LATVIJAS UNIVERSITĀTE
DATORIKAS FAKULTĀTE

LATVIJAS ATVĒRTO DATU KVALITĀTES ANALĪZE UN IDENTIFICĒTO KVALITĀTES TRŪKUMU KLASIFICĒŠANA

BAKALAURA DARBS

Autore: **Ieva Baldere**
Studenta apliecības Nr.: ib17111
Darba vadītāja: docente, Dr. sc. comp. Anastasija Ņikiforova

RĪGA 2021

ANOTĀCIJA

Bakalaura darbā “Latvijas atvērto datu kvalitātes analīze un identificēto kvalitātes trūkumu klasificēšana” tiek analizētas un salīdzinātas atvērto datu kopas, noskaidrojot datu kvalitātes trūkumus.

Darba mērķis ir izpētīt un apkopot literatūru, veidojot pārskatu par datu kvalitātes dimensijām, problēmu veidiem, standartiem un vadlīnijām un izveidot pārbaudāmo pozīciju apkopojumu, uz kuru balstoties veikt SQL-balstītu analīzi vairākām Latvijas atvērto datu kopām, rezultātā secinot, kuras ir visizplatītākās datu kvalitātes problēmas un trūkumi.

Atslēgvārdi: datu kvalitāte, atvērtie dati, datu kvalitātes dimensijas, atvērto datu portāls

ABSTRACT

ANALYSIS OF LATVIAN OPEN DATA AND CLASSIFICATION OF THE IDENTIFIED QUALITY PROBLEMS

The Bachelor's Thesis "Analysis of Latvian open data and classification of the identified quality problems" analyses and compares open data sets by determining the data quality problems.

The aim of the work is to study and summarize literature by creating an overview of data quality dimensions, types of problems, standards and guidelines, and to create a list of the positions to be tested by a SQL-based analysis on several Latvian open datasets thereby concluding which are the most common data quality problems.

Keywords: data quality, open data, data quality dimensions, open data portal

SATURA RĀDĪTĀJS

APZĪMĒJUMU SARAKSTS

API – funkciju un procedūru kopums, kas ļauj veidot lietojumprogrammas, kas piekļūst operētājsistēmas, lietojumprogrammas vai cita pakalpojuma funkcijām vai datiem.

CSV – komatu atdalītu vērtību fails ir norobežots teksta fails, kurā vērtību atdalīšanai tiek izmantots komats.

ISO – starptautiskā Standartizācijas organizācija ir nacionālo standartizācijas organizāciju federācija, kuras galvenās funkcijas ir starptautisko standartu izstrādāšana, publicēšana un izplatīšana.

PDF - pārnesams dokumentu formāts ir elektronisku dokumentu datņu formāts.

RDF – sistēma resursu aprakstīšanai tīmeklī

SQL – vaicājumu valoda, kas paredzēta datu manipulēšanai relāciju datubāžu pārvaldības sistēmās.

URI - vienotais resursu identifikators ir unikāla rakstzīmju secība, kas identificē loģisko vai fizisko resursu, ko izmanto tīmekļa tehnoloģijas.

W3C – vispasaules Tīmekļa konsorcijs veido standartus jeb "rekomendācijas" vispasaules tīmeklim.

XML – paplašināmā iezīmēšanas valoda jeb XML ir W3C rekomendācija speciālas nozīmes iezīmēšanas valodu veidošanai.

IEVADS

Datu kvalitāte atvērto datu kontekstā ir svarīga, jo pieprasījums pēdējo gadu laikā pēc atvērtajiem datiem ir tikai audzis. Atvērto datu kvalitāte ir svarīga, lai jebkura ieinteresēta persona, gan fiziskā, gan juridiskā persona varētu gūt ieguvumus, atvērtos datus izmantojot saviem nolūkiem.

Darba mērķis ir izpētīt un apkopot literatūru, veidojot pārskatu par datu kvalitātes dimensijām, to metrikām, trūkumu veidiem un pārbaudāmo pozīciju apkopojumu, uz kuru balstoties veikt SQL-balstītu analīzi vairākām Latvijas atvērto datu kopām, rezultātā secinot, kuras ir visizplatītākās datu kvalitātes problēmas un trūkumi, un kā attiecīgo datu kopu kvalitāti varētu uzlabot.

Mērķa sasniegšanai izvirzīti sekojoši uzdevumi:

1. izpētīt datu kvalitātes dimensijas, atvērto datu kvalitātes problēmas un tās apkopot;
2. izpētīt Latvijas Atvērto datu portāla pamatprincipus;
3. izpētīt datu kvalitātes nozīmīgumu;
4. izpētīt, vai eksistē datu kvalitātes standarti, vadlīnijas;
5. no apkopotās informācijas izveidot atvērto datu kopu pārbaudāmo kvalitātes aspektu sarakstu;
6. balstoties uz identificēto pārbaužu sarakstu, veikt datu kopu analīzi;
7. salīdzināt savā starpā vairāku datu kopu veiktās datu kvalitātes analīzes rezultātus, nosakot, kā datu kvalitāte varētu tikt uzlabota;
8. pieredzes rezultātā secināt, kas ir izplatītākie atvērto datu kvalitātes trūkumi un, kura pieeja datu kvalitātes analīzei ir galalietotājam piemērotāka.

Darba teorētiskajā daļā tiek apskatīta un analizēta literatūra par datu kvalitāti, tās dimensijām, datu kvalitātes nozīmīgumu un tās standartiem un vadlīnijām. Balstoties uz analizēto literatūru tiek izveidots datu kvalitātes pārbaužu saraksts, kas tiek izmantots darba praktiskajā daļā.

Darba praktiskajā daļā autore analizē datu kopas no Latvijas Atvērto datu portāla kategorijas “Valsts pārvalde”, tādējādi nodrošinot iespēju nepieciešamības gadījumā veikt dažādu datu kopu kontekstuālo pārbaudi, portāla datu kopu analizējot, izsecinot, ka dotā

kategorija ir vispiemērotākā dotajam nolūkam. Papildus tas ļauj iekļaut analīzē datu kopu "Valsts informācijas sistēmu reģistrs", jo autore kursa darbā attiecīgās datu kopas datu kvalitāti pārbaudīja izmantojot datu kvalitātes analīzes rīkus, tādējādi arī rezultātā nosakot priekšrocības un trūkumus datu kvalitātes analīzi veicot ar gataviem rīkiem vai SQL-balstītu analīzi.

Darbā izmantotas pētniecības metodes:

1. analītiskā metode – vairāku avotu analīze, lai izpētītu datu kvalitātes dimensijas, datu kvalitātes nozīmīgumu, t.sk. atvērto datu, atvērto datu kopu analīze;

2. aprakstošā metode – izpētīt problēmu, lai atrisinātu to, apkopojot izlasītu literatūru, aprakstot darba gaitu un iegūtus rezultātus.

3. salīdzinošā metode – salīdzināt autores kursa darbā izmantotos datu kvalitātes analīzes rīkus ar SQL-balstītu analīzi;

4. eksperimentālā metode – pielietot literatūras avotu apkopoto informāciju, lai analizētu datu kopas.

Darbs sastāv no ievada, 4 nodaļām, izmantotās literatūras saraksta, kas sastāv no 25 avotiem. Darbā iekļautas 8 tabulas, 7 attēli, 2 pielikumi.

1. ATVĒRTO DATU UN DATU KVALITĀTES JĒDZIENI

1.1. Datu kvalitāte, dimensijas

Datu kvalitāte ir datu piemērotība lietošanai. Datu kvalitāte literatūrā ir definēta kā piemērotība lietotāju vajadzībām un lietojumam (angl. use-case) [1]. Pētījumi apstiprinājuši, ka datu kvalitāte ir daudzdimensiju koncepcija [11]. Tomēr tiek atzīts, ka datu kvalitāti ir grūti definēt. Tas ir pamatojams ar to, ka atšķirībā no ražotajiem izstrādājumiem datiem nav fizisku īpašību, kas ļautu viegli novērtēt to kvalitāti.

Piemērotība lietošanai nozīmē, ka atbilstošais datu kvalitātes līmenis ir atkarīgs no konteksta. Ir grūti noteikt vajadzīgo kvalitāti, ja dažādiem lietotājiem ir atšķirīgas vajadzības. Var rasties vēlme lietot viskvalitatīvāko, ar milzīgu daudzumu datu kvalitātes prasībām, kas var šķist vislabākais variants. Taču minētais lietojums organizācijai var būt visai maznozīmīgs un nesvarīgs, ja visas kvalitātes prasības nemaz neattiecas uz konkrētās organizācijas datiem. Tādējādi ir jālīdzsvaro konfliktējošās prasības attiecībā uz datu kvalitāti [19].

Datu kvalitātes dimensija tiek definēta kā datu kvalitātes atribūtu kopums, kas pārstāv vienu datu kvalitātes aspektu vai struktūru. Literatūrā [24] izmanto trīs pieejas, lai pētītu datu kvalitāti:

1. Intuitīvā pieeja;
2. teorētiskā pieeja;
3. empīriskā pieeja.

Intuitīvā pieeja tiek izmantota, ja datu kvalitātes atribūtu izvēle jebkuram konkrētam pētījumam ir balstīta uz pētnieku pieredzi vai intuitīvu izpratni par to, kādi atribūti ir svarīgi. Lielākā daļa datu kvalitātes pētījumu ietilpst šajā kategorijā [24]. Arī darba autore izmanto šo pieeju, lai veidotu datu kvalitātes pārbaudes sarakstu datu kopu analīzei.

