LATVIJAS UNIVERSITĀTE
DATORIKAS FAKULTĀTE

LATVIJAS ATVĒRTO DATU PORTĀLA UN TĀ DATU KOPU KVALITĀTES ANALĪZE

BAKALAURA DARBS

Autors: Beāte Beizaka
Studenta apliecības Nr.: bb17016
Darba vadītāja: docente, Dr. dat. Anastasija Ņikiforova

RĪGA 2021

ANOTĀCIJA

Darbā tiek apskatīts un analizēts Latvijas atvērto datu portāls, tajā esošās datu kopas.

Darba mērķis ir izpētīt Latvijas atvērto datu portālu un tā populārākās datu kopu kvalitāti. Darbā tiek apskatīts atvērto datu jēdziens, to kvalitātes analīzes un noteikšanas paņēmieni, kā arī sīkāk izpētīta informācija par Latvijas atvērto datu portālu, tā kvalitātes noteikšanu. Tāpat tiek veikts pētījums, kurā tiek noskaidrotas būtiskākās datu kvalitātes problēmas Latvijas atvērto datu portālā un tā saturošajās datu kopās.

Atslēgvārdi: dati, atvērtie dati, Latvijas atvērto datu portāls, datu kopa, datu kvalitātes novērtēšana.

ABSTRACT

ANALYSIS OF THE LATVIAN OPEN DATA PORTAL AND QUALITY OF ITS DATA SETS

This work looks at and analyzes Latvia's open data portal and its existing data sets.

The aim of the work is to explore the quality of Latvia's open data portals and its most popular data sets. The work looks at the concept of open data, the methods of analysis and determination of their quality, as well as more detailed information on Latvia's open data portal, its quality determination. A study is also being carried out in which the most important data quality problems are identified in the Latvian Open Data Portal and its data sets.

Keywords: data, open data, Latvia open data portal, data set, assessment of data quality.

SATURS

APZĪMĒJUMU SARAKSTS

LADP – Latvijas atvērto datu portāls

CSV – ir norobežots teksta fails, kas vērtību atdalīšanai izmanto komatu.

XLSX – faila formāts, kas tiek izmantoti Microsoft Excel, izklājlapu lietojumprogrammā, kas izmanto tabulas, lai kārtotu, analizētu un saglabātu datus.

WMS – ir standarta protokols ģeotelpisko datu (karšu) apmaiņai internetā, ko veido izmantojot ģeogrāfiskās informācijas sistēmas (ĢIS).

XLS – izklājlapas fails, ko izveidojis Microsoft Excel vai cita izklājlapu programma, piemēram, OpenOffice Calc vai Apple Numbers. Tajā ir viena vai vairākas darbalapas, kurās dati tiek glabāti un parādīti tabulu formātā.

Odata – atvērts protokols, kas ļauj izveidot un patērēt vaicājamus.

PDF – ir faila formāts, kurā dokumentus, ieskaitot teksta formatēšanu un attēlus, parāda nemainīgā veidā, kas nav atkarīgas no lietojumprogrammas, aparatūras vai operētājsistēmas.

SHP – ģeotelpisku vektoru datu formāts ģeogrāfiskās informācijas sistēmas (ĢIS) programmatūrai.

DOCX – ir faila formāts, Microsoft Word dokuments, kas parasti satur tekstu.

XML – Paplašināmā iezīmēšanas valoda.

VAS – Valsts akciju sabiedrība.

HTML – standarta iezīmēšanas valoda dokumentiem, kas paredzēti parādīšanai tīmekļa pārlūkprogrammā.

ODS – operatīvā datu krātuve.

TXT – teksta faila paplašinājums, ko izmanto dažādi teksta redaktori.

JSON – datu apmaiņas formāts, balstīts uz teksta formātu.

ZIP – arhīva faila formāts, kas atbalsta bezzudumu datu saspiešanu.

API – lietojumprogrammas saskarne ir iepriekš definētu klašu, procedūru, funkciju, struktūru un konstanšu kopums, kas tiek pasniegts kā pielikums, kuru iespējams izmantot ārējiem programmatūras produktiem.

URL – Vienotais resursu vietrādis ir adrese, kas pārlūkprogrammā norāda tīmekļa lappuses vai cita interneta resursa atrašanās vietu.

SQL – vaicājumu valoda, kas paredzēta datu manipulēšanai relāciju datu bāžu pārvaldības sistēmā

IEVADS

Mūsdienās arvien lielāka loma ir dažādām tehnoloģijām, kas nepārtraukti attīstās un, to izmantošanas dēļ, tiek radīts liels datu un informācijas apjoms elektroniskā veidā. Bieži vien dati tiek izmantoti tikai vienu reizi un glabāti tā, ka neviens cits tiem nevar piekļūt, tā neizmantojot to pilno potenciālu un dažreiz iegūstot vienus un tos pašus datu vairākkārt. Šī iemesla dēļ dati, kas potenciāli ir spējīgi, kādam indivīdam vai uzņēmumam, ir jāpadara publiski jeb atkal-izmantojami. Datu atkalizmantošana ir ļoti svarīga, lai veicinātu dažādu jomu attīstību, tāpēc daudzās pasaules valstīs, arī Latvijā, ir izveidots atvērto datu portāls, kurā var publicēt datus un kāds cits šos datus pēc tam var izmantot, lai uz to pamata iegūtu informāciju vai radītu jaunus produktus un pakalpojumus. Kaut gan datu publicēšanas ideja nav tik jauna, tā vēl aizvien netiek pilnvērtīgi izmantota, jo bieži cilvēki un/vai iestādes nevēlas publicēt datus, vai arī uzskata, ka tas nav nepieciešams vai izdevīgi.

Tāpat ir zināms, ka jau publicētiem atvērtajiem datiem bieži ir raksturīgas kvalitātes problēmas, jo, ievērojot likumu, dati nedrīkst norādīt uz konkrēto indivīdu vai citādi būt identificējami, tāpēc tie tiek dažādos veidos mainīti un apstrādāti, bieži zaudējot gan datu kvalitāti, gan jēgu. Bieži šiem datiem ir arī citas kvalitātes problēmas, par kurām iespējams pat nezina datu uzkrājēji, jo šīs problēmas ir pašās datu bāzēs vai informācijas sistēmās. Šīs datu problēmas traucē izmantot datus atkārtoti, vai arī noved pie nekorektiem rezultātiem, ja dati tiek izmantoti, nezinot par datu kvalitātes problēmām tajos. Tāpēc ir svarīgi veidot un publicēt kvalitatīvus datus, kurus pēc tam ir iespēja un jēga atkalizmantot.

2021. gada maijā Latvija atvērto datu portālā ir publicētas 472 datu kopas 14 dažādās kategorijās un tās ir iegūtas no 85 dažādiem datu kopu publicētājiem. Pasaulē Latvijas atvērto datu portāls ieņem diezgan augstu pozīciju, kaut gan pēdējā gada laikā pozīcija ir kritusies, tomēr LADP datu kvalitātes ziņā Eiropā ieņem 19. vietu [11]. Tas nozīmē, kaut gan pārsvarā dati ir kvalitatīvi un izmantojami, tomēr ir novērojamas nepilnības, kuras būtu vēlams labot. Bet jāņem vērā, ka šie rādītāji ir tik augsti arī tāpēc, ka ar datu kvalitāti šādos pētījumos parasti nav domāta pašu datu kvalitāte, kā arī tiek virspusēji skatīts, kādi dati ir datu kopā, dažreiz lielāku uzmanību veltot portāla funkcionalitātei un saturam.

Darba autore analizē Latvijas atvērto datu portāla kvalitāti un tajā pieejamo datu kopu kvalitāti, vispirms veicot Latvijas atvērto datu portāla kvalitātes analīzei, tai sekojot datu kvalitātes

analīzei, datu kopas izvēloties no visām kategorijām pēc to popularitātes principa, tādējādi nodrošinot iespēju veikt objektīvāku analīzi un secinājumus.

Darba **mērķis**: izanalizēt Latvijas atvērto datu portāla un tajā pieejamo datu kopu kvalitāti, un veikt secinājumus par Latvijas atvērto datu portāla kopējo kvalitāti un Latvijas atvērto datu kvalitāti.

Darba **uzdevumi**:

1) aplūkot "atvērto datu" jēdzienu un ieguvumus, ko tie ir potenciāli spējīgi sniegt sabiedrībai un ekonomikai;
2) aplūkot "datu kvalitātes" jēdzienu un datu kvalitātes analīzes tehnikas;
3) aplūkot "atvērto datu kvalitātes" jēdzienu;
4) noteikt populārākas datu kvalitātes problēmas, uz kuru esamību pārbaudīt Latvijas atvērtos datus;
5) izpētīt, kā nosaka kvalitāti atvērto datu portālam;
6) izpētīt Latvijas atvērto datu portālu un tā kvalitāti;
7) izanalizēt Latvijas atvērto datu portāla atvērto datu kopu kvalitāti, izmantojot atvērto datu kvalitātes noteicošos raksturlielumus, kā arī SQL balstītu kvalitātes analīzi.

Darba struktūra: darbs sastāv no 2 daļām – teorētiskās daļas un praktiskās daļas. Teorētiskajā daļā darba autore pēta informāciju par datiem, atvērtiem datiem un to kvalitātes noteikšanas iespējām, kā arī pēta Latvijas atvērto datu portālu un tajā esošās datu kopas Praktiskajā daļā tiek veikta Latvijas atvērto datu portāla un tā saturošo datu kopu analīze, secinot, kādas nepilnības piemīt šiem datiem un vai tie ir pietiekami kvalitatīvi, lai tos varētu atkalizmantot.

1. DATI UN ATVĒRTIE DATI

Atvērto datu jēdziens nav jauns, jau sen pastāvot idejai, ka dati ir jāatver un jāpadara atkalizmantojami, lai sabiedrība un uzņēmumi var gūt labumu no jau iegūtiem datiem un nav jātērē resursi, lai atkal iegūtu datus, ko kāds cits jau ir uzkrājis.

Bet tomēr atvērto datu jēdziena formalizēšana un izmantošana ir relatīvi jauna, jo tikai pēdējos gados tiek īpaši mudināts datus atvērt un padarīt tos koplietojamus, gan veidojot dažādus atvērto datu portālus, gan mudinot organizācijas un indivīdus, kas uzkrāj datus, tos publicēt, lai sabiedrība un citas organizācijas varētu gūt labumu no šiem datiem.

1.1. Dati un informācijas jēdzieni

Dati ir neapstrādātā veidā ierakstīts notikuma momentuzņēmums. Dati ir tikai fakti vai skaitļi-informācijas biti, bet ne pati informācija [14]. Tas nozīmē, ka dati ir mums visapkārt, bet, nezinot to kontekstu, šie dati ir vienkārši skaitļi vai tekstuālas virknes bez jēgas.

Lai dati iegūtu jēgu un tie būtu izmantojami, datiem jāpievieno kādi aprakstoši faktori. Ja datus apstrādā, interpretē, organizē, strukturē vai uzrāda tā, lai tie būtu jēgpilni vai noderīgi, tos sauc par informāciju. [13]. Tas nozīmē, ka, lai datus varētu izmantot, tiem ir jāpievieno konteksts jeb apraksts, ko tie apzīmē, tādējādi iegūstot informāciju, ko var arī tālāk izmantot un apstrādāt.

1.2. Atvērto datu definīcija

Pastāv dažādas atvērto datu definīcijas un skaidrojumi, jo šis jēdziens nav viennozīmīgs, bet Eiropas atvērto datu portālā tie ir definēti, kā dati, kuriem ikviens var piekļūt, tos izmantot un koplietot [2]. Tas nozīmē, lai datus uzskatītu par atvērtiem, datiem, jāvar piekļūt ērti un salīdzinoši vienkārši jebkuram indivīdam, kā arī šos datus jāvar izmantot, t.i. tiem nevar būt autortiesības, kā arī tiem ir jābūt viegli koplietojamiem.

Tāpat ir nodefinēti 8 atvērto datu principi, kuriem jābūt ievērotiem, lai datus varētu klasificēt kā atvērtus datus [7]. Atbilstoši šiem principiem, datiem jābūt:

1. pilnīgiem – visiem datiem ir jābūt pieejamiem un pēc iespējas pilnīgiem (likumu robežās). Ir jābūt pieejamiem arī metadatiem, kas definē un apraksta publicēto datu kopu, skaidrojot arī kā

konkrēto lauku, vērtības tika iegūtas vai aprēķinātas, sniedzot datu lietotājiem iespēju pētīt datus pēc iespējas dziļākā detalizācijas līmenī;

2. primārajiem – publicētājiem datiem ir pilnībā jāatbilst datu avotam, no kura tie tika iegūti;

3. laicīgiem – datiem ir jābūt pieejamiem pēc iespējas ātrāk, nodrošinot pēc iespējas laicīgāku datu nodošanu galalietotājam;

4. pieejamiem – datiem ir jābūt pieejamiem pēc iespējas plašākam lietotāju lokam visiem iespējamajiem nolūkiem;

5. mašīnlasāmiem – datiem ir jābūt strukturētiem, lai tie varētu tikt automatizēti apstrādāti. Piemēram, dati, kas ir sniegti .pdf formātā ir grūti apstrādājami un netiek uzskatīti par atvērtajiem datiem. Datiem ir jābūt pieejamiem plaši lietojama mašīnlasāma formātā (piemēram, .csv, .xls utt.);

6. nediskriminējošiem – datiem ir jābūt pieejamiem visiem bez ierobežojumiem, izslēdzot nepieciešamību reģistrēties to iegūšanai;

7. atvērtajā datu formātā – datiem ir jābūt pieejamiem tādā formātā, par kuru nevienam nav īpašas kontroles;

8. bez licences – uz datiem nedrīkst būt attiecinātām nekādām autortiesībām, preču zīmju vai patentu likumiem [12].

Ļoti svarīga atvērtiem datiem ir to anonimitāte. Datu anonimizācija ir privātas vai slepenas informācijas aizsardzības process, dzēšot vai šifrējot identifikatorus, kas savieno personu ar uzkrātajiem datiem [8]. Tas nozīmē, ka pirms datu atvēršanas, ir obligāti jāveic datu anonimizācija, kas jāveic, lai pēc šiem datiem nevarētu būt identificējams konkrēts indivīds vai indivīdu grupa. Bieži dati zaudē jēgu pēc to anonimizācijas, tāpēc tie ir jāpārveido uzmanīgi, lai tie nezaudētu jēgu un tiem nerastos datu kvalitātes problēmas.

Atvērtos datus veido gan valdības, padarot savus iekšienē uzkrātos datus atvērtus, gan uzņēmumi un privātpersonas, kas veido un apkopo kādus datus. Ja kādiem ievāktiem datiem ir jēga un tos potenciāli var atkalizmantot, tiek mudināts organizācijas un privātpersonas šos datus atvērt, lai sabiedrība un citas organizācijas varētu gūt sociālus, ekonomiskus un vides ieguvumus [2]. Protams, ir jāatver tikai tādi dati, no kuriem ir kāda jēga un no kuriem var kāds gūt labumu, necenšoties atvērt visus datus.

1.3. Ieguvumi no atvērtajiem datiem

Atvērtie dati sniedz daudzveidīgus ieguvumus, ieskaitot valsts pārvaldes efektīvāku darbību, ekonomisko izaugsmi privātajā sektorā un plašāku sociālo nodrošinājumu [2]. Tas nozīmē, ka datu atvēršana spēj palīdzēt ne tikai kādam indivīdam, tiem piemītot daudz lielākiem, globālākiem ieguvumiem- kā sākotnēji varētu šķist.

Eiropas atvērto datu portālā ir atzīmēti trīs galvenie ieguvumi no datu atvēršanas:

- darbības uzlabojumi – atvērtie dati var palīdzēt uzlabot sabiedriskos pakalpojumus, kā arī pakalpojumu sniegšanas efektivitāti, kas tiek panākts ar labāku datu apmaiņu starp sektoriem;
- ieguvumi ekonomikai - vieglāka piekļuve informācijai, saturam un zināšanām veicina inovatīvu pakalpojumu izstrādi un jaunu uzņēmējdarbības modeļu radīšanu;
- sociālā nodrošinājuma uzlabojumi — sabiedrībai ir pārredzamāka un vieglāk pieejamāka informācijas. Atvērtie dati veicina sadarbību, līdzdalību un sociālo inovācijas [2].

Tas nozīmē, ka, kaut gan varētu šķist, ka datu atvēršana tikai aizņem laiku, nesniedzot ievērojamus, būtiskus labumus, tas ir maldīgs priekšstats, datu atvēršana var palīdzēt ne tikai kādam indivīdam vai indivīdu grupai, bet lielākā mērogā pat valstij.

