LATVIJAS UNIVERSITĀTE
DATORIKAS FAKULTĀTE

LATVIJAS ATVĒRTO IZGLĪTĪBAS DATU KVALITĀTES ANALĪZE

BAKALAURA DARBS

Autors: **Arta Kristapšteina**

Studenta apliecības Nr.: ak18174

Darba vadītājs: profesors Dr. dat. Jānis Bičevskis

RĪGA 2022

ANOTĀCIJA

Darbā tiek analizēta Latvijas Atvērto datu portāla izglītības datu kopu kvalitāte. Tiek dots pārskats par atvērto datu un to kvalitātes jēdzieniem, aktualitāti un problēmām. Novērtējot atvērto izglītības datu kopu kvalitāti, tiek pielietota jauna Latvijas Universitātes zinātnieku piedāvāta datu kvalitātes novērtēšanas pieeja. Tā būtiski atšķiras no tradicionālās datu kvalitātes novērtēšanas, kas pielieto dimensijas jēdzienus. Jaunā metode piedāvā datu kvalitāti vērtēt datu lietojumu kontekstā. Tiek parādītas izglītības atvērto datu konkrētas kvalitātes problēmas, to ietekme uz šo datu lietojumiem un piedāvātas datu kvalitātes uzlabošanas metodes.

Atslēgvārdi: atvērtie dati, datu kvalitāte, datu kvalitātes novērtēšana, Latvijas Atvērto datu portāls

ABSTRACT

ANALYSIS OF LATVIAN OPEN EDUCATION DATA QUALITY

The aim of this work is to analyse the quality of the education datasets of the Latvian Open Data portal. An overview of the concepts, relevance and problems of open data and their quality is given. When assessing the quality of the open educational data sets, a new data quality assessment approach proposed by scientists at the University of Latvia is applied. It differs significantly from traditional data quality assessment which uses dimensional concepts. The new method offers to assess data quality in the context of use-cases. Specific quality problems of open-education data, their impact on the use-cases, and methods of improving data quality are presented.

Keywords: open data, data quality, data quality assessment, Latvia Open Data portal

SATURA RĀDĪTĀJS

APZĪMĒJUMU SARAKSTS

Apzīmējums	Skaidrojums
ADSL	Domain Specific Language, domēnspecifiskā valoda
CSV	Teksta faila formāts ar atdalītājiem, kas vērtību atdalīšanai izmanto komatu
DAMA UK Working Group	Data Management Association International UK Working Group
DQ	Data quality, datu kvalitāte
IT	Information technology, informācijas tehnoloģija
Odata	Atvērts protokols, kas ļauj izveidot un patērēt vaicājamus
PIM	Platform independent model, platformas neatkarīgs modelis (PNM)
PSM	Platform specific model, platformas atkarīgs modelis (PAM)
SQL	Vaicājumu valoda, kas paredzēta datu manipulēšanai relāciju datubāžu pārvaldības sistēmās
TDQM	Total Data Quality Management
VIIS	Valsts izglītības informācijas sistēma
VIRSIS	Valsts informācijas sistēmu savietotājs
XLSX	Faila formāts, kas izmanto tabulas *Microsoft Excel* izklājlapās, lai kārtotu, analizētu un saglabātu datus.

IEVADS

Mūsdienās pieaugot datu apjomiem un lietojumiem, nozīmīgāks kļūst datu kvalitātes jēdziens un prasības. Tā vietā, lai ierobežotu piekļuvi datiem, arvien vairāk tie tiek atklāti plašākai sabiedrībai, tādējādi paaugstinot prasības datu kvalitātei.

Atvērto datu jēdziens apraksta datus, kas, atkarībā no konkrētās valsts likumiem, ir brīvi pieejami sabiedrībai izmantošanai, pārpublicēšanai un izplatīšanai. Rezultātā atvērtie dati ir nozīmīgs informācijas tehnoloģiju instruments ekonomikas, sociālās un sabiedrības izaugsmes veicināšanai. Daudzās valstīs, tostarp Latvijā, ir izveidots atvērto datu portāls, kurā datus var publicēt un izmantot informācijas iegūšanai, vai jaunu risinājumu un pakalpojumu izstrādei.

Lai gan ir daudzi inovācijas panākumi, kuru pamatā ir atvērtie dati, pastāv nenoteiktība par datu kopu datu kvalitāti. Tādēļ ir svarīgi ne tikai veidot un publicēt datus, bet arī rūpīgi pārbaudīt un izvērtēt to kvalitāti, lai turpinātu uzlabot šos procesus visā pasaulē.

Darba mērķis ir veikt Latvijas Atvērto datu portāla izglītības datu kopu kvalitātes analīzi, pielietojot jauno unikālo datu kvalitātes novērtēšanas pieeju, rezultātā veicot secinājumus par izglītības datu kvalitāti, piedāvājot tās uzlabošanas metodes.

Darba gaitā tika izvirzīti sekojošie uzdevumi:

1) izpētīt “datu kvalitātes” un “atvērto datu” jēdzienus;
2) izpētīt datu kvalitātes nozīmīgumu un to ietekmi;
3) izpētīt atvērto izglītības datu nozīmi;
4) aplūkot Latvijas Atvērto datu portālu;
5) aplūkot un novērtēt esošos datu kvalitātes novērtēšanas risinājumus, nosakot to priekšrocības un trūkumus;
6) veikt izglītības datu kopu kvalitātes analīzi, pielietojot lietojumvirzīto datu kvalitātes novērtēšanas pieeju;
7) veikt secinājumus par atvērto izglītības datu kopu kvalitāti, piedāvājot ieteikumus kvalitātes uzlabošanai.

Darba teorētiskajā daļā tiek apskatīta un analizēta literatūra par datu kvalitāti, tās aktualitāti, atvērto izglītības datu aktualitāti, kā arī aplūkoti eksistējošie risinājumi datu kvalitātes analīzei un novērtēšanai.

Darba praktiskajā daļā tiek veikta Latvijas Atvērto datu portāla izglītības datu kopu kvalitātes analīze 3 kārtās – ortogrāfijas, sintakses un konteksta līmeņos.

Darbā tika izmantotas pētniecības metodes:

1. analītiskā metode – informācijas avotu analīze, lai izpētītu datu kvalitātes un atvērto datu jēdzienus, to nozīmīgumu, veikt atvērto izglītības datu analīzi;

2. salīdzinošā metode – datu kvalitātes dimensiju un lietojumorientētās pieejas salīdzināšana;
3. eksperimentālā metode – lietojumu orientētas datu kvalitātes pieeja atvērto izglītības datu kopu analīzei;
4. aprakstošā metode – literatūras izpēte, tās apkopošana un aprakstīšana, darba gaitas un iegūto rezultātu aprakstīšana.

Darbs sastāv no ievada, pamatdaļas ar 6 sadaļām un 18 apakšsadaļām, un secinājumiem. Darbā ir iekļauti 39 literatūras avoti, 9 attēli un 5 tabulas.

1. DATU KVALITĀTE

1.1. Datu kvalitātes jēdziens

Datu kvalitāte ir datu kopas un tās īpašību piemērotība konkrētam jautājumam, uzdevumam jeb lietošanas piemēram, kas ir atkarīgs no datu lietotāja jeb datu patērētāja [1].

Pētījumos datu kvalitāte ir norādīta kā daudzdimensiju koncepcija. Bieži tiek minētas īpašības, kā precizitāte, pilnīgums, konsekvence un savlaicīgums. Šo īpašību izvēle, galvenokārt, balstās uz intuitīvu izpratni, nozares pieredzi vai literatūras pārskatiem [2]. Tomēr Vanga (Wang) un Storeja (Storey) literatūras pārskats rāda, ka nav vispārējas vienošanās par datu kvalitātes dimensijām [3]. Tāpat jēdzienam "datu kvalitāte" nav konkrētas definīcijas, tāpēc tiek piedāvātas vairākas jēdzienu raksturojošas alternatīvas.

Tayi terminu "datu kvalitāte" raksturo kā "piemērotību lietošanai", kas nozīmē, ka datu kvalitāte ir relatīvs jēdziens. Rezultātā dati, kas tiek uzskatīti par piemērotiem vienam lietošanas piemēram, var nebūt pietiekami kvalitatīvi citam [4].

Vēl viena datu kvalitātes definīcija, ko piedāvā Orr, ir "attālums starp informācijas sistēmas sniegtajiem datu skatiem un tiem pašiem datiem reālajā pasaulē" [5]. Tomēr datu kvalitātes novērtēšana, pamatojoties uz salīdzinājumu ar reālo pasauli, ir ļoti grūts uzdevums [6].

Lee pētījumā kopā ar Pipino, Vangu un Funku uzskata, ka datiem ir jābūt kļūdu nesaturošiem, kā arī pieejamiem atbilstošā daudzumā. Kvalitatīvus datus raksturo īpašības, kā datu pilnīgums, nepretrunīgums, savlaicīgums, ticamība, kā arī piekļūstamība [7].

Tiek atzīts, ka datu kvalitāte ir grūti definējams jēdziens, jo datiem nav fizisko īpašību, kas ļautu tos viegli novērtēt, kā arī prasības, lai dati tiktu uzskatīti par kvalitatīviem, atšķiras atkarībā no lietojuma. Tas nozīmē, ka, atkarībā no datu lietojuma, iespējams, ir nepieciešams definēt dažādas datu kvalitātes prasības [4].

1.2. Atvērtie dati

Atvērtie dati ir publiski pieejama bezmaksas informācija bez atkalizmantošanas ierobežojumiem, kuru var automatizēti apstrādāt un rediģēt ar brīvi pieejamām lietojumprogrammām [9].

Pieeja atvērtajiem datiem ir īpaši nozīmīga publiskajā sektorā, kā, piemēram, valsts un pašvaldību iestādēs, kultūras iestādēs, kur tiek radīts, apstrādāts un izmantots liels apjoms dažādas sabiedrībai aktuālas informācijas – publiski reģistri un valsts informācijas sistēmu publiskās daļas, ģeotelpiskā informācija, pētījumi, statistika, tabulas. Ieteicamais veids, kā

nodrošināt šīs informācijas atkalizmantošanu, ir to publicēt atvērto datu formā [9]. Informācija sabiedrībai ir jānodod tādā formā, lai to varētu apstrādāt un brīvi lietot [10].

Lai dotais darbs, t.i. datu kopa vai saturs, tiktu uzskatīts par atvērtu, tam būtu jābūt [11]:

- publiskam vai nodrošinātam ar atvērto licenci. Darbam pievienotie papildu nosacījumi, kā, piemēram, lietošanas noteikumi, nedrīkst būt pretrunā ar darba publiskā īpašuma statusu vai atvērtās licences noteikumiem;
- pilnībā publicētam, lejupielādējamam bezmaksas. Darbam jāpievieno arī jebkura papildu informācija, kas nepieciešama, lai nodrošinātu licences nosacījumu izpildi;
- mašīnlasāmā formā, kas ir viegli apstrādājama ar datoru un kurā var viegli piekļūt un pārveidot atsevišķus darba elementus;
- atvērtā formātā, kurā nav noteikti nekādi monetāri vai citādi ierobežojumi lietošanai, un kuru var pilnībā apstrādāt ar vismaz vienu bezmaksas, atvērtā pirmkoda programmatūras rīku.

Lai dati tiktu atzīti par atvērtiem, tiem ir jāatbilst 8 principiem [12]. Sekojoši datiem ir jābūt:

1. pilnīgiem – visi publiskie dati ir brīvi pieejami. Tie ir dati, uz kuriem neattiecas privātuma, drošības vai privilēģiju ierobežojumi;
2. primāriem – publicētie dati pilnībā atbilst datu avotam, no kura tie tika izgūti ar lielāko iespējamo detalizācijas pakāpi;
3. laicīgiem – dati galalietotājam ir pieejami pēc iespējas ātrāk;
4. pieejamiem – dati ir pieejami pēc iespējas plašākam lietotāju lokam un iespējamajiem nolūkiem;
5. mašīnlasāmiem – dati ir pietiekami strukturēti, lai tos varētu automatizēti apstrādāt;
6. nediskriminējošiem – dati ir pieejami ikvienam, bez nepieciešamības reģistrēties to iegūšanai;
7. atvērtā datu formātā – dati ir pieejami brīvā datu formātā, par kuru nevienam nav īpašas kontroles;
8. bez licences – uz datiem netiek attiecināts autortiesību, patentu, preču zīmju vai komercnoslēpumu regulējums.

1.3. Atvērto datu un to kvalitātes nozīmīgums

Atvērto datu pieejamība ir strauji pieaugusi, palielinoties spiedienam uz visu veidu sabiedriskajām organizācijām, lai tās publiskotu savus uzkrātos datus [13]. Organizācijas ir

motivētas publicēt datus, jo brīva piekļuve publiski finansētiem datiem var nodrošināt lielāku peļņu no publiskiem ieguldījumiem, var radīt labklājību, tāpat nodrošināt politikas veidotājiem datus, kas nepieciešami sarežģītu problēmu risināšanai [14]. Atvērtie dati bieži ir neaizstājami valsts politikas veidošanai un pakalpojumu sniegšanai, taču tos var izmantot arī citiem mērķiem, piemēram, satiksmes informācijai [13].

Mūsdienās tiek veikti dažādi pētījumi, lai noskaidrotu atvērto datu ietekmi. Pētījumā, kuru veica Janssen ar līdzautoriem [13], tika noskaidrots, ka atvērtajiem datiem ir daudz priekšrocību, kā caurspīdīgums valsts pārvaldē, ekonomikas izaugsmes un inovāciju veicināšana, atbildības stiprināšana, uzticības veidošana un iedzīvotāju apmierinātības uzlabošana. Taču vairāki šķēršļi tika konstatēti arī uzdevumu sarežģītības, izmantošanas, legalizācijas, datu kvalitātes un iesaistes jomās. Piemēram, ir jābūt pieejamiem resursiem, lai nodrošinātu, ka datu kopas tiek ne tikai publicētas, bet ir arī tiešām lietotājam draudzīgas. Šie šķēršļi bieži ir savstarpēji saistīti, kas palielina vispārējo sarežģītību [13].

Zema datu kvalitāte dažādos veidos ietekmē organizācijas, izraisot neapmierinātību gan klientu, gan paša uzņēmuma darbinieku vidū, veicinot neefektīvu lēmumu pieņemšanu, kā arī palielinot izmaksas. Organizācijas zema datu kvalitāte palielina izdevumus, jo tiek tērēts laiks un citi resursi kļūdu atrašanai un novēršanai [15].

2021. gada Gartner pētījuma rezultāti liecina, ka katru gadu datu kvalitātes problēmu dēļ uzņēmumi zaudē vidēji 12.9 miljonus ASV dolāru. Papildus tūlītējai ietekmei uz ieņēmumiem, nekvalitatīvie dati ilgtermiņā palielina datu ekosistēmu sarežģītību un noved pie sliktas lēmumu pieņemšanas [16].

Tomēr datu kvalitātes problēmas ne tikai attiecas uz precizitāti, bet arī ietver tādus elementus kā pilnīgums un pieejamība. Liels ražošanas uzņēmums atklāja, ka tas nevarēja piekļūt visiem viena klienta pārdošanas datiem, jo tika piešķirti daudzi atšķirīgi klientu numuri, viena un tā paša klienta pārstāvēšanai. Kopsavilkumā, zemai datu kvalitātei ir būtiskas sociālās un ekonomiskās sekas [17].

Atvērto datu radīto vērtību var definēt kā tiešus ieguvumus no to izmantošanas, kā arī netiešus ieguvumus, kā, piemēram, jaunas informācijas radīšana, preču un pakalpojumu izstrāde, kā arī procesu uzlabošana, atkārtoti izmantojot datus. Atvērtie dati var palīdzēt uzņēmumiem un valdībām gūt lielākus ienākumus, kā arī samazināt izdevumus, sniedzot jaunus pakalpojumus, ietaupīt laiku, saudzēt vidi un uzlabot zināšanu nodošanu, izmantojot valodu pakalpojumus, kā arī pat palīdzēt glābt dzīvības. 2020. gada ziņojumā par atvērto datu ekonomisko vērtību Eiropā, tika noskaidrots, ka, izmantojot atvērtos datus, tikušas izglabātas līdz pat 202 tūkstošiem dzīvību, jo tādējādi tika nodrošināta iespēja ātrāk reaģēt uz ārkārtas situācijām [18].

1.4. Atvērtie izglītības dati

Atvērtie izglītības dati ir ļoti svarīgi izglītības vienlīdzības nodrošināšanā un mūžizglītības veicināšanā. Atbilstoša izglītības sistēma ir viens no priekšnoteikumiem veiksmīgai valsts attīstībai un konkurētspējīgai tautsaimniecības izaugsmei.

Tomēr atvērtie izglītības dati ir salīdzinoši jauna interešu joma. Definīcija "atvērtie izglītības dati" vēl joprojām ir brīvi definēta, bet to var izmantot, lai norādītu [19]:

- atvērtie dati, ko izdod izglītības iestādes,
- visi pieejamie dati, ko var izmantot izglītības nolūkā.

Atvērto izglītības datu kopas mēdz saturēt informāciju par iekšējiem datiem, kā darbavietām akadēmiskajās aprindās, mācību programmām, novērtējumiem, administratīvajiem datiem, piemēram, akadēmisko iestāžu atrašanās vietām, utt. Tos mēdz sagatavot, piemēram, vispārējās, profesionālās izglītības iestādes un universitātes, pašvaldības, domes, kā arī Izglītības un zinātnes ministrija [19] [20].

Pasaules ekonomikas foruma ziņojumā par izglītību un prasmēm [21] ir norādīts, ka ir divu veidu izglītības dati: tradicionālie un jaunie. Tradicionālajā datu kopā ietilpst identitātes dati un sistēmas dati, piemēram, informācija par apmeklētību, bet jaunās datu kopas ir izveidojušās lietotāju mijiedarbības rezultātā, piemēram, tās ietver informāciju par tīmekļa vietņu statistiku. [21]

Neatkarīgi no izmantotās klasifikācijas atklātās izglītības datu kopas noteikti interesē plašu personu loku, tostarp pedagogus, studentus, iestādes, valdību, vecākus un plašu sabiedrību. [19] Atenas, Javiera, Havemann, Leo ir norādījuši svarīgākos veidus, kā atvērtie dati saistās ar izglītības nozari un kā to lietošana krustojas ar trim galvenajām izglītības ieinteresēto personu kategorijām: politiķiem, vecākiem un izglītojamajiem, kā arī pedagogiem [22]:

- atvērtie dati kā izglītības ainava – tos izmanto kā pierādījumus izglītības un līdzdalības uzlabošanas stratēģiju izstrādei;
- atvērtie dati kā izglītības rādītāji – vecāki un izglītojamie tos izmanto, lai informētu izglītības programmu un pakalpojumu sniedzējus par savām vēlmēm un izvēlēm;
- atvērtie dati kā atvērts izglītības resurss – tos izmanto atvērto izglītības resursu izstrādei un attīstībai.

Atvērtie dati jau ilgāku laiku tiek reklamēti, kā lieliska iespēja augstākās izglītības iestādēm ļaut saviem studentiem mācīties patstāvīgi, izmantojot reālās dzīves piemērus. Atenas,

Havemann un Priego ir izveidojuši sarakstu ar kompetencēm, kuras studenti var apgūt, veicot ar pētniecību balstītas mācīšanās aktivitātes, kuru pamatā ir atvērto datu kopas [23].

