Alat za osiguravanje kvalitete meta podataka na portalima otvorenih podataka

Arelić, Josip

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 276

ALAT ZA OSIGURAVANJE KVALITETE META PODATAKA NA PORTALIMA OTVORENIH PODATAKA

Josip Arelić

Zagreb, veljača 2024.

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 276

ALAT ZA OSIGURAVANJE KVALITETE META PODATAKA NA PORTALIMA OTVORENIH PODATAKA

Josip Arelić

Zagreb, veljača 2024.

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

Zagreb, 2. listopada 2023.

DIPLOMSKI ZADATAK br. 276

Pristupnik:	**Josip Arelić (0036515476)**
Studij:	Računarstvo
Profil:	Programsko inženjerstvo i informacijski sustavi
Mentor:	izv. prof. dr. sc. Igor Čavrak

Zadatak: **Alat za osiguravanje kvalitete meta podataka na portalima otvorenih podataka**

Opis zadatka:

Za pronalaženje i korištenje otvorenih skupova podataka objavljenih na portalima otvorenih podataka nužan je kvalitetan opis skupa podataka i njegovih karakteristika. Ovaj opis, u cilju strojne čitljivosti i analize, definira se korištenjem meta podataka i, eventualno, odgovarajućim normama za opis skupova podataka. Zadatak ovog rada proučiti je problematiku osiguravanja kvalitete meta podataka skupova otvorenih podataka objavljivanih na portalima otvorenih podataka. Potrebno je istražiti i opisati metode osiguravanja kvalitete meta podataka u procesu objavljivanja skupova otvorenih podataka, s naglaskom na alate za pomoć u definiranju meta podataka po relevantnim normama. Nastavno, potrebno je predložiti i implementirati alat za pomoć u osiguranju kvalitete meta podataka tijekom procesa objavljivanja skupa otvorenih podataka.

Rok za predaju rada: 9. veljače 2024.

Sadržaj

Uvod

Projekti portala otvorenih podataka imaju za cilj učiniti podatke dostupnima što većem broju korisnika. Razvojem tehnologija i digitalizacijom usluga javne uprave portali otvorenih podataka utvrdili su se kao korisno rješenje za objavljivanje javnih podataka. Rastom broja portala otvorenih podataka povećava se i količina objavljenih podataka, međutim ne nužno i sama iskoristivost podataka te interoperabilnost portala. Upravo na tome području ulažu se napori za poboljšanje kvalitete tih podataka ili preciznije metapodataka. U ovom radu, nakon pružanja postojeće definicije kvalitete metapodataka, istražit će se relevantne specifikacije koje se koriste za opisivanje metapodataka skupova podataka. U tom kontekstu razmotrit će se trenutno stanje kvalitete portala otvorenih podataka u Hrvatskoj. Potom će se identificirati postojeći zahtjevi na području poboljšanja kvalitete metapodataka i pružiti pregled dosadašnjih programskih rješenja na području osiguravanja kvalitete metapodataka. Završno, implementirat će se programsko rješenje koje zadovoljava neke od identificiranih zahtjeva. Navedeni će se ciljevi obraditi s naglaskom na specifičnostima europskih portala otvorenih podataka i zahtjeva definiranih od strane organizacija i tijela Europske Unije, koja je jedan od najvećih promicatelja otvorenosti podataka u svijetu. Pri određivanju zahtjeva i mjerila kvalitete metapodataka prvenstveno će se referencirati relevantni dokumenti i izvješća europskih agencija i tijela odgovornih za portale otvorenih podataka.

1. Kvaliteta metapodataka

U ovom poglavlju pružit će se analiza kvaliteta koje metapodatci trebaju zadovoljiti u kontekstu portala otvorenih podataka. Glavni predvodnik u polju otvorenih podataka je Europska komisija koja preko svojih radnih tijela definira zahtjeve i usmjerava provedbu objavljivanja otvorenih podataka zemalja članica EU i drugih europskih zemalja. Opisat će se metodologija središnjeg europskog portala otvorenih podataka, te će se navesti pojedine specifičnosti različitih vrsta podataka i njihovih odgovarajućih metapodataka.

1.1. FAIR načela

Skupina znanstvenika potaknutih stvarnom potrebom za poboljšanjem upotrebljivosti podataka 2016. je godine zajednički predložila jezgrovit skup mjerljivih načela za objavljivanje podataka skraćeno nazvanih FAIR načela [1]. Ta načela s vremenom su postala široko prihvaćene smjernice za mjerenje kvalitete objavljivanja podataka i specifično metapodataka. Upravo ta četiri načela koriste se u metodologiji za procjenu kvalitete metapodataka europskog portala za otvorene podatke [2].

Načelo mogućnosti pronalaženja (Findable) je primarni korak u iskorištavanju objavljenih podataka. Podaci se trebaju moći lako pronaći od strane ljudskih korisnika i strojeva. Ovo načelo ostvaruje se pokazateljima koji omogućavaju lakše pronalaženje prilikom pretrage nad određenim skupom podataka. Korištenje ključnih riječi koje opisuju tematiku skupa podataka te dodjeljivanje kategorija korisni su prilikom pretraživanja teksta i tematskog filtriranja podataka. Također, dodavanje vremenskog perioda ili lokacije nad kojom je skup podataka relevantan dodatno obogaćuje mogućnosti pronalaženja.

Jednom kad je određeni skup podataka pronađen, korisniku je važno omogućiti vezu na same podatke, eventualne dodatne informacije o skupu podataka i upute za dodatnu autentifikaciju i autorizaciju ako je ona potrebna. Načelo dostupnosti (Accessible) obuhvaća upravo takve zahtjeve. Konkretno može se provjeravati postojanje pristupnog URL-a i URL-a za direktno preuzimanje podataka. Dostupnost također podrazumijeva i vremensku komponentu tako da, osim provjere ispravnosti pristupnih URL-a u određenom trenutku, dostupnost resursa se može provjeravati više puta kroz određeni vremenski period.

Nakon što se pronađe traženi skup podataka i preuzme distribucija skupa podataka u određenoj datoteci, korisnik mora biti sposoban pristupiti podacima i u potpunosti ih iskoristiti. Tu dolazi do značaja načelo interoperabilnosti (Interoperable), odnosno potreba za tim da je naveden neki od definiranih formata datoteke i vrste medija (tekst, slika, video itd.). Pri tome format datoteke mora biti otvoren i strojno čitljiv kako bi svim korisnicima bilo moguće korištenje datoteke, bez potrebe za određenim vlasničkim programima. Za osiguravanje strojne čitljivosti potrebno je ispravno definirati metapodatke po određenim specifikacijama. Zbog toga se provodi validacija pruženih metapodataka po DCAT-AP specifikaciji. DCAT-AP glavna je specifikacija za opisivanje povezanih javnih podataka na europskim portalima.

Posljednje načelo tiče se mogućnosti ponovne upotrebe (Reusable). Ostvaruje se omogućavanjem prava pristupa i objavljivanjem podataka pod odgovarajućom licencijom. U slučaju potrebe za pitanjima u vezi podataka pružaju se i kontaktna točka s daljnjim informacijama te izdavač skupa podataka kojeg se koristi.

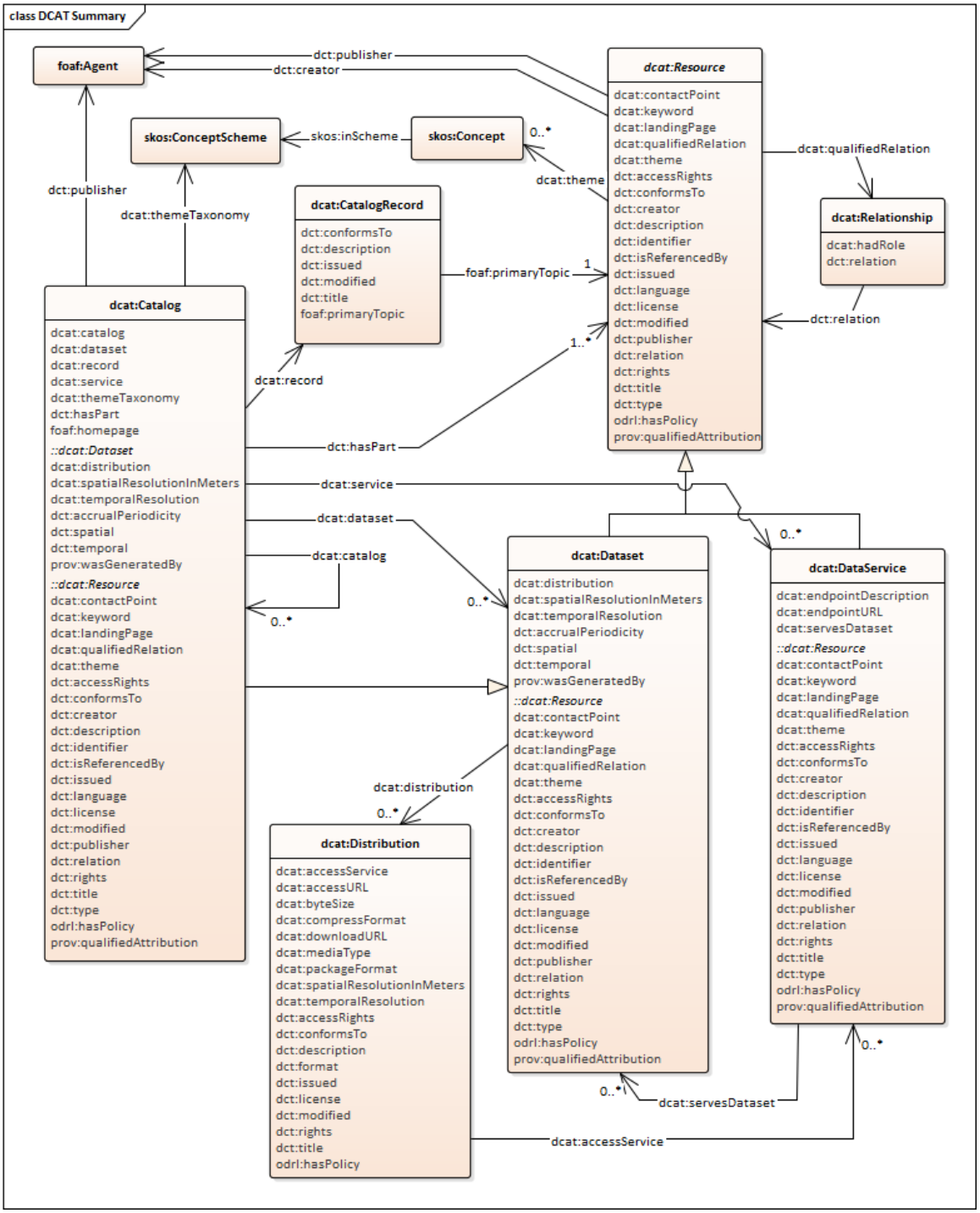
Europski portal otvorenih podataka uz navedena načela vrednuje i kontekstualnost skupa podataka. Bolja kontekstualnost postiže se navođenjem standardnih pokazatelja koji su općenito korisni kod korištenja bilo kojih podataka odnosno datoteka. To su metapodatci poput datuma izdanja, datuma zadnje izmjene, veličine datoteke i poveznice na prava korištenja.