Atvērto datu popularizēšana un lietošana arī veicina jaunu darba vietu rašanos, jo tiek izveidotas jaunas darbavietas saistībā ar atvērtajiem datiem, kas nodrošina ekonomikas stimulēšanu. Ir novērots, ka Eiropā līdz 2020. gadam tika izveidotas 25 000 jaunas darbavietas, tieši saistītas ar atvērtajiem datiem [2].

Tiek lēsts, ka Eiropas valstīs kopā, atvērto datu izmantošana sabiedrisko pakalpojumu sniegšanas procesa efektivitāti uzlabošanā, kas nodrošina datu un informācijas apmaiņai starp sektoriem, kopējais izmaksu ietaupījums Eiropas valstīs līdz ir EUR 1,7 miljardi [2]. Tas nozīmē, ka tiek ietaupīts milzīgs naudas apjoms, kas citādāk būtu jāiztērē indivīdiem, organizācijām un uzņēmumiem, tātad var secināt, ka atvērtie dati arī palīdz ietaupīt ļoti lielas naudas summas.

Ekonomiskā pētījumā par atvērto datu priekšrocībām rezultātā tika iegūti sekojoši dati:

- atvērtie dati palīdz pieņemt labākus lēmumus, kas ir skaidrojams ar to, ka ir pieejams vairāk informācijas, kā vajadzētu darīt un kā nē, kā arī dažāda cita saistoša informācija, kas ļauj, izmantojot kritisko domāšanu, pieņemt labākus un efektīvākus datu virzītus lēmumus;
- 7000 dzīvības ir izglabātas, pateicoties atvērtajiem datiem, jo varēja iegūt ātrāku atbildes reakciju, t.i. atvērtie dati ir arī ļoti svarīgi slimnīcās un citās medicīniskās iestādēs, kas palīdz glābt

cilvēkus veselību un pat dzīvību, kā arī, izmantojot un analizējot atvērtos satiksmes datus, samazinājās bojāgājušie uz ceļiem par 5,5 %;

- 629 miljoni stundu ir ietaupītas, kas ir pielīdzināmas 27,9 miljardiem eiro, kas nozīmē, ka atvērtie dati ļāva ietaupīt ļoti daudz laika, kas savukārt nozīmē, ka arī tika ietaupīts liels daudzums naudas;
- 16% mazāk enerģijas izmantotas, t.i. izmantojot atvērtos datus, tiek ietaupīta arī enerģija [2].

1.4. Atvērtie valdības dati

Tāpat kā daudzi Eiropas atvērto datu portāli, arī Latvijas atvērto datu portāls ir veidots kā atvērtais valdības datu portāls, lai pamatā valdības un valsts iestādes varētu atvērt datus un dalīties ar tiem ar citām organizācijām un sabiedrību kopumā.

Atvērtie valdības dati ir filozofija, politiku kopums, kas veicina pārredzamību, pārskata atbildību un vērtību radīšanu, padarot valdības datus pieejamus visiem [18]. Valdības un citas publiskā sektora iestādes savā ikdienā rada un iegūst ļoti lielu datu apjomus, tādus datus kā pensiju un pabalstu maksājumu pārvaldības, nodokļu iekasēšanas datus, satiksmes datu reģistrēšana un izsniegšana oficiālie dokumenti [23]. Tas nozīmē, ka šis milzīgais datu daudzums ir kaut kur jāuzkrāj un arī jāpadara pieejams un atkalizmantojams, šos datus atverot, tāpēc arī tiek radīti atvērtie valdības datu portāli, kuros visi interesenti var datus apskatīt, lejupielādē kā arī atkalizmantot. Šāda valdības datu atvēršana un popularizēšana padara valsts iestāžu darbu caurredzamāku, kā arī veicina dažādu inovatīvu risinājumu izveidi.

2. ATVĒRTO DATU UN DATU PORTĀLU KVALITĀTE

Nodaļas mērķis ir izveidot zināšanu bāzi, uz kuru balstīt praktisko darba daļu.

Datu un īpaši atvērto datu kvalitātes jēdziens ir ļoti daudzpusīgs, ar kuru mēdz saprast gan metadatu, gan datu kopas satura jeb pašu datu korektumu, gan datu kopas formāta atbilstību vispārpieņemtajiem principiem. Tie ir aspekti, ar kuriem ir jārēķinās gan datu publicētājiem, gan portāla turētājiem, gan arī reizēm lietotājiem, jo datu publicēšana un atvēršana ne vienmēr nozīmē, ka dati ir kvalitatīvi un izmantojami, reizēm ir novērojamas datu kvalitātes pārvaldības problēmas, jo tās ir informācijas sistēmās vai datu bāzēs, par kurām datu publicētājs var pat nenojaust.

Lai datiem būtu jēga un tos varētu izmantot, svarīga ir to kvalitāte, savukārt atvērtie dati kļūst izmantojami un kvalitatīvi, kad cilvēks tos var saprast un mašīna jeb dators var tos apstrādāt [2]. Tas nozīmē, ka atvērtajiem datiem ir ļoti svarīgs **formāts**, kurā tos publicē.

Ne mazāk svarīgi datu kvalitātes noteikšanai ir **datu savlaicīgums un atjaunošanas biežums** [2], kas nosaka datu vērtību, t.i., cik aktuālie ir dati. Tas ir arī viens no būtiskajiem atvērto datu principiem.

Tradicionāli atvērto datu kvalitātes analīze sākas ar atvērto datu 8 principiem (aprakstīti 1.1. nodaļā), tas nozīmē, ka datiem jābūt pilnīgiem, primāriem, laicīgiem, pieejamiem, mašīnlasāmiem, nediskriminējošiem, atvērtajā datu formātā, kā arī bez licences. Tas parasti ir cieši saistīts ar atvērto datu kopu novērtēšanu pēc **5-zvaigžņu klasifikācijas**, tāpēc 2.2 nodaļā tiek aplūkoti tās pamati.

Lai noteiktu atvērto datu kvalitāti, ir jāatrod atbildes uz dažādiem jautājumiem, piemēram, kā dati tika apstrādāti, vai tie ir neapstrādāti vai kopsavilkuma formā, kāds ir to **formāts**, vai tas ir saderīgs ar citām datu kopām, vai datu kopai ir **pieejams apraksts, parametru apraksts** [2], kā arī uz citiem jautājumiem. Tas nozīmē, ka atvērtajiem datiem ne tikai vērtē pašu **datu kvalitāti**, bet arī kādā veidā tie ir iegūti, kad tie tika iegūti, vai tie ir saprotami, kā arī vai tie tiek regulāri atjaunoti, veids, kurā tie tiek sniegti.

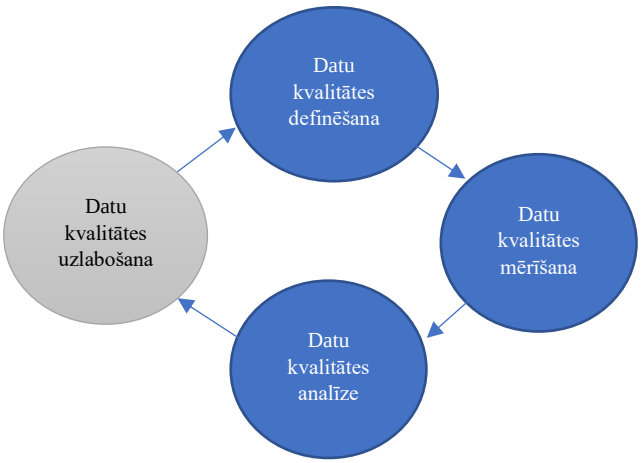
2.1. Datu kvalitātes novērtēšanas metodoloģijas

Lai varētu saprast un novērtēt datu kvalitāti, ir izveidotas dažādas datu kvalitātes noteikšanas metodoloģijas, kas mēdz atšķirties, tām esot dažādiem fokusiem. Šīs metodoloģijas palīdz iegūt padziļinātas zināšanas par datu kvalitāti un kā to noturēt tādā līmenī, lai tie ir kvalitatīvi un lietojami.

2.1.1 TDQM datu kvalitātes metodoloģija

TDQM (Total Data Quality Management) metodoloģija izmanto cilvēkresursus un kvantitatīvos resursus, lai uzlabotu ražojumus un pakalpojumus [20]. Metodoloģijas mērķis ir apgādāt informācijas lietotāju ar augstas kvalitātes datiem. Pamatprincips metodoloģijai ir procesu vienkāršošana un automatizēšana. Pamatkomponentes ir trīs: cilvēks, tehnoloģijas, procesi, kas iesaistās un mijiedarbojas, lai nodrošinātu metodoloģijas darbību. TDQM atbalsta datu bāzu migrāciju, veicina datu standartu izmantošanu un uzņēmējdarbības noteikumu izmantošanu, lai uzlabotu datubāzēs esošo datu kvalitāti [20]. TDQM veido cikls, kuru savukārt veido četras fāzes (sk. att. 3.1.):

- datu kvalitātes definēšana – datu kvalitātes raksturojumu formulēšana;
- datu kvalitātes mērīšana – analizējamo datu atlase un mērījumu veikšana;
- datu kvalitātes analīze – mērīšanas fāzē saņemto datu kvalitātes pārbaudes rezultātu analīze ar nolūku atrast datu kvalitātes problēmas un nepilnības;
- datu kvalitātes uzlabošana – datu kvalitātes uzlabošanas mehānisma izvēle un realizācija [20].



2.1.att. **TDQM cikliskās fāzes (autores [20] tulkojums)**

Kā arī visas ciklā iekļautās fāzes ir sistemātiski jāatkārto, lai

1. pārbaudītu, kā notiek datu kvalitātes uzlabošanas mehānisma realizācija
2. nodrošinātu jaunu vai modificētu datu kvalitātes pārbaudi, jo dati datu krātuvēs pastāvīgi mainās, kas savukārt var izraisīt jaunas datu kvalitātes problēmas vai arī izraisīt nepieciešamību jaunu datu kvalitātes prasību definēšanai [20].

2.1.2. ISO 8000-61 datu kvalitātes metodoloģija

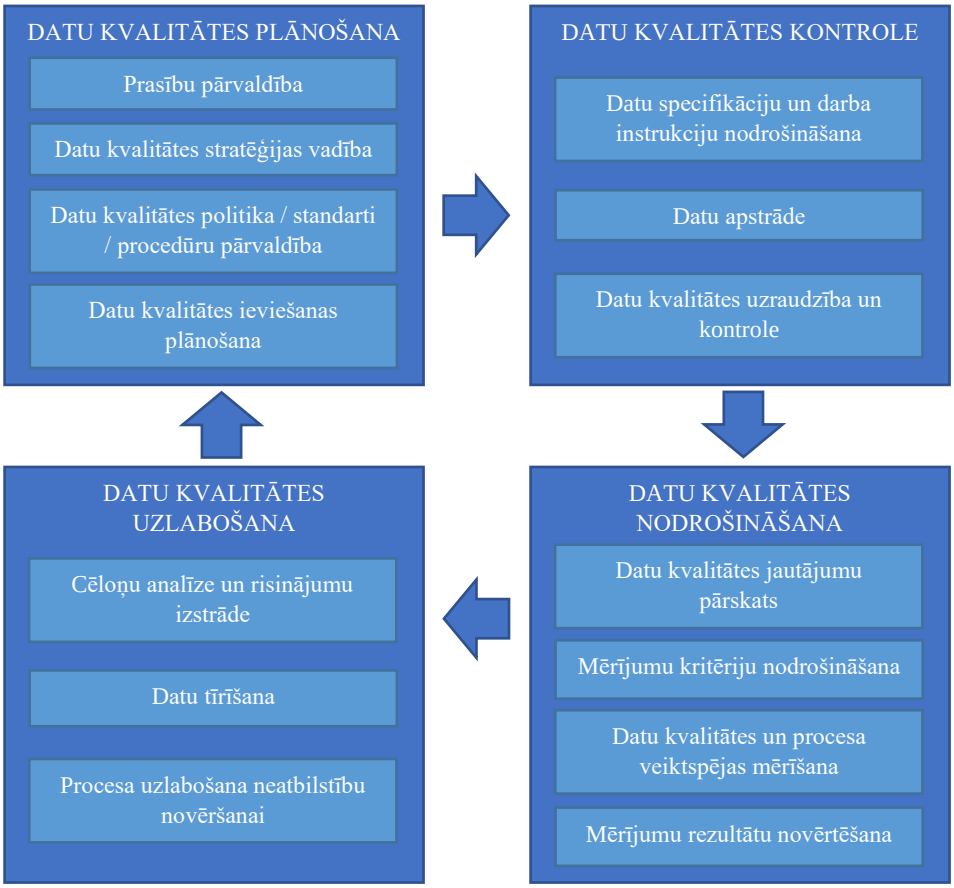
ISO 8000-61 nosaka datu kvalitātes pārvaldībai nepieciešamos procesus. Procesi tiek izmantoti, lai uzlabotu datu kvalitāti un novērtētu procesa spēju vai organizācijas gatavību datu kvalitātes pārvaldībai [10]. Tas nozīmē, ka šo metodoloģiju var izmantot ne tikai, lai uzlabotu datu kvalitāti, bet arī novērtētu, vai organizācijas ir gatava kvalitātes pārvaldībai. ISO 8000-61 darbības pamatā ir cikls, kas sastāv no 4 soļiem - plānošanas, izpildes, pārbaudes, darbības (angl. *Plan, Do, Check, Act*), kas definēts ISO 9001[10]. Pats ISO 8000-61 cikls sastāv no 4 daļām:

1) datu kvalitātes plānošana - vispārējo prasību, mērķu un plāna izstrāde un saskaņošana;
    a. prasību pārvaldība - dažādu ar organizāciju un datiem saistītu prasību noteikšana, definēšana;
    b. datu kvalitātes stratēģijas vadība - izveidot, novērtēt un uzlabot organizācijas datu kvalitātes stratēģiju;
    c. datu kvalitātes politika / standarti / procedūru pārvaldība - tādu politikas, standartu un procedūru izstrāde, kas atbalsta datu kvalitātes stratēģiju;
    d. datu kvalitātes ieviešanas plānošana - plāna izstrāde, kurā noteiktas lomas, pienākumi, secības noteikšana, finansēšana un tehnoloģijas, lai veiktu visas citas ar datu kvalitātes pārvaldību saistītas darbības.
2) datu kvalitātes kontrole - procesi, lai nodrošinātu, ka darbības rezultātā iegūtie dati atbilst prasībām;
    a. datu specifikāciju un darba instrukciju nodrošināšana - tiek noteiktas procesu prasības, kā arī procesa darbības;
    b. datu apstrāde - pārbaude, vai no procesiem iegūtie dati atbilst datu prasībām;
    c. datu kvalitātes uzraudzība un kontrole - identificēt gadījumus, kad datu apstrāde neatbilst prasībām un reaģēt uz tiem;
3) datu kvalitātes nodrošināšana - šajā procesā tiek novērtēti datu kvalitātes līmeņi un ar datu kvalitāti saistīto procesu veiktspēja;
    a. datu kvalitātes jautājumu pārskats - ziņoto datu kvalitātes jautājumu novērtējums, lai saprastu to būtību un apjomu;
    b. mērījumu kritēriju nodrošināšana - mērījumu metrikas un metožu izstrāde, lai atbalstītu datu kvalitātes mērīšanu;

c. datu kvalitātes un procesa veiktspējas mērīšana - resursu piesaistīšana datu kvalitātes līmeņa mērīšanai un mērīšanas procesa novērtēšana;
d. mērījumu rezultātu novērtēšana - analizēt datu kvalitātes mērīšanas rezultātus un novērtēt sliktas datu kvalitātes un mērīšanas procesa ietekmi;

4) datu kvalitātes uzlabošana - nodrošināt ilgtspējīgus datu kvalitātes uzlabojumus;
a. cēloņu analīze un risinājumu izstrāde - datu kvalitātes problēmu pamatcēloņus identificēšana un risinājumus, lai novērstu atkārtotu parādīšanos izstrāde;
b. datu tīrīšana - datu kvalitātes problēmu novēršana, izmantojot automatizētus rīkus un / vai cilvēku iejaukšanos;
c. procesa uzlabošana neatbilstību novēršanai - risinājumu ieviešana, lai novērstu datu atkārtotu parādīšanos un apstrādātu neatbilstības [26].

