Metodologije za kvalitativno vrednotenje kakovosti odprtih podatkov

Klara Žniderščič[1], Matija Marolt[1], Aleš Veršič[2], Matevž Pesek[1]
[1]Fakulteta za računalništvo in informatiko Univerze v Ljubljani, Večna pot 113, Ljubljana
[2]Ministrstvo za digitalno preobrazbo
klara.znidersic@fri.uni-lj.si, matija.marolt@fri.uni-lj.si, aversic@gov.si, matevz.pesek@fri.uni-lj.si

Izvleček

Članek predstavlja pregled metodologij za kvalitativno vrednotenje odprtih podatkov. Različni pristopi so bili razviti tako za ocenjevanje metapodatkov, ki opisujejo posamezne zbirke podatkov, kot tudi za ocenjevanje samih podatkov. Metodologije pogosto temeljijo na standardih in širše sprejetih smernicah za zagotavljanje kakovosti podatkov, posamezne vidike (meta)podatkov pa obravnavajo kot dimenzije, združene v kategorije. Slednje med različnimi metodologijami niso poenotene, kar otežuje primerjavo ali združevanje rezultatov, vendar poznavanje različnih pristopov pomembno pripomore pri razvoju novih, učinkovitejših in bolj interoperabilnih rešitev. V primeru razvoja novih orodij za vrednotenje odprtih vladnih podatkov v Sloveniji je potrebno upoštevati še dodatna priporočila. Trg odprtih podatkov v državah Evropske unije namreč regulirajo zakonodaje in smernice tako na nacionalni, kot na ravni Unije. V tem prispevku predstavljamo pregled smernic in standardov ter izpostavljamo izzive za izdelavo poenotene metodologije, ki bi pripomogla k dvigu ravni kakovosti in dostopnosti odprtih podatkov.

Ključne besede: metodologije kvalitativnega vrednotenja, odprti podatki, pregled področja

Methodologies for qualitative assessment of open data

Abstract

The article provides an insight into the methodologies for the qualitative assessment of open data. Various approaches have been developed both for the assessment of metadata and for the data itself. The methods are often based on standards and generally accepted guidelines for ensuring data quality, whereby individual aspects of the (meta)data are treated as dimensions grouped into categories. However, these categories are not standardized across different methodologies, making it difficult to compare or integrate the results. Nevertheless, understanding different approaches contributes significantly to the development of more effective and interoperable solutions. When developing new tools for open government data in Slovenia, additional recommendations need to be considered. Laws and directives regulate the open data market in European Union countries at both national and EU level. We provide an overview of guidelines and standards, and highlight the challenges in creating a uniform methodology to improve the quality and accessibility of open data.

Keywords: quality assessment methodologies, open data, overview of the field

1 UVOD

Pojem odprti podatki predstavlja podatke, do katerih lahko vsakdo svobodno dostopa, jih uporablja, modificira in deli za kakršenkoli namen (»Open Definition 2.1«, 2015). Ocena skupne vrednosti trga odprtih podatkov v Evropski uniji z vsakim letom izrazito narašča, o čemer poroča tudi najnovejša uradna študija evropskega podatkovnega trga, ki napoveduje, da bo trg do leta 2030 presegel vrednost 118 milijard €. To rast lahko razumemo kot posledico kopičenja podatkov na različnih evropskih podatkovnih portalih, s tem pa se hkrati dvigujejo tudi pričakovanja in zahteve potencialnih uporabnikov ponujenih podatkov. Podjetja, ki bi z rabo odprtih podatkov spodbudila inovacije in ekonomsko rast, se za to pogosto ne odločajo, saj se kljub navidez preprosti uporabi

brezplačno dostopnih odprtih podatkov neredko srečujejo z znatnimi ovirami pri identificiranju primernih nizov podatkov in pripravi le-teh za uporabo, ovire pa se pogosto navezujejo na kakovost podatkov (Krasikov & Legner, 2023). Raziskave preteklih let, ki naslavljajo to problematiko, so predstavile že več različnih pristopov k bolj in manj celostnem vrednotenju kakovosti zbirk, vendar se med področjem metapodatkov in podatkov pojavlja vrzel, ki ni zanemarljiva. Medtem ko je za vrednotenje metapodatkov, ki opisujejo objavljene zbirke podatkov, pripravljenih več uspešno evalviranih metodologij, je vrednotenje podatkov kompleksnejše narave in so zato metode pogosteje prilagojene specifičnim podatkovnim zbirkam.

V pričujočem članku so predstavljeni glavni izsledki pregleda področja kvalitativnega vrednotenja (meta)podatkov, ki so pomembni za razvoj novih metodologij ter vzpostavitev čim bolj celostnega in prilagodljivega pristopa k podajanju vrednostne ocene podatkovnih zbirk, predvsem v kontekstu odprtih podatkov. Cilj prispevka je pregled metodologij, ki bodo služile kot osnova za izdelavo celostne metodologije kot rezultat projekta *Izdelava metodologije za določanje kakovosti podatkov ter ocena kakovosti posameznih podatkovnih zbirk na nacionalnem portalu odprtih podatkov Slovenije - portalu OPSI*.