1.2. Model (meta)podataka

Model pod kojim se podaci objavljuju na portalima otvorenih podataka temelji se na DCAT-u (Data Catalog Vocabulary). Kako bi se opisalo značenje jednih podataka u odnosu na druge u strojno čitljivom obliku, glavna organizacija za standardizaciju World Wide Web usluge, W3C (World Wide Web Consortium) 1999. godine predložila je RDF (Resource Description Framework) model podataka. DCAT je RDF rječnik namijenjen za definiranje podatkovnih kataloga na način koji omogućava izvršavanje upita nad cijelim skupom kataloga definiranih tim rječnikom. Radna verzija DCAT 3 dokumenta objavljena je i javno dostupna, međutim DCAT 2 još uvijek je službeno preporučena verzija. Stoga će se kroz ovaj rad referencirati DCAT 2 dokument [3]. Za opis RDF objekata koriste se izrazi klasa (class) i svojstvo (property) slično kao u određenim objektno-orijentiranim sustavima, međutim ti izrazi nemaju u potpunosti isto značenje u drugim kontekstima.

DCAT rječnik sastoji se od šest glavnih klasa:

- dcat:Catalog
  - predstavlja kolekciju resursa, skup čiji su elementi metapodatci o skupovima podataka (datasets), podatkovnim uslugama (data services), i drugim katalozima
- dcat:Resource
  - nadklasa koja predstavlja apstraktni element kataloga, klase skupova podataka, podatkovnih usluga i kataloga nasljeđuju ovu klasu
- dcat:Dataset
  - skup podataka, podaci mogu biti u bilo kojoj vrsti medija (tekst, slika, zvuk itd.)
- dcat:Distribution
  - predstavlja datoteku, odnosno konkretni skup podataka, jedan skup podataka može biti dostupan u nekoliko distribucija (npr. različitih formata datoteke)
- dcat:DataService
  - podatkovna usluga, skup operacija dostupnih kroz aplikacijsko sučelje (API) preko kojih se pružaju podaci ili funkcije za obradu podataka (npr. strujanje podataka u stvarnom vremenu)
- dcat:CatalogRecord
  - neobavezna klasa vezana uz klasu resursa, pruža informacije o samom unosu određenog resursa, npr. tko je ga je dodao, vrijeme unosa itd.



Sl. 1.1 UML dijagram DCAT klasa i svojstva [4]

Na dijagramu je vidljivo da se za definiciju DCAT klasa ne koriste isključivo klase i svojstva DCAT rječnika, već i drugih standardnih rječnika koji se koriste u širem kontekstu weba (FOAF, ODRL itd.). Također, DCAT 3 uz navedene klase dodaje novu klasu dcat:DatasetSeries koja predstavlja kolekciju skupova podataka koji dijele određenu

karakteristiku po kojoj se grupiraju, npr. popisi željezničkih postaja iz različitih godina ili različitih zemalja.

1.3. DCAT profili

Specifikacija DCAT rječnika proširuje se radi primjene u različitim domenama, od korištenja geoprostornih i statističkih podataka do prilagodbe određenim specifičnostima nacionalnih kataloga otvorenih podataka. U tu se svrhu npr. definiraju skupovi obaveznih polja metapodataka i dozvoljenih vrijednosti određenih polja te uvode nove podklase ili svojstva. Pritom se zadržavaju postojeći osnovni DCAT zahtjevi. Tako definirana proširenja DCAT specifikacije nazivaju se DCAT profilima. Također, DCAT profili mogu biti proširenja drugih DCAT profila kao što su određeni europski nacionalni profili proširenja općeg europskog DCAT-AP profila. Pritom takvi „podprofili“ moraju zadovoljavati sva ograničenja i strukturu klasa i svojstva nadređenog im profila, u ovom slučaju DCAT-AP-a. Pri definiranju ograničenja mogu biti jedino stroži odnosno restriktivniji.

U ovom radu obradit će se nekoliko DCAT profila [5]. Osim navedenog glavnog europskog profila DCAT-AP (Application Profile), postoje i poopćena proširenja DCAT-AP-a statističkim (StatDCAT-AP) i geoprostornim (GeoDCAT-AP) podacima. Uz to određene zemlje su izradile svoje nacionalne profile koji su prilagođeni specifičnostima nacionalnih kataloga otvorenih podataka, npr. DCAT-AP_IT (Italija), DCAT-AP.de (Njemačka) itd. Za primjere nacionalnih profila, odabrani su navedeni profili zato što su njihove specifikacije javno objavljene i redovito se ažuriraju. Osim toga u sklopu njihovih nacionalnih projekata za poboljšanje kvalitete metapodataka portala otvorenih podataka objavljene su i programska rješenja za osiguravanje kvalitete metapodataka.

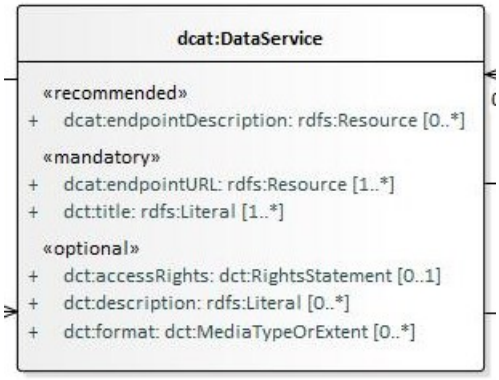
U profilima, osim definiranja novih klasa i svojstva, mogu se dodatno definirati razine obveze postojećih svojstva i klasa, te neka druga ograničenja. Razine obveze (obavezne/preporučene/neobavezne) odnose se isključivo na objavljivača takvih skupova podataka. Primatelji moraju uvijek moći obraditi (ali ne nužno analizirati, pretvoriti, pohraniti, učiniti dostupnima za pretraživanje ili prikaz, itd.) informacije o svim instancama svih klasa kao i svim njihovim svojstvima. Pri opisivanju profila i razina obveza klasa i svojstva koriste se izrazi "obavezno", "preporučeno" ili "opcionalno". Ovi izrazi imaju sljedeće značenje:

- Obavezna klasa: Objavljivači moraju pružiti informacije o instancama ove klase. Primatelji moraju moći obraditi informacije o instancama ove klase.
- Preporučena klasa: Objavljivači bi trebali pružiti informacije o instancama ove klase ako su dostupne. Primatelji moraju moći obraditi informacije o instancama ove klase.
- Neobavezna klasa: Objavljivači mogu, ali nisu obavezni, pružiti informacije o instancama ove klase.
- Obavezno svojstvo: Objavljivači moraju pružiti informacije o ovom svojstvu.
- Preporučeno svojstvo: Objavljivači trebaju pružiti informacije o ovom svojstvu ako su dostupni.
- Izborna značajka: Objavljivači mogu, ali nisu obavezni, pružiti informacije o ovom svojstvu.

1.3.1. DCAT-AP

Objavljivanjem, uređivanjem i održavanjem DCAT-AP profila upravlja SEMIC (Semantic Interoperability Community) zajednica i platforma koja pruža podršku za dijeljenje sredstava i interoperabilnost na području javnih usluga i e-uprave [6]. SEMIC je jedan od projekata veće inicijative Europske komisije zvane Interoperable Europe koja promiče prekograničnu suradnju i slobodni protok informacija u cilju eksponencijalnog razvoja inovacija [7].

DCAT struktura klasa definira skup osnovnih klasa bez određivanja obveze pojedinih svojstva. Upravo u tome DCAT-AP proširuje, odnosno specificira DCAT na već spomenuti način da sva svojstva i klase kategorizira pod obavezna, preporučena i opcionalna.



Sl. 1.2 Primjer DCAT-AP klase DataService

Također, osim rangiranja obveze, za svako se svojstvo navode i njegove kardinalnosti, odnosno koliko vrijednosti toga svojstva može biti pridruženo jednoj instanci te klase.