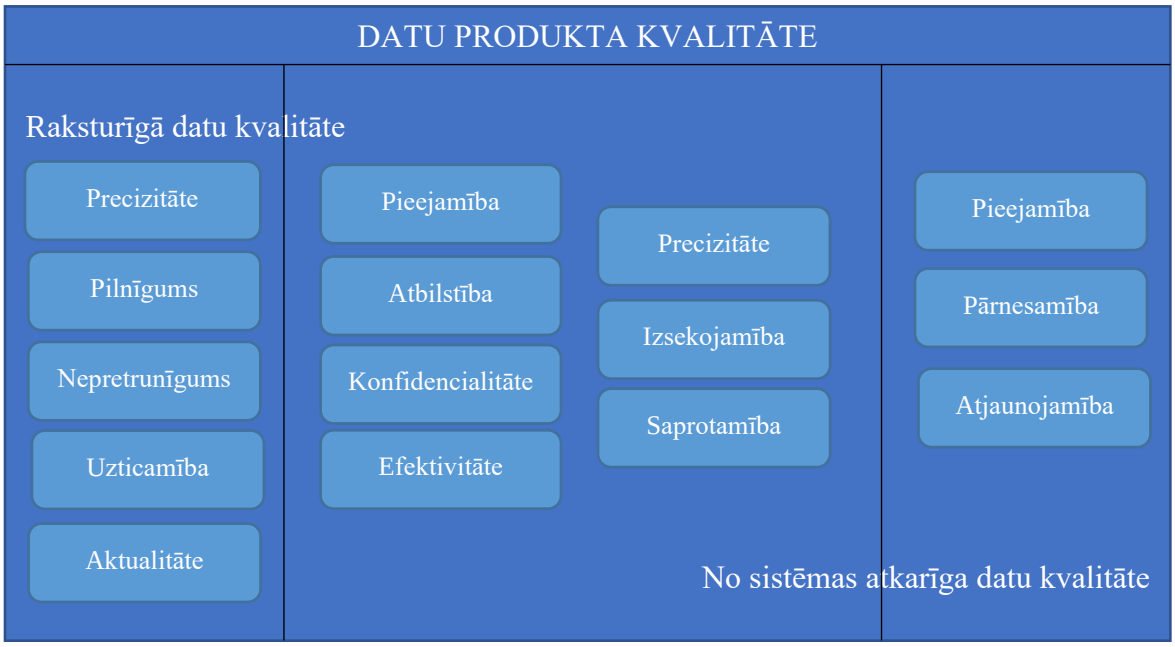
Visas ciklā iekļautās fāzes ir jāatkārto sistemātiski, lai nodrošinātu datu kvalitāti un to nepārtrauktu uzraudzīšanu.



2.2.att. **ISO 8000-61 cikliskās fāzes (autores [26] tulkojums)**

2.1.3. ISO 25012 datu kvalitātes metodoloģija

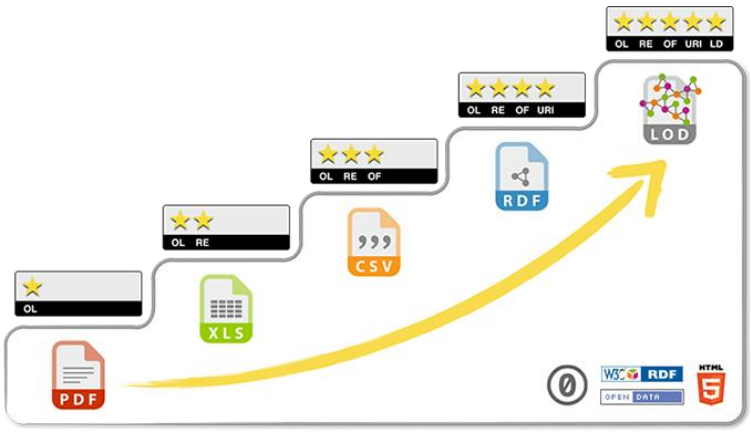
Datu kvalitātes modelis, kas definēts standartā ISO / IEC 25012, sastāv no 15 raksturlielumiem. Tie iedalās trīs daļās – iedzimtā, piemītošā (angl. *inherent*) datu kvalitāte, no sistēmas atkarīgā (angl. *system-dependent*) datu kvalitāte, kā arī kopīgie, kam piemīt gan iedzimta, gan no sistēmas atkarīgā datu kvalitāte. Pirmais raksturlielums ir precizitāte, kas satur sintaktisko precizitāti, semantisko precizitāti, kā arī metadatus. Otrais raksturlielums ir datu pilnīgums, tad seko nepretrunīgums, uzticamība, aktualitāte, pieejamība, atbilstība, konfidencialitāte, efektivitāte precizitāte, izsekojamība, saprotamība pieejamība, pārnesamība, atjaunojamība [22].



2.3. att. **ISO 25012 datu kvalitātes raksturlielumi (autores [22] tulkojums)**

2.2. Atvērto datu 5 – zvaigžņu klasifikācija

Kā tika minēts iepriekš, ir svarīgi, lai datus varētu saprast gan cilvēks, gan mašīna, kas nozīmē, ka dati jāatspoguļo saprotamā datu formātā. Tāpēc 2010. gadā Sers Tims Berners Lī ierosināja 5 – zvaigžņu novērtēšanas shēmu saistītajiem atvērtajiem datiem (2.4. attēls) [4].



2.4. att. **Piecu zvaigžņu atvērto datu klasifikācija [4]**

Galvenās iezīmes šīm zvaigznēm ir tāda, ka vienas zvaigznes dati ir pieejami visiem, bet nav mašīnlasāmi (piemēram, .pdf formātā pieejami dati), divas zvaigznes nozīmē, ka dati ir mašīnlasāmi (piemēram, .xls), trīs zvaigznes nozīmē, ka tie ir mašīnlasāmi un nepatentētā formātā (piemēram, .csv formāts), četru zvaigžņu datu pamatā ir trīs iepriekšējo līmeņu priekšrocības plus standarti (RDF, orResource apraksta struktūra), kas ļauj precīzi norādīt uz datiem tiešsaistē. Piecas zvaigznes jeb vispieejamākais atklātais datu līmenis ietver visu pārējo un spēju saistīt datus ar citiem datiem [5].

2.3. Populārākās datu kvalitātes problēmas

Datu kvalitātes problēmas ir raksturīgas lielākajai daļai datu. Datu kvalitātes problēmas kritiski ietekmē datu izmantošanu un to saprašanu. Lai uzlabotu datu kvalitāti, šīs problēmas ir jāapzinās un jāmēģina mazināt, tāpēc ir apzinātas populārākās datu kvalitātes problēmas, ar ko visbiežāk sastopas datu lietotāji. MIT institūtā ir apkopotas populārākās datu problēmas un ir izveidots saraksts, lai varētu tās klasificēt un atrast dažādās datu kopās [21] (2.1. tabula).

2.1. tabula

Vispopulārākās datu problēmas [21]

Nr.p.k.	Datu kvalitātes problēma	Paskaidrojums
1.	Trūkst vērtības/ pilnīgums	Vērtība ir obligāta, bet tā ir tukša.

2.	Sintakses pārkāpums	Sintaktiskas kļūdas, piemēram, atribūtā datums ir vērtība 2020/12/13, nevis 13/12/2020, kā visos citos ierakstos.
3.	Nepareiza vērtība	Vērtība neatbilst atļautajām atribūtā vērtībām.
4.	Domēna pārkāpums	Vērtība pārkāpj domēna noteikumus, piemēram, kādai precei pasūtītais daudzums ir negatīvs.
5.	Domēna ierobežojuma pārkāpums	Vērtība pārkāpj domēna ierobežojumus, piemēram, atribūtā klienta vārds jābūt vismaz diviem vārdiem, tomēr kādā vērtībā ir tikai viens vārds, vai, piemēram, datiem jābūt sakārtotiem pēc datuma, bet tā nav.
6.	Nederīga apakšvirkne	Vērtība nav atbilstoša, tajā ir nederīga apakšvirkne. Piemēram, atribūts *Klienta_vards* glabā arī akadēmisko grādu (piemēram, Dr. John Taylor).
7.	Rakstības kļūda	Vērtībā ir pieļauta kāda rakstības kļūda, pamērām, Atribūts *Pilseta* ir vērtība Rga, nevis Rīga
8.	Neprecīza vērtība	Vērtība ir neprecīza, piemēram viens saīsinājums var apzīmēt divas vai vairākas lietas
9.	Unikāls vērtības pārkāpums	Vērtības, kurām jābūt unikālām, tāda nav, piemēram, diviem dažādiem klientiem ir viens un tas pats nodokļu maksātāja identifikācijas numurs.
10.	Sinonīmu esamība	Viena atribūta ietvaros ir sastopami sinonīmi, kas apgrūtina datu apstrādāšanu, piemēram, SIA un “Sabiedrība ar ierobežotu atbildību”.
11.	Pustukša datu kopa	Datu kopā vairāk nekā 60% no datu kopas atribūtiem ir tukši

12.	Darbības atkarības pārkāpums	Tiek pārkāptas vērtību atkarības, piemēram, pasta indekss LV-5001, un norādīta pilsēta ir Ogre, bet citā vērtībā pasta indekss ir LV-5001, bet norādīta pilsēta Rīga.
13	Atsauces integritātes pārkāpums	Tiek izmantoti kādi ārēji avoti, piemēram, pasta indekss un tiek norādīts neeksistējošs pasta indekss.
14	Sintakšu neviendabīgums	Netiek ievērotas vienādas sintakses, piemēram, atribūtam pasutijuma_datums sintakse ir dd/mm/gggg, bet atribūta rekina_datums sintakse ir gggg/mm/dd.
15	Mērvienību neviendabīgums	Netiek ievērotas vienādas mērvienības. Viens atribūts ir pārstāvēts eiro, otrs dolāros.
16	Reprezentācijas neviendabīgums	Vienas un tā pati vērtība tiek attēlota dažādi, piemēram, lai attēlotu atribūtu dzimums, tiek lietotas vērtības S un V, bet citur ir izmantotas vērtības 0 un 1.
17	Homonīma esamība	Datu kopā ir atrodami homonīmi. Piemēram, uzņēmums ‘Pele’, bet cits uzņēmums arī pārdod mājdzīvnieku pele.

Izmantojot tabulu 2.1. darba autore veica pētījumu, kura kopsavilkums ir pieejams 4.2 nodaļā. Tika izvēlētas populārākās datu kopas un analizētas tās, lai noskaidrotu vai populārākās datu kvalitātes problēmas ir arī Latvijas atvērtā portāla datu kopās.

2.4. Datu tīrīšana

Lai nodrošinātu kvalitatīvus un pilnīgus datu, tiem ir nepieciešama datu tīrīšana. Datu tīrīšana ir laikietilpīgākā darbība datu zinātnes projektos, kuru mērķis ir piegādāt augstas kvalitātes datu kopas [9]. Tas nozīmē, ka datu tīrīšana ir ļoti laikietilpīga, bet ļoti svarīga, lai dati būtu kvalitatīvi un izmantojami. Lai datus varētu iztīrīt, vispirms ir jāsaprot, kādas ir datu kvalitātes problēmas (2.1 tabula) konkrētajiem datiem, un tad jāveic datu tīrīšana.

Datu tīrīšanu parasti veic ar datu kvalitātes rīkiem, piemēram, “OpenRefine”, “MS Data Quality Services”, “Talend” vai kādu citu automātisku datu tīrīšanas rīku, bet datus tīrīt var arī izmantojot SQL bāzētu datu tīrīšanu, kas arī tika izmantota darba praktiskajā daļā.

2.5. Atvērto datu portāla kvalitātes noteikšana

Lai noteiktu Atvērto datu portāla kvalitāti, nepietiek tikai noteikt tikai tā saturošo datu kvalitāti un to problēmas, kas aprakstītas 2.3. nodaļā, bet jāpēta arī pats portāls un dažādi portāla raksturlielumi, lai saprastu to kvalitāti. Darba autore veica literatūras analīzi, lai noskaidrotu, kā tiek veikta atvērto datu portālu analīze, vispiemērotākais šķita Máchová, R. un Lněnička, M. izveidotais atvērto datu portālu kvalitātes novērtēšanas piedāvātais satvars, jo tas bija visaptverošākais un piemērotākais LADP, kas ir balstīts uz ļoti daudziem citiem pētījumiem. Šis satvars ļauj analizēt atvērtā datu portāla vispārējo kvalitāti. Tā pamataspekti ir sniegti tabulā 2.2., aprakstot katru atvērto datu portāla analizējamo aspektu. Piedāvātais satvars tika pielietots Latvijas atvērtajam datu portālam, veiktas analīzes, kopsavilkumu sniedzot 4.1. nodaļā.

2.2. tabula

Novērtēšanas sistēma atvērto datu portālu kvalitātes novērtēšanai (autores [23] tulkojums)

I.Atvērto datu portāla vispārīgās īpašības	
Metriku saraksts	Kvalitātes novērtēšanas prasību apraksts
1. Tehniskā dimensija	
1.1 Iestāde un atbildība	Portāliem būtu jāsniedz informācija par iestādi, kas uztur portālu un pārvaldes modelis vai institucionālā sistēma, kas atbalsta datu sniegšanas modeļus
1.2 Datu vadības sistēma	Portāliem jāsniedz informācija par datu pārvaldības sistēmu, kas ir izmanto portāla darbināšanai
1.3 Valodas	Portāliem jāpiedāvā iespēja izvēlēties vairākas valodas, lai iegūtu vairāk lietotāju, un uzlabot portāla vispārējo kvalitāti
1.4 Bez maksas	Portāliem jānodrošina, ka visas datu kopas un pakalpojumi ir pieejami bez maksas un bez jebkādiem ierobežojumiem saskaņā ar atvērtajām licencēm
2. Pieejamības un piekļuves dimensija	

2.1 Datu kopu skaits	Portāliem jānorāda tajos iekļauto datu kopu skaits
2.2 Atkārtotās izmantošanas skaits	Portālos jānorāda to lietojumprogrammu skaits, kas izstrādātas, pamatojoties uz atvērtajiem datu atkalizmantošanas
2.3 Meklētājprogrammas (filtrs)	Portāliem jānodrošina spēcīgas datu kopas meklēšanas iespējas un atlases rīkus, izmantojot dažādus kritēriji kategoriju pārlūkošanai un filtru pārlūkošanai
2.4 API	Portāliem jānodrošina API ieinteresētajām personām lietojumprogrammu izstrādei, izmantojot atvērtos dati
2.5 Lietotāja konti	Portāliem jāatbalsta lietotāja konta izveide, lai personalizētu skatus un parādīto informāciju
2.6 Tematiskās kategorijas	Portāliem jānodrošina portāla sniegto datu kopu tematiskās kategorijas. Portālā skaidri jānošķir kategorijas (tēmas) no birkām (atslēgas vārdi)
2.7 Birkas (atslēgvārdi)	
3. Komunikācijas un līdzdalības dimensija	
3.1 Forums (atsauksmes)	Portāliem jānodrošina iespēja iesniegt atsauksmes par lietotāju datiem, pakalpojumu sniedzējiem un forumiem, lai apspriestu un apmainītos ar idejām starp lietotājiem.
3.2 Pieprasījuma veidlapas	Portālos jābūt pieejamai veidlapai, lai pieprasītu vai ieteiktu jaunu atvērto datu tipu vai formāta tipu.
3.3 Palīdzība (lietojamība)	Portālos būtu jāiekļauj augsta kvalitātes dokumentācijas un palīdzības funkcionalitāte, lai apgūtu, kā izmantot portālu un uzlabotu tā lietojamību.
3.4 Biežāk uzdotie jautājumi	Portāliem ir jānodrošina biežāk uzdoto jautājumu sadaļa, lai palīdzētu atrisināt iespējamās problēmas.
3.5 Sociālie mediji	Portāliem būtu jāpieslēdzas pie sociālo mediju platformas, lai radītu sociālo izplatīšanas kanālu atvērtiem datiem. Datu lietotāji un pakalpojumu sniedzēji var informēt citus par to, ko viņi izdarīja vai uzzināja no datu kopas.
II. Datu kopas vispārīgās īpašības	
1. Nosaukums un apraksts	Jāiesniedz datu kopas kopā ar to aprakstu, kā arī kādam nolūkam dati tika savākti.