2 TEMELJI ZA ZAGOTAVLJANJE KAKOVOSTI PODATKOV

Dolgoletna zgodovina zbiranja podatkov in vse bolj zgoščenega kopičenja informacij izpostavlja velik pomen zagotavljanja in vzdrževanja kar najvišje stopnje kakovosti podatkov. S tem namenom so bili razviti standardi in druga pravila, ki ponudnikom podatkov predstavljajo smernice za doseganje kakovosti in so usmerjena v doseganje optimalne stopnje urejenosti na danem področju.

2.1 Standard in aplikacijski profil evropskih podatkovnih portalov

Besednjak podatkovnih katalogov DCAT (»Data Catalog Vocabulary - Version 2«, 2020) za podatkovni model RDF (Resource Description Framework oz. ogrodje za opis virov), razvit v okviru organizacije W3C, ponuja smernice in specifikacije za opisovanje naborov podatkov v katalogih podatkov z namenom izboljšanja prepoznavnosti in interoperabilnosti naborov podatkov na spletu. DCAT je standard, ki se lahko uporablja globalno in ni povezan z nobenim regulativnim okvirjem, a je za specifične potrebe različnih podatkovnih portalov lahko dopolnjen z aplikacijskimi profili. Kot specifikacija za opis povezanih javnih podatkov v Evropi je v uporabi aplikacijski profil za evropske podatkovne portale DCAT-AP (Slika 1). Aplikacijski profil DCAT-AP, ki je bil zasnovan za rabo v evropskih javnih upravah, ponuja posebne smernice za zajemanje pravnih in licenčnih informacij v zvezi z nabori podatkov, ki so v skladu z evropskimi pravnimi zahtevami za odprte podatke ter zaradi svoje specializacije pripomore k večji interoperabilnosti med državami članicami EU.

Podatkovni model RDF za predstavitev informacije uporablja trojice subjekt–predikat–objekt, pri čemer se za opis podatkovnih katalogov uporablja nabor razredov in lastnosti, ki jih definira DCAT.

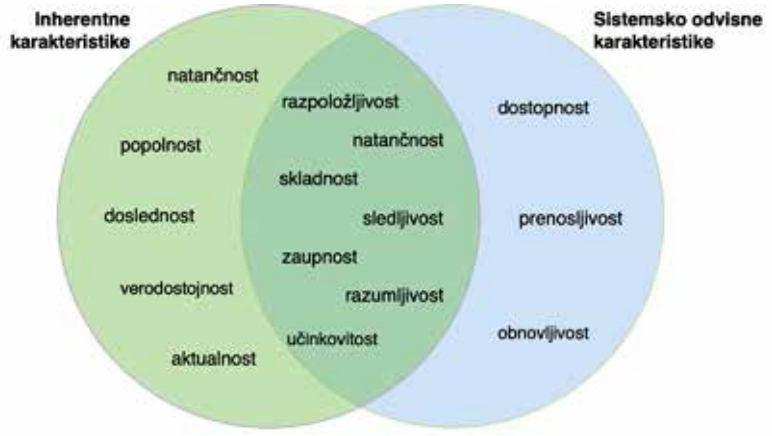
2.2 Standard ISO 25012

Serija standardov Mednarodne organizacije za standardizacijo (ISO), katerih glavni cilj je usmerjati razvoj programskih izdelkov s specifikacijo kakovostnih zahtev in kriterijev za evalvacijo, je združena pod kratico SQuaRE (System and Software Requirements and Evaluation). Za podatke, shranjene v strukturirani obliki v računalniškem sistemu, je splošni model kakovosti podatkov opredeljen s standardom ISO/IEC 25012:2008. Uporablja se lahko za

```
@prefix dcat: <http://www.w3.org/ns/dcat#> .
@prefix dct: <http://purl.org/dc/terms/> .
@prefix example-ds: <https://data.gov.gr/id/dataset/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .

example-ds:BeePopulation a dcat:Dataset;
    dct:title "Bee population"@en;
    dct:description "A dataset about bee population in Greece"@en;
    dct:publisher <https://agencies.gov.gr/id/GreekEnvironmentAgency> .
```

Slika 1: **Primer DCAT-AP opisa podatkovne zbirke o populaciji čebel (Van Nuffelen, 2023)**



Slika 2: **Tabela karakteristik kvalitete podatkov ISO/IEC 25012 (»ISO 25012«, 2022)**

opredeljevanje kvalitativnih meril ali načrtovanje in izvajanje ocenjevanja kakovosti podatkov. Kakovost podatkovnega izdelka je lahko interpretirana kot stopnja, do katere podatki izpolnjujejo zahteve petnajstih karakteristik modela, te pa so razvrščene v dve glavni kategoriji - inherentna in sistemsko odvisna kakovost podatkov (Slika 2). Inherentna kakovost podatkov se nanaša zlasti na podatkovne vrednosti, razmerja med temi vrednostmi ter metapodatke, kategorija sistemsko odvisnih kakovosti podatkov pa določa stopnjo, do katere je kakovost dosežena in ohranjena v računalniškem sistemu. V tej kategoriji na kakovost vplivajo zmogljivosti komponent računalniških sistemov (strojna in programska oprema).