Balstoties uz kādu noteiktu kontekstu, datu kvalitāte var būt gan **subjektīvi** noteikta, gan objektīvi mērīta [11]. Objektīvas pārbaudes, piemēram, nozīmē to, ka tās var veikt jebkurai datu kopai. Atvērto datu kontekstā tas būtu pārbaudīt failu tipus (CSV, XML, PDF u.c.), taču subjektīvas pārbaudes būtu vairāk attiecināmas uz katru datu kopu individuāli, piemēram, pieļaujamas vērtības, ‘NULL’ pieļaujamība atribūtam u.c.

Objektīvie novērtējumi var būt gan neatkarīgi, gan atkarīgi no uzdevumiem. No uzdevumiem neatkarīga metrika atspoguļo datu stāvokli bez konteksta zināšanām par lietojumprogrammu, un to var izmantot jebkurai datu kopai neatkarīgi no uzdevumiem. No uzdevumiem atkarīgā metrika, kas ietver organizācijas uzņēmējdarbības noteikumus, uzņēmuma un valdības noteikumus un ierobežojumus, ko nodrošina datu bāzes administrators, tiek izstrādāta īpašos lietojumprogrammas kontekstos [11].

Tabulā 1.1.1. apskatāms datu kvalitātes vadošo pētnieku Pipino, Lee, Wang dimensiju un to definīciju apkopojumu.

*Tabula Nr. 1.1.1.. **Datu kvalitātes dimensijas [11]***

Dimensijas	**Definīcijas**
Pieejamība	Cik daudz datu ir pieejami vai viegli un ātri iegūstami
Atbilstošs datu apjoms	Cik lielā mērā datu apjoms ir piemērots attiecīgajam uzdevumam
Ticamība	Apjoms, kādā dati tiek uzskatīti par patiesiem un ticamiem
Pilnīgums	Apjoms, kādā dati neiztrūkst, un ir pietiekamā apjomā veicamajam uzdevumam
Reprezentācija	Apjoms, kādā dati tiek sniegti vienā un tajā pašā formātā
Manipulēšanas ērtība (angl. ease of manipulation)	Cik viegli ir manipulēt ar datiem un tos piemērot dažādiem uzdevumiem
Pareizība	Cik lielā mērā dati ir pareizi un ticami
Interpretējamība	Datu apjoms piemērotās valodās, simbolos un vienībās, un to definīcijas ir skaidras
Objektivitāte	Apjoms, kādā dati ir objektīvi, bez aizspriedumiem
Atbilstība	Apjoms, kādā dati ir piemēroti un noderīgi veicamajam uzdevumam
Reputācija	Apjoms, cik augstu dati tiek vērtēti to avotu un satura ziņā
Drošība	Apjoms, kādā piekļuve datiem tiek atbilstoši ierobežota, lai saglabātu to drošību
Savlaicīgums	Cik lielā mērā dati ir pietiekami atjaunināti
Saprotamība	Cik viegli dati ir saprotami
Pievienotā vērtība	Apjoms, kādā dati ir izdevīgi un sniedz priekšrocības no to izmantošanas

1.2. Atvērtie dati un kvalitāte

Atvērtie dati ir dati, kurus ikviens var brīvi izmantot, modificēt un koplietot jebkādiem mērķiem, un tos atļauts izplatīt bez ierobežojumiem. Salīdzinot ar datiem, kas pakļauti īpašumtiesībām, tie atšķiras ar to, ka tie tiek publicēti tādā veidā, kas atļauj to tālāku izmantošanu - raksturīgi mazāki ierobežojumi, kas tiek piemēroti to apritei un atkārtotai izmantošanai. Atvērtie dati ir apstrādājami – tos iespējams lejupielādēt, lai lietotāji datus var rediģēt un analizēt.

Saskaņā ar atvērto datu definīciju [13] paredzēts, ka:

- datiem jābūt **publiskiem vai nodrošinātiem ar atvērto licenci**. Visi datiem pievienotie papildu noteikumi (piemēram, lietošanas noteikumi) nedrīkst būt pretrunā ar datu publiskā domēna statusu vai licences noteikumiem;
- datiem jābūt **brīvi pieejamiem**, lejupielādējamiem bezmaksas. Tiem jāpievieno arī papildu informācija, kas nepieciešama, lai nodrošinātu licences atbilstību;
- datiem jābūt **mašīnlasāmiem** – tādā formā, kas ir viegli apstrādājama ar datoru un kurā var viegli piekļūt atsevišķiem darba elementiem un pārveidot tos;
- datiem jābūt **atvērtā formātā** - tajā nav noteikti nekādi monetāri vai citādi ierobežojumi lietošanai, un kuru var pilnībā apstrādāt ar vismaz vienu bezmaksas, atvērtā pirmkoda programmatūras rīku.

2006. gadā Tims Bērnerss-Lī publicēja atvērto datu izvietošanas shēmu, kuras pamatā ir piecas pieaugošas un pakāpeniski prasīgas prasības, kas apzīmētas kā zvaigznes. 5 zvaigžņu atvērtai datu kopai jāatbilst visām šīm prasībām:

1. Pieejams tīmeklī, jebkura formāta nodrošinātajiem datiem ir atvērta licence;
2. Pieejami kā mašīnlasāmi strukturēti dati (piemēram, programma Excel, nevis attēlu skenēšana);
3. Pieejamais nepatentētais formāts (piemēram, CSV, nevis Excel);
4. Tiek izmantots W3C (RDF) un URIs atvērtie standarti;
5. Lai nodrošinātu kontekstu, datus sasaista ar citu pakalpojumu sniedzēju datiem.

Lai gan šī piecu zvaigžņu sistēma plaši minēta, tā tiek attiecināta tikai uz konkrētu kvalitātes aspektu, t.i., formātu vai kodējumu, ko izmanto datu publicēšanai [23].

1.3. Latvijas Atvērto datu portāls

Latvijas atvērto datu centralizētais sniedzējs ir Latvijas Atvērto datu portāls (https://data.gov.lv) [22], kas ir vienota platforma piekļuvei valsts pārvaldes atvērtajiem datiem, kas tika publicēta 2017.gadā. Portālā ir datu katalogs, kas nodrošina iespēju aprakstīt atvērto datu metadatus, kā arī pievienot failus vai norādīt saites uz atvērto datu resursiem. Pašlaik portāla datu katalogā atrodamas 483 datu kopas.

Saskaņā ar Latvijas valsts pārvaldes vērtības un ētikas principiem, viena no valsts vērtībām ir tāda valsts pārvalde, kas ir atklāta un sabiedrībai pieejama. Valsts apzinās un atbalsta to, ka sabiedrībai nepieciešama informācija par valsts pārvaldes darbu un pakalpojumiem, viegli pieejamā un saprotamā veidā. Ja informācija ir pieejama, tā dod sabiedrībai iespēju analizēt datus, no datiem veidot jaunus produktus, pakalpojumus un pētījumus, kā arī iesaistīties valsts pārvaldības procesu uzlabošanā, mazinot korupcijas riskus un veicinot uzticēšanos valsts pārvaldei [22].

Arī Latvijas Atvērto datu portāla vadlīnijās [22] iekļauti pamatprincipi, ko vajadzētu ievērot publicējot datus:

- Datiem jābūt **pilnīgiem** - dati tiek publicēti tādi, kādi tie oriģināli tiek iegūti ar lielāko iespējamo detalizācijas pakāpi. Tie netiek modificēti, mainīti, tādējādi nodrošinot, ka dati netiek pakļauti iespējamām nepilnībām un neprecizitātēm. Laba prakse būtu datus portālā eksportēt un publicēt automatizēti;
- datiem jābūt **saprotamiem** – datu kopai jānodrošina metadatu pievienošana un papildus informācija par datu struktūru un to saturu saprotamā terminoloģijā, lai ikvienam lietotājam būtu skaidra datu nozīme un spētu tos izmantot;
- jānodrošina datu **atjaunošana un uzturēšana** – dati ir jāatjauno tik bieži, cik ir norādīts paredzētais atjaunošanas biežums. Saskaņā ar Ministru Kabineta noteikumu Nr. 445 "Kārtība, kādā iestādes ievieto informāciju internetā" 41.punktu [12], iestādes pienākums ir saskaņā ar iestādes norādīto datu atjaunošanas biežuma klasifikatoru aktualizēt atvērto datu portālā ievietotos datus un nodrošināt to atbilstību metadatiem.

1.4. Datu kvalitātes nozīmīgums

Plašā literatūrā ir identificētas zemas datu kvalitātes augstās izmaksas, atzīstot, ka šādu datu dēļ uzņēmumi var zaudēt vairāk nekā 10% ieņēmumu, kā arī citas nopietnas sekas, kas saistītas ar taktisko lēmumu pieņemšanu un stratēģiju veidošanu. Gartner (viena no pasaules vadošajām IT pētniecības organizācijām) 2011. gada ziņojumā tika teikts, ka "datu kvalitātes nenodrošināšanas dēļ 75% organizāciju ievērojami samazināsies ieņēmumu pieauguma potenciāls un palielināsies izmaksas" [1].

Balstoties uz pētnieku darbiem (Friedman, Redman, Laney) pētījumā [7] norādīts, ka zema datu kvalitāte ir galvenais iemesls, kāpēc 40% no visām uzņēmējdarbības iniciatīvām nespēj sasniegt savu mērķa peļņu. Noteikts arī, ka datu kvalitāte ietekmē vispārējo darba ražīgumu par 20%, un, tā kā vairāk biznesa procesu kļūst automatizēti, datu kvalitāte kļūst par vispārējo procesu kvalitātes ierobežojošo faktoru, ja datu kvalitātes līmenis nav pietiekami augsts.