2. Datu kopas autors	Ir jābūt parādītam datu kopas autoram, lai pārbaudītu to autentiskumu.
3. Izlaišanas datums un līdz datums	Datu kopas ir jāsaista ar konkrētu laiku vai periodu. Visi informācijai datu kopā jābūt atjauninātai
4. Licence	Datu kopām jāsniedz licences informācija, kas saistīta ar publicētās informācijas izmantošanu. Datu kopas, kurām nav tieši atvērtas licences, nav atvērti dati.
5. Ģeogrāfiskais pārklājums	Jāapraksta, vai datu kopā dati attiecas uz valsts, reģionālo vai vietējo līmeni.
6. Datu kopas URL	Datu kopu URL jābūt pieejamam katrai datu kopai.
7. Datu kopas (faila) lielums	Jābūt pieejamiem datu kopu (failu) lielumiem.
8. Skatījumu (apmeklējumu) skaits	Datu kopai ir pieejams kopējais tiešsaistes skatījumu skaits.
9. Lejupielāžu skaits	Kopējais datu kopas lejupielādes reižu skaits.
10. Mašīnlasāmi formāti	Datu kopas ir jānodrošina formātos, kas ir tikpat ērti un viegli analizējami un pārvietojami kā arī lejupielādējami.
11. Vizualizācijas	Datu kopām jānodrošina vizualizācijas.
12. Lietotāju vērtējums un diskusijas ziņojums	Datu kopām jānodrošina iespējas, kas ļauj apkopot lietotāju vērtējumus un komentārus par datu kopu vai apspriest secinājumus, pamatojoties uz datu izmantošanu.

3. LATVIJAS ATVĒRTO DATU PORTĀLS UN TĀ KVALITĀTE

Pēdējos gados arvien vairāk valstis veido savus atvērto datu portālus. Arī Latvijai ir izveidots savs atvērto datu portāls, kurš ir salīdzinoši jauns, t.i., tas tika izveidots 2017. gadā. Latvijas Atvērto datu portāla mērķis ir sniegt Latvijas sabiedrībai pieejamu atvērto datu platformu, kas ir vienota sadarbības vide Latvijas atvērto datu publicētājiem un izmantotājiem. Portāla datu publicētāji ir valsts un pašvaldību iestādes vai organizācijas, kas pilda valsts funkcijas. Saskaņā ar normatīvajiem aktiem atvērtos datus iestādes publicē pēc savas iniciatīvas. Savukārt datu izmantotājs ir ikviena persona (iedzīvotājs, uzņēmējs u.tml.), kurš apmeklē portālu, lai atrastu un izmantotu datus. Datu izmantošana portālā ir bezmaksas [15]. Tas nozīmē, ka LADP var izmantot ļoti daudz un dažādi cilvēki dažādiem mērķiem.

Lai gan Latvijas atvērtais datu portāls ir salīdzinoši jauns, tas ir salīdzinoši augsti novērtēts gan Eiropā, gan pasaulē. Eiropā portāls ieņem 19. vietu (pagājušo gadu, tā bija 10.) [6], kaut gan portālā kopējās tendences novērtējumā no pagājuša gada ir uzlabojušās, bet tās nav uzlabojušās pietiekami daudz, lai saglabātu savu pozīciju pirmajā desmitā. Nākamajā nodaļā ir sniegts darba autores veiktais Latvijas atvērto datu portāla analīzes pētījums.

3.1. Latvijas atvērto datu portāla datu kopas

Latvijas atvērto datu portāls tika pētīts 2021. gada maijā. Uz to brīdi portālā bija pieejamas 472 datu kopas, kuras ir pievienojuši 85 datu kopu publicētāji un kas ir sadalītas 14 dažādās kategorijās [1].

Populārākās kategorijas ir "ekonomika un uzņēmējdarbība", "reģioni un pašvaldības", kā arī "iedzīvotāji un sabiedrība" un "valsts pārvalde", kas arī ir saprotami, jo šīs jomas ir nozīmīgas valstij un par tām tiek ievākts daudz datu, kurus arī dažādas organizācijas ir izvēlējušās atvērt, lai sabiedrība un citas organizācijas varētu iegūt labumu no tām. Pilns populārāko kategoriju pārskats pēc to popularitātes ir pieejams 1. pielikumā.

Saskaitot datu kopu skaitu pa kategorijām, tas ir lielāks nekā kopējais datu kopu skaits, jo datu kopas var piederēt vairāk nekā vienai kategorijai, kā arī dažām datu kopām kategorija var nebūt norādīta. Tas apgrūtina meklēšanu, ja, piemēram, interesē visas medicīnas kategorijas datu

kopas, bet dažas iespējams nemaz nav norādītas šajā kategorijā, tad, iespējams, tās atrast tikai meklējot pēc atslēgvārdiem, navigējot pēc birkām vai vienkārši izpētot visu datu kopu sarakstu.

Latvijas atvērto datu portālā datu kopas ir pievienojuši 85 publicētāji. Tas nozīmē, ka ir salīdzinoši daudz dažādu publicētāju, tāpēc arī tiek nodrošināts salīdzinoši liels datu pārklājums par dažādām tēmām. 2. pielikumā var redzēt populārākās datu kopu publicētājus. Populārākie datu publicētāji ir tādas organizācijas kā ĢEOLatvija.lv, Centrālā statistikas pārvalde, Valsts reģionālās attīstības aģentūra. Tāpat var ievērot, ka populārākās organizācijas, kas atver datus, ir valsts iestādes, kas ir likumsakarīgi, jo Latvijas atvērto datu portāls ir atvērto pārvaldes datu portāls jeb atvērtais valdības datu portāls, kurā galvenokārt atver datu tieši valsts iestādes.

Apskatot Latvijas atvērto datu portāla datu kopu formātus, kuru pilnais pārskats ir pieejams 3. pielikumā, tiek secināts, ka visplašāk tiek izmantots CSV formāts, kas ir ērts gan izdevējam, gan lietotājam, kā arī tas atbilst 3 zvaigžņu līmenim 5 – zvaigžņu atvērto datu klasifikācijā. Tāpat populāri formāti ir XLSX, JSON, WMS un XLS, tas ir izskaidrojams ar to, ka šādi attēloti dati ir ērti lietošanai, gan datu patērētājam, tos viegli un ērti var mainīt, savienot, lasīt, gan arī tos ir salīdzinoši viegli izveidot un publicēt. Daudzas datu kopas ir pieejamas dažādos formātos, kā arī dažām datu kopām apraksti vai parametru apraksti tiek pievienoti DOCX vai PDF formātos, tāpēc šo formātu skaits ir tik liels. Pielikumā var redzēt, autores noteiktu katra datu formāta atbilstību 5-zvaigžņu atvērto datu kvalifikācijai. Kaut gan sākotnējie datu formāti, kas iekļauti šajā klasifikācijā, bija maz (aprakstīti 2.2. nodaļā), nespējot aprakstīt mūsdienās izmantotos formātus, kuru skaits un raksturs ir mainījies, ir pieejami avoti, kuros arī citus formātus, piemēram, SHP, JSON un citus mēģina klasificēt pēc šo 5- zvaigžņu formāta, tiem piemītošo īpašību dēļ [19], ko pārņem arī autore, sniedzot pilnīgāku novērtējumu. Lielākā dala no izmantotajiem formātiem atbilst 3 zvaigznēm 5 – zvaigžņu atvērto datu klasifikācijā, tas ir skaidrojams ar to, ka šāda līmeņa dati ir samērā viegli publicējami, kā arī viegli apskatāmi un izmantojami lietotājiem.

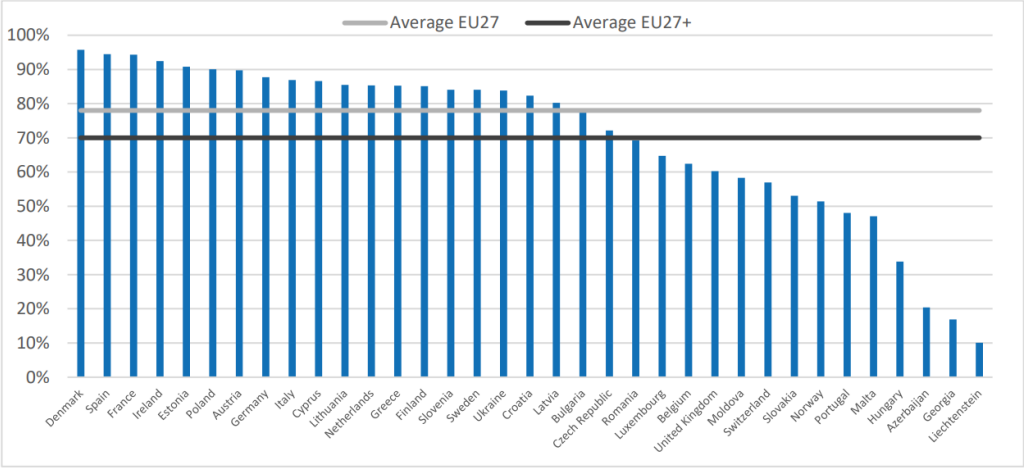
Lai ērti un viegli varētu pārvietoties pa datu kopām un atrast sev saistošās, ir Latvijas atvērto datu portālā ir izveidotas birkas, kas vienai datu kopai var būt arī vairākas, šis aspekts arī tiek aplūkots un vērtēts datu portālu kvalitātes novērtēšanas satvarā **(**2.2. tabula**)**. Populārākās birkas ir "statistika", "2020", "pilsēta", "2019" un "administratīvā teritorija", 4. pielikumā var apskatīt populārākās birkas Latvijas atvērto datu portālā.

3.2. Eiropas oficiālā datu portāla Latvijas atvērto datu portāla kvalitātes analīze

Katru gadu Eiropas oficiālais datu portāls pēta Eiropas Savienības un Eiropas ekonomiskās zonas dalībvalstu atvērto datu portālus, lai noskaidrotu kopējo portālu kvalitāti un sarindotu tos pēc kopējā iegūtā punktu skaita. Portālus vērtē pēc to:

- atvērto datu politikas – vērsta uz konkrētu politiku un stratēģiju īstenošanu, lai veicinātu atvērtos datus valsts līmenī;
- atvērto datu ietekme – apskata darbības, kas veiktas, lai uzraudzītu un mērītu atvērto datu atkalizmantošanu, un ietekmi, ko rada šī atkalizmantošana. Pēta arī lietojumprogrammas, produktus un pakalpojumus, kas ir izstrādāti balstoties uz šiem atvērtajiem datiem;
- atvērto datu portāls – koncentrējas uz uzlabotām portāla funkcijām, kas veicina mijiedarbību starp datu izdevējiem un atkalizmantotājiem. Turklāt dimensija novērtē, cik lielā mērā portāla vadītāji izmanto tīmekļa analīzes rīkus, lai labāk izprastu lietotāju vajadzības, kā arī ieviestos pasākumus, lai nodrošinātu portāla ilgtspēju;
- atvērto datu kvalitāte – koncentrējas uz portāla vadītāju pieņemtajiem pasākumiem, lai nodrošinātu sistemātisko metadatu ieguvi no avotiem visā valstī, kā arī pēta, lai dati būtu atvērto datu formātos, mašīnlasāmi, augstas kvalitātes, kā arī tos varētu saistīt ar citiem datiem [16].

Latvijas atvērto datu portāls šajā reitingā ieņemt 19. vietu, kas ir aptuveni pa vidu, bet tāpat tas ir mazliet augstāks starp Eiropas Savienības valstīm, bet krietni augstāks par vidējo vērtējumu Eiropas Savienības kopā ar Eiropas ekonomiskās ietekmes valstu vidējo vērtējumu (3.1 att.).

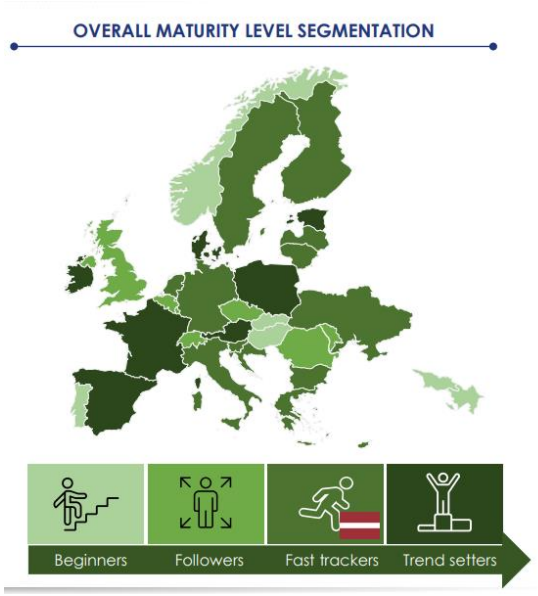


3.1. att. **Eiropas atvērto datu portālu vidējais kvalitātes mērījums [16]**

Reitingā valstis tiek iedalītas četras grupās pēc to portālu brieduma (angl. *maturity*):

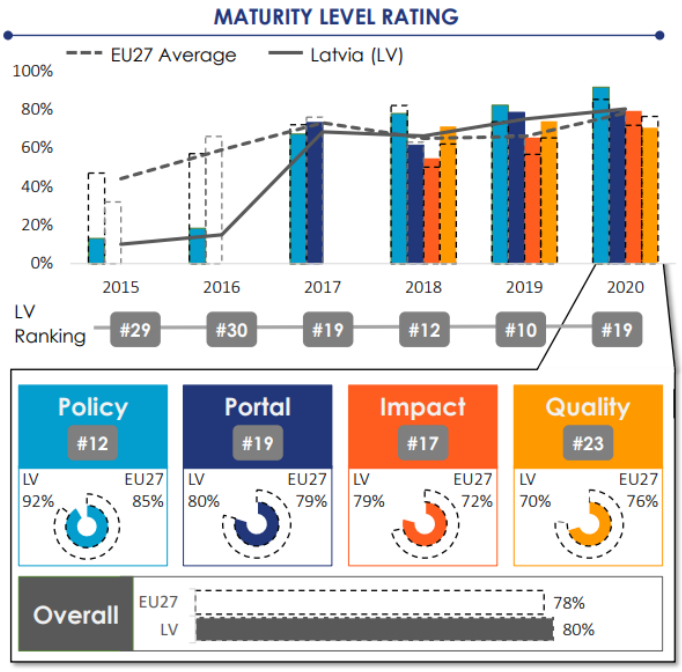
- iesācēji (angl. *beginner*) – valsts atvērto datu portāls nespēj attīstīties tādā pašā tempā kā citas valstis. Iespējams, valstij nav atvērta datu portāla, vai, ja tāds ir, portālam ir ierobežots datu kopu skaits, salīdzinot ar valsts potenciālu. Attiecībā uz datu kvalitātes, valsts maz rīkojas, lai varētu publicēt datus augstākā kvalitātē;
- sekotāji (angl. *follower*) – valstij jau ir atvērta datu politika, un tā veic darbības, lai nodrošinātu taisnīgu atklāto datu darbību koordinācijas līmeni. Portālā ir pieejamas standarta funkcijas, bet ierobežotā skaitā, kas neatbilst pieredzējušu lietotāju vajadzībām. Ir veiktas dažādas, darbības, lai tiktu, veicināta kvalitatīvu datu publicēšana;
- ātri izsekotāji (angl. *fast-tracker*) – valsts parāda labu portāla brieduma līmeni visās dimenijās. Kopumā valsts datu portāls ir sasniedzis augstu standartu līmeni. Portāls nodrošina labu funkcionalitāti, lai apmierinātu pieredzējušu un pamatlietotāju vajadzības;
- tendenču noteicēji (angl. *trend-setter*) – valsts ir ieviesusi progresīvu atvērto datu politiku ar stingru koordināciju par atvērtām datu darbībām visos valdības līmeņos. Valsts portāls nodrošina plašu līdzekļu un pakalpojumu klāstu pieredzējušu lietotāju vajadzībām, kā arī datu izdevējiem. Atvērto datu kvalitātes līmenis valstī ir ļoti augsts [16].

Latvija ierindojas ātro izsekotāju grupā, kas nozīmē, kaut gan ir redzamas dažādas nepilnības atvērto datu portālā, kopumā atvērto datu portāls ir labā kvalitātē un tam ir pieejamas dažādas funkcijas gan parastam lietotājam, gan arī pieredzējušam lietotājam. (sk. 3.2. att.).



3.2. att. **LADP brieduma līmeņu segmentācija [17]**

Kaut gan Latvijas atvērto datu portāls kopš pagājušā gada, kurā tas ierindojās 10. vietā, šogad ir noslīdējis līdz 19. vietai, vidējais datu portāla brieduma reitings ir pakāpies. Kā redzams 3.3. att., visas dimensijas kopš 2019. gada ir uzlabojušās, izņemot kvalitāti, kas ir samazinājusies un tā ir arī vienīgā dimensija, kas portālā ir sliktāka par vidējo vērtību Eiropas Savienības valstīs. Attēlā var arī redzēt, ka LADP politika un ietekme ir daudz augstāka nekā vidējā vērtība Eiropas Savienības valstu vidū.