2.3 5-zvezdični sistem za ocenjevanje

Leta 2006 je ustanovitelj svetovnega spleta Tim Berners-Lee predstavil sistem za ocenjevanje povezanih podatkov (Berners-Lee, b.d.), s katerim je želel spodbuditi predvsem upravljavce vladnih podatkov k dvigu kakovosti podatkov in boljšo povezljivost. Sistem temelji na standardnih spletnih tehnologijah, kot so HTTP, RDF in URI, ki jih uporablja za deljenje informacij na strojno berljiv način in s tem omogoča povezovanje podatkov iz različnih virov ter razširja možnosti semantičnih poizvedb. Za vpeljavo odprtih podatkov je Berners-Lee predlagal petzvezdično shemo, pri kateri je zadostovanje pogojem vsake predhodne stopnje predpogoj za izpolnjevanje zahtev naslednje stopnje:

1. Podatki so dostopni na spletu v formatu z odprto licenco.
2. Podatki so na voljo v strojno berljivi obliki (npr. tekstovna datoteka namesto digitalno preslikane razpredelnice).
3. Podatki so dostopni v nelastniškem formatu (npr. CSV namesto Excel).
4. Za identificiranje so uporabljeni odprti standardi organizacije W3C (RDF in SPARQL) ter URI.
5. Podatki so za namen kontekstualizacije povezani z ostalimi dostopnimi podatki.

2.4 Načela FAIR

Potreba po izboljšavi infrastrukture, ki podpira ponovno uporabo znanstvenih podatkov, je bila naslovljena z načeli FAIR (»GO FAIR: FAIR Principles«, b. d.), ki poseben poudarek usmerjajo v izboljšanje strojnih zmožnosti za samodejno iskanje in uporabo podatkov. Utemeljena so na štirih osrednjih kriterijih – najdljivosti (F - findability), dosegljivosti (A - accessibility), interoperabilnosti (I -interoperability) in ponovni uporabnosti (R – reusability). FAIR opisuje jedrnata, od področja neodvisna načela, ki jih je mogoče uporabiti za širok spekter znanstvenih rezultatov.

3 PRISTOPI H KVALITATIVNEMU VREDNOTENJU (META)PODATKOV

V začetku je potrebno izpostaviti pomen razlikovanja med izrazoma profiliranje podatkov in ocenjevanje kakovosti podatkov. Medtem ko je prvo zgolj sistematizirano zbiranje informacij o podatkih, drugo predstavlja preverjanje ustreznosti podatkov določenim kakovostnim merilom. Profiliranje torej lahko razumemo kot komplementarno zgodnjo fazo vrednotenja podatkov (Debattista in sod., 2016).

3.1 Vrednotenje metapodatkov

Velika večina raziskav o kvalitativnem vrednotenju podatkovnih zbirk je bila osredotočena na vrednotenje metapodatkov. Različne metode in kazalniki za oceno kakovosti metapodatkov so bili preizkušeni na številnih zbirkah podatkov, med drugim tudi na vladnih portalih odprtih podatkov. Metode so pogosto razčlenjene tako, da obravnavajo posamezne aspekte podatkov - dimenzije, ki so razdeljene v kategorije. V posamezno kategorijo so zbrane kvalitativne dimenzije, v katerih je kot kazalnik uporabljena skupna vrsta informacije (npr. dostopnost, ki vključuje dimenzije razpoložljivosti, varnosti in delovanja), saj takšno združevanje pri velikem številu dimenzij omogoča boljši pregled vseh aspektov kvalitete. Posamezna dimenzija ima lahko eno ali več metrik, t.j. konkretnih meril kakovosti za posamezni kazalnik, ki so običajno povezani z merilnim postopkom in vrnejo numerično ali boolean oceno.

Ena obsežnejših metapodatkovnih evalvacij preteklih let je bila izvedena v okviru projekta *Open Data Portal Watch* (Neumaier in sod., 2016), kjer so predstavili orodje za avtomatizirano vrednotenje kvalitete metapodatkov s pretvorbo v standardizirano shemo DCAT. Na podlagi izbranih 29 dimenzij iz besednjaka DCAT in izračunu metrik za posamezno dimenzijo je bilo ovrednotenih preko milijon podatkovnih zbirk, kar je služilo tudi nadzoru in razvrščanju portalov. Nadgradnjo projekta je predstavljala implementacija analitičnega hierarhičnega procesa, ki je omogočil primerjavo preko 250 portalov, na katerih so bile objavljene ovrednotene zbirke (Kubler in sod., 2018).

Po številu primerljiv obseg zbirk dosega tudi Uradni portal za evropske podatke, ki za preučevanje kakovosti metapodatkov objavljenih zbirk uporablja orodje Metadata Quality Assessment (MQA). Metrike tega orodja so predstavljene v Tabeli 1. MQA naslavlja osnovni vprašanji o kakovosti metapodatkov za podatke iz javnega sektorja v vseevropski regiji in identificira največje ovire pri doseganju boljše kakovosti. Poleg ocene kakovosti metapodatkov zbirk javnega sektorja v vseevropski regiji, MQA prav tako pomaga pri iskanju preprek pri doseganju višje kakovosti. Metodologija MQA preverja:

- skladnost z aplikacijskim profilom DCAT in derivati aplikacijskega profila DCAT;
- razkritje informacij, ki niso obvezne za aplikacijski profil DCAT;
- dostopnost podatkov, na katere se sklicujejo metapodatki prek URL povezave za dostop in URL povezave za prenos;
- strojno berljivost referenčnih podatkov;
- uporabo licenc.