Zema datu kvalitāte mēdz ietekmēt tipisko uzņēmumu dažādos veidos. Tie tiešā veidā noved pie klientu neapmierinātības, pieaugošām izmaksām un pazemina darbinieku apmierinātību ar darbu. Zema datu kvalitāte palielina darbības izmaksas, jo laiks un citi resursi tiek tērēti kļūdu noteikšanai un to labošanai. Ja vien uzņēmums nav pielicis pūles, lai nodrošinātu augstu datu kvalitāti, tam vajadzētu sagaidīt, ka datu kļūdu līmenis ir aptuveni 1–5%, kur kļūdu līmenis tiek aprēķināts kā kļūdainu lauku skaits pret kopējo lauku skaitu [1].

Rezultātā rodas situācija, ka datu kvalitāte ir zemāka par maksimāli sasniedzamo jeb absolūto kvalitāti. Ieviešot profilakses pasākumus, sākotnējo situāciju var ievērojami uzlabot, un kvalitātes līmeni var būtiski paaugstināt. Taču, lai gan tiek apgalvots, ka būtu iespējams sasniegt maksimālo datu kvalitāti, tiek uzskatīts, ka absolūtā kvalitāte nav sasniedzama. Tas ir cieši saistīts ar atvērto datu jēdzienu, kad dati ir pieejami ikvienai ieinteresētai personai un datiem var tikt nodefinēti vairāki lietošanas piemēri, tostarp tie, par kuriem datu turētājs nav iedomājies, uzkrājot un uzturot datus savā sistēmā. Taču tas ļautu būtiski uzlabot kvalitāti vismaz organizācijas nolūkiem, t.i. ikdienas darbiem, kas konkrētai organizācijai būtu svarīgākais aspekts [1].

Zemas kvalitātes dati var būtiski negatīvi ietekmēt organizācijas efektivitāti, savukārt augstas kvalitātes datiem bieži ir izšķiroša nozīme uzņēmuma panākumos [23, 1].

Jau tagad iespējams norādīt uz lielu skaitu jomu, kur atvērtie dati sniedz vērtību. Dažas no šīm jomām ir šādas [14]:

- Caurspīdīgums valsts pārvaldē;
- ekonomikas izaugsme (vairāki pētījumi ir aprēķinājuši atvērto datu ekonomisko vērtību vairākos desmitos miljardu eiro gadā Eiropas Savienībā);
- līdzdalība un sadarbība;
- patstāvība;
- jauni vai uzlaboti produkti un pakalpojumi (Google Translate izmanto milzīgo Eiropas Savienības dokumentu apjomu, kas parādās visās Eiropas valodās, lai apmācītu tulkošanas algoritmus, tādējādi uzlabojot to kvalitāti);
- inovācijas;
- uzlabota valsts dienestu efektivitāte (piemēram, tas var palielināt valdības efektivitāti. Nīderlandes Izglītības ministrija visus ar izglītību saistītos datus ir publicējusi tiešsaistē atkārtotai izmantošanai. Kopš tā laika saņemto jautājumu skaits ir samazinājies, samazinoties darba slodzei un izmaksām, un atlikušie jautājumi ir vieglāk atbildami arī ierēdņiem, jo ir skaidrs, kur var atrast attiecīgos datus).

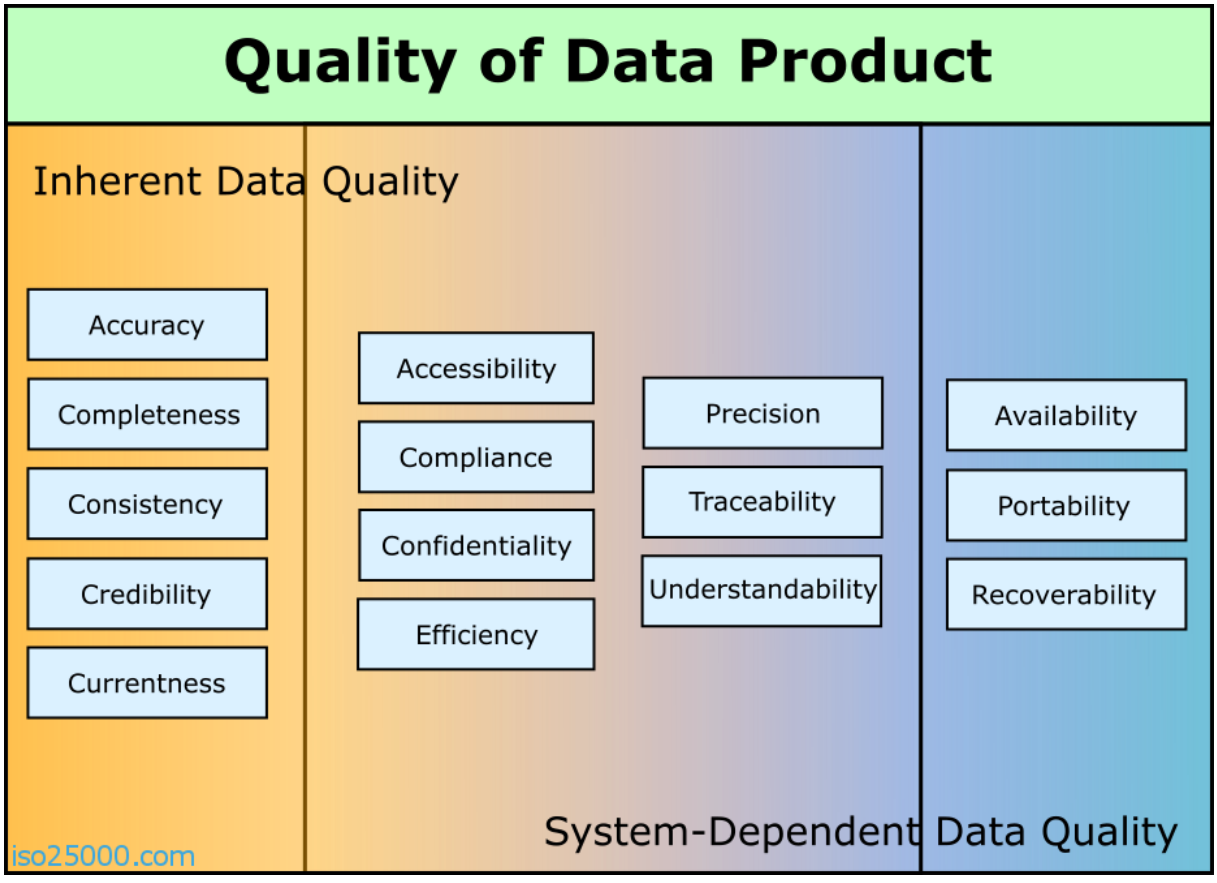
2. VADLĪNIJAS, STANDARTI DATU KVALITĀTES NODROŠINĀŠANAI

2.1. ISO standarts

ISO/IEC 25012 datu kvalitātes modelis [9], kas ietverts ISO/IEC 25000 standartu sērijā ir pamats, uz kura balstās datu produktu kvalitātes novērtēšanas sistēma. Datu kvalitātes modelī ir noteiktas galvenās datu kvalitātes īpašības, kas jāņem vērā, novērtējot paredzētā datu produkta īpašības. Datu produkta kvalitāti var uzskatīt par pakāpi, kādā dati atbilst produkta īpašnieka organizācijas noteiktajām prasībām. Šīs prasības ir tās, kas ir atspoguļotas datu kvalitātes modelī, izmantojot dimensijas.

Datu kvalitātes modelis, kas definēts standartā ISO/IEC 25012, sastāv no 15 dimensijām, kas parādīti attēlā 2.1.1.. Arī šeit dimensijas iedalītas divās kategorijās – iekšējā datu kvalitāte un datu kvalitāte, kas atkarīga no sistēmas. Iekšējā datu kvalitāte tiešā veidā attiecas uz pašiem datiem, it īpaši uz datu vērtībām, metadatiem. Datu kvalitāte, kas atkarīga no sistēmas attiecas uz pakāpi, kādā datu kvalitāte ir sasniegta datorsistēmā un ir atkarīgs no datora aparatūras, datorsistēmas programmatūras un citas programmatūras.