1.3.2. GeoDCAT-AP

GeoDCAT-AP je proširenje DCAT-AP profila za opisivanje geoprostornih skupova podataka i drugih struktura podataka poput serije skupova podataka (DatasetSeries) i servisa podataka (DataService) [8]. Poput DCAT-AP-a objavljuje ga i verzionira SEMIC zajednica. Glavna svrha ovog profila je učiniti geoprostorne skupove podataka i ostale prostorne strukture podataka pretraživima na širem skupu portala otvorenih podataka. Sukladno tome GeoDCAT-AP koristi postojeće norme, odnosno standarde za geografsko i prostorno opisivanje podataka, prvenstveno ISO 19115:2003 normu i INSPIRE tehničke smjernice.

INSPIRE (INfrastructure for SPatial Information) je direktiva Europske Unije o uspostavljanju infrastrukture za prostorne informacije u Europskoj zajednici. Detaljnije tehničke odredbe definiraju se provedbenim pravilima i tehničkim specifikacijama. INSPIRE se zasniva na postojećim infrastrukturama prostornih podataka zemalja članica i ne zahtijeva novo prikupljanje podataka, ali zahtijeva harmonizaciju postojećih podataka. Osnovne komponente INSPIRE direktive su:

- metapodatci,
- interoperabilnost prostornih podataka i usluga,
- mrežne usluge (otkrivanje, pregledavanje, preuzimanje, transformacija i pozivanje),
- zajedničko korištenje prostornih podataka i usluga,

- koordinacija te mjere nadzora i izvještavanja. [9]

ISO 19115 norma definira shemu za opis geografskih informacija i servisa korištenjem metapodataka [10]. Trenutno važeća verzija norme je ISO 19115-1:2014. Koristi se upravo pri katalogiziranju raznih resursa i definiranju obaveznih i izbornih polja geografskih metapodataka.

DCAT-AP profil proširuje se uvođenjem dodatnih klasa i svojstva [11]. Pritom se koriste klase poput *locn:Adress* i *vcard:Adress* iz Location Core rječnika za opisivanje imena, adrese i geometrije lokacija te vCard rječnika za opis osoba i organizacija. Također, iz Data Quality rječnika koriste se klase *dqv:Metric* i *dqv:QualityMeasurement* za opisivanje svojstva geografskih lokacija korištenjem različitih vrsta metrika.

1.3.3. StatDCAT-AP

StatDCAT-AP proširenje je DCAT-AP-a usmjereno na omogućavanje statističke analize i obrade otvorenih skupova podataka. Definiran je u suradnjom Eurostata i SEMIC organizacije [12]. Ovo proširenje definira nekoliko dodataka DCAT-AP modelu podataka koji su relevantni za statističke skupove podataka. Klasama su dodana svojstva poput *stat:statMeasure* i *stat:numSeries* definirana u SDMX (Statistical Data and Metadata eXchange) rječniku. Također, iz navedenog rječnika dodani su metapodatci *stat:dimension* za definiranje dimenzija skupa podataka poput vremena, spola ili starosti te *stat:attribute* za definiranje statističkih atributa poput mjerne jednice ili faktora skaliranja.

Sl. 1.3 Prikaz Dataset klase StatDCAT-AP-a [13]

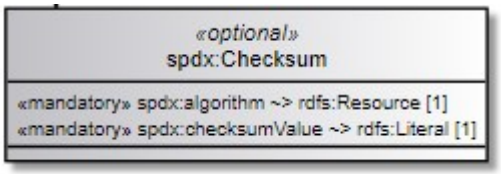
1.3.4. DCAT-AP.de

DCAT-AP.de njemačka je specifikacija DCAT-a. Njemačko državno tijelo zaduženo za provođenje europskih direktiva na području IT-a, IT-Planungsrat, 2018. godine razvilo je ovaj standard kojim se propisuje način za objavljivanje metapodataka portalima otvorenih podataka na području Njemačke.

Propisani standard DCAT-AP.de međutim ne sastoji se isključivo od specifikacije modela metapodataka, sastoji se od tri komponente [14]:

- specifikacija: Specifikacija donosi semantičke propise za komunikaciju sa središnjim njemačkim portalom (GovData.de) kao i za komunikaciju s europskim podatkovnim portalom te usvaja pravila europske sheme razmjene metapodataka DCAT-AP s pojedinačnim ograničenjima i proširenjima.
- koncept URI-ja : Koncept URI-ja ograničava slobodu komunikacijskih partnera u pogledu imenovanja URI-ja i nudi URI-je u prostoru imena „dcat-ap.de“ i „govdata.de” kako bi se omogućile reference na vokabulare koji idu izvan referenci koje pružaju DCAT i DCAT-AP.
- priručnik o konvencijama: Kako bi se dodatno povećala interoperabilnost, izrađen je priručnik o konvencijama DCAT-AP.de za GovData, koji dodatno standardizira komunikaciju s GovData putem pravila. Ovdje su definirani dodatni popisi vrijednosti i URI-ja te ostale tehničke upute za objavljivače i primatelje koji se ne odnose na sam model metapodataka.

Ono što nas u kontekstu ovoga rada zanima jest sami model metapodataka. Model DCAT-AP.de temelji se na DCAT-AP i GeoDCAT-AP profilima. Shema klasa ne razlikuje se značajno od GeoDCAT-AP sheme, ne dodaju se značajne nove klase. Jedna od dodanih pomoćnih klasa koju ne sadrže GeoDCAT-AP i DCAT-AP je opcionalna klasa spdx:Checksum. Software Package Data Exchange (SPDX) je standard za opisivanje licenci, autorskih prava, sigurnosnih značajki itd.

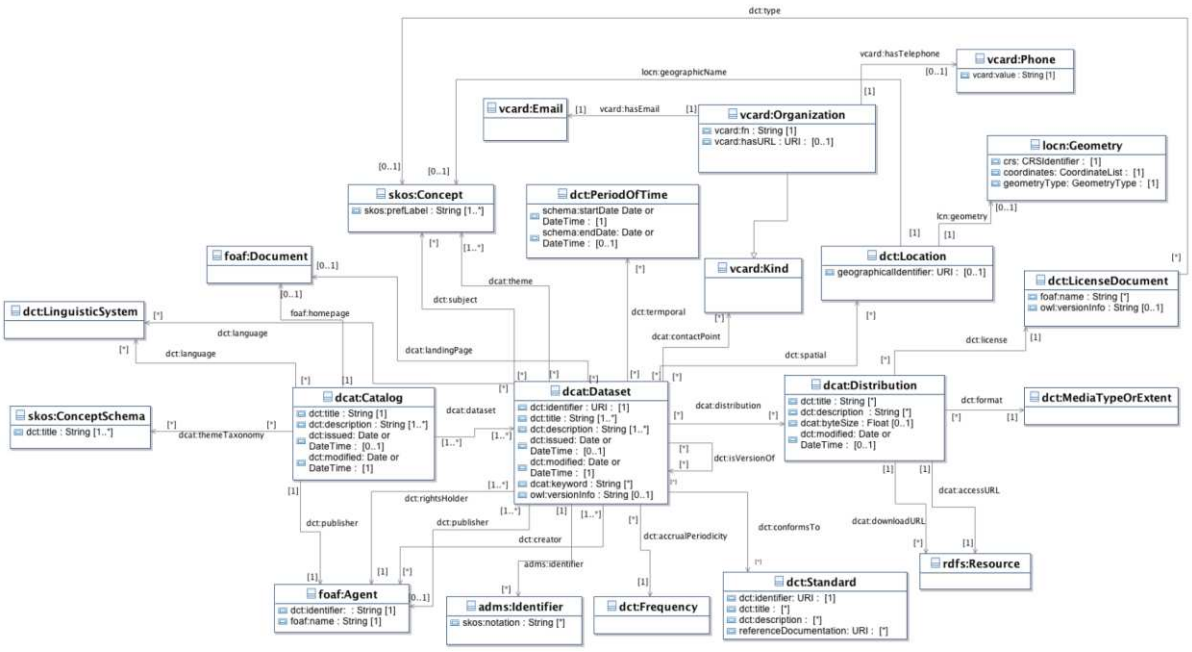


Sl. 1.4 Klasa spdx:Checksum [15]

1.3.5. DCAT-AP_IT

DCAT-AP_IT je talijanski nacionalni DCAT profil. Definiran je slično kao i njemački profil pod vodstvom talijanske nacionalne agencije, Agenzia per l'Italia digitale (AgID) 2016. godine [16].

Model ovog profila koristi GeoDCAT-AP i DCAT-AP klase i svojstva te se ne razlikuje značajno shematski od tih profila. Zato je upravo DCAT-AP_IT dobar primjer toga da se sami model nacionalnog DCAT profila ne mora nužno značajno razlikovati od izvornog europskog profila. Definiranje nacionalnog profila prije svega služi uspostavljanju standardnog načina objavljivanja na portalima otvorenih podataka nižih administrativnih jedinica kako bi se središnjem nacionalnom portalu omogućilo kvalitetno adresiranje i referenciranje metapodataka.