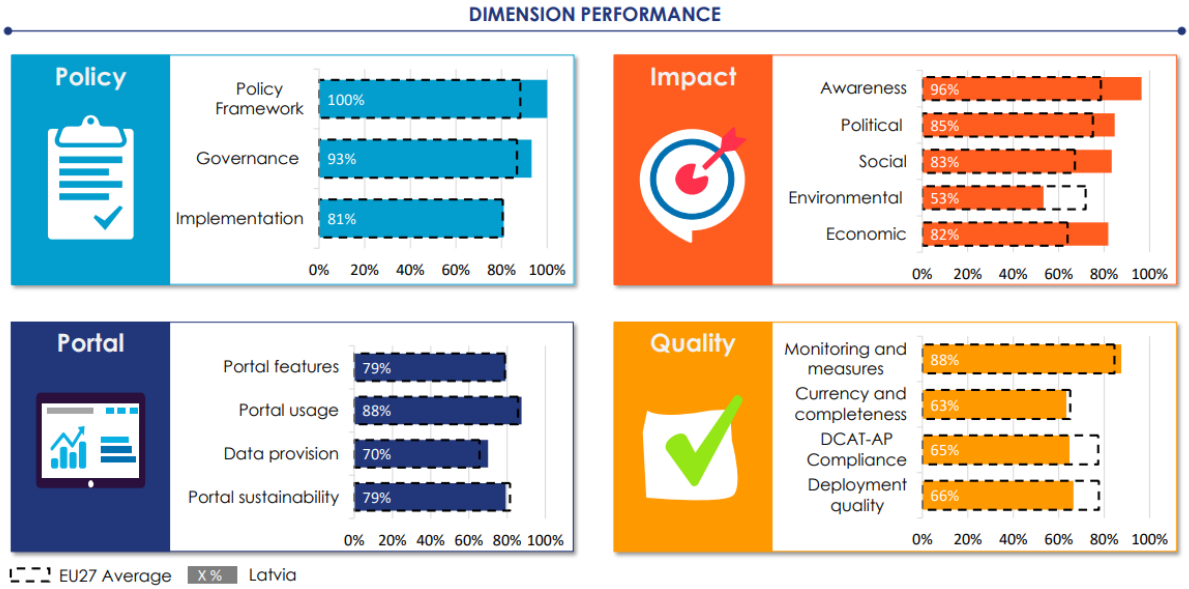
3.3 att. **Latvijas atvērto datu portāla brieduma reitings [17]**

Pētot sīkāk Latvijas atvērto datu portālu pētītās dimensijas, var redzēt, ka "politika" portālam ir ļoti augstā līmenī, tā sastāv no politikas satvara, kas ir ieguvis visaugstāko iespējamo vērtējumu 100%, pārvaldība, kas ir ieguvusi 93% un tā ir, virs vidējā vērtējuma Eiropā, kā arī realizācija, kas ir 81% un tā ir vidējā Eiropas līmenī.

Otrā pētītā dimensija ir "portāls". Šī dimensija sastāv no portāla funkcijām, portāla lietojamības, datu sniegšanas, kā arī portāla ilgtspējība, kuras visas ir vai Eiropas līmenī vai mazliet augstākas, izņemot portāla ilgtspējību, kas ir novērtēta mazliet zemāk par vidējo vērtību Eiropā.

Pētījumā tika pētīta arī atvērto datu ietekmes dimensija, kas sastāv no apzināšanās, politiskā, sociālā, vides, kā arī ekonomiskiem aspektiem. Var redzēt, ka LADP parādīja ļoti augstu vērtējumus visās šajās iezīmes, izņemot vidi, kas bija krietni zemāks rādījums nekā Eiropas vidējais rādījums.

Ceturtā dimensija ir "kvalitāte". Ar kvalitāti tiek saprasta uzraudzība un pasākumi, savlaicīgums (angl. currency) un (meta) datu pilnīgums, DCAT-PA atbilstība, kā arī izvēršanas kvalitāte (angl. *deployment),* kuras pamatā ir datu atbilstība 5 – zvaigžņu klasifikācijai. Šajā dimensijā Latvijas atvērto datu portāla sniegums bija vissliktākais, tikai uzraudzība un pasākumi bija augstāki par vidējo līmeni, bet citas sadaļas bija zem vai krietni zem vidējā vērtējuma Eiropā.



3.3 att. **Latvijas atvērto datu portāla brieduma reitings [17]**

3.3. Latvijas atvērto datu kvalitātes analīze

Kursa darba ietvaros [12] darba autore izpētīja atvērto datu kvalitāti Latvijas atvērto datu portāla specifiskai datu kategorijai, precīzāk veselības kategoriju datu kopu kvalitāti. Pasaulē esošās situācijas dēļ, tā ir ļoti svarīga lietotājiem un tā pieder arī pie augstās vērtības datu kopām. Īsi apkopojot kursa darba rakstīšanas laikā iegūto pieredzi - tika pētītas tādas atvērto datu kvalitātes noteicošās vērtības, kā:

1. vai dati tiek atjaunoti noteiktajā datu atjaunošanas biežumā;
2. vai ir pieejams detalizētāks datu kopas apraksts, tas nozīmē, ne tikai nokopēts datu kopas nosaukums, bet arī tās paskaidrojums, tādējādi sniedzot datu lietotājam skaidrību par datu kopas saturu, datu raksturu, tāpat arī kādiem nolūkiem dati tika vākti utt.;
3. vai ir aprakstīti datu kopas parametri, dota detalizētāka informācija, ko nozīmē katrs parametrs (piemēram, skaitlisko vērtību mērvienība);

4. datu formāts un kuram līmenim 5 – zvaigžņu klasifikācijā atbilst konkrētais formāts. Gadījumā, ja konkrētā datu kopa bija pieejama vairākos formātos, tika vērtēts augstākais vērtējums);

5. datu priekšskatījuma esamība, lai lietotājs datus varētu ātri un ērti apskatīt. [12]

Kopā tika pētītas 24 veselības kategorijā esošās datu kopas, no iegūtajiem datiem, var secināt, ka datu kopas atbilst solītajai datu atjaunošanai 13 datu kopām no 24, bet datu kopas apraksts ir pieejams 22 datu kopām, tas nozīmē, ka datu kopu apraksti bija pieejami gandrīz visām datu kopām. Ar parametru aprakstu bija sliktāk, jo tas bija pieejams tikai 6 datu kopām, kas ir tikai ceturtā daļa no visām pieejamajām datu kopām. Priekšskatījumu arī varēja apskatīt diezgan lielam datu kopu skaitam, kas vija 19 datu kopas no 24 (5. pielikums).

Populārākie datu formāti ir veselības kategorijas datu kopām bija CSV un XLSX, JSON, kas ir viegli apskatāmi un izmantojami. Tie arī atbilst 2 un 3 zvaigznēm 5 – zvaigžņu klasifikatorā, kas ir ērti gan datu publicētājam, gan arī datu patērētājam. Divas datu kopas atbilda 1 zvaigznei 5 – zvaigžņu klasifikatorā, jo ir pieejami HTML formātā (sk. 5. pielikumu). Šie dati nav ērti apskatāmi un pārvirza datu patērētāju uz citu tīmekļa vietni.

Darba autore arī veica pētījumu, nosakot, kādas ir atvērto datu kvalitātes noteicošās vērtības darbā pētījumā izmantotajām datu kopām (sk. 6. pielikumu). Tika iegūti līdzīgi dati, kas aprakstīti 4.2. nodaļā.

4. LATVIJAS ATVĒRTO DATU PORTĀLA UN TĀ DATU KVALITĀTES ANALĪZE

Lai noteiktu Latvijas atvērto datu portāla un tā datu kvalitāti, darba autore izvēlējās to vērtēt divos posmos, pirmajā posmā pētot Latvijas atvērto datu portāla kvalitāti, tā nosakot nepilnības pašā atvērto datu portālā un atvērto datu piekļūšanā, otrajā daļā pētot atvērto datu kopu saturošo datu kvalitāti, izmantojot atvērto datu kvalitātes noteicošos raksturlielumus, kā arī izmantojot SQL analīzi, pētot, vai datiem piemīt vispopulārākās datu kvalitātes problēmas.

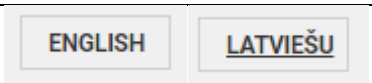
4.1. Latvijas atvērto datu portāla kvalitātes analīze

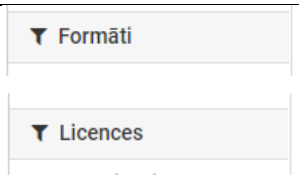
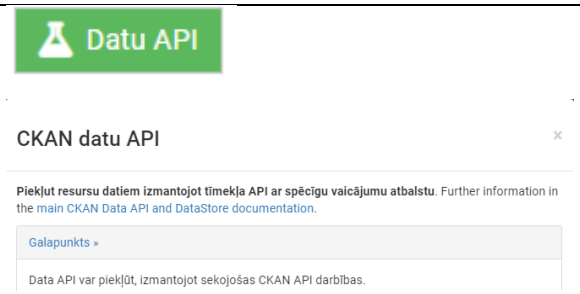
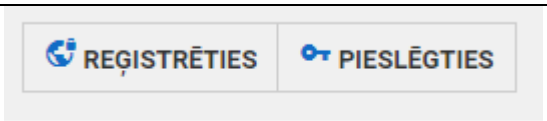
Latvijas atvērto datu portāla kvalitāte tika analizēta pēc 2.5 nodaļā aprakstītajām datu portāla kvalitātes noteikšanas vadlīnijām. Darba autore arī katram punktam sniedza vērtējumu no 1-3, kur 1 ir, viszemākā atzīme, šī datu metrika nav ievērota, 2 nozīmē, ka šī datu metrika ir ievērota, bet ir redzamas problēmas, bet 3 visaugstākā, kas nozīmē, ka šis punkts ir izpildīts perfekti.

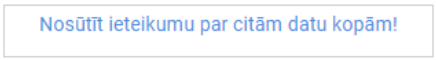
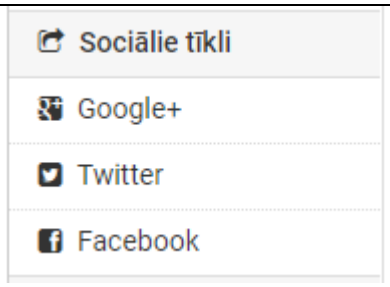
4.1. tabula

Latvijas atvērto datu portāla kvalitātes analīze

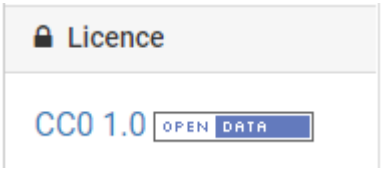
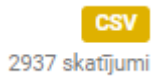
I.Atvērto datu portāla vispārīgās īpašības		Vērtējums
Metriku saraksts	Kvalitātes novērtēšanas prasību apraksts	
1. Tehniskā dimensija		
1.1 Iestāde un atbildība	© 2017. - 2019. Valsts Reģionālās attīstības aģentūra. Visas tiesības aizsargātas. Par datu kataloga saturu atbildīgi datu publicētāji. dati@varam.gov.lv LADP apakšā ir redzama iestāde, kas uztur šo datu portālu, kā arī tiek parādīts e-pasts, kam rakstīt, lai sazinātos ar šo iestādi.	3
1.2 Datu vadības sistēma	Portālā izmantotas atvērtā koda tehnoloģijas t.sk. CKAN atvērto datu kataloga platforma. Izstrādātie papildinājumi ir pieejami atvērtā koda veidā vietnē: https://github.com/dpp-dev LADP sniedz informāciju par tā datu vadības sistēmu.	3

1.3 Valodas	 Kaut gan datu portālā tiek piedāvātas divas valodas, latviešu un angļu, nomainot valodu uz angļu valodu, nomainās tikai portāla izvēles, bet ne kategoriju nosaukumi, ne birkas, ne pašu datu kopu nosaukumi nemainās. Kā arī, vajadzētu pievienot arī citas valodas, kas ir populāras Latvijas teritorijā.	2
1.4 Bez maksas	Portāliem visas datu kopas un pakalpojumi ir pieejami bez maksas, kā arī nav pat jāreģistrējas vai jāpieslēdzas, lai apskatītu un lejupielādētu datu kopas.	3
2. Pieejamības un piekļuves dimensija		
2.1 Datu kopu skaits	 Ieejot portālā, sākumlapā ir redzamas, cik kopas ir pieejamas, bet tās nesakrīt ar portālā esošo datu kopu skaitu, kas ir mazliet lielāks, to var apskatīt sadaļā "datu katalogs".	2
2.2 Atkārtotās izmantošanas skaits	Portālā nav redzams lietojumprogrammu skaits, kas izstrādātas, pamatojoties uz portālā esošajiem atvērtajiem, bet caur "European Data Maturity" portālu var atrast saiti (https://data.gov.lv/lv/piemeri), kur var apskatīt dažus piemērus, kur ir izmantoti Latvijas atvērtie dati, bet šo saiti nav iespējams atrast portālā.	1
2.3 Meklētājprogrammas (filtrs)		2

	 Portālā ir iespējams datus filtrēt pēc organizācijas, kategorijas, birkas, formāta, kā arī licences (kas ir tikai viena visiem datiem, kaut gan filtrēšanas mehānismi ir ļoti labi, varētu tos papildināt.	
2.4 API	 Ne visām datu kopām ir pieejams API, bet aptuveni pusei datu kopu ir pieejams API.	2
2.5 Lietotāja konti	 Piereģistrēties portālam var ātri un ērti, bet diemžēl portāls neatbalsta ārējās reģistrēšanās, kā Google vai Facebook. Ja nospiež sekot kādai no datu kopai, tad pie profila var redzēt to atjaunošanu un citu informāciju par abonētajām datu kopā. Reģistrētam lietotājam ir diezgan maz papildus funkcionalitātes portālā.	2
2.6 Tematiskās kategorijas	Datu portālā ir pieejamas 14 dažādas, kas nepārklājas ar birkām.	3
2.7 Birkas (atslēgvārdi)	Datu portālā ir pieejamas birkas un tās nepārklājas ar kategorijām.	3
3. Komunikācijas un līdzdalības dimensija		
3.1 Forums (atsauksmes)	Nav pieejams lietotāju forums.	1

3.2 Pieprasījuma veidlapas	 Portālā var pievienot ieteikumu par datu kopām, kurām būtu jābūt pieejamām, kā arī var redzēt, kuras no ietiktajām datu kopām ir ieviestas, un kuras nav, var balsot par tām.	3
3.3 Palīdzība (lietojamība)	Ir pieejamas vadlīnijas, kā piereģistrēties, kā jāsaglabā CSV datu kopas to publicēšanai, kā arī kā jāaizpilda datu aprakstīšanas veidlapa, bet cita palīdzība, kā lietot portālu nav pieejama.	2
3.4 Biežāk uzdotie jautājumi	 Portālā ir apkopoti biežākie jautājumi par atvērtajiem datiem un to lietošanu, kā arī ir iesējams uzdot savu jautājumu.	3
3.5 Sociālie mediji	 Ir pieejamai daži sociālie mēdiji, bet tie ir ļoti maz un brīžiem ir problēmas ar to darbību.	2
II. Datu kopas vispārīgās īpašības		
1. Nosaukums un apraksts	Jāiesniedz datu kopas kopā ar to aprakstu, kā arī kādam nolūkam dati tika savākti. Datu kopas apraksts ir pieejams pārsvarā visām datu kopām, bet dažām nav, to var redzēt 5. pielikumā, kur ir pieejams mans kursa darba pētījums par veselības kategorijas datu kopām, bet arī ir pieejams 6. pielikumā, kur ir pētītas LADP populārākās datu kopas.	2