Sarakstā iekļautās kompetences ietver [23]:

- kritisko domāšanu;
- prasmes datu apstrādē un pētniecībā;
- statistisko pratību, t.i. prasmju kopums, kas, pamatojoties uz datiem, ļauj pieņemt pārdomātus lēmumus;
- komandas darba prasmes;
- spēju domāt kritiski, izvērtējot gan vietējas, gan globālas problēmas.

Atvērtajiem izglītības datiem daudziem ir milzīgs potenciāls, un to izpēte ir ne tikai nepieciešama, bet arī neizbēgama. [19]

1.5. Latvijas atvērto datu portāls

Latvijas Atvērto datu portāls (*https://data.gov.lv*) ir vienota platforma piekļuvei valsts pārvaldes atvērtajiem datiem, kurš tika izveidots Eiropas reģionālās attīstības fonda līdzfinansētā projekta ietvaros ar Vides aizsardzības un reģionālās attīstības ministrijas atbalstu. 2022. gada maijā Latvijas Atvērto datu portāla katalogā ir atrodamas 615 datu kopas no 93 publicētājiem. Portālā datu kopas ir pieejamas dažādos formātos, tomēr lielākā daļa datu kopu ir pieejamas *.csv* un *.xlsx* formātos [20].

Latvijas Atvērto datu portāla vadlīnijās ir iekļauti datu publicēšanas pamatprincipi, kurus ir nepieciešams ievērot [24]:

- datiem jābūt pilnīgiem – dati tiek publicēti tādi, kādi tie oriģināli tiek iegūti ar lielāko iespējamo detalizācijas pakāpi, nevis apkopotā vai pārveidotā formā. Labākā prakse ir datus automatizēt un regulāri eksportēt, izveidot procesu, tādējādi izslēdzot cilvēciskās kļūdas iespējamību;
- datiem jābūt saprotamiem – datu kopai tiek nodrošināta metadatu pievienošana un papildu informācijas publicēšana par datu struktūru un tā saturu saprotamā terminoloģijā, lai lietotājam nebūtu problēmu datus interpretēt un tos izmantot;
- jānodrošina datu atjaunošana un uzturēšana – dati tiek regulāri atjaunoti atbilstoši to paredzētajam atjaunošanas biežumam. Iestādes pienākums ir saskaņā ar iestādes norādīto datu atjaunošanas biežuma klasifikatoru aktualizēt datu portālā ievietotos datus un nodrošināt to atbilstību metadatiem [25].

Šī darba ietvaros tiks vērtēta Latvijas Atvērtā portālā izglītības datu kopas patiesā kvalitāte, izmantojot lietojumu orientēto pieeju.

2. DATU KVALITĀTES NOVĒRTĒŠANAS RISINĀJUMI

2.1. Datu kvalitātes novērtēšana ar dimensiju palīdzību

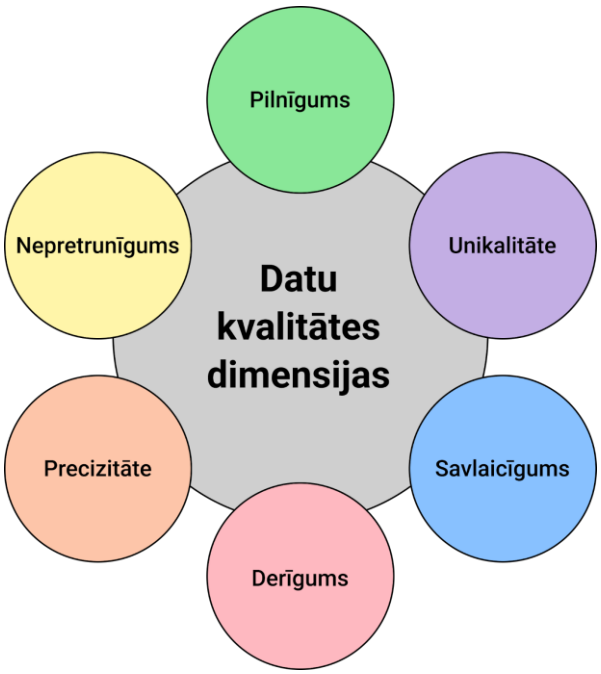
Datu kvalitātes (*DQ*) dimensija ir atzīts termins, ko izmanto datu pārvaldības speciālisti, lai aprakstītu datu iezīmes, ko var izmērīt vai novērtēt saskaņā ar noteiktiem standartiem, lai noteiktu datu kvalitāti [26]. Lielāko daļu teorētisko pētījumu raksturo plašs datu kvalitātes dimensiju klāsts. Tomēr dimensiju skaits mēdz atšķirties, pētījumos autoriem piedāvājot no dažām dimensijām līdz pat piecpadsmit. Saskaņā ar Batini un Scannapieco, datu kvalitātes teorētiskie pētījumi vēl nav nodrošinājuši vienotu to kvalitātes novērtēšanas sistēmu [27].

Tādēļ tiek piedāvāta datu lietojumu virzīta pieeja datu kvalitātes analīzei, kura tiks izmantota ierastās datu kvalitātes novērtēšanas ar dimensiju palīdzību vietā. Turpmāk tā tiks saukta par lietojumu virzītu pieeju datu kvalitātes analīzei.

2.1.1. DAMA UK Working group datu kvalitātes dimensijas

2013. gadā *DAMA UK Working Group* izstrādaja dokumentu, lai palīdzētu lietotājiem novērtēt un raksturot datu kvalitāti savās organizācijās. Lietotājiem tiek piedāvāts mazāk detalizēts dimensiju saraksts, lai mazinātu nenoteiktību un neskaidrības, kuras var rasties, novērtējot datu kvalitāti. Tiek paredzēts, ka izmantos piedāvātās 6 dimensijas, lai novērtētu sliktās datu kvalitātes ietekmi, piemēram, attiecībā uz organizācijas izmaksām un reputāciju. *DAMA UK Working Group* piedāvā mūsdienās visplašāk lietotās 6 datu kvalitātes dimensijas (skat. att. 2.1.) [26] :

1. pilnīgums (angl. *completeness*) – attiecība starp pieejamiem un visiem potenciāli iespējamiem datiem;
2. unikalitāte (angl. *uniqueness*) – reālās pasaules objekta identificēšana pēc konkrēta parametra;
3. savlaicīgums (angl. *timeliness*) – pakāpe, kādā dati atspoguļo realitāti norādītajā laika periodā;
4. derīgums (angl. *validity*) – datu atbilstība definētajām sintakses prasībām, piemēram, pieļaujamajam tipam, formātam un diapazonam;
5. precizitāte (angl. *accuracy*) – pakāpe, kādā dati pareizi apraksta reālās pasaules objektu un notikumu;
6. nepretrunīgums (angl. *consistency*) – sakritība starp divām vai vairākām reālās pasaules objektu reprezentācijām.



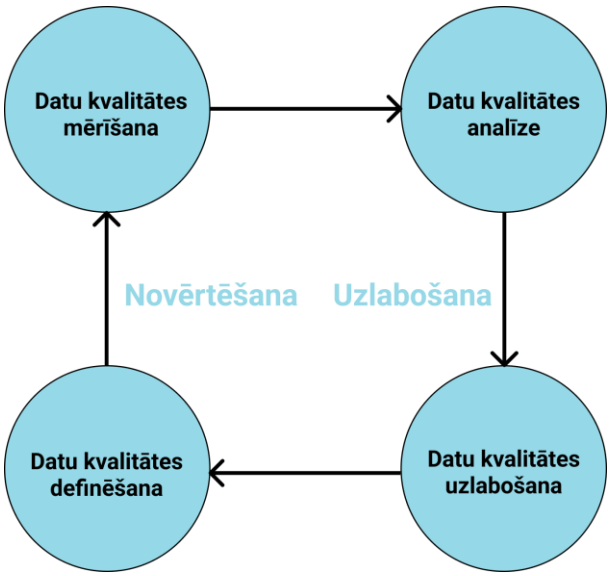
2.1. att. **DAMA UK Working Group datu kvalitātes dimensijas (autores [26] tulkojums)**

2.1.2. TDQM datu kvalitātes metodoloģija

Viena no visplašāk zināmākajām datu kvalitātes "programmām" jeb metodoloģijām ir Masačūsetsas Tehnoloģiju institūta izstrādātā *Total Data Quality Management* (*TDQM*). Tiek izmantoti cilvēkresursi un kvantitatīvie resursi, lai uzlabotu produktus un pakalpojumus. *TDQM* veicina datu standartu lietošanu, kā arī atbalsta datu bāžu migrāciju un biznesa noteikumu izmantošanu to uzlabošanā [28].

TDQM veido dzīvescikls, kurš sastāv no četrām savstarpēji saistītām fāzēm (skat. att. 2.2.) [28]:

- datu kvalitātes definēšanas fāze – tiek apkopotas biznesa prasības un analizēti dati;
- datu kvalitātes mērīšanas fāze – tiek noteikti informācijas kvalitātes rādītāji, kā arī tiek atklāti trūkumi;
- datu kvalitātes analīzes fāze – tiek novērtētas iepriekšējās fāzes iegūtās informācijas kvalitātes problēmas un noteikti kļūdu cēloņi;
- datu kvalitātes uzlabošanas fāze – tiek atlasītas galvenās datu kvalitātes uzlabošanas jomas, kā arī stratēģijas un metodes.



2.2. att. **TDQM fāzes (autores [28] tulkojums)**

Lai sasniegtu un uzturētu augstu datu kvalitāti, visas cikla fāzes ir jāatkārto sistemātiski, lai:

1. pārbaudītu, kā notiek datu kvalitātes uzlabošanas mehānisma realizācija. To parasti ir iespējams veikt, atkārtojot mērīšanas un analīzes fāzes;
2. nodrošinātu jaunu vai modificētu datu kvalitātes pārbaudi, lai dati krātuvē paliktu nemainīgi. Dati krātuvēs pastāvīgi mainās, tie var izraisīt jaunas datu kvalitātes problēmas. Jaunā cikla iterācijā var izvēlēties citus kritērijus, kā arī rast nepieciešamību datu kvalitātes prasību definēšanai [29].

2.1.3. Vanga, Strongas un Guarascio datu kvalitātes dimensijas

1994. gadā Vangs (Wang), Stronga un Guarascio veica pētījumu, apkopojot lietotāju viedokļus aptaujas veidā. Pētījuma rezultātā uzzinot, ka datu kvalitāti raksturo 15 datu kvalitātes dimensijas, kā, piemēram, savlaicīgums (angl. *timeliness*), reputācija (angl. *reputation*), atbilstība (angl. *relevance*). Dimensijas iedalās 4 grupās – iekšējā (angl. *intrinsic*), kontekstuālā (angl. *contextual*), attēlošanas (angl. *representation*) un pieejamības (angl. *accessibility*) datu kvalitāte [17].

Daudzi veselības aprūpes, finanšu un patēriņa preču uzņēmumi ir izmantojuši anketu, kas izstrādāta, lai novērtētu datu kvalitāti, izmantojot 2.1. tabulā minētās datu kvalitātes dimensijas [30].

2.1. tabula

Datu kvalitātes dimensijas [17]

Dimensija	Definīcija
Atbilstība	Apjoms, kādā dati ir piemēroti un noderīgi veicamajam uzdevumam
Atbilstošs datu apjoms	Cik lielā mērā dati ir piemēroti attiecīgajam uzdevumam
Drošība	Apjoms, kādā piekļuve datiem tikusi ierobežota, lai saglabātu to drošību
Interpretējamība	Datu apjoms atbilstošās valodās, simbolos un vienībās, un to definīcijas ir skaidras
Manipulāciju iespējamība	Cik vienkārši ir modificēt datus un piemērot tos dažādām darbībām
Objektivitāte	Cik lielā mērā dati ir objektīvi un bez aizspriedumiem
Pareizība	Cik lielā mērā dati ir pareizi un ticami
Pieejamība	Cik daudz datu ir viegli pieejami un ātri iegūstami
Pievienotā vērtība	Apjoms, kādā dati ir noderīgi un kādas priekšrocības var gūt no to izmantošanas
Pilnīgums	Apjoms, kādā dati neiztrūkst, un ir pietiekamā apjomā veicamajam uzdevumam
Reprezentācija	Apjoms, kādā dati tiek sniegti vienā un tajā pašā formātā
Reputācija	Apjoms, kādā pakāpē dati tiek vērtēti to avotu un satura ziņā
Saprotamība	Cik lielā mērā dati ir viegli uztverami
Savlaicīgums	Cik lielā mērā dati ir pietiekami atjaunināti
Ticamība	Apjoms, kādā dati tiek uzskatīti par ticamiem un patiesiem

2.1.4. Bantini dimensiju definīciju salīdzinošā analīze

Iepriekšējo apakšnodaļu minētie piemēri norāda, ka dažas datu kvalitātes dimensijas ir kopīgas vairākām pieejām, taču ir arī tādas, kuras ir unikālas katrai pieejai. Tāpat pastāv pētījumi, kuros apkopo un salīdzina datu kvalitātes dimensiju klasifikāciju.

Tabulā nr. 2.2. ir redzams, kā Batini apkopojis dažādu autoru savlaicīguma (angl. *timeliness*), izplatības (angl. *currency*) un svārstīguma (angl. *volatility*) definīcijas, tādējādi secinot, ka septiņu autoru darbos uz savlaicīguma dimensiju ir 3 dažādi nosaukumi, kā arī definīcijas [27].

Vanda (Wand) un Redmans sniedz ļoti līdzīgas definīcijas dažādām dimensijām, t.i., attiecīgi savlaicīgumam (angl. *timeliness*) un izplatībai (angl. *currency*). Vangs (Wang) un Liu uzskata to pašu par savlaicīgumu, bet Naumanns ierosina tam ļoti atšķirīgu definīciju. Bovee definētā izplatība atbilst Vanga un Liu definētajam savlaicīgumam. Svārstīgumam ir līdzīga nozīme Bovee un Jarke definētajam [27].

Batini savā darbā secina, ka dimensijām nav saskaņoti nosaukumi. Autori uztver un definē vienādas dimensijas pēc nosaukuma ļoti dažādi, tāpat vienai dimensijai piešķir dažādus nosaukumus [27].

2.2. tabula

Bantini dimensiju definīciju salīdzinošās analīzes rezultāti [27]

Atsauce	Definīcija
Wand 1996	Savlaicīgums (angl. *timeliness*) attiecas tikai uz aizkavēšanos starp reālās pasaules stāvokļa un no tā izrietošo informācijas sistēmas stāvokļa maiņu
Wang 1996	Savlaicīgums (angl. *timeliness*) ir apjoms, kādā datu vecums ir piemērots attiecīgajam uzdevumam.
Redman 1996	Izplatība (angl. *currency*) ir stāvoklis, līdz kuram dati ir atjaunināti. Datu vērtība ir atjaunināta, ja tā ir pareiza, neraugoties uz iespējamām neatbilstībām, ko izraisījušas ar laiku saistītas pareizas vērtības izmaiņas.
Jarke 1999	Izplatība (angl. *currency*) apraksta, kad informācija ir ievadīta avotos un/vai datu noliktavā. Svārstīgums (angl. *volatility*) raksturo laika periodu, kurā informācija ir derīga reālajā pasaulē.

Bovee 2001	Savlaicīgumam (angl. *timeliness*) ir divi aspekti: vecums un svārstīgums. Vecums vai izplatība (angl. *currency*) norāda, cik sena ir informācija, pamatojoties uz to, kad tā tika reģistrēta. Svārstīgums (angl. *volatility*) ir informācijas nestabilitātes mērs — entītijas atribūta vērtības izmaiņu biežums.
Naumann 2002	Savlaicīgums (angl. *timeliness*) ir avotā esošo datu vidējais vecums.
Liu 2002	Savlaicīgums (angl. *timeliness*), cik lielā mērā dati ir pietiekami atjaunināti uzdevuma veikšanai.

2.2. Datu kvalitātes novērtēšana ar Latvijas Universitātes zinātnieku radītās lietojumvirzītās pieejas palīdzību

Pati datu kvalitātes teorija ir radusies no datu kvalitātes pārvaldības procesa modeļa, kurš sākas ar datu ieguvi pārbaudei no datu avotiem (skat. 2.3. att). Operācija ietver datu izgūšanu no daudziem avotiem, kas var būt dažādi datu formāti, kas atbalsta datu filtrēšanu un formātu pārveidošanu. Iegūtie rezultāti tiek ierakstīti datu objektā [31].

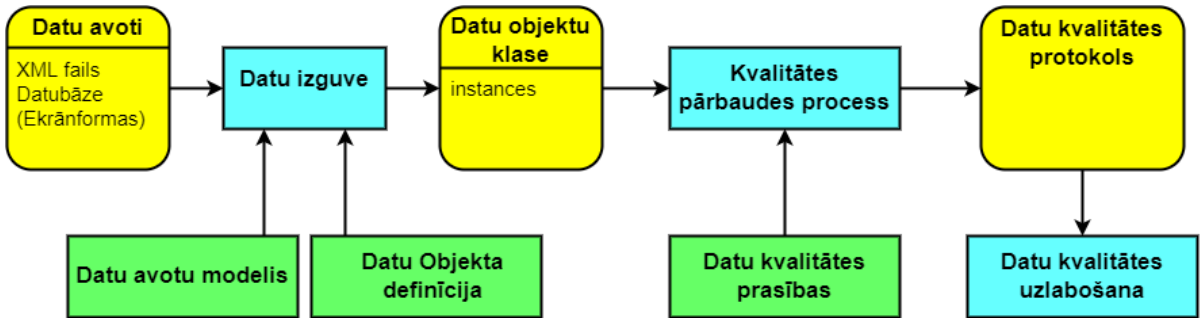
Kad dati tiek iegūti, tiek sagatavota datu objekta kvalitātes specifikācija. Tajā ir ietverti nosacījumi, kuriem jāatbilst pareizajiem datiem, kas ir definēti, izmantojot loģiskās izteiksmes. Pēc tam tiek veikts kvalitātes pārbaudes process, tostarp pārbaude, vai iegūtie dati atbilst kvalitātes prasībām. Tiek ziņots par pārbaudes procesa rezultātiem, lai uzlabotu datu kvalitāti [31].

Lietojumvirzītā pieeja un datu kvalitātes modelis sastāv no trīs galvenajiem komponentiem [32]:

1. **datu objekts** – parametru vērtību kopa, kas raksturo konkrētu reālās pasaules objektu, definē tos datus, kuru kvalitāte tiks analizēta. Datu objekts var būt **primārais, sekundārais**. Gan primārajam, gan sekundārajam objektam mēdz būt **apakšobjekti** – vienādas struktūras objektu kolekcija veido datu objektu klasi;
2. **datu kvalitātes specifikācija** – visi nosacījumi, kuri jāievēro, lai datus atzītu par kvalitatīviem;

3. **datu kvalitātes pārbaudes process** – visas veicamās aktivitātes, lai novērtētu datu atbilstību izvirzītājām kvalitātes prasībām, secinot par to kvalitāti.

Visi trīs galvenie datu kvalitātes modeļa komponenti tiek aprakstīti ar valodu metamodeļiem, kas papildu sintaksei uzdod arī modeļa grafisku reprezentāciju. Uzdodot grafisko diagrammu izpildes likumus, tiek aprakstīta modeļa semantika, tādējādi datu kvalitātes modeļiem kļūstot izpildāmiem un praktiski pielietojamiem [1].