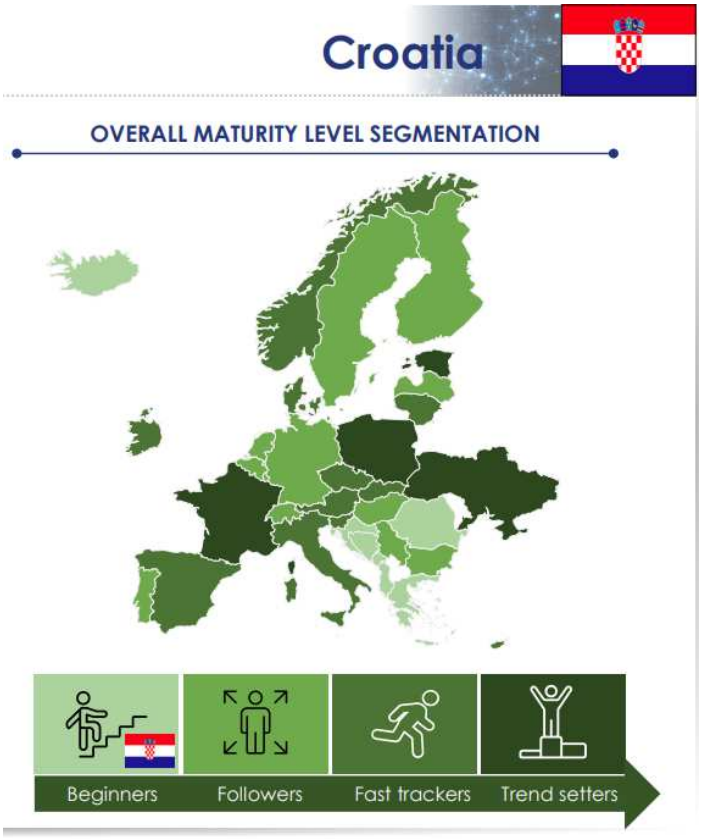


Sl. 1.5 UML dijagram DCAT-AP_IT modela metapodataka [17]

1.4. Stanje u Hrvatskoj

Trenutno nema uspostavljenog nacionalnog DCAT-AP profila na području Republike Hrvatske. Izradom takvoga prilagođenog nacionalnog profila ili korištenjem već postojećeg npr. osnovnog DCAT-AP profila kao osnove predstavljalo bi konkretan korak prema povećanju usklađenosti sheme metapodataka s europskim standardima.

Središnji europski portal (data.europa.eu) u svojem izvješću za 2023. godinu o stanju portala otvorenih podataka Hrvatsku stavlja u nižu kategoriju zemalja.



Sl. 1.6 Ocjena zrelosti podataka Hrvatske [18]

Navedeni portal također ocjenjuje katalog glavnog hrvatskog portala za otvorene podatke negativno u mnogim kategorijama kvalitete metapodataka [19]. Tako je na primjer kontekstualnost, koja se postiže metapodatcima poput datuma izdanja, podataka o izmjenama i pravima korištenja, u potpunosti nezadovoljena. Od ostalih dimenzija kvalitete kao što su interoperabilnost, mogućnost ponovne upotrebe i mogućnost pronalaženja ni jedna nije na zadovoljavajućoj razini.

2. Osiguravanje kvalitete metapodataka

Jednom kad su definirane mjere kvalitete metapodataka na objavljivačima i vlasnicima portala otvorenih podataka je da određenim praksama, procedurama i alatima za validaciju osiguraju traženu razinu kvalitete metapodataka. Data.europa.eu inicijativa u svojoj studiji poboljšanja objavljivanja podataka bavi se istraživanjem nedostataka postojećih portala otvorenih podataka [20]. Na temelju te analize pružaju se smjernice za poboljšanje postojećih i razvoj budućih rješenja za kvalitetnije objavljivanje podataka.

U studiji je sudjelovalo dvadeset voditelja portala otvorenih podataka rješavajući upitnik i pružajući odgovore slobodnog oblika na postavljena pitanja. Kroz upitnik, pružene odgovore i dodatna intervjuiranja identificirano je nekoliko područja za poboljšanje kvalitete podataka i metapodataka:

- upravljanje i metode
- znanje i informacije
- usklađivanje i standardizacija
- automatizacije i alati

U području upravljanja i korištenja određenih metoda navedena je potreba za korištenjem odgovarajuće sheme i osiguravanjem točnosti metapodataka. Prijenos znanja i informacija u svrhu poboljšanja kvalitete portala ostvaruje se povećanjem razine stručnosti djelatnika i dijeljenjem znanja među portalima. Područja u kojima alat za poboljšanje kvalitete metapodataka prvenstveno doprinosi jesu usklađivanje i standardizacija te automatizacije i alati.

Pruženi prijedlozi za poboljšanje usklađivanje i standardizacije odnose se na korištenje zajedničkih standarda i rječnika (poput npr. DCAT-AP) prilikom definiranja sheme metapodataka te standardiziranje postojećih podataka i metapodataka po zadanom standardu. Također, preporučuje se korištenje dobro podržanog alata za upravljanje metapodatcima i navođenje svih korištenih tehnologija u procesu objavljivanja podataka. Konačno, kao preporuka navodi se poboljšanje mehanizama za osiguravanje kvalitete

metapodataka kroz korištenje MQA usluge (poput one pružene na data.europa.eu) i automatiziranje toga procesa.

Uz navedene poopćene smjernice za rad na poboljšanju kvalitete metapodataka, korisno je analizirati i postojeće projekte na tome području te iz njih izvući zahtjeve i metode kojima su zadovoljeni ti zahtjevi.

2.1. Postojeća programska rješenja za osiguravanje kvalitete metapodataka

U okviru nekoliko nacionalnih portala za otvorene podatke, uz definiranje zasebnog nacionalnog DCAT profila, razvijena su i programska rješenja za osiguravanje usklađenosti s novodefiniranim profilom metapodataka. Pri tome se koristi temeljna CKAN + DCAT ekstenzija za CKAN portale kao osnova za usklađivanje mapiranja metapodataka između CKAN i DCAT shema.

2.1.1. CKAN + DCAT

Ovom se CKAN ekstenzijom pružaju značajke kojima se omogućuje objavljivanje CKAN skupova podataka u obliku RDF serijalizacija te uvoz tako objavljenih RDF serijalizacija iz drugih kataloga među CKAN skupove podataka [21]. Omogućava se i uvoz JSON objekata koji su definirani u skladu s DCAT rječnikom.

Navedene su funkcionalnosti ostvarene prvenstveno definiranjem mapiranja klasa i svojstva između CKAN i DCAT struktura podataka. Na temelju toga obavljaju se operacije RDF parsiranja i serijalizacije. U tu svrhu pruža se i mogućnost dodatnog definiranja i korištenja prilagođenih DCAT profila. Zadani profil po kojem se izvode parsiranje i serijalizacija je DCAT-AP. Proširenje je objavljeno pod otvorenom licencom, a pod autorom se navodi sama CKAN organizacija. U sklopu projekta dostupan je repozitorij otvorenog koda s preko trideset doprinositelja.

Predložena ekstenzija za poboljšanje kvalitete metapodataka od navedenih funkcionalnosti ponajprije bi koristila RDF parser i serializer.

2.1.2. DCAT-AP.de ekstenzija

Pod nazivom DCAT-AP.de u Njemačkoj je definirano proširenje DCAT-AP profila prilagođeno objavi podataka njemačkih portala otvorenih podataka. U sklopu toga razvijena je CKAN ekstenzija za prilagodbu sheme CKAN skupova podataka novodefiniranom DCAT profilu. Prilagođeno je sučelje za stvaranje novih skupova podataka uključivanjem potrebnih polja metapodataka. Postojeće funkcionalnosti CKAN + DCAT ekstenzije za objavljivanje i uvoz skupova podataka izmijenjene su prema DCAT-AP.de profilu metapodataka. Također, implementirano je sučelje prema vanjskom SHACL validatoru. SHACL (Shapes Constraint Language) je W3C standard za validaciju RDF grafova definiranjem skupa svojstava sheme metapodataka, njihovih referenci u RDF rječnicima te drugih vrsta ograničenja. Primjer takvog validatora je SHACL validator pružen u okviru data.europa.eu MQA usluge.

Izvorni kod projekta dostupan je u GitHub repozitoriju kojega vodi središnji njemački portal za otvorene podatke GovData [22].

2.1.3. DCAT-AP_IT ekstenzija

Poput prethodne ekstenzije, i ova ekstenzija razvijena je za potrebe implementacije novodefiniranog nacionalnog DCAT profila, DCAT-AP_IT za talijanske portale otvorenih podataka [23]. Ovaj projekt također vodi središnji nacionalni portal za otvorene podatke, u Italiji je to dati.gov.it.

Ova ekstenzija ispunjava slične zahtjeve kao i DCAT-AP.de ekstenzija; prilagođava uvoz i objavu RDF reprezentacija skupova podataka te izgled korisničkog sučelja za izradu i uređivanje novih skupova podataka. Za validaciju generiranih RDF dokumenata kao također se koristi vanjska usluga kojoj se pristupa preko API poziva.

2.2. Zahtjevi alata za osiguravanje kvalitete metapodataka

Sukladno trenutno korištenim tehnologijama i standardima na razini europskih i hrvatskih portala otvorenih podataka najprikladnije bi bilo programsko rješenje ostvariti u obliku

CKAN ekstenzije koja se temelji na DCAT-AP profilu metapodataka za definiranje kvalitete.

Do inicijalnog skupa zahtjeva našega programskog rješenja može se doći analizom studije poboljšanja objavljivanja podataka i usporedbom identificiranih potreba sa značajkama postojećih rješenja za poboljšanje kvalitete metapodataka. Studija navodi različite načine za poboljšanje vođenja i objavljivanja podataka.