2. Datu kopas autors	 **Datu publicētāja struktūrvienība** — Valsts ieņēmumu dienesta Nodokļu pārvalde **Saziņas e-pasts datu jautājumiem** — dati@ur.gov.lv Ir pieejams datu kopas autors un dažās datu kopās ir arī redzama šīs organizācijas struktūrvienība, kas izdevumi datus, lai varētu pārliecināties par datu autentiskumu, kā arī ir pieejams saziņas e-pasts, lai varētu sazināties par datu atbildīgo, ja rodas kādi jautājumi.	3
3. Izlaišanas datums un līdz datums	**Datu izdošanas datums** — 2020-03-25 **Datu pēdējo izmaiņu datums** — 2020-04-07 **Atjaunošanas biežums** — reizi nedēļā **Datu izdošanas datums** **Datu pēdējo izmaiņu datums** Kā redzams ekrānuzņēmumos, ir pieejami datu kopas aprakstošie logi, kā datu izdošanas datums, datu pēdējo izmaiņu datums, kā arī atjaunošanas biežums, kaut gan ir pieejami šie raksturlielumi, daudzām kopām šīs vērtības nav norādītas, kā arī netiek ievērots atjaunošanās biežums. Nodaļā 4.2. var redzēt, ka vienas kategorijas ietvaros,	2

	atjaunošanās datums tiek ievērots tikai aptuveni pusei no pieejamajām datu kopām.	
4. Licence	 Atverot datu kopu, kreisajā malā pēdējais raksturlielums ir licence, kura parāda, ka šī datu kopa ir atvērta.	3
5. Ģeogrāfiskais pārklājums	Ģeogrāfiskais pārklājums ir noteikts datu kopu nosaukumā vai arī tās aprakstā, kā arī, ja nav precizēts, ir saprotams, ka tas attiecas uz Latvijas teritoriju, jo tas ir Latvijas atvērto datu portāls, bet portālā vajadzētu pievienot raksturlielumu, kur apraksta kāds ģeogrāfisko pārklājums atteicas uz konkrēto datu kopu, lai arī būtu pēc tam būtu vieglāk filtrēt datus.	2
6. Datu kopas URL	 Datu kopām ir pieejami URL.	3
7. Datu kopas (faila) lielums	Nav pieejams un redzams datu kopu (failu) lielums pirms to lejupielādēšanas.	1
8. Skatījumu (apmeklējumu) skaits	 Apskatot datu kopas sarakstā, labajā pusē zem formāta, ir redzams datu kopas kopējais tiešsaistes skatījumu skaits.	3
9. Lejupielāžu skaits	Nav redzams datu kopas lejupielāžu reižu skaits.	1
10. Mašīnlasāmi formāti	Lielākā daļā datu kopas ir mašīnlasāmā datu formātā, kā var redzēt 3. pielikumā, bet ir arī datu kopas, kuras nav mašīnlasāmā formātā.	2
11. Vizualizācijas	Ja dati ir mašīnlasāmā formātā, tad tiek nodrošināta datu vizualizācija, bet ne visi dati ir tādā formātā, tāpēc vizualizācija netiek nodrošināta visiem datiem, kā arī, ja ir	2

	pieejami kartes dati, tiek nodrošināta kartes vizualizācija, bet brīžiem šī funkcionalitāte darbojas nepilnīgi un nekorekti	
12. Lietotāju vērtējums un diskusijas ziņojums	Ir pieejama sadaļa, kurā var vērtēt ieteiktās datu kopas, kuras vajadzētu pievienot portāla, bet esošās datu kopas nav iespējams novērtēt. Nav iespējama diskusija ar citiem datu lietotājiem, ir iespējams tikai nosūtīt ziņu datu publicētājam.	1

Izpētot Latvijas atvērto datu portālu, pēc iepriekš minētajām metrikām, tika iegūti tādi vērtējumi, ka 1 balle, kas ir viszemākais vērtējums, un tas nozīmē, ka šī metrika vispār nav vai netiek ievērota, bija 5 metrikām. Portālā:

1) nav redzams lejupielāžu skaits jeb, cik reizes konkrētā datu kopa ir lejupielādēta;
2) portālā nav pieejams uzzināt datu kopas (faila) lielumu pirms tā lejupielādes;
3) nav pieejams arī atkārtotās izmantošanas skaits;
4) nav pieejams forums lietotājiem;
5) nav pieejams lietotāju vērtējums un diskusijas ziņojums.

Vidējo vērtējumu, kas ir 2, ieguva 12 metrikas, tas nozīmē, ka, kaut gan ir redzams, ka tiek ievērota konkrētā metrika, ir arī novērojamas nepilnības.

Visaugstāko vērtējumu ieguva 11 metrikas, kas ir ļoti labs rezultāts. To starpā ir skatījumu (apmeklējumu) skaits, licence, datu kopas autors kā arī citas metrikas.

Apvienojot visus iegūtos rezultātu, vidējais vērtējums Latvijas atvērto datu portālam bija 2,25, kas ir ļoti augsts, jo tas nozīmē, ka ir novērojamas dažas nepilnības, bet kopumā portāls tiecas uz vislabāko vērtējumu.

4.2. Latvijas atvērto datu portāla datu kopu saturošos datu kvalitātes analīze

Lai veiktu Latvijas atvērto datu portāla datu kopu saturošos datu kvalitātes analīzi, tika izvēlētas 14 datu kopas, no katras kategorijas izvēloties populārāko, tā kā datu kopa var būt vairākās kategorijās, ja viena datu kopa bija populārākā divās kategorijās, vienā no kategorijām tika ņemta nākama populārāka.

Pirms tika veikta SQL balstīta datu kvalitātes analīze, darba autore veica atvērto datu kopu kvalitātes noteikšanu pēc atvērto datu kvalitātes raksturlielumiem:

1. vai datu kopa tiek atjaunota regulāri;
2. vai ir pieejams detalizētāks datu kopas apraksts;
3. vai ir aprakstīti datu kopas parametri;
4. datu formāts un kuram līmenim 5 – zvaigžņu klasifikācijā;
5. datu priekšskatījuma esamība (sk. 3.2. nodaļu).

Tika iegūti diezgan līdzīgi dati kā kursa darbā apskatītājām datu kopām, bet bija dažas atšķirības. Datu kopu atjaunošana tika ievērota 10 datu kopās, 4 netika ievērota. Datu apraksts bija gandrīz visām datu kopām, tikai vienai datu kopai datu kopas apraksts nebija pieejams. Ar parametru apraksta esamību bija vissliktāk, jo tas bija pieejams tikai vienai datu kopai. Tā kā visi dati bija mašīnlasāmā formātā, datu kopu priekšskatījums bija pieejams visām datu kopām. Kā arī visas datu kopas atbilda 3 zvaigznēm 5-zvaigžņu klasifikācijai, kas ir ļoti labs rezultāts (sk. 6. pielikumu). Šie rādījumi ir mazliet augstākie nekā kursa darba pētītajā veselības kategorijā (sk. 3.2. nodaļu), jo šīs ir populārākās datu kopas no katras kategorijas, tātad šīs datu kopas tiek lietotas un apskatītas biežāk.

4.2. tabula

Latvijas atvērto populārāko datu kopu kvalitātes analīze

Datu kvalitātes vērtība	Atbilst/Ir (datu kopu skaits)	Neatbilst/Nav (datu kopu skaits)
Datu kopu atjaunošana	10	4
Datu kopas apraksta esamība	13	1
Parametru apraksta esamība	1	13
Priekšskatījums datiem	14	0

Darba autore arī veica datu kvalitātes analīzi pēc 2.3. nodaļā aprakstītajām populārajām datu kvalitātes problēmām. Pārsvarā datu kvalitātes analīze tika veikta, izmantojot SQL vaicājumus, bet dažās problēmas ir ļoti grūti precīzi nosakāmas, piemēram, homonīma esamība, sinonīmi un citas, bet darba autore mēģināja novērtēt arī šī datu kvalitātes problēmas, izdarot dažādus secinājumus no esošajiem datiem.

Kā 4.2. tabulā ir redzams, gandrīz visām datu kopā arī ir grūti novērtēt esošo parametrus, jo ļoti bieži trūkst parametru apraksti un darba autorei bija jāizanalizē, kur drīkst būt tukšas vērtības

un kur nē. Parametru apraksti ir izšķiroši un svarīgi plašākai izpratnei un datu, piemēram, ģeodatu vai tabulu, pareizu izmantošanu, taču kā rāda pētījums 82% datiem Austrijas datu portālā trūkst parametru aprakstu [24]. Tas nozīmē, ka ne tikai Latvijas atvērto datu portālā ir novērojama šāda problēma, bet arī citā valsts atvērto datu portālā.

Lai datus apstrādātu, tie tika lejupielādēti, kā arī atvērti "SQL Server Management Studio", kurā pēc tam tika izveidoti konkrētie SQL vaicājumi, lai noteiktu vai datu kopai ir esošās datu kvalitātes problēmas vai nav.

Datu kopas, kuras tika pētītas un kādas ir to galvenās datu kvalitātes problēmas:

1) veselības kategorija – "COVID19 vakcinācijas", datu kopas datu kvalitātes problēmas var apskatīt 7. pielikuma 1. tabulā;
    - galvenās problēmas datu kopas datu kvalitātē bija, ka atribūtā "Vakcinētās personas dzimums" 28 vērtības ir tukšas, kā arī atribūtā "Vakcinētās personas vecums" 41 vērtībās ir kļūdas, jo vakcināciju neveic personām zem 16 gadu vecuma, kā arī atribūtā "Vakcinētās personas dzimums" 9 personām dzimums ir norādīts kā N.
2) ekonomikas un uzņēmējdarbības kategorija – "Traktortehnikas un tās piekabju ikgadējās valsts tehniskās apskates grafiki (plānotās)" , datu kopas datu kvalitātes problēmas var apskatīt 7. pielikuma 2. tabulā;
    - dati ir ļoti labā kvalitātē un nav novērojamas datu kvalitātes problēmas.
3) reģionu un pašvaldību kategorija – "Administratīvo teritoriju, teritoriālo vienību un statistisko (NUTS 3) reģionu klasifikators", datu kopas datu kvalitātes problēmas var apskatīt 7. pielikuma 3. tabulā;
    - datu kopas datiem ir novērojams trūkstošas vērtības vairākos atribūtos, kā "TV kods", "TV nosaukums" un citās.
4) iedzīvotāju un sabiedrības kategorija – "Ienākumu un dzīves apstākļu apsekojuma (EU-SILC) individuālie dati mācībām", datu kopas datu kvalitātes problēmas var apskatīt 7. pielikuma 4. tabulā;
    - liels pluss datu kopai bija tas, ka bija pievienots pdf fails, kurā bija aprakstīti parametri, lai saprastu, ko katrs nozīmē. Atribūtā "HH070", kaut gana aprakstā nav norādīts, ka nevar būt -1, tas atkārtojas 286 reizes.
5) valsts pārvaldes kategorija – "Vakances", datu kopas datu kvalitātes problēmas var apskatīt 7. pielikuma 5. tabulā;

- trūkst 217 atribūta "Slodzes tips" vērtības. Kā arī 661 vērtība trūkst atribūtā "Darba stundas nedēļā". Kā arī, ir novērojams atribūts "attēli", kur nav nevienas vērtības.

6) zemkopības, pārtikas un mežsaimniecības kategorija – "Traktortehnikas reģistrācijas dati Latvijas Republikā" apakškopa "Pirmo reizi reģistrēta lietota traktortehnika, tās piekabes un to markas", datu kopas datu kvalitātes problēmas var apskatīt 7. pielikuma 6. tabulā;
    - galvenā novērojamā datu kvalitātes problēma ir pustukša datu kopa, jo vairāk kā 60% no datu kopas ir tukšas vērtības.
7) izglītības un sporta kategorija - "Medikamenti, kas satur dopinga vielas", datu kopas datu kvalitātes problēmas var apskatīt 7. pielikuma 7. tabulā;
    - dati ir ļoti labā stāvoklī, un nav novērotas datu kvalitātes problēmas.
8) vides kategorija – "Dati par ūdens saimnieciskajiem iecirkņiem un ūdensteču garuma kategorijām", datu kopas datu kvalitātes problēmas var apskatīt 7. pielikuma 8. tabulā;
    - galvenās datu problēmas ir tādas, ka vairākos atribūtos ir sastopami tukšumi, kur vajadzētu būt vērtībai, kā arī 10 datu rindas ir nobīdījušās pa labi, tā vērtībām atrodoties nepareizajā vietā un jaucot visu datu kopu.
9) tieslietu, iekšlietu un drošības kategorija – "Darbinieku prasījumi", datu kopas datu kvalitātes problēmas var apskatīt 7. pielikuma 9. tabulā;
    - dati ir labā stāvoklī un nav novērotas datu kvalitātes problēmas.
10) transporta kategorija – "Traktortehnikas vadītāju izglītības iestādes", datu kopas datu kvalitātes problēmas var apskatīt 7. pielikuma 10. tabulā;
    - dati kopumā ir ļoti labi, bet vienīgā datu problēma ir tāda, ka trūkst viena e-pasta adrese.
11) zinātnes un tehnoloģiju kategorija – "Ar IKT jomu saistītās programmās studējošie Latvijā laika posmā no 2009.-2019.gadam", datu kopas datu kvalitātes problēmas var apskatīt 7. pielikum11. tabulā;
    - datu kopā ir novērojamas tukšās vērtības, "Imatrikulēti pavisam" ir rādījumi, kur nav norādīts nekas, kaut gan citur ir 0, tas pats attiecas uz atribūtu "Imatrikulēti budžetā" kā arī daudziem citiem atribūtiem.

12) kultūras kategorija – "Bibliotēku statistika", datu kopas datu kvalitātes problēmas var apskatīt 7. pielikuma 12. tabulā;
    - datu kopa ir ļoti haotiska un tajā ir ļoti daudz nepilnības, sākot ar to, ka nav aprakstīti parametri un tie nav pašsaprotami, tāpēc ir grūti uztvert informāciju. Lielākajai daļai atribūtu, kur nedrīkstētu būt nulles vērtības, tādas ir. Ir arī datu kvalitātes problēma nepareiza vērtība, 12 vērtībās atribūtā "Darbojās 2019. gadā", ir nepareizas, tās sastāv nevis no vien "jā", bet vairākiem "jā" pēc kārtas. Ir arī nederīga apakšvirkne atribūtā "Sākums" dažreiz ir norādīts datums, dažreiz apraksts.
13) ārlietu kategorija – "Saimniecisko darbību statistiskā klasifikācija Eiropas Kopienā, 2. redakcija", datu kopas datu kvalitātes problēmas var apskatīt 7. pielikuma 13. tabulā;
    - vienīgā atrastā datu problēma datu kopā ir 21 tukša vērtība atribūtā "vecaka_kods", un aprakstā tukšas ir 199 vērtības, bet šo nevarētu uzskatīt par datu problēmu, jo ir atļautas nulles vērtības.
14) enerģijas kategorija – "Valsts informācijas sistēmas, valsts informācijas resursi un IKT starpiestāžu pakalpojumi, to apraksti", datu kopas datu kvalitātes problēmas var apskatīt 7. pielikuma 14. tabulā;
    - galvenās datu problēmas šajā datu kopā bija tādas, ka apraksts nebija pieejams 4 vērtībām, kā arī tika lietoti sinonīmi VUGD un Valsts ugunsdzēsības un glābšanas dienests, kaut gan tas ir saīsinājumi.

Datu analīzē iegūtais rezultātu apkopojums ir redzami 4.2 tabulā. Šie dati arī ir tik kvalitatīvi, jo tika pētītas populārākās datu kopas, tas nozīmē, ka tās ir salīdzinoši labā kvalitātē, jo ļoti daudz cilvēki tās apskata un izmanto. Kā arī no iegūtajiem datiem, var redzēt, ka datu kopas, kurās nebija nekādas problēmas, bija tikai 3.

4.2. tabula

Latvijas atvērto datu portāla datu kopu kvalitātes problēmu apkopojums

N.p.k.	Datu kvalitātes problēma	Datu kopu skaits, kas to satur pret kopējo analizējamo datu kopu skaitu	Piemērs

1.	Trūkst vērtības	8/14	Atribūtā "Imatrikulēti pavisam" ir tukšās vērtības, kur nav norādīts nekas, kaut gan citur, ja ir 0 studenti imatrikulēti ir 0.
2.	Sintakses pārkāpums	0/14	-
3.	Nepareiza vērtība	4/14	Datu kopai "Ienākumu un dzīves apstākļu apsekojuma (EU-SILC) atribūtā "HH070", kaut gana atribūtu aprakstā ir norādīts, ka nevar būt -1, tas atkārtojas 286 reizes.
4.	Domēna pārkāpums	0/14	-
5.	Domēna ierobežojuma pārkāpšana	0/14	-
6.	Nederīga apakšvirkne	1/14	Datu kopai "Bibliotēku statistika" atribūtā "Sākums" dažreiz ir norādīts datums, dažreiz apraksts.
7.	Rakstības kļūda	0/14	-
8.	Neprecīza vērtība	0/14	-
9.	Unikāls vērtības pārkāpums	0/14	-
10.	Sinonīmu esamība	1/14	Datu kopai "Valsts informācijas sistēmas, valsts informācijas resursi un IKT starpiestāžu pakalpojumi, to apraksti" atribūtā "Iestāde" ir vērtība "Valsts ugunsdzēsības un

			glābšanas dienests", bet aprakstā minēts saīsinājums "VUGD"
11.	Pustukša datu kopa	1/14	Datu kopai "Traktortehnikas reģistrācijas dati Latvijas Republikā" Datu kopā ir vairāk nekā 60% tukšas vērtības.
12.	Darbības atkarības pārkāpums	0/14	-
13	Atsauces integritātes pārkāpums	0/14	-
14	Sintakšu neviendabīgums	0/14	-
15	Mērvienību neviendabīgums	0/14	-
16	Reprezentācijas neviendabīgums	0/14	-
17	Homonīma esamība	0/14	-

Izanalizējot iegūtos datus, galvenās datu kvalitātes problēmas pētītajās datu kopas ir, ka tām trūkst vērtības, ir nepareiza vērtība, kā arī ir nederīga apakšvirkne, ir sinonīmi un pustukša datu kopa. Tas ir skaidrojams arī ar to, ka šīs datu kvalitātes problēmas ir visvieglāk atrast un noteikt. Kā arī šīs datu kopas ir salīdzinoši mazas, tām pārsvarā bija ~10 atribūti, tas nozīmē, ka tās ir mazas datu kopas ar maz atribūtiem un tādās ir mazāk problēmas un datu kļūdas, nekā lielās datu noliktavās.