2.3. att. **Datu kvalitātes vadības arhitektūra(autores [31] tulkojums)**

Šī datu objektu virzītā pieeja izmanto vispārīgāku jēdzienu "datu kvalitātes prasība" datu kvalitātes dimensiju jēdziena vietā, tādējādi visas vispārpieņemtās datu kvalitātes un ar to saistīto jēdzienu definīcijas tiek ievērotas citādi. "Datu kvalitātes prasība" jēdziena izmantošana "datu kvalitātes dimensiju" jēdziena vietā [33]:

- ietaupa datu kvalitātes pieeju izstrādātāju un lietotāju laiku, atvieglojot gan to izstrādes, gan izmantošanas procesu, t.i. nav jāveic vairākas resursietilpīgas darbības, kas ir saistītas ar datu kvalitātes dimensijām;
- ļauj lietotājiem bez padziļinātām IT un datu kvalitātes zināšanām piedalīties datu kvalitātes analīzes procesā;
- veicina vairāku lietotāju savstarpējo sadarbību;
- neierobežo izvirzāmo kvalitātes prasību raksturu [33].

Tā kā konkrētam datu objektam kvalitātes prasības nosaka pats lietotājs, katru komponentu ir iespējams definēt, izmantojot grafiskas blokshēmām līdzīgas diagrammas. Tas ir iespējams katram komponentam, izstrādājot grafisko domēnspecifisko valodu (DSL) [1].

Viens no pieejas pamatjēdzieniem ir datu objekts. Datu objekta un kvalitātes specifikācijas ir balstītas uz lietošanas piemēru. Tātad atkarībā no konkrētā gadījuma, tikai tie reālu objektu raksturojošie lauki, kuri būs svarīgi lietotājam, būs nepieciešami datu kvalitātes analīzei. Rezultātā datu objekta, kuru reprezentē viens reāls objekts, raksturojošo parametru skaits un struktūra var atšķirties atkarībā no lietošanas piemēra [1].

Datu kvalitātes modelis var tikt definēts un izmantots divos veidos [31]:

1. neformāli (līdzīgi PIM) – nepieciešamās pārbaužu darbības tiek aprakstītas izmantojot dabisko valodu;
2. formāli (līdzīgi PSM) – neformālus tekstus aizstājot ar izpildāmiem, piemēram, programmkodu vai *SQL* vaicājumiem.

Atšķirībā no lielākās daļas esošo datu kvalitātes risinājumu, pieeja ir paredzēta plašam lietotāju lokam. Lietotājiem procesa veikšanai nav nepieciešamas iepriekšējas zināšanas IT vai datu kvalitātes jomās, lai definētu datu objektu un to kvalitātes prasības. Tomēr IT speciālistu iesaistei vajadzētu būt tikai papildinošai un pieļaujamai vēlākos datu kvalitātes analīzes posmos, lai neformālas prasības pārveidotu par izpildāmām [32].

Lai neformāli aprakstītu datu kvalitātes prasības, tiek izveidots datu kvalitātes modulis, kas sastāv no grafiskajiem modeļiem, kur katrs konkrēts datu kvalitātes pārbaudes posms tiek aprakstīts ar diagrammas palīdzību. Katra diagramma sastāv no virsotnēm un lokiem, kur [29]:

- virsotnes attēlo elementārās datu kvalitātes vadības darbības,
- loki savieno virsotnes, norādot uz veicamo darbību secību.

Diagrammās ir iespējams iekļaut arī citas darbības, piemēram, kļūdu ziņojumu sagatavošanu, kas ir paredzēti datu kvalitātes problēmu reģistrēšanai, pēc tam izmantojot tos datu labošanai [29].

Dotā pieeja ļauj veikt konteksta pārbaudes, lai realizētu padziļinātu datu kvalitātes analīzi – analizējot nepieciešamās datu kopas kvalitāti pret citām datu kopām. Par primāro datu objektu kļūst datu objekts, kura kvalitāte tiek analizēta. Pārējie datu kvalitātes analīzē izmantotie datu objekti, pret kuriem tiek pārbaudīta primārā datu objekta kvalitāte, kļūst par sekundārajiem datu objektiem. [1] Sekundārais datu objekts parasti ir datu kopa, kas tika apstrādāta un uzkrāta ar citu no primārā datu objekta neatkarīga datu sniedzēja. Tādējādi iespējams pārbaudīt primārā datu objekta kvalitāti attiecībā pret citu neatkarīgu datu objektu. Tāpat nav ierobežots iesaistīto sekundāro datu objektu skaits, ko var izmantot viena primārā datu objekta kvalitātes analīzē [1].

Lietotājiem ir iespēja veikt datu kvalitātes analīzi, nezinot, kā datu sniedzēji apstrādāja un glabāja datus, izmantojot lietojumu orientēto pieeju datu kvalitātes definēšanai un novērtēšanai [1]. Tas ietver arī to, ka pieeja var tikt pielietota "trešo pušu" datiem. Tie var būt gan "atvērtie", gan "slēgtie" dati [32].

Nākamajā nodaļā piedāvātā metodoloģija tiek pielietota reālu datu analīzē.

3. METODOLOĢIJA

Lai veiktu Latvijas atvērto izglītības datu kvalitātes analīzi, darba autore izvēlējas to pārbaudīt trīs kārtās, pirmajā kārtā pārbaudot ortogrāfijas kļūdas datu kopās, otrajā kārtā veicot sintakses pārbaudi un trešajā kārtā veicot konteksta pārbaudi, izmantojot lietojumu orientēto datu kvalitātes novērtēšanas pieeju, kuru izstrādājis prof. J. Bičevskis kopā ar līdzautoriem Ph.D. A. Ņikiforovu, prof. Z. Bičevsku, asoc. prof. I. Odīti un prof. Ģ. Karnīti.

3.1. Pirmā kārta – ortogrāfijas pārbaude

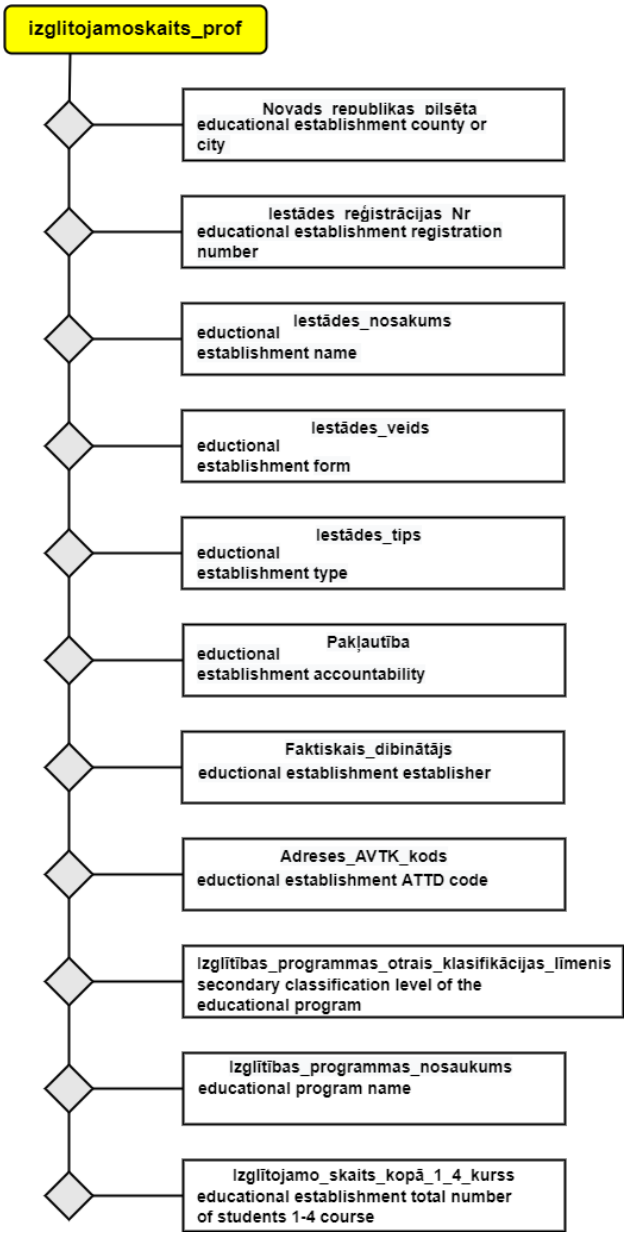
Pirmais datu kvalitātes analīzes posms ir ortogrāfijas pārbaude. Lai pārbaudītu iespējamās ortogrāfijas problēmas Latvijas Atvērto datu portāla "Izglītības" datu kopās, darba autore izvēlējas manuāli pārbaudīt tās, izmantojot gan *Microsoft Excel* programmatūru, gan *SQL* vaicājumus.

Sākumā tiek izvēlēti pārbaudāmie datu objekti, nosakot tiem lietošanas piemērus. Tālāk tiek nolasīta informācija no datu avota un ierakstīta *Microsoft Excel* programmatūras darblapās, vai arī tiek veikta datu ielāde *Microsoft SQL Server Management Studio 2018*. Atkarībā no izvēlētās programmatūras bija nepieciešams veikt papildsoļus. *Microsoft Excel* gadījumā vienkāršākais veids, kā iegūt nepieciešamos datus, ir izmantojot *.xlsx* formāta datnes, bet *.csv* un *.odata* formāti pieprasa papildsoļus informācijas ievadīšanai. Datu ievadei datubāzē ir tieši pretēji – *.xlsx* formāta datnes vēlams pārveidot *.csv* formātā, lai atvieglotu datu ielādes un pārbaudes procesu. Pētījumā *.csv* un *.odata* formātiem tiek izvēlēts *UTF-8* kā faila pirmavots. Atklātās ortogrāfijas kļūdas tiek manuāli pārbaudītas, salīdzinot ar "*Tēzaurs*" (*https://tezaurs.lv/*) vārdnīcas pieejamo informāciju.

3.2. Otrā kārta – sintakses pārbaude

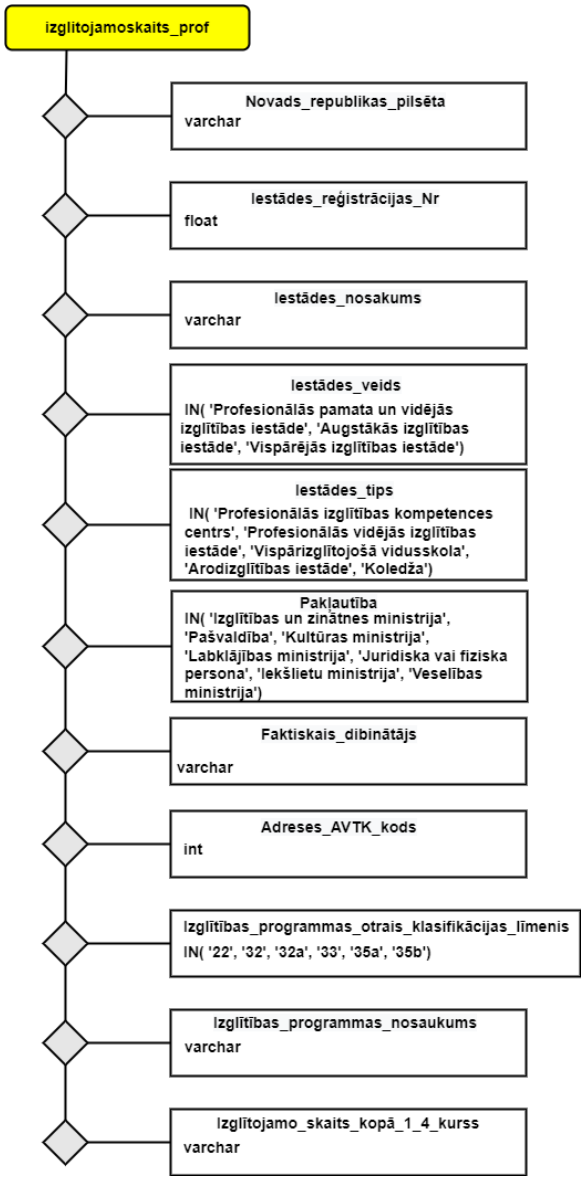
Otrais datu kvalitātes analīzes posms ir kvalitātes specifikācijas definēšana jeb sintakses pārbaude. Lai dati tiktu uzskatīti par kvalitatīviem, tiem ir jāatbilst konkrēta datu objekta datu kvalitātes specifikācijas nosacījumiem, kā, piemēram, vērtību tips, formāts, minimālās vai maksimālās vērtības ierobežojumi [34].

Pirmkārt, tiek izvēlēts datu objekts, kura kvalitāte tiks analizēta, kā arī tiek definētas datu kvalitātes prasības datu objekta klasei. Ar grafisku diagrammu palīdzību neformāli tiek aprakstītas datu kvalitātes prasības, kur grafa virsotnes attēlo izpildāmās pārbaudes un operācijas, bet to izpildīšanas secību parāda loki, tādējādi izveidojot platformneatkarīgu modeli jeb PIM modeli [1]. Pētījuma ietvaros neformāls datu apraksts tiek iegūts gan pētot datu kopas saturu, gan no parametru nosaukumiem (skat. att. 3.1.)



3.1. att. **Datu objekta "izglitojamoskaits_prof" PIM modelis**

Pēc PIM modeļa izveides, tiek izveidots platformatkarīgs modelis jeb PSM modelis, kurā, atšķirība no PIM modeļa, ir iekļautas tehniskas detaļas, tādējādi neformālus aprakstus ir iespējams aizstāt ar izpildāmiem (skat. att. 3.2.). Pētījuma ietvaros tiek iegūta informācija par laukiem, veicot datu kopas vērtību analīzi.



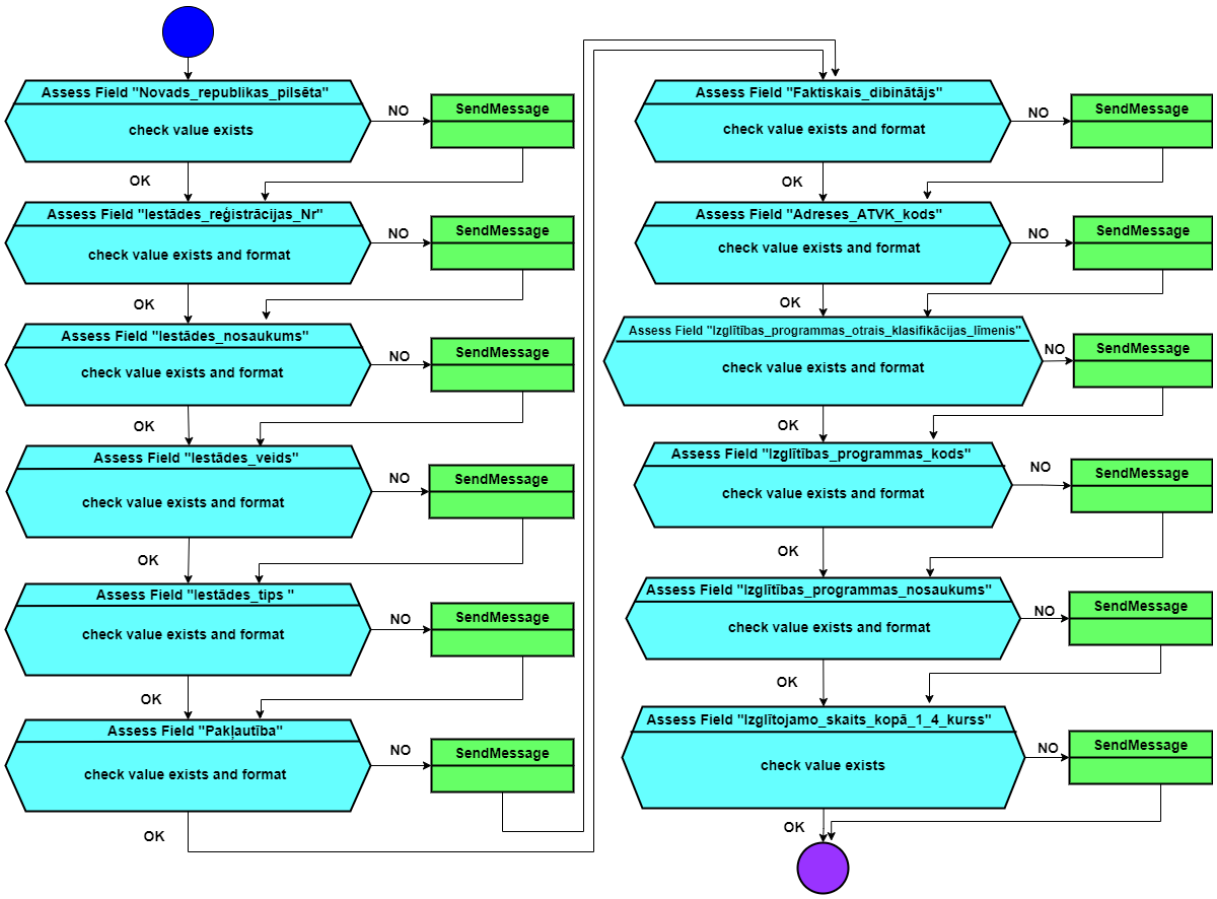
3.2. att. **Datu objekta "izglitojamoskaits_prof" PSM modelis**

Iepriekš definētajam datu objektam ir nepieciešama kvalitātes specifikācijas definēšana, kur šī specifikācija sastāv no nosacījumiem, lai dati tiktu novērtēti un atzīti par kvalitatīviem, piemēram:

1. vai simbola virkne atbilst kā izglītības iestādes nosaukums,
2. vai reģistrācijas numurs atbilst iestāžu reģistrācijas numura paraugiem,
3. vai AVTK kods eksistē un atbilst norādītajai atrašanās vietai.