Proces je omejen z metapodatki, ki jih orodje lahko pregleda in ki so dostopni na portalu data.europa.eu. V izogib prenosu napak iz izvirnih metapodatkov v proces vrednotenja, MQA že pred objavo zbirk zagotavlja storitev, ki jo lahko uporabijo ponudniki podatkov za potrjevanje veljavnih formatov in skladnost z DCAT-AP, preden svoje metapodatke vključijo v postopek vrednotenja. Integriran je tudi mehanizem za ponovno preverjanje dostopnosti objavljenih spletnih povezav po določenem časovnem obdobju. V postopku vrednotenja je upoštevanih pet dimenzij – štiri iz principov FAIR ter dimenzija kontekstualnosti – z vnaprej določenimi metrikami in največjim možnim številom točk, ki so pretvorjene v opisne ocene (odlično, dobro, zadostno ali slabo). V večini primerov metodologija ocenjuje zgolj Boolovo vrednost metrike (da/ne).

Pogosteje so bile izvedene manj obsežne raziskave, ki so podlago za izbiro dimenzij poiskale v drugih smernicah, a so kljub temu sorodne zgoraj opisanim. Primer je Bogdanović-Dinić in sod. (2014), ki je 7 portalov odprtih podatkov ovrednotila na podlagi 8 principov (»The 8 Principles of Open Government Data«, b.d.): odprtost (opis, možnost prenosa, strojna berljivost, povezljivost), primarnost (podatki v izvirni obliki), pravočasnost (časovno obdobje, frekvenca posodobitev, zadnja posodobitev), dostopnost, možnost strojne obdelave (PDF/XLS, CSV/HTML/XTT, XML/RDF), nediskriminatornost dostopa, objava v nelastniških formatih, brez licence oz. licenca za odprti dostop.

3.2 Vrednotenje kakovosti podatkov v podatkovnih zbirkah

Izrazito manj zastopane so študije kakovosti dejanske vsebine podatkovnih zbirk. Teoretično izhodišče tem raziskavam je v več primerih predstavljalo znanstveno delo Data Quality: Concepts, Methodologies and Techniques (Batini & Scannapieca, 2006). Principi, ki jih obravnava delo, so formulirani neodvisno od konteksta uporabe in med raziskovalci zanimivi zaradi svoje univerzalnosti, predvsem zato, ker je v novejši literaturi pogosteje izpostavljeno vrednotenje kako-

Tabela 1: **Metrike v posameznih kategorijah MQA dimenzij**

Dimenzija	Metrika
Najdljivost	ključne besede kategorije geografsko iskanje časovno iskanje
Dostopnost	dostopnost URL-ja za dostop URL za prenos dostopnost URL-ja za prenos
Interoperabilnost	format vrsta medija vrsta medija/formata iz slovarja strojna berljivost skladnost z DCAT-AP
Ponovna uporabnost	licenca slovar licence omejitve dostopa slovar omejitev dostopa kontaktna točka izdajatelj
Kontekstualnost	pravice velikost datoteke datum izdaje datum spremembe

vosti, prilagojeno specifičnim kontekstom in nadaljnji rabi. Dimenzije kakovosti, definirane v knjigi, so povzele številne metodologije, ki so nabor dopolnile glede na potrebe raziskav in razvoj tehnologij. Knjiga ob tem ponuja tudi praktične rešitve in metodologije za lokalizacijo in odpravljanje podatkovnih napak, identifikacijo objektov ter integracijo podatkov.

Med raziskavami velja izpostaviti nabor dimenzij kakovosti, ki je nastal na podlagi pregleda 30 različnih pristopov in 12 orodij (Zaveri in sod., 2016). Za vsako izmed 18 dimenzij je ponujena natančna definicija, več metrik (ki so opredeljene kot kakovostne ali kvantitativne) in primer uporabe. Kvalitativna oz. kvantitativna narava metrike je pomembna predvsem za proces avtomatizacije vrednotenja, saj lahko slednja zagotavlja zanesljivejše rezultate pri objektivnih kvantitativnih metrikah (npr. št. sintaktično pravilnih vrednosti), medtem ko so kvalitativne ocene lahko odvisne od subjektivne ocenjevalčeve percepcije (npr. ugled podatkovne zbirke).