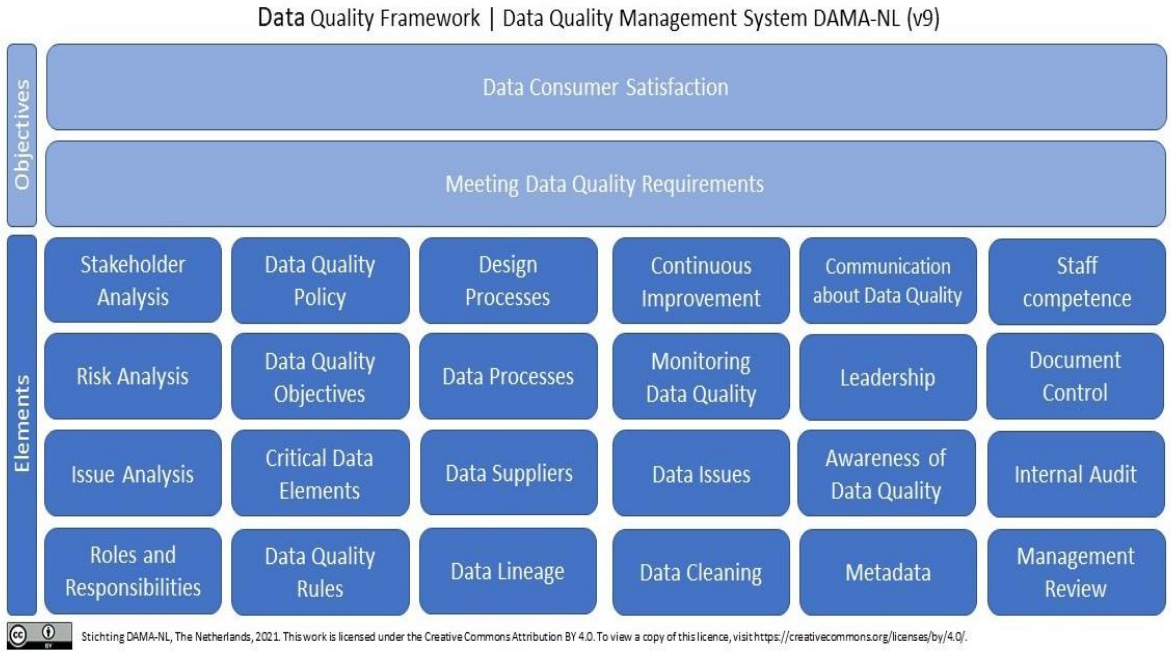
Attēls Nr. 2.1.1. ***ISO/IEC 25012 datu kvalitātes dimensijas [9]***



2.2. DAMA NL datu kvalitātes ietvars

Pētnieku grupa DAMA NL izstrādājusi datu kvalitātes vadības sistēmu jeb datu kvalitātes ietvaru [3]. Tajā iekļauti daudz dažādu procesu, piemēram, datu kvalitātes monitorēšana, nepārtraukti uzlabojumi, datu kvalitātes noteikumi, kuru galvenie mērķi ir datu lietotāju apmierinātība un datu kvalitātes prasību izpilde (skat. att. 2.2.1.).

Attēls Nr. 2.2.1. ***DAMA NL datu kvalitātes ietvars [3]***



Pētnieku grupa ir veikusi pētījumu par datu kvalitātes dimensiju definīcijām. Tā ir apkopojusi definīcijas no dažādiem avotiem un salīdzinājusi tās savā starpā. Pārbaudītas arī definīcijas, salīdzinot ar kritērijiem, kas atvasināti no konceptu un definīciju standarta: ISO 704 (nosaka pamatprincipus un metodes terminoloģiju sagatavošanai un apkopošanai gan standartizācijas ietvaros, gan ārpus tās [8]).

Pateicoties DAMA NL publicētajam literatūras klāstam, kā, piemēram, "Datu kvalitātes dimensijas, "Kā izvēlēties piemērotākās datu kvalitātes dimensijas", "Datu kvalitātes dimensiju vārdnīca", kas pieejams [3], datu publicētājam tiek sniegtas plašas iespējas izstrādāt savām vajadzībām pielāgotas datu kvalitātes vadlīnijas, kas būtiski uzlabos datu kvalitāti.

2.3. The Open Data Institute vadlīnijas

Atvērto datu institūts (The Open Data Institute jeb The ODI) izstrādājis vadlīnijas atvērtam standartam, kas paredzēts datiem [21]. Atvērtie datu standarti ir atkārtoti izmantojamas vienošanās, kas cilvēkiem un organizācijām atvieglo labākas kvalitātes datu publicēšanu, piekļuvi tiem, koplietošanu un izmantošanu. To vizualizācija apskatāma attēlā 2.3.1..

Attēls Nr. 2.3.1. ***Standartizēšanas iespēju veidi [21]***

Izmantojot atvērtos datu standartus iespējams koplietot vārdu krājumus jeb **koplietojamas vārdnīcas** un kopēju valodu, izmantojot kopīgus modeļus, atribūtus un definīcijas, ar tādiem izmantojamiem rezultātiem, kā reģistri, taksonomijas, vārdu krājumi, tādējādi iespējams apmainīties ar datiem (**datu apmaiņa**) organizācijās, uzņēmumos un sistēmās starp tām. Ar standartu palīdzību arī var sniegt **vadlīnijas**, norādījumus un ieteikumus augstākas kvalitātes datu veidošanā un publicēšanā izmantojot ceļvežus, protokolus, modeļus [21]. Attēlā 2.3.1. apskatāmie standartizēšanas iespēju veidi plašāk aprakstīti tabulā 2.3.2.

Tabula Nr. 2.3.2. ***Atvērtie standarti [21]***

Grupējuma tips	Standartizējamās vienības	Skaidrojums
Koplietojama vārdnīca	**Vārdi**	Vienošanās par vārdu definīcijām, ko mēs izmantojam, piemēram, "skola", "tiesa" vai "līgums"
Koplietojama vārdnīca	**Modeļi**	Saskaņots veids, kā domāt par to datu tipiem, ar kuriem mēs vēlamies apmainīties, piemēram, koncepcijas karte vai entītijas attiecību modelis
Koplietojama vārdnīca	**Taksonomijas**	Kā klasificējam un aprakstām lietas, piemēram, kodus un kategorijas
Koplietojama vārdnīca	**Identifikatori**	Identifikatori, kurus mēs piekrītam izmantot, lai palīdzētu aprakstīt datus par cilvēkiem, vietu, lietām un jēdzieniem, piemēram, uzņēmuma numurs
Datu apmaiņa	**Failu formāti**	Failu formāti, ko izmantojam informācijas glabāšanai un publicēšanai, piemēram, JSON, CSV vai XML
Datu apmaiņa	**Shēmas**	Noteikumi par to, kā izmantot faila formātu, lai apmainītos ar datiem, piemēram, kā izmantot CSV failu, lai koplietotu datus
Datu apmaiņa	**Datu tipi**	Kā mēs aprakstām dažādus datu tipus, piemēram, datuma, laika un valūtas formātus
Datu apmaiņa	**Datu pārsūtīšana**	Metodes, ar kurām mēs apmaināmies ar informāciju vai sniedzam piekļuvi tai, piemēram, API, lai atrastu datus
Vadlīnijas	**Kā mēs ievācam, apkopojam datus**	Kā mēs mērām vērtības un novērojam datus, piemēram, kā reģistrēt temperatūras nolasījumu vai izmērīt vērtību eksperimentā
Vadlīnijas	**Mērvienības**	Mērvienības un mērījumi, ko mēs izmantojam, lai palīdzētu mums apkopot datus, piemēram, centimetri utt.
Vadlīnijas	**Prakses kodekss**	Labā prakse, ieteikumi un citi norādījumi, kas atbalsta labu datu praksi

Atvērto datu institūta vadlīnijas vēl plašāk izklāstītas organizācijas veidotajā vietnē [20], kur tās var skatīt jebkurš, lai, iespējams, ar vadlīniju palīdzību izveidotu noteiktu standartu, kas būtu derīgs sev nepieciešamajā veidā, it īpaši, ja tā izstrādē ņemtu vērā arī iepriekš minēto DAMA NL datu kvalitātes ietvaru.

3. POPULĀRĀKO DATU KVALITĀTES PRASĪBU IDENTIFICĒŠANA

3.1. Literatūras datu dimensiju apkopojums

Lai būtu iespējams veikt darba praktisko daļu – datu kopu analīzi, nepieciešams apkopot literatūras avotos minētās datu kvalitātes dimensijas (skat. tabulu 3.1.1.). Novērojams, ka dažos darbos ir izvēlētas tikai dažas dimensijas, kas darbu autoriem šķitušas visplašāk izmantojamās vai svarīgākās, kā arī citos darbos šis dimensiju uzskaitījums ir daudz plašāks, tātad atkal tiek pierādīts, ka nav īsti iespējams visos gadījumos piemērot vienu noteiktu datu kvalitātes dimensiju kopumu.

Tabula Nr. 3.1.1. ***Datu kvalitātes dimensijas autores pētītajā literatūrā***

Pieminētās dimensijas	**Avots**
Precizitāte, pilnīgums, lasāmība, pieejamība, konsekvence	[16]
Nepretrunīgums, pilnīgums, unikalitāte, precizitāte, derīgums un savlaicīgums	[6]
Pieejamība, pilnīgums, precizitāte, reprezentācija	[18]
Pilnīgums, saprotamība, precizitāte, aktualitāte, izsekojamība, derīgums, atbilstība	[23]
Pieejamības dimensijas (pieejamība, licencēšana, drošība, savstarpēja sasaiste, veiktspēja), iekšējās dimensijas (precizitāte, konsekvence, kodolīgums), uzticamības dimensijas (reputācija, ticamība, pārbaudāmība, objektivitāte), kontekstuālās dimensijas (pilnīgums, datu daudzums, atbilstība), reprezentatīvās dimensijas (reprezentācijas kodolīgums, saprotamība, interpretējamība, daudzpusība)	[25]
Pilnīgums (tukši lauki, tukšo vērtību apstrāde), precizitāte (atbilstība rakstam, atbilstība pieļaujamu vērtību sarakstam), unikalitāte, kontekstuālā atbilstība	[1]
Pieejamība, atbilstošs datu apjoms, ticamība, reprezentācija, manipulācijas vieglums, pareizība, intrepretējamība, objektivitāte, atbilstība, reputācija, drošība, savlaicīgums, saprotamība, pievienotā vērtība	[11]

Pieejamība, izmantojamība, uzticmība, atbilstība, prezentācijas kvalitāte	[2]

Pētot literatūru, autore secina, ka joprojām tiek izmantotas datu kvalitātes definētās dimensijas no pagājušā gadsimta beigu un šī gadsimta sākuma literatūras (Redman, Wang, Strong, Pipino, Lee), kas liecina par šo pētnieku nenoliedzami lielo nozīmi datu kvalitātes pētīšanas jomā, un, ka pēdējo gadu pētījumi, jaunumi jomā balstās uz šo pētnieku darbiem. Taču, autores prāt, pašlaik visplašāk apkopotais un aprakstītais datu kvalitātes dimensiju, to kategoriju saraksts ir pētnieku grupas DAMA NL pētījumā [4]. Tajā kopumā ir 60 dimensijas, kur to definīcijas un aprakstus iespējams salīdzināt no dažādiem avotiem.