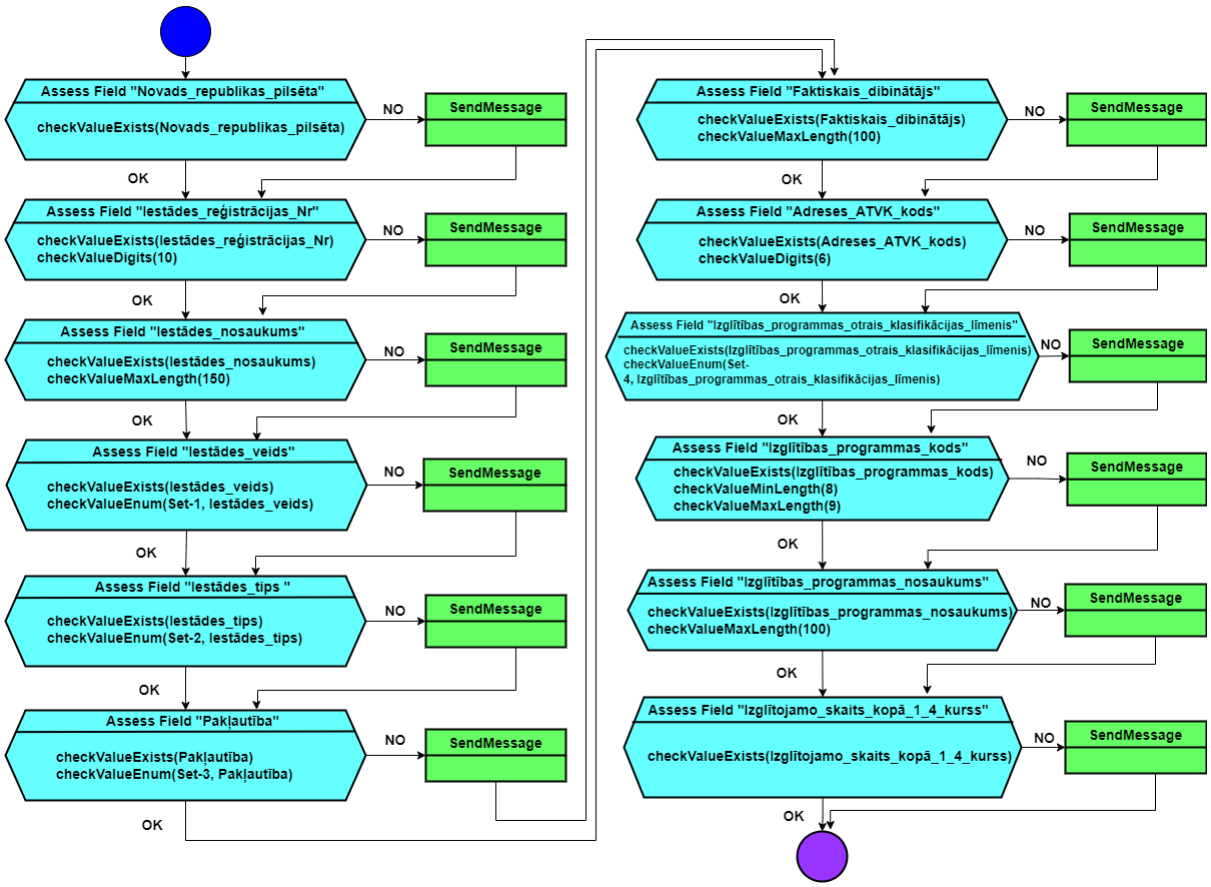
Pētījuma ietvaros datu prasības tiek definētas ar lietotāju, atkarībā no definētā lietošanas piemēra. Tad izveidotais PIM modelis tiek pārveidots PSM, datu kvalitātes prasības aizstājot ar formālām prasībām, izmantojot loģiskās izteiksmes, piemēram:

1. vai parametra "reģistrācijas numurs" vērtības atbilst 10 simbolu virknei,
2. vai parametrs "iestādes tips" vērtība eksistē un atbilst kādai no esošajām sarakstā pieļaujamajām vērtībām (skat. att. 3.3.).



3.3. att. **Datu kvalitātes specifikācija datu objektam "izglitojamoskaits_prof" PIM modelis**

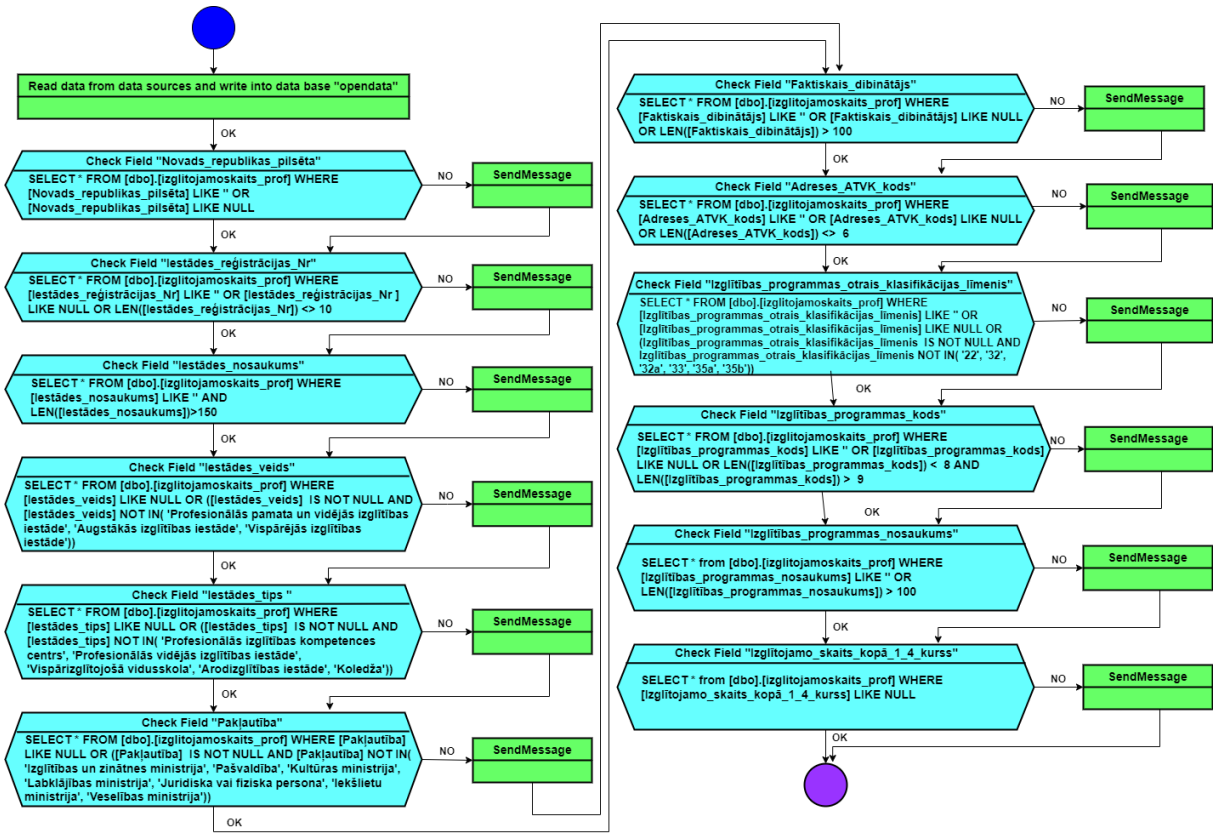
Tāpat pētījumā datu objektu parametru definētās prasības tiek norādītas, kā – vērtību eksistence; formāta pārbaude; atbilstība noteiktam datu tipam; atbilstība noteiktam paraugam; vērtību atbilstība pieļaujamo vērtību sarakstam; vērtību derīguma pārbaude (skat. att. 3.4.) [1].



3.4. att. **Datu kvalitātes specifikācija datu objektam "izglitojamoskaits_prof" PSM modelis**

Tad dati tiek nolasīti no datu avota un ierakstīti datubāzē. Kā jau iepriekš tika minēts, pētījuma nolūkos tiek izmantots *Microsoft SQL Server Management Studio 2018*. Datu ielādes sarežģītība datubāzē ir atkarīga no datu formāta, jo *.csv* formāta datnes problēmas nesagādā, bet *.xlxs* datnes formāta ielādei ir nepieciešamas papildus darbības, kā, piemēram, datnes formāta pārveidošana *.csv* formātā, kā arī datu lauku automātiski noteiktā formāta labošana.

Pēc datu ielādes datubāzē, seko datu kvalitātes pārbaudes process. Attēlā nr. 3.5. ir redzams datu objekta "izglitojamoskaits_prof" datu kvalitātes pārbaudes procesa modelis. Modelī pirmais elements ir datu nolasīšana un ierakstīšana datubāzē, pēc kā tiek veikta vērtību kvalitātes pārbaude, izpildoties *SQL* vaicājumiem.



3.5. att. **Datu kvalitātes pārbaudes process datu objektam "izglitojamoskaits_prof"**

Izpildot *SQL* vaicājumus, tiek iegūti rezultāti, kas satur atrastās neatbilstības noteiktajām datu kvalitātes prasībām. Kvalitātes novērtēšanas procesa rezultātā iegūtie rezutāti tiek analizēti, lai noskaidrotu iespējamās problēmas, kā arī to galvenos cēloņus.

3.3. Trešā kārta – kontekstuālā pārbaude

Trešais datu kvalitātes analīzes posms ir kontekstuālā jeb semantiskā pārbaude. Tiek veikta pārbaude, vai datu objekts ir atbilstoši saistīts ar citiem datu objektiem, vai tas ir saderīgs ar citām datu avotā ievadītām datu objektu vērtībām, kas ir ievadītas datu avotā, tādējādi nosakot, vai dati nav pretrunīgi.

Vispirms tiek atrasti visi ieraksti abās datu kopās, pēc tam datu kopas tiek saistītas, pamatojoties uz konkrētiem parametriem, un tiek pārbaudīta katra saistītā pāra vērtību atbilstība.

Datu objekts tiek iedalīts primārajā un sekundārajā. Centrālais datu kvalitātes analīzes objekts ir tieši primārais datu objekts, kura kvalitāte tiek analizēta. Datu objekts tiek uzskatīts par sekundāro datu objektu, ja tas veido analizējamā jeb primārā datu objekta kontekstu.

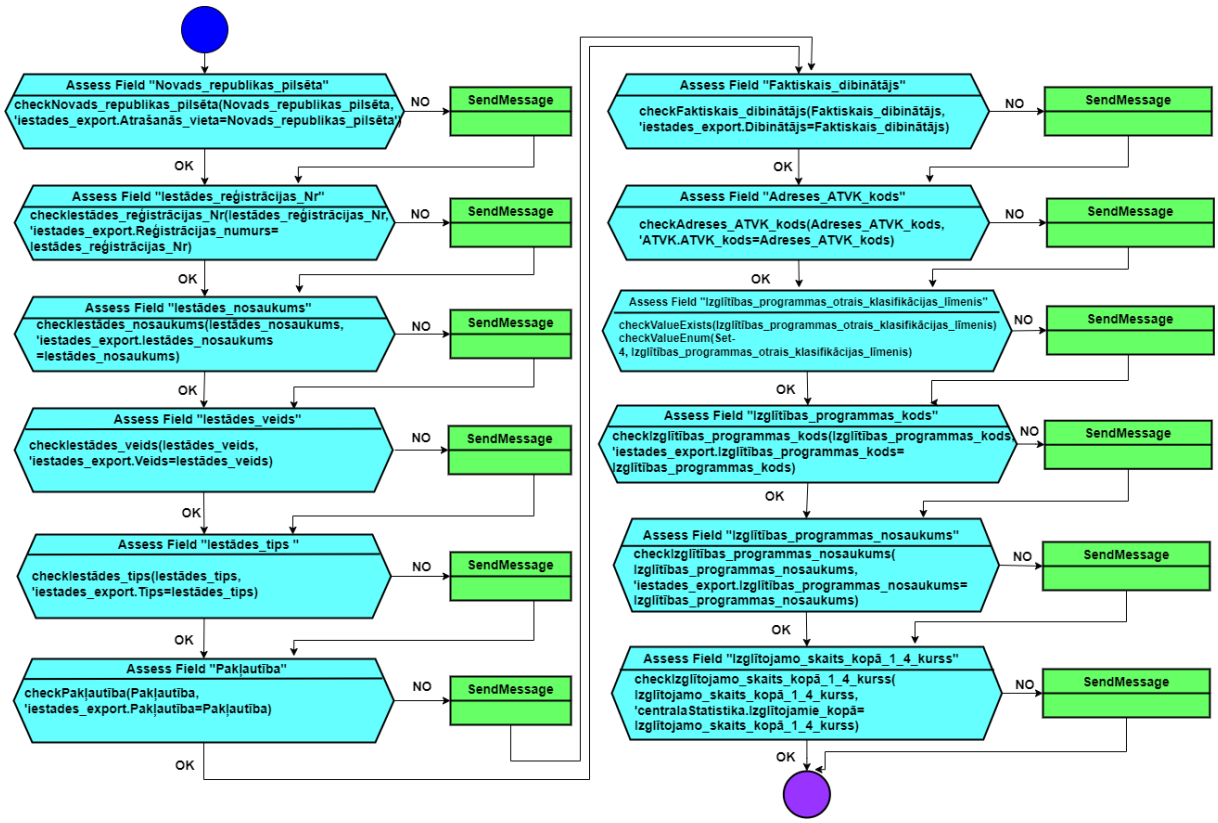
Galalietotājs definē gan primāro, gan sekundāro datu objektu, tādējādi visi primārā datu objekta raksturīpašības un tā izveides principi attiecas arī uz sekundāro datu objektu.

Primārā datu objekta analīzei pret sekundāro datu objektu var noteikt vairākas kvalitātes prasības, piemēram:

1. reģistrā esošajam iestādes nosaukumam ir jāatbilst informācijai par iestāžu nosaukumiem valsts izglītības informācijas sistēmas reģistrā.
2. reģistrā esošajam iestādes reģistrācijas numuram ir jāatbilst iestādes reģistrācijas numuram valsts izglītības informācijas sistēmas reģistrā.

Sekundārajam datu objektam nav paredzēta kvalitātes pārbaude, jo tiek pieņemts, ka tā kvalitāte ir iepriekš pārbaudīta vai arī tiek uzskatīts, ka datu objekts ir pietiekami kvalitatīvs, lai veiktu pārbaudes pret to.

Pēc primārā un sekundārā datu objektu izvēles, iepriekšējā posmā izveidotā PMS diagramma tiek papildināta ar datu kvalitātes prasībām primārā datu objekta attiecīgo parametru pārbaudei pret sekundāro datu objektu parametriem (skat. att. 3.6.). Datu kvalitātes pārbaude pret sekundāro objektu tiek veikta, izmantojot *SQL* vaicājumus, un iegūtie rezultāti tiek manuāli pārbaudīti.



3.6. att. **Datu kvalitātes prasību izpildes pārbaude datu objektam "izglitojamoskaits_prof"**

4. RISINĀJUMS

Lai veiktu datu kvalitātes analīzi Latvijas Atvērto datu portāla izglītības datu kopā, darba autore pirmās kārtas ortogrāfijas pārbaudei izvēlējas 26 no 28 datu kopas, kopumā 48 datnes, no kurām 13 datnēs no 10 datu kopām tika atklātas ortogrāfijas kļūdas. Apkopotie rezultāti ir aplūkojami 1. pielikumā.

Darba autore izvēlējas 4 datu kopas, kurām veikt sintakses un kontekstuālo pārbaudi. Sintakses un kontekstuālās jeb semantiskās pārbaudes tika veiktas kopumā 6 datnēm, kuras tika salīdzinātas ar datu objektiem no 4 informācijas avotiem, kā, piemēram, no *VIIS* jeb Valsts izglītības informācijas sistēmas, Oficiālās statistikas portāla, vienotajiem interneta datiem jeb *vid.lv* un Latvijas Atvērto datu portāla. Darba autore pieņēma, ka šo datu objektu kvalitāte ir pietiekami kvalitatīva, lai veiktu datu kvalitātes pārbaudi.

Lai samazinātu darba apjomu, pielikumos redzama tikai daļa no veiktā pētījuma rezultātiem. Tabulās 4.2. – 4.3., kā arī pielikumos, attēloto lauku skaits ir samazināts, lai izvairītos no mākslīga darba apjoma palielināšanas. Kļūdu koeficients tiek aprēķināts, attiecinot identificēto kļūdu skaitu pret analizēto ierakstu skaitu. Tāpat parametru kļūdu koeficients tiek aprēķināts, attiecinot kļūdu saturošo parametru skaitu pret analizēto parametru skaitu.

4.1. Izglītojamo skaits sadalījumā pa vispārējās izglītības programmām

Datu kopā "Izglītojamo skaits sadalījumā pa vispārējās izglītības programmām" ir aplūkojama informācija par izglītojamo skaitu vispārējās izglītības programmās, t.i., pirmsskolas izglītības programmās, vispārējās pamata un vidējās izglītības programmās. Tā satur divas datnes, kā "Izglītojamo skaits uz 01.05.2020." un "Izglītojamo skaits uz 01.04.2021.", kuras pirmo reizi publicējusi Izglītības un zinātnes ministrija 2020. gada 15. maijā.

Kopumā 2020. gada datne satur 27 kolonnas un 3719 vērtības, kopā 100413 pārbaudāmus datu laukus. Tāpat arī 2021. gada datne satur 27 kolonnas, bet 4072 vērtības, kuri kopā veido 109944 pārbaudāmus laukus.

Lai veiktu datu kopas kvalitātes analīzi, tika noteikti trīs lietošanas piemēri:

1. lietotājs var atrast izglītības iestādi pēc tā nosaukuma, reģistrācijas numura, tās veida, tipa, pakļautības, kā arī faktiskā dibinātāja;
2. lietotājs var atrast izglītības iestādes atrašanās vietu, izmantojot vietu un AVTK kodu;
3. lietotājs var atrast izglītības iestādes pieejamās izglītības programmas nosaukumus un to kodus.

Darba autore, pēc lietošanas piemēru identificēšanas, veica datu kvalitātes pārbaudi 13 parametriem. Veicot ortogrāfijas pārbaudi, tika novērots, ka 2020. gada pārbaudāmajā datnē ir 10 ortogrāfijas kļūdas, bet 2021. gada datnē ir 12 kļūdas (skat. 4.1. tabulu). Pārbaudāmie datu objekti atrodas kopīgā datu kopā, tādēļ tika sagaidīts, ka sastopamās kļūdas ir gandrīz identiskas. Bet 2021. gada datnē tiek novērotas ortogrāfijas kļūdas, kā, piemēram, vārds "izglītibas", kurš otrā datnē ir atrodams divas reizes, kā arī vārdu salikums "pamatizglītībaspirmā".

4.1. tabula

Ortogrāfijas pārbaudes rezultāti "Izglītojamo skaits sadalījumā pa vispārējās izglītības programmām" datu kopai

Nr.	Publicētājs	Datu kopa	Dati un resursi	Datu formāts	Kopējais kļūdu saturošais atribūtu/ vērtību skaits	Piemēri
1.	Izglītības un zinātnes ministrija	Izglītojamo skaits sadalījumā pa vispārējās izglītības programmām	Izglītojamo skaits uz 01.05.2020.	.xlsx	2/13 10/48347 (15.4%/0.021%)	"mācīšnās" - 1 "izgītības" - 1 "programa" - 1 "mācīšānās" - 1 "Evanģēliski" - 2 "mazākumatutību" - 1 "pimsskolas" - 1 "izglītibas" - 1 "vispārizgītojošā" - 1

2.	Izglītības un zinātnes ministrija	Izglītojamo skaits sadalījumā pa vispārējās izglītības programmām	Izglītojamo skaits uz 01.04.2021.	.xlsx	3/13 12/52936 (23.1%/0.023%)	“mācīšnās” - 1 “izgītības” - 1 “programa” - 1 “mācīšānās” - 1 “Evanģēliski” - 2 “mazākumatutību” - 1 “pimsskolas” - 1 “izglītibas” - 2 “vispārizgītojošā” - 1 “pamatizglītībaspirmā” - 1

Veicot sintakses pārbaudes kārtu, abās datnēs tika atrasta viena kļūda parametrā “Izglītības_programmas_kods” – Meirānu Kalpaka pamatskolas norādītais izglītības programmas kods (0101111) nesatur 8 simbolus (skat. 4.2. tabulu).

4.2. tabula

Sintakses pārbaudes rezultāti “Izglītojamo skaits uz 01.04.2021.”