Na podlagi sorodne metodologije LANG (Zhang in sod., 2019) je bil zasnovan dvostopenjski model za vrednotenje sintaktičnih (enolična razpoznavnost podatkov, konsitentnost formata, referenčna integriteta, skladnost metapodatkov in skladnost s poslovnimi pravili) in semantičnih (popolnost atributov, semantična konsistentnost, konsistentnost vrednosti, natančnost in neredundanca/nepodvojenost) vidikov kvalitete. Rezultati evalvacije na več kot 20 podatkovnih zbirkah so prikazali potencial, ki ga ima LANG za pospeševanje in razširjanje procesov raziskovanja podatkov, študija možnosti za avtomatizacijo tega procesa pa je dala rezulate, skoraj povsem enake rezultatom ročnega vrednotenja.

sSQuaRE-Aligned Portal Data Quality Model (SPDQM) (Moraga in sod., 2009) je bil utemeljen na 30 karakteristikah starejšega modela PDQM in dopolnjen s petimi karakteristikami, izbranimi na podlagi dodatnega sistematičnega pregleda literature ter sedmimi ISO standarda, ki še niso bili del modela. Navezuje se tako na podatke same kot tudi na kakovost sistema, saj obravnava vlogo sistema, kontekst in reprezentacijo podatkov. Za vrednotenje posamezne kategorije so bile določene naslednje karakteristike:

- Inherentna kakovost: natančnost, kredibilnost (objektivnost, ugled), sledljivost, aktualnost, zastaranost, popolnost, doslednost, dostopnost, skladnost, zaupnost, učinkovitost, natančnost, razumljivost;
- Operativna kakovost: razpoložljivost, dostopnost (interaktivnost, enostavnost uporabe in pomoč strankam), preverljivost, zaupnost, prenosljivost, popravljivost;
- Kontekstualna kakovost: veljavnost (zanesljivost, obseg), dodana vrednost (aplikabilnost, prilagodljivost, novost), relevantnost (novost, pravočasnost), specializacija, uporabnost, sledljivost, skladnost, preciznost;
- Kakovost predstavitve: jedrnatost, doslednost, razumljivost (interpretabilnost, količina podatkov, dokumentacija, organizacija), privlačnost, berljivost, učinkovitost, uspešnost.

Podjetniški vidik uporabe odprtih podatkov je naslavljala novejša študija (Krasikov & Legner, 2023), ki se je osredotočila na vprašanje, kako podjetjem pomagati pri sistematičnem pregledu, oceni in pripravi odprtih podatkov za uporabo. Pri iterativnem razvoju metode so avtorji sodelovali s podjetji, končni izdelek pa omogoča sistematično analizo ter integracijo odprtih podatkov in s tem pojmuje pripravo odprtih podatkov kot smiseln proces ustvarjanja dodane vrednosti. Tristopenjski pristop k ocenjevanju kakovosti je zajemal

1. pregled metapodatkov po predlogi starejših raziskav;
2. oceno popolnosti sheme (prisotnost potrebnih atributov);

Dimenzija	Definicija
Aktualnost	Kdaj so bili podatki vneseni v vir in/ali podatkovno skladišče.
Dodana vrednost	V kolikšni meri so podatki koristni in zagotavljajo prednosti zaradi njihove uporabe.
Doslednost	V kolikšni meri so podatki predstavljeni v enaki obliki.
Dosegljivost	V kolikšni meri so podatki na voljo oz. jih je mogoče enostavno in hitro pridobiti.
Enostavnost uporabe in vzdrževanja	V kolikšni meri je mogoče podatke uporabljati, posodabljati, vzdrževati in upravljati.
Jedrnatost	V kolikšni meri so podatki jedrnato predstavljeni.
Kredibilnost	V kolikšni meri se podatki štejejo za resnične in verodostojne.
Natančnost	V kolikšni meri so podatki pravilni, zanesljivi in zagotovljeno brez napak.
Objektivnost	V kolikšni meri so podatki nepristranski.
Popolnost	V kolikšni meri podatki ne manjkajo ter so dovolj obsežni in poglobljeni za zadevno nalogo.
Pravilnost	V kolikšni meri so podatki pravilni in zanesljivi.
Pravočasnost	V kolikšni meri so podatki dovolj posodobljeni za zadevno nalogo.
Razložljivost	V kolikšni meri so podatki v ustreznih jezikih, simbolih in enotah, opredelitve pa so jasne.
Razumljivost	V kolikšni meri so podatki jasni, brez dvoumnosti in zlahka razumljivi.
Relevantnost	V kolikšni meri so podatki uporabni in koristni za zadevno nalogo.
Sledljivost	V kolikšni meri so podatki dobro dokumentirani, preverljivi in zlahka pripisani viru.
Spremenljivost	Časovno obdobje, v katerem so informacije veljavne v resničnem svetu.
Učinkovitost	V kolikšni meri lahko podatki hitro zadovoljijo potrebe po informacijah za zadevno nalogo.
Ugled	V kolikšni meri so podatki visoko cenjeni glede na njihov vir ali vsebino.
Uporabnost	V kolikšni meri so podatki jasni in enostavni za uporabo.
Ustrezna količina podatkov	V kolikšni meri je obseg podatkov primeren za zadevno nalogo.
Varnost	V kolikšni meri je dostop do podatkov ustrezno omejen, da je zagotovljena njihova varnost.

Tabela 2: **Definicije pogosto uporabljenih dimenzij**

3 na nivoju vsebine uporabo tradicionalnih meril za kvaliteto podatkov (popolnost, edinstvenost, veljavnost).