Među navedenim preporukama one koje se više odnose na samu programsku potporu su [20]:

- na području usklađivanja i standardizacije:
  - koristiti otvorene i dobro poznate formate i usluge
  - slijediti norme poput DCAT-AP što je bliže moguće
  - ne stvarati nove standarde, umjesto toga ponovno koristiti i kombinirati postojeće standarde
  - implementirati procedure za provjeru i ako je primjenjivo, odbaciti podatke i metapodatke koji nisu usklađeni sa standardima, takvi mehanizmi u načelu mogu biti automatizirani, polu-automatizirani ili funkcioniraju ručno
- na području alata i automatizacije:
  - uspostaviti alate i mehanizme za osiguravanje provjere kvalitete podataka/metapodataka
  - implementirati procese, inicijative i kampanje u cilju osiguravanje korištenja alata za osiguravanje kvalitete podataka
  - koristiti zajedničke alate za upravljanje metapodatcima
  - automatizirati proces objavljivanja podataka
  - objaviti svoja rješenja sa zajednicom

Odabirom CKAN-a i DCAT-AP-a za temeljnu platformu i referentni profil metapodataka rješava se i olakšava ispunjavanje dijela navedenih zahtjeva. CKAN je sustav za skladištenje i objavljivanje podataka otvorenog koda koji se najviše koristi diljem svijeta za potrebe sustavnog objavljivanja otvorenih podataka. Koriste ga brojna državna, regionalna i lokalna tijela i organizacije te iza njega stoji velika stručna zajednica. CKAN podupire dodavanje novih funkcionalnosti portalu otvorenih podataka kroz razvoj CKAN

ekstenzija. Zbog navedenih činjenica, korištenjem CKAN-a kao temeljne platforme ispunjavaju se zahtjevi za korištenjem zajedničkih i dobro poznatih alata. Također, time se povećava dostupnost i iskoristivost rješenja među širom zajednicom. DCAT-AP otvoren je i dobro poznat standard korišten diljem europskih portala otvorenih podataka te zadovoljava prethodno navedene zahtjeve.

Koristeći primjere CKAN ekstenzija i njihovih funkcionalnosti, može se dobiti jasnija slika značajki kojima se može ispuniti preostale zahtjeve. Zahtjev za provjerom i kontrolom usklađenosti metapodataka sa standardima moguće je ispuniti implementiranjem promjene opisa skupova podataka dodavanjem potrebnih polja metapodataka. Takve bi izmjene zahvaćale korisničko sučelje te postupak stvaranja i uređivanja novih skupova podataka. U sklopu primjene određenog metapodatkovnog standarda nad shemom skupova podataka, javlja se potreba za validacijom. Validacija nad shemama skupova podataka ili cijelim katalogom uspostavlja se kao servis kojega koriste svi portali s određenim metapodatkovnim profilom. Tako npr. portal data.europa.eu preko API sučelja pruža uslugu validacije po DCAT-AP profilu metapodataka s ciljem usvajanja navedenoga profila od strane što većeg broja europskih portala otvorenih podataka. Time se i dugoročno olakšava verzioniranje i usvajanje trenutne najnovije verzije profila metapodataka jer više portala koristi istu uslugu validacije.

Što se tiče omogućavanja automatiziranja validacije i objavljivanja skupova podataka, CKAN ekstenzijom se u tu svrhu može implementirati jednostavna provjera usklađenosti skupa podataka ili kataloga. Klikom na jedan gumb može se ostvariti komunikacija između korisnikova web preglednika i usluge za validaciju te generirati informativna poruka o nepravilnostima u shemi metapodataka u slučaju nezadovoljavanja određenih zahtjeva zadanog standarda.

3. Implementiranje alata za osiguravanje kvalitete metapodataka

U ovom poglavlju opisat će se izrada programskog rješenja za poboljšanje kvalitete metapodataka te sva korištena tehnologija. Prije svega objasnit će se arhitektura CKAN-a za bolje razumijevanje procesa razvoja novih CKAN ekstenzija.

3.1. Postavljanje CKAN-a

CKAN je sustav koji se sastoji od nekoliko servisa. U svojoj biti to je web poslužitelj usmjeren na pružanje korisnicima upravljanja i pretrage nad velikim skupovima podataka. Web portal prezentira podatke u grupama skupova podataka, a skupovi podataka mogu biti dostupni u više distribucija. Korisnici objavljuju podatke kao članovi određenih organizacija.

Osnovu pozadinskog sustava ili backend CKAN-a čine Python web poslužitelj, podržan Redis bazom podataka u memoriji za sami rad poslužitelja te PostgreSQL bazom podataka za pohranu podataka. Također, za izvršavanje pretraga nad PostgreSQL bazom podataka koristi se Apache Solr servis.

Sučelje sustava ili frontend dostupno je preko korisničkog sučelja ostvarenog kombinacijom HTML-a i JavaScript-a. Uz to podaci su izloženi preko strojnog-čitljivog sučelja u obliku RPC API sučelja.

Vrijedno je napomenuti da je CKAN open-source projekt te su sve navedene korištene tehnologije također otvorenog koda u cilju što veće pristupačnosti i dostupnosti. CKAN platforma prvenstveno je dostupna na Ubuntu operacijskom sustavu, međutim za potrebe ovog diplomskog rada platforma je pokrenuta u kontejneriziranom okruženju koristeći besplatnu bazičnu verziju Docker alata.

Glavni repozitorij CKAN projekta objavljen je na GitHub stranicama i vođen od preko dvadeset kontributora koji sačinjavaju CKAN organizaciju. Uz izvorni kod samog CKAN-a dostupan je i ckan-docker repozitorij koji sadržava Docker skripte i ostale konfiguracijske datoteke potrebne za podizanje CKAN sustava na Dockeru [24].

Projekt se koristi službenim CKAN slikama (*images*) CKAN web poslužitelja. Verzija CKAN-a korištena za izradu ovog rada je CKAN 2.10. Ostale slike iz kojih su podignuti kontejneri ovog sustava su:

- DataPusher: CKAN-ova konfigurirana DataPusher slika
- PostgreSQL: službena PostgreSQL slika
- Solr: CKAN-ova konfigurirana Solr image
- Redis: standardna Redis slika
- NGINX: najnovija nginx slika

DataPusher je CKAN pomoćni servis koji služi za učitavanje datoteka i podataka iz CKAN baze podataka na stranice CKAN-a u svrhe pregleda i uređivanja datoteka u sklopu korisničkog sučelja CKAN-a bez potrebe za preuzimanjem datoteka i otvaranja u nekom drugom programu. Pritom se resursi spremaju u privremenu bazu podataka zvanu DataStore za potrebe prikazivanja u sklopu određenih pretpregleda ili uređivanja. NGINX je web servis za *load balancing* i općenito optimiziranje dostupnosti i skalabilnost neke web aplikacije.

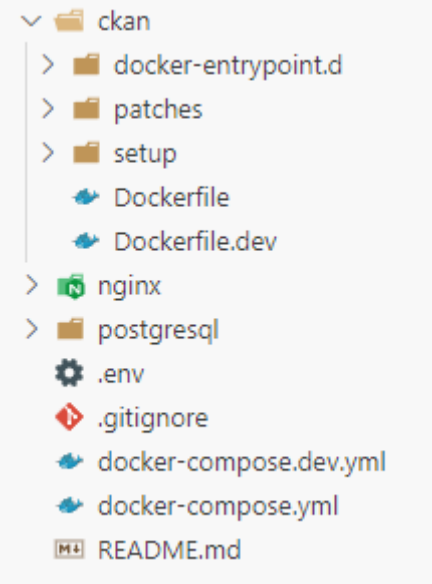
Prije pokretanja ovog Docker projekta potrebno je bilo i definirati konfiguracijske varijable navedene u .env datoteci. Tu se nalaze varijable poput verzija baze podataka i drugih usluga, korisničkog imena i varijable zadanih korisnika odnosno administratora, adrese i portovi pokrenutih web usluga, lista uključenih ekstenzija itd.

Naredbom *docker compose* izvršava se docker skripta za pokretanje navedenih slika. CKAN-ov Python poslužitelj ima zasebnu skriptu koja se izvršava pri pokretanju kontejnera *docker compose* naredbom. U navedenoj skripti navode se naredbe poput pip install koje se prve izvršavaju unutar novopokrenutog CKAN kontejnera. Na primjer, za svrhe ovoga projekta korisno je bilo iskoristiti određene funkcije iz CKAN + DCAT ekstenzije, pa su se u toj skripti dodale linije:

```
RUN pip install -e git+https://github.com/ckan/ckanext-dcat.git#egg=ckanext-dcat
RUN pip install -r /srv/app/src/ckanext-dcat/requirements.txt
```

Njima se dohvaćaju datoteke CKAN ekstenzije i ostalih proširenja koji su potrebni za rad CKAN ekstenzije.

Konačno, naredbom *docker compose -f docker-compose.dev.yml build* pokreću se svi kontejneri po prethodno konfiguriranim postavkama. Organizacija korijenskog direktorija u kojem su datoteke projekta i iz kojeg se pokreću naredbe prikazana je na sljedećoj slici.