Nr.	**Lauka nosaukums**	**Lauka formāts**	**Tukši lauki**	**Kļūdu skaits**	**Kļūdu komentāri**
1.	Novads_republikas_pilsēta	varchar(50), NOT NULL	0	0	-
2.	Iestādes_reģistrācijas_Nr	float, NOT NULL	0	0	-
3.	Iestādes_nosaukums	varchar(100), NOT NULL	0	0	-
4.	Iestādes_veids	varchar(50), NOT NULL	0	0	-
5.	Iestādes_tips	varchar(50), NOT NULL	0	0	-

6.	Pakļautība	varchar(50), NOT NULL	0	0	-
7.	Faktiskais_dibinātājs	varchar(100), NOT NULL	0	0	-
8.	Adreses_ATVK_kods	int, NOT NULL	0	0	1235 rezultāti, bet pēc darba autores manuālas izpētes *Microsoft SQL Server Management Studio 2018* ignorē norādītās 0 sākumā, tādēļ tiek attēloti rezultāti ar 5 cipariem.
9.	Izglītības_programmas_kods	varchar(50), NOT NULL	0	1 1/52936 0.002%	Viena vērtība (Meirānu Kalpaka pamatskola), kods bija norādīts kā 0101111, kurš nesatur 8 simbolus.
10.	Izglītības_programmas_nosaukums	varchar(200), NOT NULL	0	0	-
11.	Izglītojamo_skaits_pirmsskolā_kopā	varchar(50), NULL	0	0	-
12.	Kopā_1_12_klasē	varchar(50), NULL	0	0	-
13.	Izglītojamo_skaits_kopā	varchar(50), NOT NULL	0	0	-

Veicot kontekstuālo pārbaudi "Izglītojamo skaits uz 01.04.2021." datnei, salīdzinot parametru "Iestādes_reģistrācijas_Nr" ar VIIS pieejamo informāciju, tika uzrādīti 98 unikāli rezultāti jeb no datu bāzes atgrieztās atšķirīgās vērtības (skat. 4.3. tabulu). Pēc manuālas pārbaudes darba autore novēroja, ka kļūdas atrodamas gan datnē (0.81%), gan VIIS reģistrā (6.73%). Darba autore nespēja noteikt konkrētus rezultātus, jo tiek pieņemts, ka informācija

starp abiem failiem atšķiras laika perioda dēļ, gan novadu, gan skolu reformu dēļ. Iegūtie rezultāti tika balstīti pēc darba autores ieskatiem – veicot norādītās informācijas izpēti citos avotos, piemēram, izglītību iestāžu mājaslapās un dokumentos. Tāpat, salīdzinot "Izglītojamo skaits uz 01.05.2020." datni ar "Izglītojamo skaits uz 01.04.2021.", tika uzrādīti 67 unikāli rezultāti, kuri nesakrīt jaunu reģistrētu iestāžu un to reģistrācijas numuru maiņas dēļ.

Pārbaudot parametru "Iestādes_nosaukums", tika parādīti 127 unikāli rezultāti, pēc kuru manuālas izpētes tika atklāts, ka lielākoties nosaukumi atšķiras novadu un skolu reformu dēļ. Tādēļ darba autorei ir grūti spriest, vai VIIS reģistrā ir pieejami kļūdaini dati – vai nosaukumi mainīti pēdējā gada laikā, vai tomēr jau senāk. Datnē divām iestādēm (piemēram, Mazsalacas pilsētas pirmsskolas izglītības iestādei) nav norādīts pilnvērtīgs iestādes nosaukums. Parametros "Izglītības_programmas_kods" un "Izglītības_programmas_nosaukums" norādītais kods (36011012) un programmas nosaukums (Vispārējās vidējās izglītības Medicīnas, vides un dizaina programma) nav sastopami VIIS reģistrā, bet pēc darba autores manuālas izpētes atklājās, ka tomēr tāda izglītības programma un kods pastāv.

Pētījuma ietvaros darba autore izvēlējās salīdzināt "Izglītojamo_skaits_pirmsskolā_kopā" un "Izglītojamo_skaits_kopā" parametrus ar pieejamajiem datiem Oficiālajā statistikas portālā, saskaitot parametru datus kopsummā, tādējādi novērojot datu nesakritību. Tā kā tika salīdzināti kopējie rezultāti, nav iespējams precīzi uzzināt, cik lauki ir kļūdu saturoši. Pēc darba autores ieskatiem, iespējamie iemesli datu nesakritībai ir laika posmu atšķirība, mācības sākušie un pametušie izglītojamie.

4.3. tabula

Kontekstuālās pārbaudes rezultāti "Izglītojamo skaits uz 01.04.2021."

Nr	Lauka nosaukums	Fails salīdzināts	Kļūdu skaits datnē	Kļūdu skaits VIIS	Komentāri	Pret 2020. gada failu
1.	Novads_republikas_pilsēta	vid.lv informācija Iestades_export (Atrašanās_vieta)	0	0	926 unikāli rezultāti, pēc kuru manuālas pārbaudes tika noskaidrots, ka kļūdu nav, jo pilsētu, novadu nosaukumi datnē tikuši vispārināti.	-

2.	Iestādes_reģistrācijas_Nr	iestades_export (Reģistrācijas_numurs)	5 33/52936 0.062 %	88 274/52936 0.52%	98 unikāli rezultāti, pēc manuālas pārbaudes noskaidrots, ka kļūdas atrodamas gan datnē, gan VIIS reģistrā.	67 unikāli rezultāti, kuri nesakrīt jaunu reģistrētu iestāžu un reģistrācijas numuru maiņas dēļ. Kļūdu nav.
3.	Iestādes_nosaukums	iestades_export (Iestādes_nosaukums)	2 2/52936 0.004 %	-	127 unikāli rezultāti, pēc manuālas izpētes atklājās, ka nosaukumi lielākoties atšķiras novadu un skolu reformu dēļ.	50 unikāli rezultāti, kuri nesakrīt, jo tās ir jaunas reģistrētas iestādes. Kļūdu nav.
4.	Iestādes_veids	iestades_export (Veids)	0	0	-	-
5.	Iestādes_tips	iestades_export (Tips)	0	0	-	-
6.	Pakļautība	iestades_export (Pakļautība)	0	0	-	-
7.	Faktiskais_dibinātājs	iestades_export (Dibinātājs)	0	0	13 unikāli rezultāti, pēc manuālas izpētes atklājās, ka 3 iestādes likvidētas. Kļūdu nav.	17 unikāli rezultāti, pēc manuālas izpētes jauni/mainīti faktiskie

						dibinātāji iestādēm. Kļūdu nav.
8.	Adreses_ATVK_kods	vid.lv informācija	0	0	370 unikāli rezultāti, pēc manuālas izpētes kļūdu nav.	1 unikāls rezultāts, pēc manuālas izpētes, kļūdu nav.
9.	Izglītības_programmas_kods	Izglitibas_programmas_export (Izglītības_programmas_kods)	1 1/52936 0.002 %	1 1/52936 0.002 %	2 unikāli rezultāti, no kuriem kļūdains izglītības programmas kods (0101111), un otrs kods nav sastopams VIIS reģistrā (36011012), bet pēc manuālas izpētes izglītības programma un kods patiesībā eksistē.	10 unikāli rezultāti, manuāli izpētot tika noskaidrots, ka izglītības programmu licences ir neaktīvas vai izveidotas jaunas. Kļūdu nav.
10.	Izglītības_programmas_nosaukums	Izglitibas_programmas_export (Izglītības_programmas_nosaukums)	2 2/52936 0.004 %	2 2/52936 0.004 %	Kopā 9 unikāli rezultāti, pēc darba autores manuālas izpētes tika atrasts lieks punkts izglītības programmas nosaukumā, kā arī neeksistējošs izglītības programmas nosaukums (piemēram,	28 unikāli rezultāti, pēc manuālas izpētes kļūdu nav.

					“Vispārējās vidējās izglītības programma (neklātienes forma)”). Pārējie rezultāti nav atrodami VIIS reģistrā.	
11.	Izglītojamo_skaits_pirmsskolā_kopā	Oficiālais statistikas portāls	-	-	99443 Centrālās statistikas portālā, bet datnē 104025.	-
12.	Kopā_1_12_klasē	Oficiālais statistikas portāls	-	-	217271 Centrālās statistikas portālā, bet datnē 215709.	
13.	Izglītojamo_skaits_kopā	Oficiālais statistikas portāls	-	-	316714 Centrālas statistikas portāls, bet datnē 319734.	-

Aptuveni visas iepriekš minētās kontekstuālās problēmas ir arī sastopamas “Izglītojamo skaits uz 01.05.2020.” datnē (skat. 4. pielikumu 1. tabulu).

Datu kopā abas datnes satur kļūdas 7/13 (53.85%) pārbaudītajās kolonnās. Datu kopa ir pietiekami kvalitatīva, lai galalietotāji to izmantotu visiem lietošanas piemēriem, tomēr tā satur kļūdas pirmajam un trešajam lietojumam. Darba autore uzskata, ka tomēr bez šīm kļūdam varētu iztikt, tādēļ datu kopas autoriem būtu vēlams salabot identificētās kļūdas, jo tas prasītu salīdzinoši maz resursu un laiku.

4.2. Izglītojamo skaits profesionālās izglītības programmās

Datu kopas “Izglītojamo skaits profesionālās izglītības programmās” izdevējs ir Izglītības un zinātnes ministrija, tā satur divus datnes – “Izglītojamo skaits uz 01.05.2020.” un “Izglītojamo skaits uz 01.04.2021.”. Latvijas Atvērto datu portālā datu kopa pirmo reizi publicēta 2020. gada 15. maijā. Tajā ir atrodama informācija par izglītojamo skaitu

profesionālās izglītības programmās pa izglītības iestādēm pēc izglītības programmām gan 2020. gadā, gan 2021. gadā.

Datne "Izglītojamo skaits uz 01.05.2020." satur 16 kolonnas, katrai kolonnai saturot 588 vērtības, kopā ir 9408 pārbaudāmi vērtību lauki. Kā arī "Izglītojamo skaits uz 01.04.2021." datne satur arī 16 kolonnas, bet 563 vērtības, kopā 9008 pārbaudāmi datu lauki.

Datu kopas analīzes veikšanai, tika noteikti divi lietošanas piemēri:

1. lietotājs var atrast izglītības iestādi pēc tā nosaukuma, reģistrācijas numura, tās veida, tipa, pakļautības, kā arī faktiskā dibinātāja;
2. lietotājs var atrast izglītības iestādes atrašanās vietu, izmantojot norādīto vietu un AVTK kodu.

Datu kvalitātes pārbaude tika veikta 12 parametriem. Veicot pirmās kārtas ortogrāfijas pārbaudi, tika secināts, ka ortogrāfijas kļūdas nav sastopamas šajā datu kopā. Pēc otrās kārtas sintakses pārbaudes, tika noskaidrots, ka arī pēc kvalitātes specifikācijas definēšanas kļūdas nav sastopamas. Pārbaudot "Adreses_ATVK_kods" parametru, tika atgriezti 251 unikāli rezultāti, bet pēc darba autores rezultātu manuālas izpētes, tika noskaidrots, ka kļūdu nav, jo *Microsoft SQL Server Management Studio 2018* ignorē norādītās nulles AVTK koda sākumā, tādēļ tiek attēloti rezultāti ar 5 nevis 6 cipariem.

Tomēr veicot trešās kārtas semantisko pārbaudi, tika noskaidrots, ka, salīdzinot 2020. gada datni gan ar 2021. gada datni, gan VIIS reģistrā pieejamajiem datiem, kā arī ar Oficiālās statistikas portāla datiem, kļūdas ir sastopamas 3/12 kolonnās (25%). Atšķirības starp pārbaudāmas datu kopas datnēm radās tieši iestāžu likvidācijas dēļ, kā arī iestāžu atrašanās vietas maiņas dēļ. Parametrā "Izglītības_programmas_kods" tika atrastas 3 kļūdas (skat. 4. pielikumu 2. tabulu), kuras ir izlabotas 2021. gada datnē.

Veicot semantisko jeb kontekstuālo pārbaudi datnē "Izglītojamo skaits uz 01.05.2020." un VIIS pieejamajai informācijai, tika uzrādīti daudz manuāli pārbaudāmi rezultāti. Pēc to izpētes, darba autore nonāca pie secinājuma, ka, iespējams, kļūdas ir VIIS reģistrā nevis pārbaudāmajā datnē. Tāpat jāpievērš uzmanība tam, ka dati tiek salīdzināti dažādos laika posmos, jo VIIS reģistrā tiek attēloti aktuālākie dati.

Lai pārbaudītu iegūtos rezultātus, darba autore veica izglītības iestāžu oficiālo mājaslapu pārbaudi, lai salīdzinātu pieejamo ar norādīto informāciju datnē un VIIS reģistrā. Tika secināts, ka Valsts izglītības informācijas sistēmā pieejamā informācija, kā, piemēram, iestādes reģistrācijas numurs, nav atrodams nevienā citā avotā. Novērojams, ka lielākoties Latvijas Atvērto datu portāla datu kopas saturs atbilst izglītības iestāžu mājaslapās pieejamajai informācijai, bet tomēr pastāv iespējamība, ka informācija izglītības iestāžu mājaslapās nav

atjaunināta. Tāpat VIIS reģistrā bieži vien nav iespējams aplūkot vēsturi par izglītības iestādi, kas apliecinātu, piemēram, izglītības iestādes reģistrācijas numura maiņu.

Tāpat kā tika minēts iepriekšējā apakšnodaļā (skat. 4.1.), salīdzinot izglītojamo skaita datus ar Oficiālās statistikas portāla datiem, tika novērotas nesakritības.

"Izglītojamo skaits uz 01.04.2020." datnē kļūdas sastopamas 3/12 (16.7%) kolonnās. Līdzīga situācija ir otrā datnē, kļūdas sastopamas 2/12 (16.7%) kolonnās, tiek novēroti līdzīgi iemesli datu neatbilstībai. Datu kopas kvalitāte ir pietiekami labā kvalitātē, lai datus atkalizmantotu, tomēr būtu vēlams identificētās kļūdas salabot.

4.3. Dati par izglītības iestādēm un programmām

Viena no izglītības datu kopām, kurā visvairāk atrodamas ortogrāfijas kļūdas, ir " SAM 8.4.1 dati par izglītības iestādēm un programmām" datu kopa. Tajā ir aplūkojama informācija par Eiropas Sociālā fonda projektā iesaistītajām mācību iestādēm, kur dati ir sadalīti sešās kārtās. Tās izdevējs ir Valsts izglītības attīstības aģentūra, šī datu kopa satur .xlsx datni "ESF projekts "Nodarbināto personu profesionālās kompetences pilnveide" – dati par izglītības iestādēm un programmām". Portālā datu kopa tika publicēta 2021. gada 28. decembrī, tās atjaunošanas biežums reizi gadā.

Datu kopa sastāv no 8 kolonnām, kur katrai kolonnai ir 45556 vērtības. Kopā ir 364448 pārbaudāmi datu lauki.

Pētījuma ietvaros datu kopas analīzes veikšanai, tika noteikts viens lietošanas piemērs:

1. lietotājs var atrast informāciju par apmācībām pēc izglītības iestādes, mācību nosaukuma, programmas veida, kvalifikācijas, ilguma, mācību datumiem un kārtas.

Datu kvalitātes pārbaude tika veikta visiem 8 parametriem. Aplūkojot iegūtos rezultātus no pirmās kārtas (skat. 1. pielikumu), darba autore secina, ka vislielākās problēmas datnes autoriem ir sagādājis vārds "mikrokontrolieris" – datu kopā vārds "Mikrokontrolleriem" ir sastopams 99 reizes, bet "Mikrokontrolleru" pat 138 reizes.

Datu kopas sintakses analīzes rezultātā, kuras ietvaros tika veikta pārbaude 8 parametriem, datu kvalitātes problēmas netika identificētas.

Semantiskās pārbaudes veikšanai darba autore izvēlējas salīdzināt tā paša izdevēja datni "8.4.1. Demogrāfiskie dati" ar VIIS reģistrā pieejamajiem datiem, kā arī ar Latvijas Atvērto datu portāla "Uzņēmumu reģistrs" datu kopas "Uzņēmumu reģistra atvērtie dati" datni. Pēc izglītības iestāžu nosaukumu parametra salīdzinājuma ar VIIS reģistrā pieejamo informāciju, tika uzrādīti 31 unikāli rezultāti (4571 lauki). Pēc autores manuālas izpētes, VIIS reģistrā tomēr ir atrodamas 21 no 31 iestādēm. Nesakritības radušās nosaukumu nianšu atšķirību dēļ, kā,

piemēram, “SIA” lietojums VIIS reģistrā un “Sabiedrība ar ierobežotu atbildību” pārbaudāmajā datnē.

Iestāžu nosaukumi tika arī salīdzināti ar “Uzņēmuma reģistra” datu kopas datiem, tā iegūstot 100 unikālus rezultātus. Manuāli izpētot iegūtos rezultātus, darba autore atklāja, ka iegūtie rezultāti ietver tās izglītības iestādes, kuras nav atrodamas uzņēmumu reģistrā. Tāpat arī uzņēmuma reģistrā ir atrodamas iestādes, kuru nosaukumi ir uzrakstīti citādāk nekā pārbaudāmajā datnē, kā arī tiek izmantots atšķirīgs atdalītājzīmju lietojums, tādējādi tika atrastas atlikušās 10 iestādes, kuras netika atrastas VIIS reģistrā.

Tādēļ tiek pieņemts, ka parametram “Iestādes_nosaukums” kļūdu nav, bet tomēr VIIS reģistrā nav atrodamas 10 iestādes (2006 lauki).

Kopumā datnē ir ortogrāfijas kļūdas, kuras ir sastopamas 2/8(25%) kolonnās. Datu kopa ir pietiekami kvalitatīva, lai galalietotāji to izmantotu noteiktajam lietošanas piemēram.

4.4. ESF projekta dalībnieku demogrāfiskie dati

Gandrīz identiskas kļūdas iepriekšējā apakšnodaļā aprakstītajai datu kopai (skat. 4.3.) ir atrodamas arī “ESF projekts "Nodarbināto personu profesionālās kompetences pilnveide" - dalībnieku demogrāfiskie dati” datu kopā, kuras izdevējs ir Valsts izglītības attīstības aģentūra. Datu kopa satur *.xlsx* datni “8.4.1. Demogrāfiskie dati”. Portālā datu kopa tika publicēta 2021. gada 28. decembrī, tad tā arī ir pēdējo reizi tikusi mainīta.

Datu kopā ir aplūkojama informācija par Eiropas Sociālā fonda projekta dalībnieku demogrāfiskajiem datiem jeb mācības pabeigušo skaitu, kuru dati ir sadalīti sešās kārtās. Projekta mērķis ir pilnveidot nodarbināto profesionālo kompetenci, lai mazinātu plaisu starp darbaspēka kvalifikāciju un darba tirgus pieprasījumu, veicinātu darbinieku konkurētspēju un darba ražīgumu [35]. Pēc šo programmu apguves, ir iespējams iegūt sertifikātu, pilnveides vai kvalifikācijas apliecību atkarībā no apmācības veida. Tas darba devējiem kalpo kā apliecinājums darbinieka profesionālajām spējām.

Pati datne sastāv no 12 kolonnām, kur katrai kolonnai ir 45556 vērtības. Kopā saskaitot ir 546672 pārbaudāmi datu lauki.

Datu kopas analīzes veikšanai, tika noteikts viens lietošanas piemērs:

1. lietotājs var atrast informāciju par apmācībām pēc mācību nosaukuma, programmas veida, kvalifikācijas, ilguma, mācību datumiem un kārtas.

Identificējot lietojumu, darba autore izvēlējās pārbaudīt kvalitāti 7 parametriem, kopumā tika pārbaudīti 318892 lauki.

Aplūkojot apkopotās kļūdas (skat. 1. pielikumu), ir iespējams secināt, ka tās ir identiskas “SAM 8.4.1 dati par izglītības iestādēm un programmām” datu kopā sastopamajām problēmām.

Abās datu kopās identificētie kļūdainie vārdi un to skaits ir vienāds. Tā kā tās ir publicējis viens izdevējs un tās attēlo datus par ESF projektu "Nodarbināto personu profesionālās kompetences pilnveide", darba autore uzskata, ka vienādas ortogrāfijas kļūdas ir pat sagaidāmas.