Kvalitativne dimenzije in merila za metapodatke, shemo in podatke so bile utemeljene na starejših raziskavah (Neumaier in sod., 2016; Vetrò in sod., 2016; Zhang in sod., 2019), definicije nekaterih pogosto uporabljenih dimenzij so zbrane v Tabeli 2, kot nezanemarljiva faza procesa pa je izpostavljena izdelava dokumentacije podatkovnih zbirk in integracija odprtih podatkov v podatkovne zbirke znotraj podjetja. Proces natančne dokumentacije je poenostavljen z vse bolj razširjeno rabo standardiziranih RDF slovarjev, kar pozitivno vpliva tudi na nadaljnje povezovanje zbirk.

Posebno področje kvalitativnega vrednotenja podatkov predstavlja ocena primernosti za uporabo (ang. *fitness for use*), ki so ga obravnavali avtorji odprtokodnega orodja Luzzu (Debattista in sod., 2016). Pristop za obteženo razvrščanje naborov podatkov temelji na vrednotenju podatkov, pri čemer so dimenzije za vrednotenje večinsko povzete po Zaveri in sod. (2016), lasten domenskospecifični jezik pa omogoča tudi definiranje metrik brez poznavanja programskih jezikov. Utež ocen posamezne metrike (ali dimenzije ali kategorije) lahko uporabnik določi glede na potrebe, saj je pripomoček primarno namenjen nadaljnjim uporabnikom, ki lahko s filtriranjem in razvrščanjem podatkov, ovrednotenih po izbranih merilih, poiščejo najprimernejši nabor podatkov. Pri procesiranju obsežnih zbirk podatkov so uporabljene verjetnostne metode za aproksimacijo, kar omogoča, da orodje potencialnim uporabnikom v doglednem času ponudi rezultate, ki so dovolj natančni za večino primerov uporabe.

3.3 Izzivi pri vrednotenju kakovosti podatkov

Pogosta opazka v raziskavah, ki obravnavajo kvaliteto podatkov ali metapodatkov, je problematika definicije *dobrih* podatkov. Presoja o kvaliteti je namreč pogosto odvisna od namena njihove uporabe ali zahtev uporabnikov in ponudnikov podatkov. Pri meritvah kakovosti dodatno nastajajo ovire zaradi pomanjkljivih, nenatančnih in nepopolnih informacij, zastarelih ali neveljavnih podatkov, nejasnih vrednosti, prevelike količine informacij za procesiranje, odsotnosti bistvenih podatkov in enake vsebine, ki v različnih sistemih ponuja različne rezultate (Janssen & Kronenburg, 2012; Krasikov & Legner, 2023). Tehnične ovire se pojavljajo pri nejasnih licenčnih pogojih, neprimerno opredeljenem ali težko dostopnem formatu, odsotnosti standardov, zastarelih sistemih za objavo podatkov, pomanjkanju standardne programske opreme za obdelavo odprtih podatkov ter razdrobljenosti programske opreme in aplikacij.

Na področju metapodatkov so se težave pojavljale tudi pri avtomatskih DCAT pretvorbah, nedelujočih HTTP zahtevah, razlikovanju med splošno definicijo in dejanskim stanjem odprtih podatkov ter nedostopnosti strojno berljivih kontaktnih podatkov skrbnikov zbirk.

Malo pristopov k vrednotenju podatkovnih zbirk je bilo dejansko podprtih s konkretnim orodjem, toda v večini teh primerov izvorna koda, ki bi bila uporabna na drugih podatkovnih zbirkah, ni dostopna. Izjema, ki jo velja izpostaviti, je Luzzu, dostopen v odprtokodni različici, vključno z dodatnimi zunanjimi metrikami.

Tudi priljubljena petzvezdična lestvica kakovosti Tima Bernersa-Leeja predstavlja težavo, saj pokriva zgolj specifičen aspekt kvalitete podatkov, kar je format oz. kodiranje, ki je v rabi za objavo podatkov. Zbirka podatkov na lestvici lahko doseže najvišjo oceno in kljub temu z drugih vidikov ne vključuje kakovostnih podatkov. Kot primer navajajo natančnost oz. nenatančnost pri ročnem vnašanju podatkov ali napake v programski opremi, pa tudi težave z vidika popolnosti, konsistentnosti podatkov in pravočasnosti objave.

Neenotnost definicij različnih kvalitativnih dimenzij se je izkazala za problematično v večih raziskavah metodologij vrednotenja kakovosti podatkov (Vetrò in sod., 2016; Zuiderwijk in sod., 2021). Zaradi različnih definicij karakteristik podatkov so bila posledično različno določena tudi merila, kar je pri uporabi metodologij na istih zbirkah podatkov dalo različne rezultate. Zuiderwijk in sod. (2021) so pri raziskovanju meril kakovosti, na podlagi katerih so ustvarjene lestvice kakovosti podatkov različnih ponudnikov (tj. držav), odkrili, da se pri uporabi raznih metodologij enake zbirke podatkov na lestvicah različno uvrščajo.