Sl. 3.1 Prikaz organizacije mapa i datoteka korijenskog direktorija

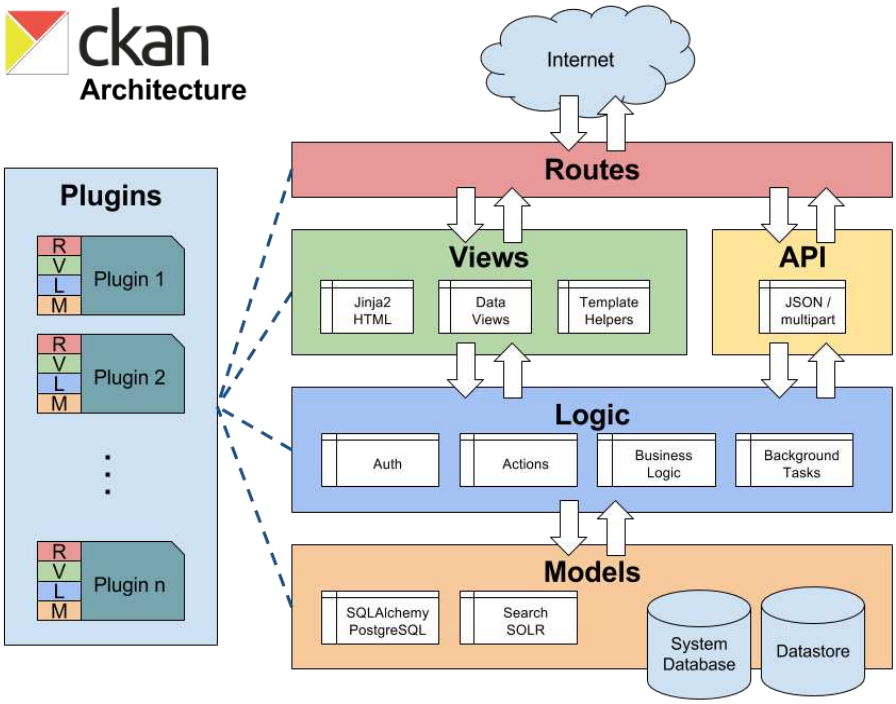
3.2. CKAN ekstenzije

Osnovne datoteke nove CKAN ekstenzije mogu se generirati naredbom *ckan generate extension --output-dir /srv/app/src_extensions* pokrenutom unutar CKAN kontejnera. Potom će se od korisnika tražiti unos imena nove ekstenzije i drugih informacija (ime autora, ime održavatelja itd.). Time se generiraju osnovne datoteke CKAN ekstenzije unutar kontejnera i izvan kontejnera u našem lokalnom direktoriju. Dva navedena direktorija biti će povezani čime će se omogućiti uređivanje izvornog koda ekstenzije koje će se preslikavati na CKAN kontejner.

CKAN ekstenzija može se sastojati od nekoliko plugin-a. Pluginovi su Python klase koje nasljeđuju *SingletonPlugin* klasu. Unutar plugin-a implementiraju se funkcije definirane u sučeljima. CKAN prilikom pokretanja izvršava sve funkcije implementiranih sučelja. Na taj način, naš se plugin izvršava u CKAN-u koji poziva metode naslijeđene iz postojećih sučelja. Za takve metode koristi se naziv nadomjestive metode (*hooks*). Tako se npr. unutar

našeg plugina implementira metoda *update_config(self, config_)* unutar koje se dodaju direktoriji resursa koje se našim pluginom želi uključiti (skripte, HTML, slike itd.).

CKAN arhitektura okvirno prati standardni arhitekturni MVC (Model View Controller) obrazac. Plugini u tom smislu proširuju funkcionalnosti na svim razinama implementirajući metode različitih sučelja koja su dostupna uvozom *ckan.plugins.interfaces* Python modula.



Sl. 3.2 Dijagram arhitekture CKAN-a [25]

Na dijagramu su prikazani svi slojevi i glavne grupe funkcionalnosti koje su ostvarene u sklopu svakog sloja. Glavni način za dohvat podataka u sloju Model je pozivom Python modula. Modul *ckan.model* namijenjen za dohvat tablica iz baze podataka (poput User, Organization, Dataset). SQLAlchemy je Python alat koji pruža upravljanje SQL bazama podataka te mapiranje SQL objekata u podatkovne strukture Python programskog jezika. Koristi se u CKAN-u za ostvarivanje veze između PostgreSQL baze podataka i Python web poslužitelja. Dostupni su također Python pozivi prema Solr serveru za izvršavanje upita nad bazom podataka.

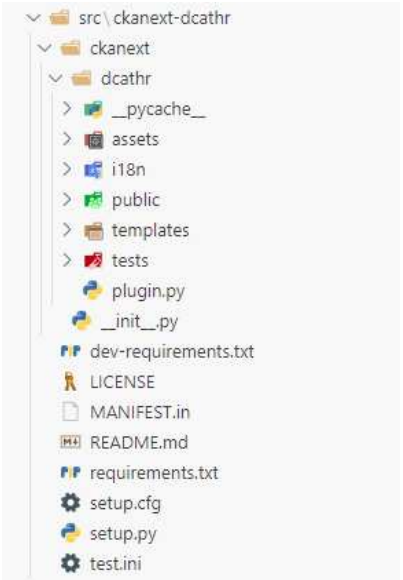
Sloj logike odgovoran je za provođenje autorizacije, validacije, stvaranje i uređivanje skupova podataka. Pritom CKAN pruža brojne pomoćne metode dostupne u modulima

ckan.logic.auth za autorizacijske metode i *ckan.logic.action* za akcijske metode (npr. create, update, get, delete itd.). Temeljni Python radni okvir kojim je ostvarena CKAN web aplikacija je Flask.

Vanjsko sučelje CKAN sustava ostvaruje se API-jem i web stranicom. Sloj Views moguće je proširivati razvojem *helper* funkcija i HTML stranica. Za izradu HTML stranica koristi se Jinja mehanizam za izradu obrazaca (*template engine*). Jinja kombinira Python i HTML sintaksu za izradu obrazaca i odsječaka. Helper funkcije definirane u Python pluginu mogu se onda pozvati unutar Jinja HTML obrazaca.

3.3. Implementacija

Prethodno navedenom ckan generate naredbom započinje se razvoj naše ekstenzije. Našoj ekstenziji dan je naziv *dcathr* u stilu opisanih postojećih ekstenzija. U lokalnom korijenskom direktoriju CKAN projekta generirana je nova mapa. Glavna datoteka ekstenzije je *setup.py*, u njoj su definirani svi pluginovi koje ekstenzija koristi, detalji poput imena, autora i licence i druge konfiguracijske varijable na razini cijele ekstenzije. Za potrebe naše ekstenzije koristit će se samo jedan plugin. Datoteka u kojoj je plugin definiran nazvana je *plugin.py*.



Sl. 3.3 Okvirni prikaz organizacije direktorija CKAN ekstenzije

3.3.1. Plugin

Plugin je nazvan *DatasetFormPlugin* jer se prvenstveno bavi izmjenama u formi za kreaciju i uređivanje skupova podataka.

```
from ckan.types import Schema
import ckan.plugins as p
import ckan.plugins.toolkit as tk

class DatasetFormPlugin(tk.DefaultDatasetForm, p.SingletonPlugin):
      p.implements(p.IConfigurer)
      p.implements(p.ITemplateHelpers)
      p.implements(p.IDatasetForm)
```

Kod 3.1 – Deklaracija klase našega plugina

Kao što je vidljivo u navedenom programskom kodu, naš plugin nasljeđuje tri CKAN plugin sučelja, *IConfigurer*, *ITemplateHelpers* te *IDatasetForm*. *IConfigurer* je sučelje je namijenjeno za konfiguriranje CKAN okoline. U našem slučaju direktoriji za HTML predloške postavljaju se pod *templates*, za statičke resurse poput slika i dokumenata pod *public* te za JavaScript i CSS datoteke za proširivanje funkcionalnosti i stila HTML stranica pod *assets*.

To se postiže implementacijom metode *update_config* IConfigurer sučelja:

```
class DatasetFormPlugin(tk.DefaultDatasetForm, p.SingletonPlugin):
...
      # IConfigurer
      def update_config(self, config_):
            tk.add_template_directory(config_, "templates")
            tk.add_public_directory(config_, "public")
            tk.add_resource("assets", "dcathr")
```

Kod 3.2 – Metoda *update_config*

Sučelje *ITemplateHelpers* i metoda *get_helpers* koriste se za dodavanje prilagođenih *helper* funkcija. Implementiranjem ovog sučelja plugini mogu definirati nove *helper* metode koje se onda pozivaju unutar HTML predložaka koristeći varijablu *h*.

```
from ckanext.dcat.processors import RDFSerializer

def serialize_dataset(dataset_id):
      serializer = RDFSerializer()
```

```
            dataset = tk.get_action('package_show')({},{'id':dataset_id})
            dataset_xml =
            serializer.serialize_dataset(dataset,_format='xml')
            return dataset_xml
...
            # ITemplateHelpers
            def get_helpers(self):
                  #Register the serialize_dataset() function above
                  #as a template helper function.
                  return {'dcathr_serialize_dataset': serialize_dataset}
```

Kod 3.3 – Metoda *get_helpers*

Metoda *serialize_dataset* koristi *RDFSerializer()*, objekt iz CKAN + DCAT proširenja za serijaliziranje CKAN skupova podataka u RDF (DCAT-AP) reprezentaciju. Ova se metoda koristi za validaciju skupa podataka, poziva se pritiskom na gumb *Validate Dataset Schema*. Pritom se rezultat koji metoda vraća, odnosno XML opis skupa podataka, predaje daljnje JavaScript modulu odgovornom za komunikaciju s validacijskim API servisom.