3.2. Pārbaudāmo dimensiju saraksts

Balstoties uz veikto pārskatu par pētījumos un standartos visbiežāk lietotajām datu kvalitātes dimensijām, tika izveidots optimālo datu kvalitātes dimensiju, metriku saraksts (skat. tabulu 3.2.1.), ko izmantot šī darba praktiskajā daļā, arī dimensiju metriku pamatā izmantojot literatūru [4, 25, 5]. Veidojot sarakstu ņemts vērā plašs literatūras apkopojums [15] par visbiežāk aprakstītajām problēmām un trūkumiem saistībā ar atvērto datu kvalitāti (skat. 1. pielikumu.), kura izstrādē ņemti vērā ap 60 dažādiem pētījumiem un rakstiem.

Sarakstā arī iekļauta informācija par to, vai attiecīgo pārbaudi iespējams pārbaudīt ar datu kvalitātes analīzes rīku vai SQL-balstītu analīzi. Prakse rāda, ka dažreiz SQL-balstīta analīze ļauj veikt padziļinātāko datu kvalitātes analīzi, ļaujot definēt sarežģītākas pārbaudes, kur datu kvalitātes rīkiem tipiski ir ierobežots predefinēto pārbaužu klāsts, ar ko vairums gadījumos nepietiek.

Tabula Nr.3.2.1. ***Pārbaudāmo dimensiju, to metriku saraksts***

Dimensija	**Metrika**	**Apraksts, piemērs**	**Potenciālais defekts**	**Iespējams pārbaudīt ar rīku/SQL analīzi?**
Pilnīgums	Tukšas vērtības	Cik lielā mērā dati ir pilnīgi, nav trūkstošu vērtību	Tukšas vērtības/rindas	Ir
Pilnīgums	Tukšu vērtību pieļaujamība	Vai kāda noteikta atribūta vērtības drīkst būt tukšas	Tukšas vērtības neatļautās vietās	Nav
Precizitāte	Atbilstība rakstam	Tālruņa numurs (8 cipari), e-pasta adrese (satur '@', '.lv' utt.), tīmekļvietne (satur 'www.', '.lv' utt.)	Vērtība neatbilst rakstam	Ir
Atbilstība	Datu atbilstība starp atribūtiem	Salīdzināti tiek divi atribūti ('sākuma datums' un 'beigu datums')	Datu nesaderība/nesaskaņotība starp atribūtiem datu kopas ietvaros	Ir
Atbilstība	Datu atbilstība starp vairākām datu kopām	Vairākas datu kopas tiek salīdzinātas savā starpā	Datu nesaderība/nesaskaņotība starp vairākām datu kopām	Ir
Reprezentācija	Tukšo vērtību reprezentācija	Vai tukša vērtība tiek atzīmēta vienotā veidā, piemēram, '-', 'nav' u.c	Tukšas vērtības tiek aprakstītas dažādos veidos	Ir

Konsekvence	Atbilstība pieļaujamu vērtību sarakstam	Vai vērtība pieder nodefinētajam pieļaujamo vērtību sarakstam (‘jā’ vai ‘nē’, ‘vīrietis’ vai ‘sieviete’)	Vērtība ir ārpus pieļaujamo vērtību sarakstam	Ir
Unikalitāte	Datu elementa unikalitāte vērtību sarakstā	Vai, piemēram, kārtas numurs ir unikāls un tas neatkārtojas (izņemot gadījumus, kad tas ir atļauts)	Nepastāv unikalitāte	Ir
Saprotamība	Paskaidrojoša informācija par atribūtiem un vērtībām	Gadījumos, kad atribūtu nosaukumi ir lietotājam nesaprotami, piemēram, izmantoti saīsinājumi, to nozīmei jābūt paskaidrotai	Lietotājam nesaprotami atribūti, to vērtības	Nav
Saprotamība	Cilvēkam saprotami (ne-mašīnlasāmi) metadati	Metadati ir aprakstīti, lai būtu lietotājam saprotami	Lietotājam nesaprotami metadati	Nav
Interpretēja mība	Mašīnlasāmi metadati	Metadati ir tādā formātā, kādā to var nolasīt dators	Nepieejami mašinlasāmi metadati	Nav
Savlaicīgums	Datu atjaunināšanas biežums	Tiek norādīts datu atjaunināšanas biežums piemēram, katru dienu, reizi mēnesī	Nav norādīts biežums	Nav
Savlaicīgums	Pēdējo izmaiņu datums	Tiek norādīts pēdējo izmaiņu datums	Nav norādīts izmaiņu datums	Nav

Savlaicīgums	Datu atjaunināšanas biežuma ievērošana	Dati tiek regulāri atjaunināti atbilstoši noteiktajam atjaunošanas biežumam	Dati netiek atjaunoti atbilstoši noteiktajam biežumam	Nav
Pieejamība	Iespējams piekļūt datiem	Datu kopu iespējams apskatīt un brīvi lejupielādēt	Dati netiek ielādēti, nevar apskatīt un lejupielādēt	Nav
Uzticamība	Licences esamība	Visām atvērtajām datu kopām jābūt atbilstošai licencei	Nav norādīta vai atbilstoša licence	Nav

3.3. Datu kvalitātes analīzes rīki

Tā kā darba ietvaros paredzēts arī salīdzināt bezmaksas datu kvalitātes rīku analīzes salīdzināšanu ar SQL-balstītu analīzi, nepieciešams īsumā aprakstīt autores kursa darbā paveikto analīzi [1].

Tabulā 3.3.1. apskatāms salīdzinājums, vai rīki ļauj izpildīt pārbaudes no definētā saraksta, iekļaujot arī, vai šajā darbā pievienotu pārbaudi būtu iespējams veikt rīkos:

- ‘+’ - ļauj,
- ‘-’ - neļauj,
- ‘+/-’ - nosacīti ļauj (nav konkrēta funkcija, taču pārbaudi iespējams veikt pašam lietotājam veidojot regulāro izteiksmi(regex), kas, visticamāk, ar IT jomu nesaistītam cilvēkam, varētu radīt problēmas).

Tabula Nr. 3.3.1. ***Nodefinēto pārbaužu saraksta izpildes iespējamība datu kvalitātes rīkos Data Cleaner un OpenRefine***

Metrika/Rīka nosaukums	**Data Cleaner**	**OpenRefine**
Tukšas vērtības	+	+
Tukšu vērtību reprezentācija	-	-

Atbilstība rakstam	+/-	+/-
Unikalitāte	+	+
Atbilstība vērtību sarakstam	+	+/-
Datu atbilstība starp vairākās datu kopām	+	-
Datu atbilstība starp atribūtiem	+	-

4. LATVIJAS ATVĒRTO DATU KOPU ANALĪZE

Veicot Latvijas Atvērto datu kopu analīzi [10] tika izmantots autores definētais pārbaužu saraksts (skat. tabulu 3.2.1.). Darbs tika veikts izmantojot Microsoft SQL Server Management Studio 2018 un veidojot SQL vaicājumus, kuru daži piemēri apskatāmi 2. pielikumā.

Datu kopas izvēlētas no kategorijas "Valsts pārvalde", lai datu kopas pēc to analīzes spētu efektīvāk salīdzināt un identificēt problēmas. Kā arī, valsts pārvaldes dati ir nozīmīgs atvērto datu tips, kas, kā jau teorētiskajā daļā minēts, nodrošina valsts pārvaldes caurspīdīgumu.

Datu kvalitātes pētīšanā un analīzē autore izmanto intuitīvo pieeju [24], ņemot vērā, ka atribūtu izvēle balstīta uz zinātnieku pieredzi un autores intuitīvo izpratni.

4.1. VISR datu kopa

Valsts informācijas sistēmu, to pārziņu un augstāko iestāžu saraksts (VISR). Darbojās saskaņā ar 2005. gada 2. augusta Ministru kabineta noteikumiem "Valsts informācijas sistēmu reģistrācijas noteikumi". Pārtrauca darbību ar 01.01.2020. līdz ar VIRSIS iedarbināšanu. Tās izdevējs ir Vides aizsardzības un reģionālās attīstības ministrija.

Datu kopā ir 38 atribūti, katram atribūtam ir 245 vērtības, tātad kopumā datu kopā ir 9310 datu lauki. Rezultāti aplūkojami tabulā Nr. 4.1.1.