Datu kopas sintakses analīzes ietvaros, datu kvalitātes problēmas netika identificētas. Lai veiktu semantisko pārbaudi, darba autore izvēlējas salīdzināt ar tā paša izdevēja datni "ESF projekts "Nodarbināto personu profesionālās kompetences pilnveide" – dati par izglītības iestādēm un programmām", bet, salīdzinot abas datnes, jaunas datu kļūdas netika novērotas.

Kopumā kļūdas sastopamas 2/7 (28.6%) kolonnās. Datu kopa ir pietiekami kvalitatīva, lai galalietotāji to izmantotu noteiktajam lietošanas piemēram.

5. ANALĪZES KOPSAVILKUMS

Veicot atvērto izglītības datu kopu kvalitātes analīzi uz reāliem datiem, darba autore novēroja, ka pārbaudītās datu kopas ir pietiekoši labā kvalitātē, lai dati tiktu atkalizmantoti, tomēr ir novērojamas datu kvalitātes problēmas.

Pēc pirmās kārtas ortogrāfijas pārbaudes, darba autore secina, ka problēmas atrodamas 10/26 (38.5%) kolonnās no visām pārbaudītajām datu kopām, bet 3/4 (75%) kolonnās no datu kopām, kurām veica arī sintakses un semantisko analīzi. Šādas ortogrāfijas kļūdas lietotājiem var atstāt nepatīkamu iespaidu par izvēlētās datu kopas kvalitāti, liekot apšaubīt informācijas patiesumu.

Otrās kārtas sintakses pārbaudes laikā pārsteidzoši netika atrastas daudzas problēmas analizētajās datu kopās. Atrastās problēmas ir saistītas ar nepareizu izglītības programmas koda ievadi.

Tomēr veicot kontekstuālās datu kvalitātes pārbaudi, darba autorei novēroja nesakritību starp pieejamajiem datiem Valsts izglītības informācijas sistēmā un atvērto izglītības datu kopās. Kopumā semantiskās problēmas ir atrodamas 50% no pārbaudītajām datu kopām. Aplūkojot VIIS mājaslapu, tika noskaidrots, ka Latvijas Atvērto datu portāla pārbaudītās datu kopas satur Valsts izglītības informācijas sistēmas operatīvās statistikas datus dažādos griezumos, tādēļ datu avotos pieejamajai informācijai būtu jābūt identiskai [34]. Veicot apjomīgu manuālu izpēti, aplūkojot izglītības iestāžu mājaslapas un oficiālos dokumentus, VIIS reģistrā pieejamā informācija bieži vien nebija atrodama citos avotos, tādējādi darba autorei secinot, ka, iespējams, kļūdas ir tieši VIIS reģistrā nevis pārbaudāmajā failā. Tomēr jāņem vērā fakts, ka dati no datu kopas tika salīdzināti ar VIIS aktuālāko informāciju, jo reģistrā nav iespējams atlasīt senākus datus, kā arī nav pieejama vēsture par datu izmaiņām sistēmā. Iespējamie iemesli datu atšķirībai ir dažādi, kā, piemēram, īstenota Administratīvi teritoriālā reforma Latvijā 2021. gada 1. jūlijā, izglītības iestāžu reformas, gan juridiskās, gan faktiskās adreses maiņa, kā arī neatjaunināta informācija oficiālajās izglītību iestāžu mājaslapās.

Tāpat datu nesakritība novērota, salīdzinot atvērto izglītības datu kopas ar pieejamajiem datiem Oficiālajā statistikas portālā. Darba autore pieļauj, ka izglītojamo skaita atšķirības datos ir atrodamas, jo izglītojamie pamet mācības vai arī maina mācību iestādi, pārvācoties uz ārzemēm, mācību gada laikā. Iespējams, dati tika salīdzināti divos dažādos laika posmos, piemēram, mācību gada sākumā un beigās. Kaut arī atšķirības kopskaitos nav milzīgas, tomēr datiem vajadzētu būt vienādiem.

Darba autorei pašlaik nav iespējams uzzināt, kuri no iegūtajiem pārbaudes rezultātiem ir patiesāki, tomēr iespējams secināt, ka atvērtajos izglītības datos ir novērojamas datu kvalitātes problēmas.

Viena no galvenajām problēmām atvērto datu izmantošanā, ir lietotāju neziņa par pieejamo datu kvalitāti. Nav iespējams noteikt pie kādiem lietošanas piemēriem konkrētas datu kopas būtu pietiekami kvalitatīvas analīzei un lēmumu pieņemšanai, tāpat arī nav iespējams noteikt, kad datu kopas ir nelietojamas, iegūstot neprecīzus vai pat nederīgus rezultātus to izmantošanas rezultātā. Novērtējot, vai datu kopa atbilst lietotāja prasībām, tās kvalitāte ir jāpārbauda jau iepriekš, lai nodrošinātu, ka tā atbilst noteikta lietošanas piemēra nosacījumiem. To ir iespējams veiksmīgi pārbaudīt, izmantojot lietojumu orientēto datu kvalitātes novērtēšanas pieeju. Pēc autores domām ar *SQL* vaicājumiem ir iespējams veikt padziļinātāku analīzi, kā arī pati izmantotā pieeja ir pietiekami vienkārša, lai to varētu izmantot plašs lietotāju loks, pat lietotāji bez padziļinātām zināšanām informācijas tehnoloģiju jomā.

Problēmas atvērtajos izglītības datos var rezultēties ar lietotāju kļūdainiem lēmumiem, nekorekti sagatavotiem dokumentiem, kā arī nevēlamiem neprecīziem meklēšanas rezultātiem. Piemēram, galalietotājam veicot tiešo meklēšanu, izmantojot kļūdainos atvērto izglītības datu kopu datus, netiks atrasti vēlamie rezultāti. Ja informācija tiks meklēta pēc līdzības, iespējams, tikai tad lietotājs atradīs vēlamo informāciju. Tāpat galalietotājiem, izmantojot atvērto izglītības kopu datus, veidojot pētījumu vai apkopojot statistiku, iegūtie rezultāti var būt neprecīzi.

Kopumā analizētajās datu kopās faktiskā datu kvalitāte ir pietiekami laba, lai dati būtu atkalizmantojami. Bet, kā jau iepriekš tika minēts, iespējamās kļūdas netika atrastas tikai atvērtajos datos, bet arī VIIS reģistrā. Nākamajā nodaļā tiek aplūkota informācija par valsts reģistru integrāciju, lai uzlabotu datu kvalitāti.

6. VALSTS REĢISTRU INTEGRĀCIJA DATU KVALITĀTES UZLABOŠANAI

Identificētās datu kvalitātes problēmas Valsts izglītības informācijas sistēmā liek aizdomāties par iespējamajām problēmām informācijas apmaiņā Valsts informācijas sistēmās. Lai novērstu pārrakstīšanās kļūdu iespējamību un uzlabotu vispārējo datu kvalitāti, valstī būtu nepieciešams reģistrēt datus tikai vienā iestādē, bet pārējām iestādēm vajadzētu saņemt to kopijas.

Viens no būtiskākajiem pārvaldes pakalpojumu kvalitātes nodrošināšanas nosacījumiem ir datu kvalitātes uzlabošana. Lai veicinātu datu kvalitāti un nodrošinātu datu apmaiņu starp iestādēm, tiek savienotas Valsts informācijas sistēmas [37].

Pašlaik, 2022. gadā, Latvijā iestāžu informācijas sistēmu centralizētu datu apmaiņu veic Valsts reģionālās attīstības aģentūras pārziņā esošais Valsts informācijas sistēmu savietotājs (VIRSIS). Tas kopā ar Valsts pārvaldes pakalpojumu portālu (www.latvija.lv) veido e-pakalpojumu koplietošanas platformu [37].

Pirms divdesmit gadiem J. Bičevskis kopā ar Ģ. Karnīti veica pētījumu par problēmām valsts nozīmes reģistru integrācijā, jau tad identificējot problēmas, kā [38]:

- tiesiska nekārtība;
- dažādu institūciju nespēja sadarboties;
- nepietiekama datu ticamība;
- slikta dokumentācija sistēmās.

Piedāvātais risinājums bija megasistēmas izveidošana, kura sastāvētu no dažādām valsts nozīmes primārajām informācijas sistēmām un tajās esošajiem datiem. Esošās informācijas sistēmas būtu jāreorganizē, tāpat megasistēmai nepieciešams universāls meklēšanas mehānisms, kas ļautu visām ieinteresētajām personām atrast vajadzīgo informāciju no visiem reģistriem [38].

Izveidojot un izmantojot megasistēmu, valsts no tā iegūtu daudzas priekšrocības [38]:

- tiktu īstenotas personu tiesības uz informāciju;
- tiktu ieviesti likumi par katras informācijas sistēmas darbību;
- tiktu precizēti datu avoti un atbildība par datu patiesumu;
- tiktu novērsta informācijas vākšanas un ierakstu dublēšanās starp reģistriem;
- tiktu uzlabota datu kvalitāte;
- tiktu uzlabota reģistru dokumentācija.

Viens no mūsdienu efektīvākajiem risinājumiem informācijas sistēmu veiksmīgai savienošanai ir Igaunijas programmatūras rīks "X-Road". Tas savieno valsts dažādās publiskā

un privātā sektora e-pakalpojumu informācijas sistēmas un ļauj tām strādāt saskaņoti. Visi izejošie dati tiek parakstīti ar ciparparakstu un šifrēti, un visi ienākošie dati tiek autentificēti un reģistrēti, lai nodrošinātu drošu datu pārsūtīšanu [39].

Tas savieno dažādas informācijas sistēmas, kas var saturēt plašu pakalpojumu klāstu. Tas ir attīstījies par rīku, kas var vienlaikus rakstīt vairākās informācijas sistēmās, pārraidīt lielus datu apjomus un veikt meklēšanu vairākās informācijas sistēmās. X-Road tika izveidots, ņemot vērā izaugsmi, lai to varētu paplašināt, kad tiešsaistē pieejami jauni e-pakalpojumi un platformas [39].

Mūsdienās rīks tiek īstenots arī Somijā, Kirgizstānā, Fēru salās, Islandē, Japānā un citās valstīs. Līdzīgas tehnoloģijas, kas balstītas uz Igaunijas sadarbspējas pieredzi, ir īstenotas arī Ukrainā un Namībijā [39].

Pēc darba autores uzskatiem, Igaunijas "X-Road" rīks, iespējams, būtu lielisks risinājums arī Latvijas informācijas sistēmu savienošanai, tādējādi palīdzot uzlabot datu kvalitāti.

SECINĀJUMI

Darba izstrādes laikā autore veica literatūras izpēti par datu kvalitātes jēdzienu, tā aktualitāti, eksistējošām pieejām datu kvalitātes novērtēšanai. Tika veikts pētījums, lai noskaidrotu Latvijas atvērto izglītības datu kvalitāti, izmantojot lietojumvirzīto datu kvalitātes novērtēšanas pieeju, tādējādi atrodot iespējamās problēmas gan atvērtajos izglītības datos, gan Valsts informācijas sistēmu reģistrā.

Darba gaitā tika izdarīti vairāki secinājumi:

1) Datu kvalitāte ir grūti definējams jēdziens, jo datiem trūkst fizisko īpašību, kas ļautu tos viegli novērtēt, un prasības atšķiras atkarībā no lietojuma, lai dati tiktu uzskatīti par kvalitatīviem. Atkarībā no paredzētā datu lietojuma var būt nepieciešams definēt dažādas datu kvalitātes prasības.
2) Lielākoties datu kvalitāte tiek pētīta izmantojot dimensijas, vairākos piedāvātajos risinājumos tiek izmantots liels skaits dimensiju, tiek paredzēta datu kvalitātes dimensiju un prasību definēšana. Kā arī tiek prasīts attiecināt datu kvalitātes prasības uz atbilstošām dimensijām, kas var būt izaicinājums pat datu kvalitātes speciālistiem.
3) Dimensijām nav saskaņoti ne nosaukumi, ne definīcijas. Pētījumu autori uztver un definē vienādas dimensijas pēc nosaukuma ļoti dažādi, tādējādi piešķirot vienai tai pašai dimensijai dažādus nosaukumus. Lai atrisinātu konstatētās problēmas, dimensiju vietā tiek piedāvāts pielietot jaunu, Latvijas Universitātes zinātnieku piedāvāto, datu kvalitātes novērtēšanas pieeju. Jaunā metode piedāvā datu kvalitāti vērtēt datu lietojumu kontekstā.
4) Datu kopās ir atrodamas vairākas ortogrāfijas kļūdas, kas lietotājiem var atstāt nepatīkamu iespaidu par izvēlētās datu kopas kvalitāti, liekot apšaubīt informācijas patiesumu.
5) Veicot kontekstuālo datu kvalitātes pārbaudi, darba autore novēroja daudz nesakritību starp pieejamajiem datiem Valsts izglītības informācijas sistēmas reģistrā un atvērto izglītības datu kopās. Pārbaudītās datu kopas satur VIIS operatīvās statistikas datus dažādos laika posmos, tādēļ datu avotos informācijai jābūt identiskai.
6) Darba autore uzskata, ka ar *SQL* vaicājumiem ir iespējams veikt padziļinātāku analīzi, kā arī pati izmantotā lietojumu orientētā pieeja ir pietiekami vienkārša, lai to varētu izmantot plašs lietotāju loks, pat bez padziļinātām zināšanām informācijas tehnoloģiju jomā.

7) Pamatojoties uz iegūtajiem rezultātiem, analizētajās Latvijas atvērto izglītības datu kopās faktiskā datu kvalitāte ir pietiekami laba, lai dati būtu atkalizmantojami, tomēr identificēto kļūdu dēļ iespējama nekorekta dokumentu sagatavošana. Atklāto kļūdu novēršana neprasītu daudz laika un resursus, bet to neviens vēl nav izdarījis.
8) Iespējamās datu kvalitātes problēmas Valsts izglītības informācijas sistēmā liek aizdomāties par potenciālajām problēmām informācijas apmaiņā Valsts informācijas sistēmās. Iespējams, Igaunijas "X-Road" programmatūras risinājums ir veiksmīgs paraugs informācijas sistēmu savienošanai.

IZMANTOTĀ LITERATŪRA

1. Nikiforova A. “Datu kvalitātes definēšana un novērtēšana”, 2020 [tiešsaiste]. [atsauce 18.03.2022.] Pieejams internetā: https://dspace.lu.lv/dspace/bitstream/handle/7/50456/298-74796-Nikiforova_Anastasija_an11093.pdf?sequence=1&isAllowed=y
2. Wand Y., Wang R.Y. “Anchoring Data Quality Dimensions in Ontological Foundations”. *Communication of the ACM*, 39(11), pp. 86-95, 1996.
3. Wang, R.Y., Storey, V.C., and Firth, C.P. "A framework for analysis of data quality research”. *IEEE Trans. on Knowl. Data Eng*. 7(4), pp. 623–640, 1995.
4. Tayi, G. K., & Ballou, D. P. “Examining data quality”. *Communications of the ACM*, 41(2), pp. 54, 1998.
5. Orr K., “Data Quality and Systems Theory”. *In Communications of the ACM*, 4(2), 1998.
6. Scannapieco, M., & Catarci, T. “Data quality under a computer science perspective”. *Archivi & Computer*, 2, pp. 1-15, 2002.
7. Lee, Y. W., Pipino, L. L., Funk, J. D., & Wang, R. Y. “Journey to data quality”. *The MIT Press*, pp. 53-63, 2009.
8. “Informācijas atklātības likums” [tiešsaiste]. [atsauce 19.03.2022.] Pieejams internetā: https://likumi.lv/doc.php?id=50601
9. Bojārs, U. “Atvērto datu vadlīnijas” [tiešsaiste]. [atsauce 02.04.2022.] Pieejams internetā: http://opendata.lumii.lv/vadlinijas/1.0/
10. Vides aizsardzības un reģionālās attīstības ministrija. “Atvērtie dati”. [tiešsaiste]. [atsauce 02.04.2022.] Pieejams internetā: https://www.varam.gov.lv/lv/atvertie-dati
11. Open data definition [tiešsaiste]. [atsauce 02.04.2022.] Pieejams internetā: http://opendefinition.org/od/2.1/en/
12. Open Government Data Principles [tiešsaiste]. [atsauce 02.04.2022.] Pieejams: https://public.resource.org/8_principles.html
13. Janssen, M., Charalabidis, Y., & Zuiderwijk, A. “Benefits, Adoption Barriers and Myths of Open Data and Open Government”. *Information Systems Management*, 29(4), pp. 258–268, 2012.
14. Arzberger, P. “SCIENCE AND GOVERNMENT: An International Framework to Promote Access to Data”. *Science*, 303(5665), pp. 1777–1778, 2004.
15. Redman, Thomas. “The Impact of Poor Data Quality on the Typical Enterprise”. *Communications of the ACM*, 1998 [tiešsaiste]. [atsauce 02.04.2022.] Pieejams

internetā: https://www.researchgate.net/publication/27295794_The_Impact_of_Poor_Data_Quality_on_the_Typical_Enterprise

16. Sakpal, M. How to Improve Your Data Quality [tiešsaiste]. [atsauce 10.04.2022.] Pieejams internetā: https://www.gartner.com/smarterwithgartner/how-to-improve-your-data-quality
17. Wang, R.Y., Strong, D. M. "Beyond Accuracy: What Data Quality Means to Data Consumers", *Journal of Management Information Systems*, 12(4), pp. 5-33, 1996.
18. The economic impact of open data, 2020 [tiešsaite]. [atsauce 13.04.2022.] Pieejams internetā: https://data.europa.eu/sites/default/files/the-economic-impact-of-open-data.pdf
19. Guy, M. "The Open Education Working Group: Bringing People, Projects and Data Together". *Lecture Notes in Computer Science*, 166–187, 2016.
20. Latvijas Atvērto datu portāls [tiešsaiste]. [atsauce 25.05.2022.] Pieejams internetā: https://www.data.gov.lv/lv
21. World Economic Forum. Education, skills 2.0: New targets and innovative approaches [tiešsaiste]. [atsauce 27.05.2022.] Pieejams internetā: https://www3.weforum.org/docs/GAC/2014/WEF_GAC_EducationSkills_TargetsInnovativeApproaches_Book_2014.pdf
22. Atenas, J., Havemann, L. "Open Data Sectors and Communities: Education". *Davies, Tim; Rubinstein, Mor; Walker, Stephen B. and Perini, Fernando eds. The State of Open Data: Histories and Horizons. Cape Town and Ottawa: African Minds and International Development Research Centre*, 2019 [tiešsaiste]. [atsauce 15.04.2022.] Pieejams internetā: http://oro.open.ac.uk/61881/1/01_06_Ch6_The_State_of_Open_Data_9781928331957%5B1%5D.pdf
23. Javiera Atenas, Leo Havemann, Ernesto Priego. "Open Data as Open Educational Resources: Towards transversal skills and global citizenship", 2015 [tiešsaiste]. [atsauce 15.04.2022.] Pieejams internetā: http://oro.open.ac.uk/56364/1/233-1107-2-PB%20%282%29.pdf
24. Vadlīnijas "atvērts pēc noklusējuma" principa ieviešanai [tiešsaiste]. [atsauce 25.04.2022] Pieejams: https://data.gov.lv/sites/default/files/2019-12/Atverts_pec_noklusejuma_1_0.pdf