4 VLADNI ODPRTI PODATKI

Trg vladnih odprtih podatkov na območju Evropske unije urejajo evropske direktive in državne zakonodaje, EU pa je zaradi strme rasti količine podatkov v okviru Uradnega portala za evropske podatke podprla tudi izvedbo številnih študij na več nivojih podatkovne krajine. Njihovi rezultati sodelujočim državam predstavljajo spodbudo in zagotavljajo osnovo za izboljšavo, popravilo in razvoj nacionalnih portalov in zbirk odprtih podatkov. Smernice v splošnem spodbujajo čim širšo dostopnost podatkov in nadaljnjo rabo, evropski podatkovni portal pa visoko kvaliteto pripisuje podatkom, ki so človeško in strojno enostavno berljivi ter dosegajo čim višjo vrednost na petzvezdični ocenjevalni lestvici. Direktiva (EU) 2019/1024 predstavlja ključni dokument Evropske unije, ki obravnava področje odprtih podatkov in ponovne uporabe informacij javnega sektorja. S tem dokumentom so bile uvedene smernice in pravila, ki spodbujajo prost dostop do podatkov, ki jih proizvaja javni sektor, ter njihovo ponovno uporabo v gospodarske in družbene namene. Hkrati spodbuja izmenjavo informacij med državami članicami EU, združljivost z načeli FAIR ter prispeva k razvoju in inovacijam v digitalni družbi. Od leta 2015 izhaja tudi vsakoletno poročilo o zrelosti podatkov na nacionalnih portalih odprtih podatkov, ki je osredotočeno na štiri glavna področja – državno zakonodajo, meritve vpliva ponovne uporabe podatkov, oceno nacionalnih portalov in mehanizme za kvaliteto (meta)podatkov.

V slovenski zakonodaji se na temo odprtih podatkov nanašata Zakon o dostopu do informacij javnega značaja in Uredba o posredovanju in ponovni uporabi informacij javnega značaja, ki prenašata evropsko direktivo na pravni red Republike Slovenije. Z zakonom je urejena definicija relevantnih pojmov in postopek, ki omogoča prost dostop in ponovno uporabo informacij javnega značaja ter obvezuje organe k objavi informacij na namenskem portalu.

5 ZAKLJUČEK

Vpogled v stanje na področju kvalitativnega vrednotenja podatkov predstavlja najvidnejše standarde, definicije in načela za delo z (odprtimi) podatki, ki so pomembni gradniki pri vzpostavitvi metodologij vrednotenja, saj uporaba poenotenih pravil predstavlja predpogoj za zanesljive in primerljive rezultate. Kljub temu pa zgolj uporaba opisanih osnov ne zadostuje za kakovosten proces vrednotenja in je za to potreben tudi podrobnejši razmislek o kriterijih, ki naj bi jih upoštevala generalizirana metodologija za vrednotenje odprtih podatkov. Pretekle raziskave so v večjem delu obravnavale vrednotenje metapodatkov, saj je to področje bolj standardizirano (npr. uporaba RDF slovarjev in različnih lestvic kriterijev), med tem ko je zaradi raznolikosti vsebine podatkovnih zbirk, raziskav, usmerjenih na polje samih podatkov, manj. Slednje pri metodologijah pogosto izpostavljajo velik pomen določanja cilja vrednotenja – prilagajanje kriterijev potrebam želene nadaljnje uporabe podatkov.

Zgolj dostop do podatkov ni dovolj za učinkovito nadaljnjo rabo, temveč je zanjo potrebno upoštevati tudi večdimenzionalni aspekt kakovosti objavljenih podatkov. Predstavljene metodologije zato pogosto temeljijo na dimenzijah podatkov in njim prilagojenim metrikah, vendar se nabor dimenzij kot tudi njihove definicije med metodologijami razlikujejo. Posledično prihaja do odstopanj v kakovostnih lestvicah podatkovnih zbirk, raziskovalci pa navajajo tudi številne druge težave pri razvoju metodologij, predstavljenih v poglavju 3.3.

V tem pregledu stanja metodologij smo identificirali nekatere izzive, ki jih bomo v prihodnosti naslavljali v kontekstu odprtih podatkov. Z ozirom na slovenski portal odprtih podatkov - OPSI - je takšna metodologija nujno potrebna zaradi raznovrstnosti objavljenih podatkov, dostopnosti in povečanih potreb deležnikov po dvigu kakovosti in količini podatkov. Kljub upoštevanju evropskih in državnih smernic ter zakonodaj, navedenih v tem članku, so realne zmožnosti državnih agencij napram vedno večjim zahtevam uporabnikov odprtih podatkov pomemben omejevalni dejavnik.

ZAHVALA

Projekt Izdelava metodologije za določanje kakovosti podatkov ter ocena kakovosti posameznih podatkovnih zbirk na nacionalnem portalu odprtih podatkov Slovenije – portalu OPSI (št. V2-2388) je sofinancirala Javna agencija za raziskovalno dejavnost Republike Slovenije iz državnega proračuna.