REZULTĀTI

Darba izstrādes laikā tika izpildīti visi nodefinētie uzdevumi: aplūkots "atvērto datu" jēdziens un ieguvumi no tā, aplūkots "datu kvalitātes" un "atvērto datu kvalitātes" jēdzieni, un datu kvalitātes analīzes tehnikas. Tika noteiktas populārākas datu kvalitātes problēmas, kā arī izpētīts, kā nosaka kvalitāti atvērto datu portālam, izanalizēts Latvijas atvērto datu portāls, tā kvalitāte, kā arī izpētīta atvērto datu kopu kvalitāte, izmantojot SQL balstītu kvalitātes analīzi.

Pētījumu veido divas daļas:

1. daļa bija Latvijas atvērto datu portāla kvalitātes analīze, kurā darba autore pētīja Latvijas atvērto datu portālu pēc citu autoru pētījumu rezultātā noteiktām 28 metrikām;
2. daļa bija Latvijas atvērto datu portāla datu kopu saturošos datu kvalitātes analīze, kurā darba autore pētīja katras kategorijas populārākās datu kopas kvalitāti.

Pētījumā tika noskaidrots, ka Latvijas atvērto datu portāls ir diezgan kvalitatīvs un tajā ir pieejama lielākā daļa no atvērto datu portāla vajadzīgajām funkcionalitātēm. Katra pētītā metrika tika novērtēta no 1-3, 1 nozīmē, ka tas nav pieejams portālā, bet 3, ka tas ir pieejams portālā un pilda savu funkcionalitāti bez problēmām. Pārsvarā viss bija augstā līmenī, bet bija arī tādas metrikas, kas nebija ievērotas, piemēram, portālā nav redzams lejupielāžu skaits jeb, cik reižu konkrētā datu kopa ir lejupielādēta, kā arī, portālā nav pieejams uzzināt datu kopas (faila) lielumu pirms tā lejupielādes. Nav arī pieejams atkārtotās izmantošanas skaits un forums lietotājiem. 4 no 28 metrikām ieguva vērtējumu 1, 13 no 28 metrikām ieguva vērtējumu 2, bet vērtējumu 3 ieguva 11 metrikas. Izrēķinātais vidējais vērtējums portālam ir 2,25, kas ir salīdzinoši augsts vērtējums.

Pētījuma otrajā daļā darba autore vērtēja 14 atvērto datu kopas un ieguva, ka tikai 3 datu kopās nebija nevienas no populārākajām datu problēmām, bet 8 no 14 datu kopām trūka kādas vērtības, 4 datu kopās bija kāda nepareiza vērtība, 1 datu kopā bija nederīga apakšvirkne, 1 datu kopā bija izmantoti sinonīmi, kā arī 1 datu kopa bija pustukša, kaut gan tika atrastas salīdzinoši daudz datu nepilnību, dati kopumā ir salīdzinoši kvalitatīvi un ir atkalizmantojami.

SECINĀJUMI

Darba izstrādes rezultātā autore secina:

1. lai datu varētu uzskatīt par atvērtiem – tiem jāatbilst 8 atvērto datu principi;
2. atvērtie dati dod ļoti daudz dažādus ieguvumus sabiedrībai, uzņēmumiem, ne tikai ietaupot naudas līdzekļus un laiku, bet glābjot pat dzīvības;
3. datiem var būt ļoti daudz datu kvalitātes problēmas, kuras ir jāapzinās un jāveic datu tīrīšana, lai nodrošinātu kvalitatīvus un izmantojamus datus;
4. lai uzlabotu atvērto datu izmantošanu, ir svarīgi, lai atvēro datu portāls ir kvalitatīvs un tajā ir pieejamas un augstā līmenī nodrošinātas visas nepieciešamās funkcionalitātes;
5. Latvijas atvērto datu portāls ir labā kvalitātes līmenī, kaut gan tam varētu veikt uzlabojumus, jo nav pieejams dažas svarīgas funkcionalitātes;
6. Latvijas atvērtā portāla saturošo datu kvalitāte ir salīdzinoši laba, bet tiem ir novērojamas kvalitātes problēmas, kuras vajadzētu izlabot;
7. visbiežāk izplatītākā Latvijas atvērto datu kopu saturošā datu problēma ir, ka dati nav pilnīgi, kā arī tie satur vērtības, kas nav atļautas;
8. datu kopām reti tiek aprakstīti parametri, kas traucē datu lietojamību, t.i. lietotājam ir pašam jāveic pieņēmumi par to, ko katrs konkrēts parametrs nozīmē, kas ir tajā glabājamās vērtības;
9. kaut gan atvērto datu kopu datiem ir novērojamas dažas nepilnības, tie ir atkalizmantojami.

Iegūtie rezultāti apkopotā veidā tiks sniegti datu sniedzējam, lai varētu uzlabot gan datu portāla, gan datu kvalitāti.

Darba sākumā nodefinētie uzdevumi ir izpildīti un uzstādītais mērķis ir sasniegts.

IZMANTOTĀ LITERATŪRA

1. Latvijs avērto datu portāls, [tiešsaiste] – [atsauce 01.05.2021.] Pieejams: https://data.gov.lv/lv
2. Eiropas avērto datu portāls, [tiešsaiste] – [atsauce 15.05.2021.] Pieejams: https://data.europa.eu/en
3. What is Open Data? ,[tiešsaiste] – [atsauce 25.04.2021.] Pieejams: https://opendatahandbook.org/guide/en/what-is-open-data/
4. 5 ★ OPEN DATA [tiešsaiste] – [atsauce 03.01.2020.] Pieejams: https://5stardata.info/en/
5. Thorsby, J., Stowers, G. N., Wolslegel, K., & Tumbuan, E. (2017). Understanding the content and features of open data portals in American cities. Government Information Quarterly, 34(1), 53-61.
6. Global open data index [tiešsaiste] – [atsauce 10.05.2021.] Pieejams: https://index.okfn.org/place/
7. Open Government Data Principles [tiešsaiste] – [atsauce 15.05.2021.] Pieejams: https://public.resource.org/8_principles.html
8. What Is Data Anonymization [tiešsaiste] – [atsauce 15.05.2021.] Pieejams: https://www.imperva.com/learn/data-security/anonymization/
9. Petrova-Antonova, D., & Tancheva, R. (2020, September). Data Cleaning: A Case Study with OpenRefine and Trifacta Wrangler. In International Conference on the Quality of Information and Communications Technology (pp. 32-40). Springer, Cham.
10. ISO 8000-61:2016. [tiešsaiste] – [atsauce 20.05.2021.] Pieejams: https://www.dpadvantage.co.uk/2020/02/05/iso-8000-61-the-data-quality-management-standard/
11. Open Data in Europe 2020. ,[tiešsaiste] – [atsauce 16.05.2021.] Pieejams: https://data.europa.eu/en/dashboard/2020
12. Beāte Beizaka (2020). Kursa darbs “Latvijas atvērto datu kvalitātes analīze”.
13. Data vs. Information [tiešsaiste] – [atsauce 05.05.2021.] Pieejams: https://www.diffen.com/difference/Data_vs_Information
14. The knowledge continuum: how data, information, and intelligence work together. [tiešsaiste] – [atsauce 07.05.2021.] Pieejams: https://www.introhive.com/resources/difference-between-data-information-intelligence/

15. Datu pārvaldība. [tiešsaiste] – [atsauce 12.05.2021.] Pieejams: https://www.varam.gov.lv/lv/datu-parvaldiba

16. Open data maturity report 2020. [tiešsaiste] – [atsauce 13.05.2021.] Pieejams: https://data.europa.eu/sites/default/files/edp_landscaping_insight_report_n6_2020.pdf

17. Open data maturity 2020, Latvia. [tiešsaiste] – [atsauce 13.05.2021.] Pieejams: https://data.europ.eu/sites/default/files/country-factsheet_latvia_2020.pdf

18. Open Government Data. [tiešsaiste] – [atsauce 18.05.2021.] Pieejams: https://www.oecd.org/gov/digital-government/open-government-data.htm

19. Formats. [tiešsaiste] – [atsauce 18.05.2021.] Pieejams: https://data.gov.ie/formats

20. Linstedt, D., & Olschimke, M. (2015). Building a scalable data warehouse with data vault 2.0. Morgan Kaufmann.

21. Oliveira, P., Rodrigues, F., & Henriques, P. R. (2005, November). A formal definition of data quality problems. In ICIQ.

22. ISO/IEC 25012. [tiešsaiste] – [atsauce 20.05.2021.] Pieejams: https://iso25000.com/index.php/en/iso-25000-standards/iso-25012

23. Máchová, R., & Lněnička, M. (2017). Evaluating the quality of open data portals on the national level. Journal of theoretical and applied electronic commerce research, 12(1), 21-41.

24. Schauppenlehner, T., & Muhar, A. (2018). Theoretical availability versus practical accessibility: The critical role of metadata management in open data portals. Sustainability, 10(2), 545.

PIELIKUMI

1. pielikums

Latvijas atvērto datu portāla datu kopas sadalītas pa kategorijām

Kategorija	Datu kopu skaits
Ekonomika un uzņēmējdarbība	108
Reģioni un pašvaldības	99
Iedzīvotāji un sabiedrība	90
Valsts pārvalde	77
Zemkopība, pārtika un mežsaimniecība	41
Veselība	25
Vide	23
Izglītība un sports	21
Tieslietas, iekšlietas un drošība	16
Transports	16
Zinātne un tehnoloģijas	15
Kultūra	13
Ārlietas	7
Enerģija	3

2. pielikums

Latvijas atvērto datu portāla datu kopu populārākie publicētāji

Organizācija	Datu kopu skaits
ĢEOLatvija.lv	42
Centrālā statistikas pārvalde	40
Valsts reģionālās attīstības aģentūra	28
Būvniecības valsts kontroles birojs	28
Iepirkumu uzraudzības birojs	24
Vides aizsardzības un reģionālās attīstības ministrija	20
VAS "Latvijas Valsts ceļi"	15
Rīgas dome	13
Cēsu novada pašvaldība	13
Labklājības ministrija	12

3. pielikums

Latvijas atvērto datu portāla datu kopu populārākie formāti

Nr.p.k.	Formāts	Datu kopu skaits	Atbilstība 5 – zvaigžņu atvērto datu klasifikācijai
1.	CSV	230	3
2.	XLSX	124	3
3.	XLS	33	2
4.	WMS	31	2
5.	JSON	26	3
6.	OData	19	3
7.	SHP	17	3
8.	DOCX	14	1
9.	ZIP	14	1
10.	XML	11	3
11.	PDF	9	1

4. pielikums

Latvijas atvērto datu portāla populārākās birkas

Birka	Datu kopu skaits
Statistika	25
2020	21
Pilsēta	20
2019	19
Administratīvā teritorija	19
2018	18
BIS	18
novads	18
lauksaimniecība	17
2017	16

5. pielikums

Latvijas atvērto datu veselības kategorijas datu kopu kvalitātes analīze

1. tabula

Latvijas atvērto datu veselības kategorijas datu kopu kvalitātes atbilstība (autores [12] apkopojums)

Datu kvalitātes vērtība	Atbilst/Ir (datu kopu skaits)	Neatbilst/Nav (datu kopu skaits)
Datu kopu atjaunošana	13	11
Datu kopas apraksta esamība	22	2
Parametru apraksta esamība	6	18
Priekšskatījums datiem	19	5

2. tabula

Izmantotie datu formāti [12]

Datu formāts	**Datu kopu skaits**
CSV	13
XLSX	10
JSON	5
ZIP	4
TXT	3
XML	2
XLS	2
HTML	2
ODS	1

6. pielikums

Latvijas atvērto datu portāla populārāko datu kopu analīze

Publicētājs	Datu kopa	Atbilstība atjaunošanas biežumam (1-atbilst, 0-neatbilst)	Datu kopas apraksta esamība (0 - nav, 1 - ir)	Parametru apraksta esamība (0 - nav, 1 - ir)	Datu formāts/i	Datu formāta atbilstība 5-zvaigžņu klasifikācijai	Priekšskatījums->datu tabula {0-nav, 1-ir}
Nacionālais veselības dienests	COVID19 vakcinācijas	1	1	0	XLSX	3	1
Valsts tehniskās uzraudzības aģentūra	Traktortehnikas un tās piekabju ikgadējās valsts tehniskās apskates grafiki (plānotās)	1	1	0	CSV, XLSX	3	1
Centrālā statistikas pārvalde	Administratīvo teritoriju, teritoriālo vienību un statistisko (NUTS 3) reģionu klasifikators	1	0	0	CSV, JSON	3	1
Centrālā statistikas pārvalde	Ienākumu un dzīves apstākļu apsekojuma (EU-SILC) individuālie dati mācībām	1	1	1	PDF, ZIP, CSV	3	1

Nodarbinātības Valsts Aģentūra	Vakances	0	1	0	CSV	3	1
Valsts tehniskās uzraudzības aģentūra	Traktortehnikas reģistrācijas dati Latvijas Republikā	1	1	0	CSV, XLSX	3	1
Latvijas Antidopinga birojs	Medikamenti, kas satur dopinga vielas	1	1	0	CSV	3	1
Vides aizsardzības un reģionālās attīstības ministrija	Dati par ūdens saimnieciskajiem iecirkņiem un ūdensteču garuma kategorijām	1	1	0	CSV, JSON, SHP	3	1
Maksātnespējas kontroles dienests	Darbinieku prasījumi	0	1	0	CSV, XLSX	3	1
Valsts tehniskās uzraudzības aģentūra	Traktortehnikas vadītāju izglītības iestādes	1	1	0	CSV, XLSX	3	1
Izglītības un zinātnes ministrija	Ar IKT jomu saistītās programmās studējošie Latvijā laika posmā no 2009.-2019.gadam	1	1	0	XLSX	3	1

Kultūras informācijas sistēmu centrs	Bibliotēku statistika	0	1	0	CSV	3	1
Centrālā statistikas pārvalde	Saimniecisko darbību statistiskā klasifikācija Eiropas Kopienā, 2. redakcija	1	1	0	CSV, TSV, JSON	3	1
Vides aizsardzības un reģionālās attīstības ministrija	Valsts informācijas sistēmas, valsts informācijas resursi un IKT starpiestāžu pakalpojumi, to apraksti	0	1	0	CSV	3	1

7. pielikums

Datu kvalitātes analīzes populārākajām datu kopām

1. tabula

Datu kvalitātes analīze datu kopai "COVID19 vakcinācijas"

N.p.k.	Datu kvalitātes problēma	Ir/nav sastopasms	Pakaidrojums
1.	Trūkst vērtības	Ir	Atribūtā "Vakcinētās personas dzimums" 28 vērtības ir tukšas
2.	Sintakses pārkāpums	Nav	
3.	Nepareiza vērtība	Ir	Atribūtā "Vakcinētās personas vecums" 41 vērtībās ir kļūdas, jo vakcināciju neveic personām zem 16 gadu vecuma. Atribūtā "Vakcinētās personas dzimums" 9 personām dzimums ir norādīts kā N.
4.	Domēna pārkāpums	Nav	
5.	Domēna ierobežojuma pārkāpšana	Nav	
6.	Nederīga apakšvirkne	Nav	
7.	Rakstības kļūda	Nav	
8.	Neprecīza vērtība	Nav	
9.	Unikāls vērtības pārkāpums	Nav	
10.	Sinonīmu esamība	Nav	
11.	Pustukša datu kopa	Nav	
12.	Darbības atkarības pārkāpums	Nav	
13	Atsauces integritātes pārkāpums	Nav	
14	Sintakšu neviendabīgums	Nav	
15	Mērvienību neviendabīgums	Nav	
16	Reprezentācijas neviendabīgums	Nav	
17	Homonīma esamība	Nav	