Tabula Nr. 4.1.1. ***Rezultātu tabula***

Metrika	**Rezultāts**	**Kopējais defektu saturošais atribūtu/ vērtību skaits**
Tukšas vērtības	Ir tukšas vērtības	28/421 (74%/4.5%)
Tukšu vērtību pieļaujamība	Nav noteikts, to iespējams tikai noprast lietotājam subjektīvi	-
Atbilstība rakstam	Daļēja atbilstība	4/97 (10.5%/1%)
Datu atbilstība starp atribūtiem	Ir atbilstība (ja atribūta 'aktīva vai slēgta sistēma' vērtība ir 'slēgts', tad atribūta 'slēgšanas	-

	datums' vērtība ir aizpildīta	
Datu atbilstība starp vairākām datu kopām	Ir atbilstība, taču dažas datu rindas nav izdevies apvienot	2/25 (5%/0.3%)
Tukšu vērtību reprezentācija	Daļēji ievērota (tukšas vērtības atzīmētas ar '-', defektu gadījumos '"-"', '.')	2/2 (5%/0.02%)
Atbilstība pieļaujamu vērtību sarakstam	Ir atbilstība	-
Datu elementa unikalitāte vērtību sarakstā	Ir	-
Paskaidrojoša informācija par atribūtiem un vērtībām	Nav (tikai mašīnlasāmā formātā)	-
Cilvēkam saprotami (ne-mašīnlasāmi) metadati	Ir	-
Mašīnlasāmi metadati	Ir (JSON fails)	-
Datu atjaunināšanas biežums	Ir	-
Pēdējo izmaiņu datums	Ir	-
Datu atjaunināšanas biežuma ievērošana	Nevar noteikt (datu kopa vairs netiek atjaunināta, jo reģistrs ir pārtraucis darbību)	-
Iespējams piekļūt datiem	Ir	-
Licences esamība	Ir (CC0 1.0)	-

4.1.1. Papildus defektu identificēšana

Nav korekti lietotas atdalītājzīmes (skat. attēlu 4.1.1.1.). Dažu atribūtu vērtības ir korekti noformētas, taču dažiem klāt ir liekas atdalītājzīmes('Turētājs Iestādes Kods'), kas būtiski apgrūtina datu kvalitātes pārbaudi, tādēļ, pirms datu kopas analīzes veikšanas, autorei nācās no atdalītājzīmēm atbrīvoties manuāli.

Attēls Nr. 4.1.1.1. ***Nekorekts atdalītājzīmju lietojums***

Pārzinis Augstākā iestāde	Turētājs Iestādes kods
Vides aizsardzības un reģionālās attīstības ministrija	;90001733697;
Finanšu ministrija	;90000014724;
Finanšu ministrija	;90000014724;
Finanšu ministrija	;90000014724;
Finanšu ministrija	;90000014724;
Labklājības ministrija	;90001669496;
Finanšu ministrija	;90000014724;
Tieslietu ministrija	;90001037264;
Tieslietu ministrija	;90001037264;
Tieslietu ministrija	;90001037264;

Vēl viens papildus defekts, kas tika atrasts ir gramatikas kļūdas atribūtu nosaukumos – ‘Atbildīgās perosnas uzvārds’, ‘Atbildīgās perosnas tālrunis’, ‘Aktrīvā vai slēgtā sistēma’ (skat. attēlu 4.1.1.2.).

Attēls Nr. 4.1.1.2. ***Gramatikas kļūdas***

Atbildīgās perosnas uzvārds	Atbildīgās perosnas tālrunis
Guds	67026525
Minkevičs	67095602
Minkevičs	67095602
Minkevičs	67095602
Minkevičs	67095602
Ziediņa	67013638
Minkevičs	67095602

4.2. Patiesā labuma guvēji

Datu kopā iekļautas ziņas par tiesību subjektu aktuālajiem patiesajiem labuma guvējiem (PLG) – fiziskajām personām. Tiesiskais pamats fizisko personu datu publiskošanai – likuma "Par Latvijas Republikas Uzņēmumu reģistru" 4.10 panta divpadsmitā daļa (spēkā ar 01.11.2020). Datu izdevējs ir Latvijas Republikas Uzņēmumu reģistrs.

Datu kopā ir 10 atribūti, katram atribūtam ir 188 075 vērtības, tātad kopumā datu kopā ir 1 880 750 datu lauki.

Tabula Nr. 4.2.1. ***Rezultātu tabula***

Metrika	**Rezultāts**	**Kopējais defektu saturošais atribūtu/ vērtību skaits**
Tukšas vērtības	Ir tukšas vērtības	3/177 (30%/0.009%)
Tukšu vērtību pieļaujamība	Ir noteiktas (ja personas kods ir 'NULL', tad personas dzimšanas diena nevar būt 'NULL' un otrādi)	0
Atbilstība rakstam	Atbilst rakstam (personas kods ir korektā formā)	0
Datu atbilstība starp atribūtiem	Līdzīgi kā punktā par tukšu vērtību pieļaujamību, tiek salīdzināti atribūti par personas kodu un dzimšanas datumu	0
Datu atbilstība starp vairākām datu kopām	Ir atbilstība, taču dažas datu rindas nav izdevies apvienot	1/4547 (10%/0.24%)
Tukšu vērtību reprezentācija	Ievērota (tukšas vērtības atzīmētas ar '-')	0
Atbilstība pieļaujamu vērtību sarakstam	Ir atbilstība (Valstspiederība un	0

	Dzīvesvietas valsts atbilst ISO 3166 Alpha-2 kodam)	
Datu elementa unikalitāte vērtību sarakstā	Ir	0
Paskaidrojoša informācija par atribūtiem un vērtībām	Ir pieejams apraksts	-
Cilvēkam saprotami (ne-mašīnlasāmi) metadati	Ir	-
Mašīnlasāmi metadati	Nav	-
Datu atjaunināšanas biežums	Ir	-
Pēdējo izmaiņu datums	Nav	-
Datu atjaunināšanas biežuma ievērošana	Nevar noteikt, skatoties portālā, taču pašā datu kopā ir atribūts ‘last_modified_at’ pēc kura redzams, ka datu kopa tiek atjaunota	-
Iespējams piekļūt datiem	Ir	-
Licences esamība	Ir (CC0 1.0)	-

4.2.1. Papildus defektu identificēšana

Pētot tukšu vērtību reprezentāciju šajā datu kopā, tika pamanīts, ka tukšas vērtības tiek piešķirtas personu uzvārdiem (skat. attēlu 4.2.1.1.), kas var šķist mulsinoši. Parasti tiek uzskatīts, ka personas uzvārds nedrīkstētu būt ar tukšu vērtību, bet, tā kā sīkāks skaidrojums nav sniegts par uzvārdu neesamību, visticamāk, ārzemju tautības personām, tad nevar nonākt pie secinājuma, vai to var uzskatīt par defektu vai nē.

Attēls Nr. 4.2.1.1. ***Iespējamie uzvārdu ievades defekti***

	id	legal_entity_registration_number	forename	surname
1	66091	40203211227	Kuldeep	-
2	28535	40203036356	Jatinder Kumar	-
3	39798	42103084822	Sukhbir Kaur	-
4	39832	40203169663	Hardeep Singh	-
5	39834	40203169663	Pritpal Singh	-
6	39835	40203169663	Gursewak Singh	-
7	257481	50203223961	Sohan Veer Singh	-
8	265068	40203230177	Faris Mohamed	-
9	268132	42103091063	Akash Thomas	-
10	273182	40203202366	Sooraj Santhosh	-
11	100184	40103180560	Vishvajeet	-
12	168926	40203047186	Sandeep Kumar	-
13	287804	40203241467	Sooraj Santhosh	-
14	287994	40203250629	Alan Varghese	-
15	84078	40003626897	Abhishek Roshan	-

4.3. Pasūtītāju datu grupa

Datu kopā "Pasūtītāju datu grupa" iekļauta informācija par Elektronisko iepirkumu reģistrētajiem pasūtītājiem. Datu izdevējs ir Valsts reģionālās attīstības aģentūra.

Datu kopā ir 8 atribūti, katram atribūtam ir 1519 vērtības, tātad kopumā datu kopā ir 15 190 datu lauki.

Tabula Nr. 4.3.1. ***Rezultātu tabula***

Metrika	**Rezultāts**	**Kopējais defektu saturošais atribūtu/ vērtību skaits**
Tukšas vērtības	Ir tukšas vērtības	0 (pēc autores subjektīvā viedokļa, jo pie būtiskākajiem datu atribūtiem nav tukšu vērtību)
Tukšu vērtību pieļaujamība	Nav noteiktas, taču subjektīvi to var noteikt	-

Atbilstība rakstam	Atbilst rakstam (reģistrācijas datums atbilst datuma rakstam, kā arī reģistrācijas numuri ir korekti)	0
Datu atbilstība starp atribūtiem	Nepastāv iespējamas atbilstības	-
Datu atbilstība starp vairākām datu kopām	Ir atbilstība, taču dažas datu rindas nav izdevies apvienot	1/141 (12.5%/0.9%)
Tukšu vērtību reprezentācija	Nav	-
Atbilstība pieļaujamu vērtību sarakstam	Ir atbilstība (Valstspiederība un Dzīvesvietas valsts atbilst ISO 3166 Alpha-2 kodam)	1/1 (12.5%/0.006%)
Datu elementa unikalitāte vērtību sarakstā	Ir	0
Paskaidrojoša informācija par atribūtiem un vērtībām	Nav	-
Cilvēkam saprotami (ne-mašīnlasāmi) metadati	Nav	-
Mašīnlasāmi metadati	Nav	-
Datu atjaunināšanas biežums	Ir	-
Pēdējo izmaiņu datums	Nav	-
Datu atjaunināšanas biežuma ievērošana	Nevar noteikt, jo nav norādīts pēdējo izmaiņu datums	-
Iespējams piekļūt datiem	Ir	-
Licences esamība	Ir (CC0 1.0)	-

4.3.1. Papildus defektu identificēšana

Arī šajā datu kopā novērojamas problēmas ar atdalītājzīmju lietošanu (skat. attēlu 4.3.1.1.), no kurām autore atbrīvojās manuāli, lai spētu analizēt datus.