25. Ministru Kabineta noteikumi Nr. 445 “Kārtība, kādā iestādes ievieto informāciju internetā”, 2020 [tiešsaiste]. [atsauce 25.04.2022] Pieejams: https://likumi.lv/ta/id/316109-kartiba-kadaiestades-ievieto-informaciju-interneta
26. Askham, N., Cook, D., Doyle, M., Fereday, H., Gibson, M., Landbeck, U., Lee, R., Maynard, C., Palmer, G., Schwarzenbach, J.. “The six primary dimensions for data quality assessment. DAMA UK Working Group”, pp. 3-13, 2013 [tiešsaiste]. [atsauce 11.04.2022.] Pieejams internetā: https://silo.tips/download/the-six-primary-dimensions-for-data-quality-assessment
27. Batini, C., & Scannapieco, M. “Data and information quality”. *Cham, Switzerland: Springer International Publishing*, pp. 42, 2016.
28. Linstedt, D., & Olschimke, M.. “Building a scalable data warehouse with data vault 2.0. Morgan Kaufmann”, 2015
29. Nikiforova, A.. “Open Data Quality Evaluation: A Comparative Analysis of Open Data in Latvia”. *Baltic Journal of Modern Computing*, 2018 [tiešsaiste]. [atsauce 10.04.2022.] Pieejams internetā: https://www.semanticscholar.org/paper/Open-Data-Quality-Evaluation%3A-A-Comparative-of-Open-Nikiforova/653612fc03cfedcc358a3e617c51c144e9e17d56
30. Leo L. Pipino, Yang W. Lee, and Richard Y. Wang, Data Quality Assessment, 2002 [tiešsaiste]. [atsauce 28.04.2022.] Pieejams internetā: https://www.researchgate.net/publication/2881159_Data_Quality_Assessment
31. Z. Bicevska, J. Bicevskis, I. Oditis, “Models of Data Quality”, *In Information Technology for Management. Ongoing Research and Development*, Springer, Champp. 194-211, 2017.
32. J. Bicevskis, A. Nikiforova, Z. Bicevska, I. Oditis and G. Karnitis, "A Step Towards a Data Quality Theory," *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, Granada, Spain, pp. 303-308, 2019.
33. Ņikiforova, A. “Izpildāmu modeļu lietojums datu kvalitātes novērtēšanai”, 2019 [tiešsaiste]. [atsauce 19.04.2022.] Pieejams internetā: https://dspace.lu.lv/dspace/bitstream/handle/7/48281/302-69386-Nikiforova_Anastasija_an11093.pdf?sequence=1&isAllowed=y
34. Bicevskis, J., Bicevska, Z., Nikiforova, A., Oditis, I. “An Approach to Data Quality Evaluation”. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pp. 196-201, 2018.
35. ESF projekts "Nodarbināto personu profesionālās kompetences pilnveide" [tiešsaiste]. [atsauce 22.03.2022.] Pieejams internetā:

https://www.macibaspieaugusajiem.lv/esfprojekts-nodarbinato-personu-profesionalas-kompetences-pilnveide?tab=collapse-79

36. Valsts izglītības informācijas sistēma. Informācija par izglītības sistēmu [tiešsaiste]. [atsauce 22.03.2022.] Pieejams internetā: https://www.viis.gov.lv/informacija-par-izglitibas-sistemu
37. VARAM. Valsts informācijas sistēmu datu apmaiņa [tiešsaiste]. [atsauce 26.04.2022.] Pieejams internetā: https://www.varam.gov.lv/lv/valsts-informacijas-sistemu-datu-apmaina
38. Bicevskis, J., Karnitis, G. "Problems in the integration of registers of state significance in Latvia", *Baltic IT review* 1(8), pp. 85, 1998.
39. e-Estonia. X-Road [tiešsaiste]. [atsauce 22.05.2022.] Pieejams internetā: https://e-estonia.com/solutions/interoperability-services/x-road/

PIELIKUMI

1. pielikums

Atvērto datu ortogrāfijas pārbaudes rezultāti

1. tabula

Atvērto datu ortogrāfijas pārbaudes rezultāti

Nr.	Publicētājs	Datu kopa	Dati un resursi	Datu formāts	Kopējais kļūdu saturošais atribūtu/ vērtību skaits	Piemēri
1.	Izglītības un zinātnes ministrija	Valsts izglītības informācijas sistēmas klasifikatoru saturs	Mācību priekšmeti	.csv	1/4 42/12420 (25%/0.34%)	● “Automātizētās” - 1 ● “vadībību” - 1 ● “automobīļu” - 3 ● “biotehnoloģikās” - 1 ● “datoriekartu” - 1 ● “datortehmikas” - 1 ● “organizacija” - 1 ● “dzelzbetona” - 1 ● “finansu” - 1 ● “fotogrāfēšana” - 3 ● “gais” - 1

						• “galdniecībasizstrādājumu” - 1 • “hidrometereoloģija” - 1 • “ilgtspejības” - 1 • “krimināprocesa” - 1 • “pamatis” - 1 • “saplakšņu” - 2 • “Latvjas” - 1 • “novietnu” - 1 • “informācojas” - 1 • “tehnololoģiju” - 1 • “griežasanas” - 1 • “multimēdiju” - 1 • “matereriāli” - 1 • “ofsetiespiešanas” - 1 • “programmavadības” - 1 • “metināšānā” - 1 • “apaklpošana” - 1 • “perspektīve” - 1 • “sanitarija” - 1 • “higiena” - 1 • “dekoratatīvo” -1

						● “standatizācija” - 1 ● “uzņemuma” - 1 ● “izpaismes” - 1 ● “uzņēmšanas” - 1 ● “zaģmateriālu” - 1
2.	Izglītības un zinātnes ministrija	Valsts izglītības informācijas sistēmas klasifikatoru saturs	Pedagoģiskā personāla amati	.csv	-	-
3.	Izglītības un zinātnes ministrija	Valsts izglītības informācijas sistēmas klasifikatoru saturs	Interešu izglītības programmas	.csv	-	-
4.	Izglītības un zinātnes ministrija	Valsts izglītības informācijas sistēmas klasifikatoru saturs	Mācību priekšmetu kursi	.csv	1/4 1/784 (25%/0.13%)	● “Kriminālprocesa” -1
5.	Izglītības un zinātnes ministrija	Valsts izglītības informācijas sistēmas klasifikatoru saturs	Sporta federācijas	.csv	-	-
6.	Izglītības un zinātnes ministrija	Valsts izglītības informācijas sistēmas klasifikatoru saturs	Stundu skaits mācību treniņu grupās	.csv	-	-

7.	Izglītības un zinātnes ministrija	Valsts izglītības informācijas sistēmas klasifikatoru saturs	Sporta veidi	.csv	-	-
8.	Izglītības un zinātnes ministrija	Valsts izglītības informācijas sistēmas klasifikatoru saturs	Vispārējās izglītības programmu nosaukumi un kodi	.csv	-	-
9.	Valsts izglītības attīstības aģentūra	ESF projekts "Nodarbināto personu profesionālās kompetences pilnveide" - dalībnieku demogrāfiskie dati	8.4.1. Demogrāfiskie dati	.xlsx	2/7 915/318892 (28.6%/0.29%)	● "Pereodisks" - 7 ● "Mešaimniecības" - 2 ● "Noformejumam" - 20 ● "Izstrādājumiu" - 12 ● "Iesgūšana" - 12 ● "Trektortehnikas" - 27 ● "Iespiedarbu" - 12 ● "Metnāšanā" - 18 ● "Papalpojumu" - 26 ● "Iztrādājumi" - 11 ● "Programējamo" - 7 ● "Apmacība" - 1 ● "Metālastrāde" - 5 ● "arAgile" - 377 ● "Mikrokontrolleriem" - 99

						● “Mikrokontrolleru” - 138 ● “Programmēšnas” -27 ● “Elektroduinertās” - 11 ● “Excell”-103
10.	Valsts izglītības attīstības aģentūra	SAM 8.4.1 dati par izglītības iestādēm un programmām	ESF projekts "Nodarbināto personu profesionālās kompetences pilnveide" - dati par izglītības iestādēm un programmām	.xlsx	2/8 915/364448 (25% /0.25%)	● “Mešaimniecības” - 2 ● “Noformejumam” - 20 ● “Pereodiskā” - 7 ● “izstrādājumiu” - 12 ● “iesgūšana” - 12 ● “Trektortehnikas” - 27 ● “Elektroduinertās” - 11 ● “Iespiedarbu” - 12 ● “metnāšanā” - 18 ● “papalpojumu” - 26 ● “iztrādājumi” - 11 ● “Programējamo” - 7 ● “Apmacība” - 1 ● “Metālastrāde” - 5 ● “arAgile” - 377 ● “Programmēšnas” - 27 ● “Mikrokontrolleriem” - 99 ● “Mikrokontrolleru” - 138

						• Excell - 103
11.	Izglītības un zinātnes ministrija	Izglītojamo skaits sadalījumā pa vispārējās izglītības programmām	Izglītojamo skaits uz 01.05.2020.	.xlsx	2/13 10/48347 (15.4%/0.02%)	• “mācīšnās” - 1 • “izgītības” - 1 • “programa” - 1 • “mācīšānās” - 1 • “Evanģēliski” - 2 • “mazākumatutību” - 1 “pimsskolas” - 1 • “izglītibas” - 1 • “vispārizgītojošā” - 1
12.	Izglītības un zinātnes ministrija	Izglītojamo skaits sadalījumā pa vispārējās izglītības programmām	Izglītojamo skaits uz 01.04.2021.	.xlsx	3/13 12/52936 (23.07%/0.02%)	• “mācīšnās” - 1 “izgītības” - 1 “programa” - 1 • “mācīšānās” - 1 “Evanģēliski” - 2 • “mazākumatutību” - 1 “pimsskolas” - 1 “izglītibas” - 2 • “vispārizgītojošā” - 1 • “pamatizglītībaspirmā” - 1

13.	Izglītības un zinātnes ministrija	Izglītojamo skaits vispārējās izglītības programmās un profesionālās pamatizglītības programmās pie speciālās izglītības iestādēm	Izglītojamo skaits uz 01.05.2020.	.xlsx	2/26 3/35334 (7.69%/0.01%)	● “izglītibas” - 1 ● “Evaņģēliski” - 2
14.	Izglītības un zinātnes ministrija	Izglītojamo skaits vispārējās izglītības programmās un profesionālās pamatizglītības programmās pie speciālās izglītības iestādēm	Izglītojamo skaits uz 01.04.2021.	.xlsx	2/26 3/34918 (7.69%/0.001%)	● “izglītibas” - 1 ● “Evaņģēliski” - 2
15.	Izglītības un zinātnes ministrija	Ar IKT jomu saistītās programmās studējošie Latvijā laika posmā no 2009.-2019.gadam	Ar IKT jomu saistītās programmās studējošie Latvijā laika posmā no 2009.-2019.gadam	.xlsx	1/20 1/27520 (5%/0.004%)	● “kēdes” - 1

16.	Cēsu novada pašvaldība	Cēsu novada pašvaldības izglītības iestāžu rādītāji	Cēsu novada pašvaldības centralizēto eksāmenu rezultāti	.csv	1/16 1/1200 (6.25%/0.08%)	● “Bioloģia” - 1
17.	Cēsu novada pašvaldība	Cēsu novada pašvaldības izglītības iestāžu rādītāji	Cēsu novada pašvaldības skolu sasniegumi	.csv	-	-
18.	Cēsu novada pašvaldība	Cēsu novada pašvaldības izglītības iestāžu rādītāji	Cēsu novada pašvaldības izglītības rādītāji	.csv	-	-
19.	Cēsu novada pašvaldība	Cēsu novada pašvaldības izglītības iestāžu rādītāji	Cēsu novada starppašvaldību norēķini	.csv	-	-
20.	Cēsu novada pašvaldība	Cēsu novada pašvaldības izglītības iestāžu rādītāji	Cēsu novada pašvaldības skolu projekti	.csv	1/19 1/133 (5.26%/0.75%)	● (Gender Equality0
21.	Rīgas dome	Dati par Rīgas privāto pirmsskolas izglītības iestāžu tāmēm	Dati par Rīgas privāto pirmsskolas izglītības iestāžu tāmēm	OData	-	-
22..	Izglītības un zinātnes ministrija	Latvijas augstākās izglītības iestāžu 2017. un 2018.gada absolventi; 2018. un 2019.	Latvijas augstākās izglītības iestāžu 2017., 2018.gada absolventi	.csv	1/58 1/128586 (1.72%/0.00008%)	● “absoventi” -1

		taksācijas/ monitoringa gados	2018. un 2019. monitoringa gados			
23.	Izglītības un zinātnes ministrija	Latvijas augstākās izglītības iestāžu 2017. un 2018.gada absolventi; 2018. un 2019. taksācijas/ monitoringa gados	Augstākās izglītības iestāžu absolventu monitoringa metadoloģija	.doc	-	-
24.	Izglītības un zinātnes ministrija	Pedagogu skaits pa amatiem pirmsskolas izglītības iestādēs, vispārējās pamata un vidējās izglītības iestādēs, profesionālās pamata un vidējās izglītības iestādēs un koledžās izglītības iestāžu griezumā	Pedagogu skaits pa iestādēm uz 20.02.2021	.xlsx	1/51 1/68085 (1.96%/0.001 %)	● “izglītibas” - 1
25.	Centrālā statistikas pārvalde	Vismaz 25 gadus vecu pastāvīgo iedzīvotāju	Vismaz 25 gadus vecu pastāvīgo iedzīvotāju vidējā izglītošanās laika	.csv	-	-

		vidējā izglītošanās laika indekss	indekss reģionos, republikas pilsētās, novados, novadu pilsētās, pagastos (atbilstoši robežām 2022. gada sākumā), apkaimēs un blīvi apdzīvotās teritorijās gada sākumā			
26.	Centrālā statistikas pārvalde	Vismaz 25 gadus vecu pastāvīgo iedzīvotāju vidējā izglītošanās laika indekss	mean_years_of_schooling_index_2011	.xlsx	-	-
27.	Centrālā statistikas pārvalde	Vismaz 25 gadus vecu pastāvīgo iedzīvotāju vidējā izglītošanās laika indekss	Vismaz 25 gadus vecu pastāvīgo iedzīvotāju vidējā izglītošanās laika indekss reģ̆a šūnās (2000. gada tautas skaitīšana)	.csv	-	-
28.	Centrālā statistikas pārvalde	Vismaz 25 gadus vecu pastāvīgo iedzīvotāju	Vismaz 25 gadus vecu pastāvīgo iedzīvotāju vidējā izglītošanās laika	.csv	-	-

		vidējā izglītošanās laika indekss	indekss režģa šūnās (2011. gada tautas skaitīšana)			
29.	Centrālā statistikas pārvalde	Vismaz 18 gadus vecu pastāvīgo iedzīvotāju īpatsvars ar augstāko izglītību vai doktora grādu	Vismaz 18 gadus vecu pastāvīgo iedzīvotāju īpatsvars ar augstāko izglītību vai doktora grādu reģionos, republikas pilsētās, novados, novadu pilsētās, pagastos (atbilstoši robežām 2022. gada sākumā), apkaimēs un blīvi apdzīvotās teritorijās	.csv	-	-
30.	Centrālā statistikas pārvalde	Vismaz 18 gadus vecu pastāvīgo iedzīvotāju īpatsvars ar augstāko izglītību vai doktora grādu	Vismaz 18 gadus vecu pastāvīgo iedzīvotāju īpatsvars ar augstāko izglītību vai doktora grādu režģa šūnās	.csv	-	-

31.	Rīgas dome	Rīgas pašvaldības finansējuma izlietojums skolēnu ēdināšanai	Rīgas pašvaldības finansējums skolēnu ēdināšanai	OData	-	-
32.	Izglītības un zinātnes ministrija	Izglītojamo skaits profesionālās izglītības programmās	Izglītojamo skaits uz 01.05.2020.	.xlsx	-	-
33.	Izglītības un zinātnes ministrija	Izglītojamo skaits profesionālās izglītības programmās	Izglītojamo skaits uz 01.04.2021.	.xlsx	-	-
34.	Izglītības un zinātnes ministrija	Pa pašvaldībām izglītojamo skaits vispārējās izglītības programmās un profesionālās pamatizglītības programmās pie speciālās izglītības iestādēm	Izglītojamo skaits uz 01.05.2020.	.xlsx	-	-

35.	Izglītības un zinātnes ministrija	Pa pašvaldībām izglītojamo skaits vispārējās izglītības programmās un profesionālās pamatizglītības programmās pie speciālās izglītības iestādēm	Izglītojamo skaits uz 01.04.2021.	.xlsx	-	-
36.	Izglītības un zinātnes ministrija	Vispārējās vidējās izglītības programmu absolventu mācību sasniegumi izglītības dokumentos	2019./2020.mācību gadā vispārējās vidējās izglītības absolventu izglītības dokumentos	.xlsx	-	-
37.	Izglītības un zinātnes ministrija	Akadēmiskā personāla skaits pa amatiem augstākās izglītības iestādēs	Akadēmiskais personāls augstākās izglītības iestādēs uz 20.02.2021.	.xlsx	-	-
38.	Izglītības un zinātnes ministrija	Pašvaldību griezumā pedagogu skaits pirmsskolas izglītības	Pedagogu skaits pa pašvaldībām uz 20.02.2021	.xlsx	-	-

		iestādēs, vispārējās pamata un vidējās izglītības iestādēs, profesionālās pamata un vidējās izglītības iestādēs un koledžās				
39.	Izglītības un zinātnes ministrija	Izglītojamo skaits ar ilgstoši neattaisnotiem kavējumiem pašvaldību pakļautības pamata un vidējās vispārējās izglītības iestādēs un kavējumu iemesli	01.09.2020 - 31.12.2020	.xlsx	-	-
40.	Jaunatnes starptautisko programmu aģentūra	Jaunatnes starptautisko programmu aģentūras 2019. gada rezultatīvie rādītāji	JSPA 2019. gada rezultatīvie rādītāji	.xlsx	1/4 1/228 (25%/0.44%)	● “apstirināto” - 1
41.	Cēsu novada pašvaldība	Skolēnu skaits septembrī pa Cēsu novada skolām	Skolēnu skaits 2018.gada septembrī	.csv	-	-

42.	Cēsu novada pašvaldība	Skolēnu skaits septembrī pa Cēsu novada skolām	Skolēnu skaits 2019.gada septembrī	.csv	-	-
43.	Cēsu novada pašvaldība	Skolēnu skaits septembrī pa Cēsu novada skolām	Skolēnu skaits 2020.gada septembrī	.csv	-	-
44.	Rīgas dome	Pieteikumi uzņemšanai Rīgas pašvaldības pirmsskolas izglītības iestādēs	Pieteikumi uzņemšanai Rīgas pirmsskolas izglītības iestādēs	OData	-	-
45.	Rīgas dome	Uzaicināto un uzņemto bērnu skaits Rīgas pašvaldības pirmsskolas izglītības iestādēs	Bērnu skaits Rīgas pirmsskolas izglītības iestādēs	OData	-	-
46.	Rīgas dome	Valsts finansējuma un Rīgas pašvaldības piemaksas par uztura korekcijas izlietojumu skolēnu ēdināšanai	Piemaksas skolēnu ēdināšanai	OData	-	-
47.	Rīgas dome	Rīgas pašvaldības piešķirtais ēdināšanas pabalsts maznodrošināto	Ēdināšanas pabalsts maznodrošināto un	OData	-	-

		un trūcīgo ģimeņu skolēniem	trūcīgo ģimeņu skolēniem Rīgā			
48.	Rīgas dome	Dati par Rīgas privāto pirmsskolas izglītības iestāžu tāmēm	Dati par Rīgas privāto pirmsskolas izglītības iestāžu tāmēm	OData	-	-