2. tabula

Datu kvalitātes analīze datu kopai "Traktortehnikas un tās piekabju ikgadējās valsts tehniskās apskates grafiki (plānotās)"

N.p.k.	Datu kvalitātes problēma	Ir/nav sastopasms	Pakaidrojums
1.	Trūkst vērtības	Nav	
2.	Sintakses pārkāpums	Nav	
3.	Nepareiza vērtība	Nav	
4.	Domēna pārkāpums	Nav	
5.	Domēna ierobežojuma pārkāpšana	Nav	
6.	Nederīga apakšvirkne	Nav	
7.	Rakstības kļūda	Nav	
8.	Neprecīza vērtība	Nav	
9.	Unikāls vērtības pārkāpums	Nav	
10.	Sinonīmu esamība	Nav	
11.	Pustukša datu kopa	Nav	
12.	Darbības atkarības pārkāpums	Nav	
13	Atsauces integritātes pārkāpums	Nav	
14	Sintakšu neviendabīgums	Nav	
15	Mērvienību neviendabīgums	Nav	
16	Reprezentācijas neviendabīgums	Nav	
17	Homonīma esamība	Nav	

3. tabula

Datu kvalitātes analīze datu kopai "Administratīvo teritoriju, teritoriālo vienību un statistisko (NUTS 3) reģionu klasifikators"

N.p.k.	Datu kvalitātes problēma	Ir/nav sastopasms	Pakaidrojums
1.	Trūkst vērtības	Ir	Novērojams trūkstošas vērtības vairākos atribūtos, kā "TV kods", "TV nosaukums" un citās.
2.	Sintakses pārkāpums	Nav	
3.	Nepareiza vērtība	Nav	
4.	Domēna pārkāpums	Nav	
5.	Domēna ierobežojuma pārkāpšana	Nav	
6.	Nederīga apakšvirkne	Nav	
7.	Rakstības kļūda	Nav	
8.	Neprecīza vērtība	Nav	
9.	Unikāls vērtības pārkāpums	Nav	
10.	Sinonīmu esamība	Nav	
11.	Pustukša datu kopa	Nav	
12.	Darbības atkarības pārkāpums	Nav	
13	Atsauces integritātes pārkāpums	Nav	
14	Sintakšu neviendabīgums	Nav	
15	Mērvienību neviendabīgums	Nav	
16	Reprezentācijas neviendabīgums	Nav	
17	Homonīma esamība	Nav	

4. tabula

Datu kvalitātes analīze datu kopai "Ienākumu un dzīves apstākļu apsekojuma (EU-SILC) individuālie dati mācībām"

N.p.k.	Datu kvalitātes problēma	Ir/nav sastopasms	Pakaidrojums
1.	Trūkst vērtības	Nav	
2.	Sintakses pārkāpums	Nav	
3.	Nepareiza vērtība	Ir	Atribūtā "HH070", kaut gana atribūtu aprakstā ir norādīts, ka nevar būt -1, tas atkārtojas 286 reizes.
4.	Domēna pārkāpums	Nav	
5.	Domēna ierobežojuma pārkāpšana	Nav	
6.	Nederīga apakšvirkne	Nav	
7.	Rakstības kļūda	Nav	
8.	Neprecīza vērtība	Nav	
9.	Unikāls vērtības pārkāpums	Nav	
10.	Sinonīmu esamība	Nav	
11.	Pustukša datu kopa	Nav	
12.	Darbības atkarības pārkāpums	Nav	
13	Atsauces integritātes pārkāpums	Nav	
14	Sintakšu neviendabīgums	Nav	
15	Mērvienību neviendabīgums	Nav	
16	Reprezentācijas neviendabīgums	Nav	
17	Homonīma esamība	Nav	

5. tabula

Datu kvalitātes analīze datu kopai "Vakances"

N.p.k.	Datu kvalitātes problēma	Ir/nav sastopasms	Pakaidrojums
1.	Trūkst vērtības	Ir	Trūkst 217 atribūta "Slodzes tips" vērtības. Kā arī 661 vērtība trūkst atribūtā "Darba stundas nedēļā".
2.	Sintakses pārkāpums	Nav	
3.	Nepareiza vērtība	Nav	
4.	Domēna pārkāpums	Nav	
5.	Domēna ierobežojuma pārkāpšana	Nav	
6.	Nederīga apakšvirkne	Nav	
7.	Rakstības kļūda	Nav	
8.	Neprecīza vērtība	Nav	
9.	Unikāls vērtības pārkāpums	Nav	
10.	Sinonīmu esamība	Nav	
11.	Pustukša datu kopa	Nav	
12.	Darbības atkarības pārkāpums	Nav	
13	Atsauces integritātes pārkāpums	Nav	
14	Sintakšu neviendabīgums	Nav	
15	Mērvienību neviendabīgums	Nav	
16	Reprezentācijas neviendabīgums	Nav	
17	Homonīma esamība	Nav	

6. tabula

Datu kvalitātes analīze datu kopai "Traktortehnikas reģistrācijas dati Latvijas Republikā"

Apakškopa "Pirmo reizi reģistrēta lietota traktortehnika, tās piekabes un to markas"

N.p.k.	Datu kvalitātes problēma	Ir/nav sastopasms	Pakaidrojums
1.	Trūkst vērtības	Nav	
2.	Sintakses pārkāpums	Nav	
3.	Nepareiza vērtība	Nav	
4.	Domēna pārkāpums	Nav	
5.	Domēna ierobežojuma pārkāpšana	Nav	
6.	Nederīga apakšvirkne	Nav	
7.	Rakstības kļūda	Nav	
8.	Neprecīza vērtība	Nav	
9.	Unikāls vērtības pārkāpums	Nav	
10.	Sinonīmu esamība	Nav	
11.	Pustukša datu kopa	Ir	Datu kopā ir vairāk kā 60% tukšas vērtības.
12.	Darbības atkarības pārkāpums	Nav	
13	Atsauces integritātes pārkāpums	Nav	
14	Sintakšu neviendabīgums	Nav	
15	Mērvienību neviendabīgums	Nav	
16	Reprezentācijas neviendabīgums	Nav	
17	Homonīma esamība	Nav	

7. tabula

Datu kvalitātes analīze datu kopai "Medikamenti, kas satur dopinga vielas"

N.p.k.	Datu kvalitātes problēma	Ir/nav sastopasms	Pakaidrojums
1.	Trūkst vērtības	Nav	
2.	Sintakses pārkāpums	Nav	
3.	Nepareiza vērtība	Nav	
4.	Domēna pārkāpums	Nav	
5.	Domēna ierobežojuma pārkāpšana	Nav	
6.	Nederīga apakšvirkne	Nav	
7.	Rakstības kļūda	Nav	
8.	Neprecīza vērtība	Nav	
9.	Unikāls vērtības pārkāpums	Nav	
10.	Sinonīmu esamība	Nav	
11.	Pustukša datu kopa	Nav	
12.	Darbības atkarības pārkāpums	Nav	
13	Atsauces integritātes pārkāpums	Nav	
14	Sintakšu neviendabīgums	Nav	
15	Mērvienību neviendabīgums	Nav	
16	Reprezentācijas neviendabīgums	Nav	
17	Homonīma esamība	Nav	

8. tabula

Datu kvalitātes analīze datu kopai "Dati par ūdens saimnieciskajiem iecirkņiem un ūdensteču garuma kategorijām"

N.p.k.	Datu kvalitātes problēma	Ir/nav sastopasms	Pakaidrojums
1.	Trūkst vērtības	Ir	Vairākos atribūtos ir sastopami tukšumi, kur vajadzētu būt vērtībai.
2.	Sintakses pārkāpums	Nav	
3.	Nepareiza vērtība	Ir	10 datu rindas ir nobīdījušās pa labi, tā vērtībām atrodoties nepareizajā vietā.
4.	Domēna pārkāpums	Nav	
5.	Domēna ierobežojuma pārkāpšana	Nav	
6.	Nederīga apakšvirkne	Nav	
7.	Rakstības kļūda	Nav	
8.	Neprecīza vērtība	Nav	
9.	Unikāls vērtības pārkāpums	Nav	
10.	Sinonīmu esamība	Nav	
11.	Pustukša datu kopa	Nav	
12.	Darbības atkarības pārkāpums	Nav	
13	Atsauces integritātes pārkāpums	Nav	
14	Sintakšu neviendabīgums	Nav	
15	Mērvienību neviendabīgums	Nav	
16	Reprezentācijas neviendabīgums	Nav	
17	Homonīma esamība	Nav	

9. tabula

Datu kvalitātes analīze datu kopai "Darbinieku prasījumi"

N.p.k.	Datu kvalitātes problēma	Ir/nav sastopasms	Pakaidrojums
1.	Trūkst vērtības	Nav	
2.	Sintakses pārkāpums	Nav	
3.	Nepareiza vērtība	Nav	
4.	Domēna pārkāpums	Nav	
5.	Domēna ierobežojuma pārkāpšana	Nav	
6.	Nederīga apakšvirkne	Nav	
7.	Rakstības kļūda	Nav	
8.	Neprecīza vērtība	Nav	
9.	Unikāls vērtības pārkāpums	Nav	
10.	Sinonīmu esamība	Nav	
11.	Pustukša datu kopa	Nav	
12.	Darbības atkarības pārkāpums	Nav	
13	Atsauces integritātes pārkāpums	Nav	
14	Sintakšu neviendabīgums	Nav	
15	Mērvienību neviendabīgums	Nav	
16	Reprezentācijas neviendabīgums	Nav	
17	Homonīma esamība	Nav	

10. tabula

Datu kvalitātes analīze datu kopai "Traktortehnikas vadītāju izglītības iestādes"

N.p.k.	Datu kvalitātes problēma	Ir/nav sastopasms	Pakaidrojums
1.	Trūkst vērtības	Ir	Viena e-pasta adrese trūkst.
2.	Sintakses pārkāpums	Nav	
3.	Nepareiza vērtība	Nav	
4.	Domēna pārkāpums	Nav	
5.	Domēna ierobežojuma pārkāpšana	Nav	
6.	Nederīga apakšvirkne	Nav	
7.	Rakstības kļūda	Nav	
8.	Neprecīza vērtība	Nav	
9.	Unikāls vērtības pārkāpums	Nav	
10.	Sinonīmu esamība	Nav	
11.	Pustukša datu kopa	Nav	
12.	Darbības atkarības pārkāpums	Nav	
13	Atsauces integritātes pārkāpums	Nav	
14	Sintakšu neviendabīgums	Nav	
15	Mērvienību neviendabīgums	Nav	
16	Reprezentācijas neviendabīgums	Nav	
17	Homonīma esamība	Nav	

11. tabula

Datu kvalitātes analīze datu kopai "Ar IKT jomu saistītās programmās studējošie Latvijā laika posmā no 2009.-2019.gadam"

N.p.k.	Datu kvalitātes problēma	Ir/nav sastopasms	Pakaidrojums
1.	Trūkst vērtības	Ir	"Imatrikulēti pavisam" ir tukšās vērtības, kur nav norādīts nekas, kaut gan citur ir 0, tas pats attiecas uz atribūtu "Imatrikulēti budžets" kā arī daudziem citiem atribūtiem.
2.	Sintakses pārkāpums	Nav	
3.	Nepareiza vērtība	Nav	
4.	Domēna pārkāpums	Nav	
5.	Domēna ierobežojuma pārkāpšana	Nav	
6.	Nederīga apakšvirkne	Nav	
7.	Rakstības kļūda	Nav	
8.	Neprecīza vērtība	Nav	
9.	Unikāls vērtības pārkāpums	Nav	
10.	Sinonīmu esamība	Nav	
11.	Pustukša datu kopa	Nav	
12.	Darbības atkarības pārkāpums	Nav	
13	Atsauces integritātes pārkāpums	Nav	
14	Sintakšu neviendabīgums	Nav	
15	Mērvienību neviendabīgums	Nav	
16	Reprezentācijas neviendabīgums	Nav	
17	Homonīma esamība	Nav	

12. tabula

Datu kvalitātes analīze datu kopai "Bibliotēku statistika"

N.p.k.	Datu kvalitātes problēma	Ir/nav sastopasms	Pakaidrojums
1.	Trūkst vērtības	Ir	Lielākajai daļai atribūtu, kur nedrīkstētu būt nulles vērtības, tādas ir.
2.	Sintakses pārkāpums	Nav	
3.	Nepareiza vērtība	Ir	12 vērtībās atribūtā "Darbojās 2019. gadā", ir nepareizas, tās sastāv nevis no vien "jā", bet vairākiem "jā" pēc kārtas
4.	Domēna pārkāpums	Nav	
5.	Domēna ierobežojuma pārkāpšana	Nav	
6.	Nederīga apakšvirkne	Ir	Atribūtā "Sākums" dažreiz ir norādīts datums, dažreiz apraksts.
7.	Rakstības kļūda	Nav	
8.	Neprecīza vērtība	Nav	
9.	Unikāls vērtības pārkāpums	Nav	
10.	Sinonīmu esamība	Nav	
11.	Pustukša datu kopa	Nav	
12.	Darbības atkarības pārkāpums	Nav	
13	Atsauces integritātes pārkāpums	Nav	
14	Sintakšu neviendabīgums	Nav	
15	Mērvienību neviendabīgums	Nav	
16	Reprezentācijas neviendabīgums	Nav	
17	Homonīma esamība	Nav	

13. tabula

Datu kvalitātes analīze datu kopai "Saimniecisko darbību statistiskā klasifikācija Eiropas Kopienā, 2. redakcija"

N.p.k.	Datu kvalitātes problēma	Ir/nav sastopasms	Pakaidrojums
1.	Trūkst vērtības	Nav	21 tukša vērtība atribūtā "vecaka_kods" un aprakstā tukšas ir 199 vērtības, bet šo nevar uzskatīt par datu problēmu, jo ir atļautas nulles vērtības.
2.	Sintakses pārkāpums	Nav	
3.	Nepareiza vērtība	Nav	
4.	Domēna pārkāpums	Nav	
5.	Domēna ierobežojuma pārkāpšana	Nav	
6.	Nederīga apakšvirkne	Nav	
7.	Rakstības kļūda	Nav	
8.	Neprecīza vērtība	Nav	
9.	Unikāls vērtības pārkāpums	Nav	
10.	Sinonīmu esamība	Nav	
11.	Pustukša datu kopa	Nav	
12.	Darbības atkarības pārkāpums	Nav	
13	Atsauces integritātes pārkāpums	Nav	
14	Sintakšu neviendabīgums	Nav	
15	Mērvienību neviendabīgums	Nav	
16	Reprezentācijas neviendabīgums	Nav	
17	Homonīma esamība	Nav	

14. tabula

Datu kvalitātes analīze datu kopai "Valsts informācijas sistēmas, valsts informācijas resursi un IKT starpiestāžu pakalpojumi, to apraksti"

N.p.k.	Datu kvalitātes problēma	Ir/nav sastopasms	Pakaidrojums
1.	Trūkst vērtības	Nav	Atribūtā "Apraksts" 4 vērtības ir tukšas
2.	Sintakses pārkāpums	Nav	
3.	Nepareiza vērtība	Nav	
4.	Domēna pārkāpums	Nav	
5.	Domēna ierobežojuma pārkāpšana	Nav	
6.	Nederīga apakšvirkne	Nav	
7.	Rakstības kļūda	Nav	
8.	Neprecīza vērtība	Nav	
9.	Unikāls vērtības pārkāpums	Nav	
10.	Sinonīmu esamība	Ir	Atribūtā "Iestāde" ir vērtība "Valsts ugunsdzēsības un glābšanas dienests", bet aprakstā minēts saīsinājums "VUGD"
11.	Pustukša datu kopa	Nav	
12.	Darbības atkarības pārkāpums	Nav	
13	Atsauces integritātes pārkāpums	Nav	
14	Sintakšu neviendabīgums	Nav	
15	Mērvienību neviendabīgums	Nav	
16	Reprezentācijas neviendabīgums	Nav	
17	Homonīma esamība	Nav	

Bakalaura darbs „ Latvijas atvērto datu portāla un tā datu kopu kvalitātes analīze” izstrādāts LU Datorikas fakultātē.

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

Autors: _____ Beāte Beizaka 31.05.2021.

Rekomendēju darbu aizstāvēšanai
Vadītāja: Dr. dat. Anastasija Ņikiforova _____ 31.05.2021.

Recenzents: asociētais profesors Dr.sc.comp. Vineta Arnicāne

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: docents Dr.sc.comp. Ivo Odītis (_____)