Attēls Nr. 4.3.1.1. ***Nekorektas atdalītājzīmes***

	Nr	Organizacija
1	="50003888871	="'Aerones' SIA"
2	="40000000001	="Agency for Support for BEREC (BEREC Office)"
3	="90000065754	="Aglonas novada dome"
4	="90009339571	="Aglonas novada domes Sociālās aprūpes centrs 'Ag...
5	="90000074456	="Aizkraukles arodvidusskola"
6	="48703000438	="'Aizkraukles KUK', SIA"
7	="90000074812	="Aizkraukles novada pašvaldība"
8	="90009620223	="Aizkraukles Profesionālā vidusskola"
9	="40003255337	="'Aizkraukles slimnīca', SIA"
10	="42103001430	="'AIZPUTES KOMUNĀLAIS UZŅĒMUMS' SIA"
11	="42103002652	="'Aizputes nami' SIA"
12	="90000031743	="Aizputes novada dome"
13	="90000031584	="Aizputes novada domes Cīravas pagasta pārvalde"
14	="90000031565	="Aizputes novada domes Kazdangas pagasta pārval...
15	="90000022632	="Aizsardzības ministrija"
16	="40203235757	="Akciju sabiedrība 'Ventas osta'"

REZULTĀTI

Vides aizsardzības un reģionālās attīstības ministrijas dati nav īsti augstu novērtējami, jo datos ir diezgan daudz dažādu problēmu. Lai arī datu kopa satur daudz mazāk datus, kā, piemēram, datu kopa 'Patiesā labuma guvēji', tās datu kvalitātes līmenis ir acīmredzami zemāks. Trūkst lietotājam saprotama informācija par datu kopas atribūtiem. Nepieciešamības gadījumā informāciju iespējams apskatīt pievienotajā JSON failā, taču nevar sagaidīt, ka visi lietotāji to pratīs paveikt. Trūkst informācija par to, vai kādi datu lauki drīkst būt tukši, lietotājam atliek pašam to subjektīvi izprast.

Novērojams, ka Latvijas Republikas Uzņēmuma reģistra dati ir ļoti augstā līmenī, taču ir arī daži mīnusi – atklātais iespējamais papildus defekts par uzvārdu nenorādīšanu rosina neizpratni, vai tas ir patiešām defekts vai arī tā tas ir paredzēts- tātad datu kvalitātes saprotamības dimensiju būtu iespējams uzlabot pie esošo datu apraksta pievienojot vēl kādu papildus skaidrojošu informāciju. Šāda informācija arī būtu noderīga par tukšo lauku pieļaujamību. Vēlams arī būtu iekļaut mašīnlasāmus metadatus.

Savukārt, Valsts reģionālās attīstības aģentūras dati ir augstā līmenī, taču jāņem vērā, ka nebija iespējams veikt tik daudz pārbaužu kā ar pārējām datu kopām. Tomēr datu kvalitāti pazemina metadatu un pēdējo izmaiņu datuma trūkums.

Gan VISR, gan Pasūtītāju datu grupas datu kopām novērojami papildus defekti saistībā ar atdalītājzīmju lietojumu, kas abos gadījumos radīja grūtības uzsākot datu analīzes veikšanu. Tas pavisam noteikti nozīmē, ka lietotāji, kas neprastu atbrīvoties no liekajām rakstu zīmēm, datus nespētu lietot, kas, pēc autores domām, ir liela problēma, jo datiem jābūt pieejamiem un izmantojamiem.

Latvijas Republikas Uzņēmuma reģistrs un Valsts reģionālās attīstības aģentūras dati pie datu kopām nenorāda datu izmaiņu datumu, kas ir pretrunā ar iepriekš darbā minētajiem Ministru Kabineta noteikumiem Nr. 445., ka iestādes pienākums ir saskaņā ar iestādes norādīto datu atjaunošanas biežuma klasifikatoru aktualizēt atvērto datu portālā ievietotos datus un nodrošināt to atbilstību metadatiem [12]. Tātad konkrētos noteikumus Valsts reģionālās attīstības aģentūra neizpilda arī saistībā ar metadatu nenorādīšanu.

Autores prāt, intuitīvā pieeja datu kvalitātes pētīšanai ir visai efektīva, jo, apvienojot zinātnieku pētījumus un lietotāja intuitīvu izpratni, var panākt plašāk pārbaudītu datu kvalitāti. Analīzi sākot ar pētnieku definētām dimensijām un beidzot ar lietotāja izpratni par attiecīgajā kontekstā papildus nepieciešamām pārbaudēm.

Kopumā atkal tiek pierādīts, ka viens datu kvalitātes pārbaužu kopums neder visām datu kopām, jo katrā ir savas individuālās nianses un defekti, kas ir tiešā veidā atkarīgs no datu kopas veidotājiem un publicētājiem, un ar kurām datu kvalitātes rīki ne vienmēr tiek galā.

Pēc autores domām, ar SQL-balstītu analīzi ir iespējams veikt padziļinātākas analīzes, ko nav iespējams veikt ar visiem datu kvalitātes rīkiem, taču, par galalietotājam piemērotāko izvēli, kas varētu nebūt saistīts ar IT jomu, apmierinošs variants būtu izmantot rīkus, nevis veikt SQL analīzi. It īpaši, ja nav nepieciešams veikt sarežģītas pārbaudes, jo ar rīku palīdzību ļoti ātri un ērti iespējams sasniegt rezultātu.

SECINĀJUMI

Veicot literatūras izpēti, apkopojot informāciju un veicot datu kopu analīzi, tika izdarīti vairāki secinājumi.

Lai arī datu kvalitātes dimensiju daudzums ir liels, tās galvenokārt netiek radītas no jauna, bet izmantotas datu kvalitātes nozīmīgāko pētnieku veidotās dimensijas pagājušā gadsimta beigās un šī gadsimta sākumā.

Eksistē vairāki datu kvalitātes standarti un vadlīnijas, ko būtu visai vienkārši izmantot datu publicētājiem, lai paaugstinātu datu kvalitāti. Pēc autores domām, būtu noderīgi Latvijas Atvērto datu portālā ieviest vadlīnijas arī datu kvalitātes nodrošināšanai, lai dati būtu pēc iespējas ar augstāku kvalitāti, tātad, vērtīgāki un noderīgāki to izmantošanai, taču nepadarot tās pārāk detalizētas, lai nerastos liekas problēmas datu veidošanā un uzturēšanā.

Datu kvalitāte atšķiras atkarībā no datu publicētāja, taču, lai pavisam noteikti par to pārliecinātos būtu nepieciešams analizēt vēl citas datu kopas no šiem pašiem publicētājiem, lai varētu pavisam pārliecināti izdarīt šādus secinājumus.

Datu kvalitātes dimensijas, ar ko visvairāk radās problēmu pētītājās datu kopās:

- saprotamība (skaidrojošie dati netiek sniegti pietiekošā daudzumā, it īpaši, vai datu kopā drīkst būt tukši lauki);
- savlaicīgums (netiek publicēts pēdējo izmaiņu datums, tātad nevar pārliecināties par to, vai dati tik tiešām tiek atjaunoti atbilstoši norādītajam atjaunošanas biežumam).

Par galalietotājam piemērotāko izvēli, kas varētu nebūt saistīts ar IT jomu, būtu ieteicams izmantot datu kvalitātes rīkus, nevis SQL-balstītu analīzi, kas varētu radīt lietojamības problēmas. Taču citos gadījumos noteikti SQL analīze būtu labākais variants, lai lietotājs spētu padziļināti veikt dažādas analīzes.

Viens datu kvalitātes definēto pārbaužu kopums nav derīgs pilnīgi visām datu kopām, jo katrā ir savas individuālās nianses un defekti, kas ir atkarīgs no datu kopas veidotājiem un publicētājiem.

IZMANTOTĀ LITERATŪRA

1. Baldere I., Atvērto datu kvalitāte: piemērotāko risinājumu meklējumos, 2021.
2. Cai L., Zhu Y., The Challenges of Data Quality and Data Quality Assessment in the Big Data Era, 2015 [tiešsaiste]. Pieejams: https://datascience.codata.org/article/10.5334/dsj-2015-002/
3. DAMA NL, Data Quality Management System [tiešsaiste]. Pieejams: http://www.dama-nl.org/data_quality/
4. DAMA NL, Dimensions of Data Quality (DDQ), 2020 [tiešsaiste]. Pieejams: http://www.dama-nl.org/wp-content/uploads/2020/09/DDQ-Dimensions-of-Data-Quality-Research-Paper-version-1.2-d.d.-3-Sept-2020.pdf
5. DAMA NL, Factsheet Data Quality Rule, 2021 [tiešsaiste]. Pieejams: https://www.dama-nl.org/wp-content/uploads/2021/05/Factsheet-Data-Quality-Rules.pdf
6. Dama UK, The Six Primary Dimensions For Data Quality Assessment, 2013 [tiešsaiste]. Pieejams: https://silo.tips/download/the-six-primary-dimensions-for-data-quality-assessment
7. Gualo F., Rodriguez M., Verdugo J., Cavallero I., Piattini M., Data Quality Certification using ISO/IEC 25012: Industrial Experiences, 2021 [tiešsaiste]. Pieejams: https://www.researchgate.net/publication/349546494_Data_Quality_Certification_using_ISOIEC_25012_Industrial_Experiences
8. ISO 704:2009 Terminology work — Principles and methods [tiešsaiste]. Pieejams: https://www.iso.org/standard/38109.html
9. ISO/IEC 25012 [tiešsaiste]. Pieejams:]https://iso25000.com/index.php/en/iso-25000-standards/iso-25012
10. Latvijas Atvērto datu portāls [tiešsaiste]. Pieejams: https://www.data.gov.lv/lv
11. Leo L. Pipino, Yang W. Lee, and Richard Y. Wang, Data Quality Assessment, 2002 [tiešsaiste]. Pieejams: https://www.researchgate.net/publication/2881159_Data_Quality_Assessment
12. Ministru Kabineta noteikumi Nr. 445 "Kārtība, kādā iestādes ievieto informāciju internetā", 2020 [tiešsaiste]. Pieejams: https://likumi.lv/ta/id/316109-kartiba-kada-iestades-ievieto-informaciju-interneta