Posljednje sučelje čija se metode implementiraju je *IDatasetForm*. Metodama create_package_schema, update_package_schema i show_package_schema upravlja se kreacijskom shemom skupova podataka.

```
            # IDatasetForm
            def create_package_schema(self) -> Schema:
                  # let's grab the default schema in our plugin
                  schema: Schema = super(DatasetFormPlugin,
                        self).create_package_schema()
                  # our custom field
                  schema.update({'ogledno_polje':
                        [tk.get_validator('ignore_missing'),
                        tk.get_converter('convert_to_extras')]
                  })
                  return schema
```

Kod 3.4 – Metoda *create_package_schema*

Na primjeru metode create_package_schema može se vidjeti pristup koji se koristi i u drugim metodama. Prvo se dohvaća važeća shema skupova podataka kojoj se ovisno o radnji (*create*, *update*, *show*) potom dodaje niz novih polja metapodataka. U ovom

oglednom primjeru to je polje *ogledno_polje*. Inače tu se navode sva polja uključena u željenu shemu metapodataka, odnosno u našem slučaju DCAT-AP profil metapodataka.

Dosad su navedene izmjene na razini poslovne logike i modela podataka. Međutim, da bi se te izmjene vidjele na posljednjoj razini, odnosno korisničkom sučelju, potrebno je implementirati ih i u obliku HTML obrazaca i JavaScript modula.

3.3.2. Predlošci

Izmjene i dodavanje HTML obrazaca omogućuju se dodavanjem datoteka u *templates* direktorij. Izmjene na već postojećim HTML stranicama postižu se imenovanjem HTML datoteka ekstenzije istim imenima kao i imenima izvornih datoteka CKAN projekta. Sintaksa takvih HTML predložaka proširena je uvođenjem Jinja izraza, čime se između ostalog omogućavaju pozivi prethodno definiranih Python *helper* funkcija.

```
{% ckan_extends %}

{% block package_basic_fields_custom %}
      {{ form.input('ogledno_polje',
            label=_('Ogledno polje'),
            id='field-ogledno_polje',
            placeholder=_('ogledno polje'),
            value=data.ogledno_polje,
            error=errors.ogledno_polje,
            classes=['control-medium']
      ) }}
{% endblock %}
```

Kod 3.5 – Isječak package_basic_fields.html

U navedenom se isječku pokazuje takva sintaksa u primjeru dodavanja novog polja obrascu za unos metapodataka skupova podataka.

U sljedećem se bloku može vidjeti upravo korištenje *helper* funkcija te poziv vanjskog JavaScript modula. Blok predstavlja gumb kojim se pokreće validacija skupa podataka, a dodaje se na stranicama za uređivanje i pregled metapodataka postojećeg skupa podataka. U HTML opisu gumba koristi se serialize_dataset funkcija prethodno definirana unutar našeg plugina. Uz to poziva se JavaScript modul dcathr-validate-dataset kao listener modul odgovoran za obrađivanje događaja vezanih uz već spomenuti gumb *Validate Dataset*

Schema. Modulu se također kao argument predaje rezultat poziva helper funkcije, odnosno serijalizirana reprezentacija skupa podataka.

```
{% block validate_button %}
      {% if data.id and h.check_access('package_delete', {'id': data.id})
            and not data.state == 'deleted' %}
      <button class="btn btn-info" type="button" name="validate"
      data-module="dcathr-validate-dataset"
      data-module-content="{{ h.dcathr_serialize_dataset(data.id)}}">
            {% block validate_button_text %}
                  {{ _('Validate Dataset Schema') }}
            {% endblock %}
      </button>
      {% endif %}
{% endblock %}
```

Kod 3.6 – Isječak HTML gumba za validaciju

Pozivi JavaScript modula ostvareni su preko *data-module* atributa elementa *button*, unutar toga atributa u dvostrukim vitičastim zagradama također je vidljiv poziv *helper* funkcije preko varijable *h*. Pozvani JavaScript modul potrebno je uvesti u HTML isječak. To je postignuto dodavanjem retka:

```
            {% asset 'dcathr/dcathr-validate-dataset' %}
```

3.3.3. JavaScript modul

CKAN za definiranje JavaScript modula koristi jQuery. JQuery je JavaScript programska biblioteka za olakšavanje uređivanja HTML dokumenata, obrade događaja, slanje mrežnih zahtjeva itd. Za upravljanje resursima CKAN web aplikacije koristi se Python biblioteka *webassets*. Sukladno tome sve je JavaScript i CSS datoteke koje se žele uključiti u ekstenziju potrebno smjestiti u *assets* mapu te navesti ih u *webassets.yml* konfiguracijskoj datoteci.

Prethodno spomenuti modul nazvan je *dcathr-validate-dataset*. Modulima ekstenzija u svrhu izbjegavanja preklapanja imena dodaje se prefiks imena ekstenzije (*dcathr-* u našem slučaju).

```
      this.ckan.module("dcathr-validate-dataset", function ($) {
```

```
return {
        options: {
                content: '',
        },
        initialize: function () {
                jQuery.proxyAll(this, /_on/);
                this.el.on('click', this._onClick);
        },
```

Kod 3.7 – Deklaracija i inicijalizacija modula

Modul se inicijalizira preusmjeravanjem svih *click* događaja dodjeljenjog mu elementa na metode definirane u modulu. Također, kao jedini atribut ovog modula definira se varijabla *content*, koja predstavlja serijalizirani oblik validiranog skupa podataka. Na korisnikov klik poziva se funkcija validateDataset. Unutar funkcije šalje se poziv validacijskom serveru te se potom rezultat poziva izveden iz odgovora servera prikazuje na dohvaćenim HTML elementima. U slučaju negativnog odgovora za svako kršenje sheme ispisuje se odgovarajuća poruka.

```
validateDataset: function () {
        var queryParameters = "?shapeModel=dcatap211level3";
        var apiUrl =
        "https://data.europa.eu/api/mqa/shacl/validation/report" +
        queryParameters;

        var request = $.ajax({
                url: apiUrl,
                method: "POST",
                data: this.options.content,
                dataType: "text",
                contentType: "text/xml",
        });
```

Kod 3.8 – Funkcija validateDataset

API korišten za validaciju ne zahtijeva autentifikaciju ili neki drugi oblik prethodne komunikacije. Jedini parametar upita je *shapeModel* tj. model sheme po kojoj se DCAT-AP skup podataka validira. Među dostupnim modelima su različite verzije DCAT-AP-a, od 1.1 do najnovije 2.1.1 verzije. Dostupne su također različite razine zahtjevnosti odnosno

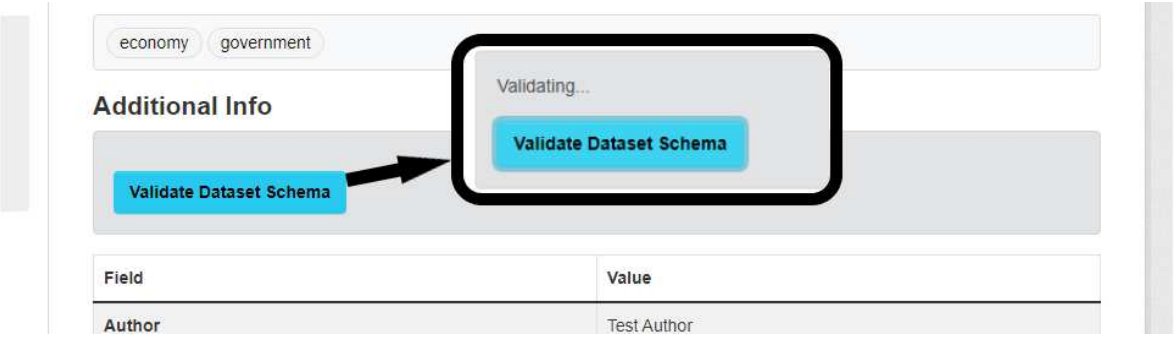
strogosti pri utvrđivanju ispravnosti skupa podataka. U sklopu SHACL validacijskog servera dostupne su tri razine usklađenosti s DCAT-AP profilom [2]:

- Svi zahtjevi (zadano): osnovna svojstva, RDF rječnici, preporučena svojstva, ograničenje raspona
- Treća razina: osnovna svojstva, RDF rječnici, preporučena svojstva
- Druga razina: osnovna svojstva, RDF rječnici
- Prva razina: osnovna svojstva

3.3.4. Validacija DCAT-AP reprezentacije skupova podataka

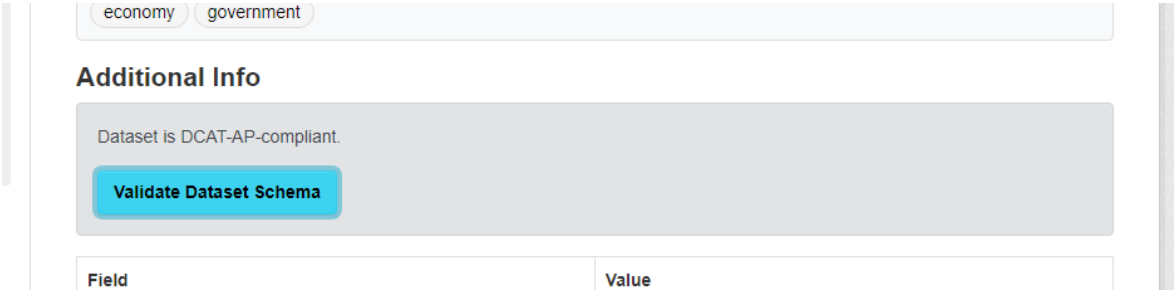
Dosad navedenim izmjenama i proširenjima programskog koda omogućena je provjera usklađenosti sheme metapodataka s DCAT-AP-om. Korisnik nakon kreacije novog skupa podataka i unosa metapodataka korištenjem gumba Validate Dataset Schema i validacijom sheme utvrđuje kvalitetu metapodataka. U slučaju određenih nepravilnosti može urediti postojeći skup podataka na zaslonu za uređivanje.

Na glavnoj stranici skupa podataka uz izdvojene metapodatke skupa podataka nalazi se gumb za validaciju.