LITERATURA

[1] The 8 Principles of Open Government Data [Dostop dne: 27. 3. 2024]. (b.d.). https://opengovdata.org/ Batini, C., & Scannapieca, M. (2006). *Data Quality*. Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-33173-5

[2] Berners-Lee, T. (b.d.). 5-star Open Data — 5stardata.info [Dostop dne: 29. 05, 2024]. https://5stardata.info/en/ Bogdanović-Dinić, S., Veljković, N., & Stoimenov, L. (2014). How open are public government data? an assessment of seven open data portals. *Public Administration and Information Technology*, 5, 25–44. https://doi.org/10.1007/978-1-4614-9982-4_3

[3] Data Catalog Vocabulary - Version 2. (2020). https://www.w3.org/TR/vocab-dcat-2/

[4] Debattista, J., Auer, S., & Lange, C. (2016). Luzzu-A Methodology and Framework for Linked Data Quality Assessment. *Journal of Data and Information Quality (JDIQ)*, 8. https://doi.org/10.1145/2992786

[5] GO FAIR: FAIR Principles [Dostop dne: 29. 05, 2024]. (b.d.). https://www.go-fair.org/fair-principles/ ISO 25012. (2022). https://iso25000.com/index.php/en/iso-25000-standards/iso-25012

[6] Janssen, K., & Kronenburg, T. (2012). EPSIP: Open Data Standardization before publication?

[7] Krasikov, P., & Legner, C. (2023). A Method to Screen, Assess, and Prepare Open Data for Use. *Journal of Data and Information Quality*, 15. https://doi.org/10.1145/3603708

[8] Kubler, S., Robert, J., Neumaier, S., Umbrich, J., & Traon, Y. L. (2018). Comparison of metadata quality in open data portals using the Analytic Hierarchy Process. *Government Information Quarterly*, 35, 13–29. https://doi.org/10.1016/J.GIQ.2017.11.003

[9] Moraga, C., Moraga, M. Á., Calero, C., & Caro, A. (2009). SQuaRE-aligned data quality model for web portals. *Proceedings - International Conference on Quality Software*, 117–122. https://doi.org/10. 1109/QSIC.2009.23

[10] Neumaier, S., Umbrich, J., & Polleres, A. (2016). Automated Quality Assessment of Metadata across Open Data Portals. *Journal of Data and Information Quality (JDIQ)*, 8. https://doi.org/10.1145/2964909

[11] Open Definition 2.1 [Version 2.1]. (2015). https://opendefinition.org/od/2.1/en/

[12] Van Nuffelen, B. (2023). DCAT-AP 3.0. https://semiceu.github.io/DCAT-AP/releases/3.0.0/

[13] Vetrò, A., Canova, L., Torchiano, M., Minotas, C. O., Iemma, R., & Morando, F. (2016). Open data quality mea- surement framework: Definition and application to Open Government Data. *Government Information Quarterly*, 33, 325–337. https://doi.org/10.1016/J.GIQ.2016.02.001

[14] Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., & Auer, S. (2016). Quality assessment for Linked Data: A Survey. *Semantic Web*, 7, 63–93. https://doi.org/10.3233/SW-150175

[15] Zhang, R., Indulska, M., & Sadiq, S. (2019). Discovering Data Quality Problems: The Case of Repurposed Data. *Business and Information Systems Engineering*, 61, 575–593. https://doi.org/10.1007/S12599- 019-00608-0

[16] Zuiderwijk, A., Pirannejad, A., & Susha, I. (2021). Comparing open data benchmarks: Which metrics and methodologies determine countries' positions in the ranking lists? *Telematics and Informatics*, 62, 101634. https://doi.org/10.1016/J.TELE.2021.101634

▪

Klara Žniderščič je od leta 2023 članica Laboratorija za računalniško grafiko in multimedije na Fakulteti za računalništvo in informatiko Univerze v Ljubljani. Ukvarja se z raziskovanjem odprtih podatkov in metodologij vrednotenja ter z digitalnimi pristopi k učenju glasbe.

▪

Matija Marolt je redni profesor in vodja Laboratorija za računalniško grafiko in multimedije na Fakulteti za računalništvo in informatiko Univerze v Ljubljani. Raziskovalno se ukvarja z analizo in prepoznavanjem multimedijskih signalov in z vizualizacijo podatkov.

▪

Aleš Veršič je sekretar na Ministrstvu za digitalno preobrazbo, kjer se ukvarja s pripravo strategij in politik za področje podatkovnega gospodarstva ter odprtih podatkov. Je tudi vodja omrežja skrbnikov podatkov (Data Stewards) v državni upravi. Kot doktorski študent je vpisan na Fakulteto za organizacijske vede, Univerze v Mariboru, kjer raziskuje področje podatkovnih prostorov (Data Spaces).

▪

Matevž Pesek je docent in raziskovalec na Fakulteti za računalništvo in informatiko Univerze v Ljubljani, kjer je diplomiral (2012) in doktoriral (2018). Od leta 2009 je član Laboratorija za računalniško grafiko in multimedije. Njegovi raziskovalni interesi so iskanje glasbenih informacij, vključno z glasbenim e-učenjem, biološko navdihnjenimi modeli in globoke arhitekture. Od leta 2014 raziskuje odprte podatke in je sorazvil več storitev, med drugim Avtolog.si, Tocen.si in Vodostaji.si.