13. Open Definition 2.1, Version 2.1, 2015 [tiešsaiste]. Pieejams: http://opendefinition.org/od/2.1/en/
14. Open Knowledge Foundation, Why Open Data? [tiešsaiste]. Pieejams: https://opendatahandbook.org/guide/en/why-open-data/
15. Rashid M., Torchiano M., Adnan M. N., Mahmud K. R., Paul B., Open Data Quality Challenges: A Survey, 2020 [tiešsaiste]. Pieejams: https://www.researchgate.net/publication/339787251_Open_Data_Quality_Challenges_A_Survey
16. Rula A., Maurino A., Batini C., Data Quality Issues in Linked Open Data, 2016 [tiešsaiste]. Pieejams: https://www.researchgate.net/publication/303602822_Data_Quality_Issues_in_Linked_Open_Data
17. Sabiedriskās politikas centrs Providus, Pašvaldības un atvērtie dati, 2012 [tiešsaiste]. Pieejams: http://providus.lv/upload_file/Projekti/Laba%20parvaldiba/faktu_lapa_final.pdf
18. Staskiewicz A., Hvam L., Haug A., Data Quality Issues When Quantifying Costs of Complexity, 2020 [tiešsaiste]. Pieejams: https://www.researchgate.net/publication/348471705_Data_Quality_Issues_When_Quantifying_Costs_of_Complexity
19. Tayi G., Examining data quality, 1998 [tiešsaiste]. Pieejams: https://www.researchgate.net/publication/27297579_Examining_Data_Quality
20. The ODI, Open Standards for Data [tiešsaiste]. Pieejams: https://standards.theodi.org/
21. The ODI, Types of open standards for data [tiešsaiste]. Pieejams: https://standards.theodi.org/introduction/types-of-open-standards-for-data/
22. Vadlīnijas "atvērts pēc noklusējuma" principa ieviešanai [tiešsaiste]. Pieejams: https://data.gov.lv/sites/default/files/2019-12/Atverts_pec_noklusejuma_1_0.pdf
23. Vetrò A., Canova L., Torchiano M., Orozco Minotas C., Iemma R., Morando F., Open data quality measurement framework: Definition and application to Open Government Data, 2016 [tiešsaiste]. Pieejams: https://www.researchgate.net/publication/295394863_Open_data_quality_measurement_framework_Definition_and_application_to_Open_Government_Data

24. Wang R.Y., Strong D.M., Beyond Accuracy: What Data Quality Means to Data Consumers, 1996 [tiešsaiste]. Pieejams: http://courses.washington.edu/geog482/resource/14_Beyond_Accuracy.pdf

25. Zaveri A., Rula A., Maurinob A., Pietrobonc R., Lehmanna J., Auer S., Quality Assessment for Linked Data: A Survey, 2016 [tiešsaiste]. http://www.semantic-web-journal.net/system/files/swj556.pdf

PIELIKUMI

1. pielikums

Pētījuma problēmas ar iespējamo risinājumu [15]

Theme Issues	Research Problems	Possible Solution (PS)	References
Quality in Use (QU)			
Legal	(QU1) Restrictive access	(PS1) Data generalization: Release data complying with the definition of openness.	11,12,13
	(QU2) Personal data and Privacy	(PS1) Data generalization: Collect the concerns of re-users and modify licenses if the barriers are too constraining.	14,15,16,17
	(QU3) Lack of uniform license	(PS2) Governance choices: Use of uniform licenses.	17,18,19,20
	(QU4) Framework concerning data in general	(PS2) Governance choices: Find a balance in the intervention of personal data and privacy policies.	11,13,15,21
	(QU5) Stacking of rights	(PS3) Cultural Shift: Favor a cultural shift in the administrations.	18,20,22,23,24
Economic	(QU6) The cost of opening data	(PS4) Realistic approach to financial risks: Assessing the costs of not opening; Share part of the costs with other Open Data platforms.	18, 19, 20, 25
	(QU7) Cost of data production	(PS4) Realistic approach to financial risks: Encourage stakeholders; promote networking between stakeholders; participate in clusters that sustain incubation of companies grounding their business model on Open Data.	16,19,20,25
Internal Quality(IQ)			
Data structure	(IQ1) Data available inheterogeneous formats	(PS5) Use of non-proprietary and machine processable format: Publish datasets in various formats.	26,27,28,29
	(IQ2) Machine readability	(PS5) Use of non-proprietary and machine processable format.	30,31,32
	(IQ3) Lack of data completeness	(PS6) Data quality assessment: Possible solution is to use a data quality standard constant to guide quality assessment process.	5,6,33,34,35
	(IQ4) Lack of data accuracy	(PS6) Data quality assessment: to identify semantic and syntactic error in data.	5,28,29,36,37
	(IQ5) Lack of validation techniques	(PS6) Data quality assessment: domain independent automated validation techniques.	28,39,40

Pētījuma problēmas ar iespējamo risinājumu[15]

External Quality(EQ)			
Publishing platform	(EQ1) Incomplete meta-data	(PS7) Metadata evaluation: Gather metadata needs from re-users; implement mechanisms to trace the provenance and timeless check.	41,42,43,44,45
	(EQ2) Risk regarding interoperability, scalability and usability	(PS8) Generic schema: Provide support using generic schema.	44,46,47,48,49,50
	(EQ3) Lack of discoverability	(PS8) Generic schema: Use of metadata evaluation to support searchability.	42,50,51,53,54
	(EQ4) Risk regarding automated data extraction	(PS8) Generic schema: Use of machine processable format.	42,47,48,50
	(EQ5) Too much vocabularies	(PS9) Data cataloging: Using controlled vocabularies.	20,51
	(EQ6) Frequency of updating data	(PS9) Data cataloging: Preserving data provenance information	44,45,55
	(EQ7) Lack of categorization facilities	(PS9) Data cataloging: Use of widely adopted data cataloging methodologies.	4,45
	(EQ8) Irregularity of deployed platform	(PS10) Publishing standards: Use of standard guideline.	43,48,51,56
	(EQ9) Lack of standard	(PS10) Publishing standards: Participate in the harmonization of various Open Data catalogs to establish accountability.	47,51,57,58
	(EQ10) Unstructured metadata	(PS10) Publishing standards: Using standards guideline.	42,43,51,57
	(EQ11) Suitability of data for release	(PS10) Publishing standards: publish data using data quality standard.	43,48,47,52
	(EQ12) Lack of API	(PS11) Use of API: Publishing platform support an API capable of reporting various metadata.	51,53,56

SQL vaicājumi datu kopu analīzes veikšanai

```
-- tukšu vērtību atrašana
SELECT * FROM dbo.visr
    WHERE Vai_sistēma_satur_personu_datus IS NULL
    OR Vai_sistēma_satur_personu_datus = '';

-- defektīvu tukšu vērtību aprakstošo vērtību atrašana
SELECT * FROM dbo.visr
    WHERE Datu_apmaiņas_protokoli IS NOT NULL
    AND LEN(Datu_apmaiņas_protokoli) < 4
    AND (Datu_apmaiņas_protokoli) NOT LIKE '-' ;

-- Atbilstība noteiktam vērtību sarakstam
SELECT * FROM dbo.visr
    WHERE Vai_sistēma_satur_personu_datus NOT LIKE 'Satur'
    AND Vai_sistēma_satur_personu_datus NOT LIKE 'Nesatur';

-- unikalitātes pētīšana
SELECT DISTINCT Turētājs_iestādes_kods;

-- atbilstība rakstam
SELECT * FROM dbo.visr
    WHERE Turētājs_adrese  NOT LIKE '%_LV__%';

-- atbilstība rakstam
SELECT Pārzinis_Augstākā_iestāde FROM dbo.visr
    WHERE Pārzinis_Augstākā_iestāde NOT LIKE '%ministrija%';

-- kontekstuālā analīze starp vairākām datu kopām
SELECT Pārzinis_iestādes_kods FROM dbo.visr
except
SELECT registrationNumber
FROM dbo.ppi
except
SELECT registrationNumber
FROM dbo.ppi_2
except
SELECT regcode
FROM dbo.register;
```

Bakalaura darbs „ Latvijas atvērto datu kvalitātes analīze un identificēto kvalitātes trūkumu klasificēšana” izstrādāts LU Datorikas fakultātē.

Ar savu parakstu apliecinu, ka pētījums veikts patstāvīgi, izmantoti tikai tajā norādītie informācijas avoti.

Autors: Ieva Baldere 31.05.2021.

Rekomendēju darbu aizstāvēšanai

Vadītāja: docente, Dr. sc. comp. Anastasija Ņikiforova 31.05.2021.

Recenzents: asociētā profesore Dr. sc. comp. Lelde Lāce

Darbs iesniegts Datorikas fakultātē 31.05.2021.

Dekāna pilnvarotā persona: vecākā metodiķe Ārija Sproģe

Darbs aizstāvēts bakalaura gala pārbaudījuma komisijas sēdē

__.06.2021. prot. Nr. ___.

Komisijas sekretārs: profesors Dr. habil. sc. comp. Juris Borzovs