopolnjujejo z atributi
Van Kleeck idr. (2016) Uvedba zagotavljanja in kontrola kakovosti zapisov	Študija primera večje visokošolske knjižnice	*(dimenzije niso navedene oz. jih ni mogoče določiti)*	Uvedba aktivnosti kontrole kakovosti in odprave napak