Sl. 3.4 Prikaz gumba na stranici skupa podataka

Pritiskom na gumb šalje se HTTP POST zahtjev validacijskom serveru. Odgovor servera na poslani zahtjev sadrži rezultat validacije. Ovisno o rezultatu, ispisuje se poruka koja sadrži sva kršenja odabranog profila metapodataka ili samo potvrdu valjanosti sheme skupa podataka.



Sl. 3.5 Potvrdna poruka validacije

U slučaju nedostataka u shemi skupa podataka kao rezultat se ispisuju nedostajuća svojstva odnosno metapodatci te poveznice na njihove izvorne RDF rječnike i definicije.

Sl. 3.6 Rezultat validacije s greškama u shemi

Na navedenim su isječcima zaslona prikazane interakcije s našom ekstenzijom.

U kontekstu ovoga rada naša ekstenzija predstavlja programsko rješenje odnosno alat kojim se rješavaju neki od prethodno navedenih zahtjeva u području osiguravanja kvalitete metapodataka portala otvorenih podataka.

Zaključak

U ovom radu objasnio se pojam kvalitete metapodataka i njegova važnost za iskorištavanje podataka. Pružio se uvid u postojeće relevantne specifikacije za definiranje metapodataka zasnovane na DCAT-u. Potom su se istražili zahtjevi za poboljšanjem kvalitete metapodataka podataka i postojeća programska rješenja za postizanje tih zahtjeva. Time se pružio skup zahtjeva koje bi alat za osiguravanje kvalitete mogao ispuniti. Završno, implementirano je jedno takvo rješenje u obliku CKAN ekstenzije koje konkretno izvršava validaciju sheme metapodataka skupa podataka. Time se pružio prototip jednog alata za osiguravanje kvalitete metapodataka na portalima otvorenih podataka te uvid u problematiku određivanja i osiguravanja kvalitete metapodataka.

Literatura

[1] Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016.)

[2] Metodologija procjene kvalitete metapodataka, (2024, veljača). Poveznica: https://data.europa.eu/mqa/methodology?locale=hr; pristupljeno 1. veljače 2024.

[3] Data Catalog Vocabulary (DCAT) - Version 2, (2024, veljača). Poveznica: https://www.w3.org/TR/2020/REC-vocab-dcat-2-20200204/; pristupljeno 1. veljače 2024.

[4] DCAT UML dijagram, (2024, veljača). Poveznica: https://www.w3.org/TR/2020/REC-vocab-dcat-2-20200204/images/schema.org-dataset.png; pristupljeno 1. veljače 2024.

[5] DCAT Profiles, (2024, veljača). Poveznica: https://www.w3.org/TR/2020/REC-vocab-dcat-2-20200204/; pristupljeno 1. veljače 2024.

[6] DCAT-AP 3.0 Context, (2024, veljača). Poveznica: https://semiceu.github.io/DCAT-AP/releases/3.0.0/#context; pristupljeno 1. veljače 2024.

[7] Interoperable Europe, (2024, veljača). Poveznica: https://joinup.ec.europa.eu/interoperable-europe; pristupljeno 1. veljače 2024.

[8] GeoDCAT-AP, (2024, veljača). Poveznica: https://semiceu.github.io/GeoDCAT-AP/drafts/latest/#context; pristupljeno 1. veljače 2024.

[9] INSPIRE, (2024, veljača). Poveznica: https://www.nipp.hr/default.aspx?id=62; pristupljeno 1. veljače 2024.

[10] ISO 19115 norma, (2024, veljača). Poveznica: https://www.iso.org/standard/26020.html; pristupljeno 1. veljače 2024.

[11] GeoDCAT Optional Classes, (2024, veljača). Poveznica: https://semiceu.github.io/GeoDCAT-AP/drafts/latest/#optional-classes; pristupljeno 1. veljače 2024.

[12] StatDCAT Application Profile for data portals in Europe, (2024, veljača). Poveznica: https://joinup.ec.europa.eu/collection/semantic-interoperability-community-semic/solution/statdcat-application-profile-data-portals-europe; pristupljeno 1. veljače 2024.

[13] Figure 2: DCAT-AP Model mapped to SDMX Model Classes, (2024, veljača). Poveznica: https://ceur-ws.org/Vol-1654/article-05.pdf; pristupljeno 1. veljače 2024.

[14] Metadatenstruktur für Daten in Deutschland, (2024, veljača). Poveznica: https://www.govdata.de/metadatenschema; pristupljeno 1. veljače 2024.

[15] DCAT-AP.de UML-Diagramm, (2024, veljača). Poveznica: https://www.dcat-ap.de/def/dcatde/2.0/spec/; pristupljeno 1. veljače 2024.

[16] DCAT-AP_IT v1.0 – Profilo italiano di DCAT-AP, (2024, veljača). Poveznica: https://dati.gov.it/onto/dcatapit; pristupljeno 1. veljače 2024.

[17] DCAT-AP_IT UML Diagram, (2024, veljača). Poveznica: https://docs.italia.it/italia/daf/linee-guida-cataloghi-dati-dcat-ap-it/it/stabile/dcat-ap_it.html; pristupljeno 1. veljače 2024.

[18] Open Data Maturity 2023 Croatia, (2024, veljača). Poveznica: https://data.europa.eu/sites/default/files/country-factsheet_croatia_2023.pdf; pristupljeno 1. veljače 2024.

[19] Kvaliteta metapodataka Portal otvorenih podataka, (2024, veljača). Poveznica: https://data.europa.eu/mqa/catalogues/portal-otvorenih-podataka/?locale=hr; pristupljeno 1. veljače 2024.

[20] Report on improving data publishing, (2024, veljača). Poveznica: https://data.europa.eu/sites/default/files/report/D2.1.4_Report_on_improving_data_publishing_en_230720.pdf; pristupljeno 1. veljače 2024.

[21] CKAN + DCAT extension, (2024, veljača). Poveznica: https://extensions.ckan.org/extension/dcat/; pristupljeno 1. veljače 2024.

[22] ckanext-dcatde repozitorij ekstenzije, (2024, veljača). Poveznica: https://github.com/GovDataOfficial/ckanext-dcatde; pristupljeno 1. veljače 2024.

[23] Ckenext-dcatapit repozitorij ekstenzije, (2024, veljača). Poveznica: https://github.com/geosolutions-it/ckanext-dcatapit; pristupljeno 1. veljače 2024.

[24] CKAN-Docker repozitorij, (2024, veljača). Poveznica: https://github.com/ckan/ckan-docker; pristupljeno 1. veljače 2024.

[25] CKAN code architecture, (2024, veljača). Poveznica: https://docs.ckan.org/en/2.9/contributing/architecture.html; pristupljeno 1. veljače 2024.

Sažetak

Alat za osiguravanje kvalitete meta podataka na portalima otvorenih podataka

Rastom broja portala otvorenih podataka povećava se i količina objavljenih podataka, međutim ne nužno i sama iskoristivost podataka te interoperabilnost portala. Upravo na tome području ulažu se napori za poboljšanje kvalitete tih podataka ili preciznije metapodataka. U ovom radu istražit će se relevantne norme koje se koriste za opisivanje skupova podataka, pružit će se pregled dosadašnjih programskih rješenja na području osiguravanja kvalitete metapodataka te završno implementirati će se programsko rješenje koje zadovoljava neke od identificiranih zahtjeva. Navedeni će se ciljevi obraditi s naglaskom na specifičnostima europskih portala otvorenih podataka i zahtjeva definiranih od strane organizacija i tijela Europske Unije, koja je jedan od najvećih promicatelja otvorenosti podataka u svijetu.

Ključne riječi: otvoreni podatci; portali otvorenih podataka; kvaliteta metapodataka; osiguravanje kvalitete metapodataka; CKAN; DCAT

Summary

Metadata Quality Assurance Tool for Open Data Portals

As the number of open data portals increases, so does the amount of published data, but not necessarily the usability of the data and the interoperability of the portals. It is precisely in this area that efforts are being made to improve the quality of this data, or more precisely metadata. In this paper, the relevant norms used to describe datasets will be examined, an overview of existing software solutions in the field of metadata quality assurance will be provided, and finally, a software solution that meets some of the identified requirements will be implemented. The stated goals will be addressed with an emphasis on the specifics of European open data portals and requirements defined by organizations and bodies of the European Union, which is one of the biggest promoters of open data in the world.

Keywords: open dana; open data portals; metadata quality; metadata quality assurance; CKAN; DCAT

Skraćenice

API	*Application programming interface*	aplikacijsko programsko sučelje
CKAN	*Comprehensive Knowledge Archive Network*	platforma portala otvorenih podataka
DCAT	*Data Catalog Vocabulary*	RDF rječnik za kataloge podataka
DCAT-AP	*DCAT Application profile*	DCAT profil
FAIR	*Findable, Accessible, Interoperable, and Reusable*	načela otvorenosti podataka
HTML	*HyperText Markup Language*	jezik za izradu web stranica
HTTP	*HyperText Transfer Protocol*	mrežni protokol za razmjenu podataka
INSPIRE	*INfrastructure for SPatial Information*	direktiva Europske Unije
ISO	*International Organization for Standardization*	standardizacijska organizacija
RDF	*Resource Description Framework*	radni okvir za modele podataka
RPC	*Remote Procedure Call*	udaljeno izvršavanje procedura
SHACL	*Shapes Constraint Language*	jezik za validaciju RDF dokumenata
URI	*Uniform Resource Identifier*	jedinstveni identifikator resursa
W3C	*World Wide Web Consortium*	web standardizacijska organizacija