2. pielikums

Atvērto izglītības datu kvalitātes pārbaudes process

1. attēls

Datu kvalitātes pārbaudes process datnei "Izglītojamo skaits uz 01.05.2020" no "Izglītojamo skaits sadalījumā pa vispārējās izglītības programmām" datu kopas

```
SELECT * FROM [dbo].[izglitojamoskvispizglprogr_pa_progr_01052020]
WHERE [Novads_republikas_pilsēta] LIKE '' OR [Novads_republikas_pilsēta] LIKE NULL

SELECT * FROM [dbo].[izglitojamoskvispizglprogr_pa_progr_01052020]
WHERE [Iestādes_reģistrācijas_Nr] LIKE '' OR [Iestādes_reģistrācijas_Nr] LIKE NULL OR LEN([Iestādes_reģistrācijas_Nr])<>10

SELECT * FROM [dbo].[izglitojamoskvispizglprogr_pa_progr_01052020]
WHERE [Iestādes_nosaukums] LIKE '' OR LEN([Iestādes_nosaukums])>100

SELECT * FROM [dbo].[izglitojamoskvispizglprogr_pa_progr_01052020]
WHERE [Iestādes_veids] LIKE NULL OR  [Iestādes_veids]  NOT IN( 'Pirmsskolas izglītības iestāde', 'Vispārējās izglītības iestāde',
'Speciālās izglītības iestāde', 'Profesionālās pamata un vidējās izglītības iestāde',
'Bērnu un jauniešu interešu izglītības iestāde', 'Profesionālās ievirzes izglītības iestāde')

SELECT * FROM [dbo].[izglitojamoskvispizglprogr_pa_progr_01052020]
WHERE [Iestādes_tips] LIKE NULL OR  [Iestādes_tips] NOT IN( 'Vispārizglītojošā vidusskola', 'Valsts ģimnāzija',
'Pirmsskolas izglītības iestāde', 'Neklātienes vidusskola', 'Internātskola - attīstības centrs',  'Vispārizglītojošā pamatskola',
'Speciālā pamatskola', 'Vispārizglītojošā sākumskola', 'Ģimnāzija', 'Speciālā vidusskola',
'Pirmsskolas izglītības konsultatīvais centrs', 'Bērnu un jauniešu interešu izglītības iestāde',
'Mākslas skola', 'Sociālās korekcijas iestāde', 'Profesionālās vidējās izglītības iestāde',
'Vispārizglītojošā vakara (maiņu) vidusskola', 'Profesionālās izglītības kompetences centrs',
'Internātskola - rehabilitācijas centrs', 'Speciālā pirmsskolas iestāde', 'Speciālā sākumskola')

SELECT * FROM [dbo].[izglitojamoskvispizglprogr_pa_progr_01052020]
WHERE [Pakļautība] LIKE NULL OR [Pakļautība] NOT IN( 'Izglītības un zinātnes ministrija', 'Pašvaldība',
'Kultūras ministrija', 'Juridiska vai fiziska persona', 'Tieslietu ministrija')

SELECT * FROM [dbo].[izglitojamoskvispizglprogr_pa_progr_01052020]
WHERE [Faktiskais_dibinātājs] LIKE '' OR [Faktiskais_dibinātājs] LIKE NULL OR LEN([Faktiskais_dibinātājs])>100

SELECT * FROM [dbo].[izglitojamoskvispizglprogr_pa_progr_01052020]
WHERE [Adreses_ATVK_kods] LIKE '' OR [Adreses_ATVK_kods] LIKE NULL OR LEN([Adreses_ATVK_kods]) <> 6

SELECT * FROM [dbo].[izglitojamoskvispizglprogr_pa_progr_01052020]
WHERE [Izglītības_programmas_kods] LIKE '' OR [Izglītības_programmas_kods] LIKE NULL OR LEN([Izglītības_programmas_kods])<>8

SELECT * FROM [dbo].[izglitojamoskvispizglprogr_pa_progr_01052020]
WHERE [Izglītības_programmas_nosaukums] LIKE '' OR LEN([Izglītības_programmas_nosaukums])>200

SELECT * FROM [dbo].[izglitojamoskvispizglprogr_pa_progr_01052020]
WHERE [Izglītojamo_skaits_pirmsskolā_kopā] LIKE NULL

SELECT * FROM [dbo].[izglitojamoskvispizglprogr_pa_progr_01052020]
WHERE [Kopā_1_12_klasē] LIKE NULL

SELECT * FROM [dbo].[izglitojamoskvispizglprogr_pa_progr_01052020]
WHERE [Izglītojamo_skaits_kopā] LIKE NULL
```

2. attēls

Datu kvalitātes pārbaudes process datnei "Izglītojamo skaits uz 01.05.2020" no "Izglītojamo skaits profesionālās izglītības programmās" datu kopas

```
SELECT * FROM [dbo].[izglitojamoskprofesionalasizglprogr_pa_progr_01052020]
WHERE [Novads_republikas_pilsēta] LIKE '' OR [Novads_republikas_pilsēta] LIKE NULL

SELECT * FROM [dbo].[izglitojamoskprofesionalasizglprogr_pa_progr_01052020]
WHERE [Iestādes_reģistrācijas_Nr] LIKE '' OR [Iestādes_reģistrācijas_Nr] LIKE NULL OR LEN([Iestādes_reģistrācijas_Nr])<>10

SELECT * FROM [dbo].[izglitojamoskprofesionalasizglprogr_pa_progr_01052020]
WHERE [Iestādes_nosaukums] LIKE '' OR LEN([Iestādes_nosaukums])>150

SELECT * FROM [dbo].[izglitojamoskprofesionalasizglprogr_pa_progr_01052020]
WHERE [Iestādes_veids] LIKE NULL OR (Iestādes_veids IS NOT NULL AND Iestādes_veids
NOT IN( 'Profesionālās pamata un vidējās izglītības iestāde', 'Augstākās izglītības iestāde', 'Vispārējās izglītības iestāde'))

SELECT * FROM [dbo].[izglitojamoskprofesionalasizglprogr_pa_progr_01052020]
WHERE [Iestādes_tips] LIKE NULL OR (Iestādes_tips IS NOT NULL AND Iestādes_tips NOT IN( 'Profesionālās izglītības kompetences centrs',
'Profesionālās vidējās izglītības iestāde', 'Vispārizglītojošā vidusskola', 'Arodizglītības iestāde', 'Koledža'))

SELECT * FROM [dbo].[izglitojamoskprofesionalasizglprogr_pa_progr_01052020]
WHERE [Pakļautība] LIKE NULL OR (Pakļautība  IS NOT NULL AND Pakļautība NOT IN( 'Izglītības un zinātnes ministrija',
'Pašvaldība', 'Kultūras ministrija', 'Labklājības ministrija', 'Juridiska vai fiziska persona', 'Iekšlietu ministrija', 'Veselības ministrija'))

SELECT * FROM [dbo].[izglitojamoskprofesionalasizglprogr_pa_progr_01052020]
WHERE [Faktiskais_dibinātājs] LIKE '' OR [Faktiskais_dibinātājs] LIKE NULL OR LEN([Faktiskais_dibinātājs])>100

SELECT * FROM [dbo].[izglitojamoskprofesionalasizglprogr_pa_progr_01052020]
WHERE [Adreses_ATVK_kods] LIKE '' OR [Adreses_ATVK_kods] LIKE NULL OR LEN([Adreses_ATVK_kods])<>6

SELECT * FROM [dbo].[izglitojamoskprofesionalasizglprogr_pa_progr_01052020]
WHERE [Izglītības_programmas_otrais_klasifikācijas_līmenis] LIKE '' OR [Izglītības_programmas_otrais_klasifikācijas_līmenis] LIKE NULL
OR (Izglītības_programmas_otrais_klasifikācijas_līmenis  IS NOT NULL
AND Izglītības_programmas_otrais_klasifikācijas_līmenis NOT IN( '22', '32', '32a', '33', '35a', '35b'))

SELECT * FROM [dbo].[izglitojamoskprofesionalasizglprogr_pa_progr_01052020]
WHERE [Izglītības_programmas_kods] LIKE '' OR [Izglītības_programmas_kods] LIKE NULL OR LEN([Izglītības_programmas_kods])<8
AND LEN([Izglītības_programmas_kods])>9

SELECT * FROM [dbo].[izglitojamoskprofesionalasizglprogr_pa_progr_01052020]
WHERE [Izglītības_programmas_nosaukums] LIKE '' OR LEN([Izglītības_programmas_nosaukums])>100

SELECT * FROM [dbo].[izglitojamoskprofesionalasizglprogr_pa_progr_01052020]
WHERE [Izglītojamo_skaits_kopā_1_4_kurss] LIKE NULL
```

3. attēls

Datu kvalitātes pārbaudes process datnei "SAM 8.4.1 dati par izglītības iestādēm un programmām"

```
SELECT * FROM [dbo].[sam-8.4.1.-pabeiguo-dalbnieku-dati-programmu-griezum]
WHERE [Izglītības_iestāde] LIKE '' OR [Izglītības_iestāde] LIKE NULL OR LEN([Izglītības_iestāde])>150

SELECT * FROM [dbo].[sam-8.4.1.-pabeiguo-dalbnieku-dati-programmu-griezum]
WHERE [Programmas_veids] LIKE '' OR [Programmas_veids] LIKE NULL OR ([Programmas_veids] IS NOT NULL
AND [Programmas_veids] NOT IN( 'Moduļu kopa', 'Neformālās izglītības programma', 'Profesionālās tālākizglītības programma',
'Profesionālās pilnveides izglītības programma', 'Modulis', 'Studiju kurss vai studiju modulis', 'Studiju kurss', 'Studiju modulis' ))

SELECT * FROM [dbo].[sam-8.4.1.-pabeiguo-dalbnieku-dati-programmu-griezum]
WHERE [Programmas_nosaukums] LIKE '' OR LEN([Programmas_nosaukums])>250

SELECT * FROM [dbo].[sam-8.4.1.-pabeiguo-dalbnieku-dati-programmu-griezum]
WHERE LEN([Programmas_kvalifikācija])>250

SELECT * FROM [dbo].[sam-8.4.1.-pabeiguo-dalbnieku-dati-programmu-griezum]
WHERE [Mācību_uzsākšanas_datums] LIKE '' OR ISDATE([Mācību_uzsākšanas_datums])=0

SELECT * FROM [dbo].[sam-8.4.1.-pabeiguo-dalbnieku-dati-programmu-griezum]
WHERE [Mācību_pabeigšanas_datums] LIKE '' OR ISDATE([Mācību_pabeigšanas_datums])=0

SELECT * FROM [dbo].[sam-8.4.1.-pabeiguo-dalbnieku-dati-programmu-griezum]
WHERE [Programmas_ilgums] LIKE NULL OR LEN([Programmas_ilgums])>3

SELECT * FROM [dbo].[sam-8.4.1.-pabeiguo-dalbnieku-dati-programmu-griezum]
WHERE [Mācību_kārta] LIKE '' OR ([Mācību_kārta] IS NOT NULL
AND [Mācību_kārta] NOT IN( '1. kārta', '2. kārta', '3. kārta', '4. kārta', 'Attālināta', '5. kārta', '6. kārta'))
```

3. pielikums

Atvērto izglītības datu sintakses pārbaudes rezultāti

1. tabula

Sintakses pārbaudes rezultāti datnei "Izglītojamo skaits uz 01.05.2020" no "Izglītojamo skaits sadalījumā pa vispārējās izglītības programmām" datu kopas

Nr.	Lauka nosaukums	Lauka formāts	Tukši lauki	Kļūdu skaits	Kļūdu komentāri
1.	Novads_republikas_pilsēta	varchar(50), NOT NULL	0	0	-
2.	Iestādes_reģistrācijas_Nr	float, NOT NULL	0	0	-
3.	Iestādes_nosaukums	varchar(100),NOT NULL	0	0	-
4.	Iestādes_veids	varchar(50), NOT NULL	0	0	-
5.	Iestādes_tips	varchar(50), NOT NULL	0	0	-
6.	Pakļautība	varchar(50), NOT NULL	0	0	-
7.	Faktiskais_dibinātājs	varchar(100),NOT NULL	0	0	-
8.	Adreses_ATVK_kods	int, NOT NULL	0	0	Tika norādīti 1057 rezultāti, bet pēc darba autores manuālas izpētes *Microsoft SQL Server Management Studio 2018* ignorē norādītās 0 sākumā, tādēļ tiek attēloti rezultāti ar 5 cipariem.

9.	Izglītības_programmas_kods	varchar(50), NOT NULL	0	1 1/48347 0.002%	Viena vērtība (Meiranu Kalpaka pamatskola), kuras kods bija norādīts 0101111, kurš nesatur 8 simbolus.
10.	Izglītības_programmas_nosaukums	varchar(200), NOT NULL	0	0	-
11.	Izglītojamo_skaits_pirmsskolā_kopā	varchar(50), NULL	-	-	-
12.	Kopā_1_12_klasē	varchar(50), NULL	-	-	-
13.	Izglītojamo_skaits_kopā	varchar(50), NOT NULL	0	-	-

2. tabula

Sintakses pārbaudes rezultāti datnei "Izglītojamo skaits uz 01.05.2020" no "Izglītojamo skaits profesionālās izglītības programmās" datu kopas

Nr.	Lauka nosaukums	Lauka formāts	Tukši lauki	Kļūdu skaits	Kļūdu komentāri
1.	Novads_republikas_pilsēta	varchar(50), NOT NULL	0	0	-
2.	Iestādes_reģistrācijas_Nr	float, NOT NULL	0	0	-
3.	Iestādes_nosaukums	varchar(150), NOT NULL	0	0	-
4.	Iestādes_veids	varchar(100), NOT NULL	0	0	-
5.	Iestādes_tips	varchar(50), NOT NULL	0	0	-

6.	Pakļautība	varchar(50), NOT NULL	0	0	-
7.	Faktiskais_dibinātājs	varchar(100), NOT NULL	0	0	-
8.	Adreses_ATVK_kods	int, NOT NULL	0	0	Tika norādīti 257 rezultāti, bet pēc darba autores manuālas izpētes *Microsoft SQL Server Management Studio 2018* ignorē norādītās 0 sākumā, tādēļ tiek attēloti rezultāti ar 5 cipariem.
9.	Izglītības_programmas_otrais_klasifikācijas_līmenis	varchar(50), NOT NULL	0	0	-
10.	Izglītības_programmas_kods	varchar(50), NOT NULL	0	0	-
11.	Izglītības_programmas_nosaukums	varchar(100), NOT NULL	0	0	-
12.	Izglītojamo_skaits_kopā_1_4_kurss	varchar(50), NOT NULL	0	0	-

3. tabula

Sintakses pārbaudes rezultāti datnei "SAM 8.4.1 dati par izglītības iestādēm un programmām"

Nr.	Lauka nosaukums	Lauka formāts	Tukši lauki	Kļūdu skaits	Kļūdu komentāri
1.	Izglītības_iestāde	varchar(150), NOT NULL	0	0	-
2.	Programmas_veids	varchar(50), NOT NULL	0	0	-
3.	Programmas_nosaukums	varchar(250), NOT NULL	0	0	-
4.	Programmas_kvalifikācija	varchar(250), NULL	0	0	-
5.	Mācību_uzsākšanas_datums	datetime, NOT NULL	0	0	-
6.	Mācību_pabeigšanas_datums	datetime, NOT NULL	0	0	-
7.	Programmas_ilgums	int, NOT NULL	0	0	-
8.	Mācību_kārta	varchar(50), NOT NULL	0	0	-

4. pielikums

Atvērto izglītības datu kontekstuālās pārbaudes rezultāti

1. tabula

Kontekstuālās pārbaudes rezultāti datnei "Izglītojamo skaits uz 01.05.2020" no "Izglītojamo skaits sadalījumā pa vispārējās izglītības programmām" datu kopas

Nr.	Lauka nosaukums	Fails salīdzināts ar	Kļūdu skaits datnē	Kļūdu skaits VIIS	Komentāri	Pret 2021. failu
1.	Novads_republikas_pilsēta	vid.lv informācija iestades_export (Atrašanās_vieta)	0	0	948 unikāli rezultāti, pēc manuālas pārbaudes kļūdu nav, jo pilsētu un novadu nosaukumi datnē tikuši vispārināti.	11 unikāli rezultāti, pēc kuru manuālas pārbaudes tika atklāts, ka kļūdu nav.
2.	Iestādes_reģistrācijas_Nr	iestades_export (Reģistrācijas_numurs)	5 20/48347 0.04%	85 243/48347 0.5%	98 unikāli rezultāti, pēc manuālas pārbaudes atklājās, ka kļūdas atrodamas gan datnē, gan VIIS reģistrā.	83 unikāli rezultāti, pēc manuālas izpētes rezultātu atšķirība ir gan likvidēto, gan reģistrēto iestāžu dēļ. Kļūdu nav.
3.	Iestādes_nosaukums	iestades_export (Iestādes_nosaukums)	2 2/48347 0.004%	0	133 unikāli rezultāti, pēc manuālas pārbaudes nosaukumi lielākoties atšķiras tieši novadu, skolu reformu dēļ.	66 unikāli rezultātu, pēc manuālas pārbaudes tika noskaidrots, ka izglītības iestādes nesakrīt, jo tikušas likvidētas. Kļūdu nav.
4.	Iestādes_veids	iestades_export (Veids)	0	0	-	-

5.	Iestādes_tips	iestades_export (Tips)	0	0	111 unikāli rezultātu, pēc manuālas izpētes atklājās, ka kļūdu nav.	3 unikāli rezultāti, pēc manuālas pārbaudes kļūdu nav. Atšķirības radušās likvidēto izglītības iestāžu un programmu dēļ.
6.	Pakļautība	iestades_export (Pakļautība)	0	0	-	-
7.	Faktiskais_dibinātājs	iestades_export (Dibinātājs)	0	0	13 unikāli rezultāti, pēc manuālas izpētes 3 iestādes likvidētas un ar pārējiem dibinātājiem viss kārtībā.	7 unikāli rezultāti, pēc kuru manuālas pārbaudes tika noskaidrots, ka iestādes ir likvidētas. Kļūdu nav.
8.	Adreses_ATVK_kods	vid.lv informācija	0	0	380 unikāli rezultāti, no kuriem 2 unikāli rezultāti (14 lauki) nebija atrodami datnē, bet pēc manuālas izpētes kļūdu nav, jo rezultāti ir vispārināti.	11 unikāli rezultāti, pēc kuru manuālas pārbaudes tika atklāts, ka kļūdu nav.
9.	Izglītības_programmas_kods	Izglitibas_programmas_export (Izglītības_programmas_kods)	1 1/48347 0.002%	1 1/48347 0.002 %	5 unikāli rezultāti, pēc manuālas apskates viens izglītības programmas kods ir kļūdains(101111) un otra iestāde nav atrodama VIIS, jo nav norādītas reģistrā izglītības programmas.	4 unikāli rezultāti, kurus manuāli pārbaudot, tika noskaidrots, ka izglītības programmu licences ir neaktīvas. Kļūdu nav.
10.	Izglītības_programmas_nosaukums	Izglitibas_programmas_export (Izglītības_programmas_nosaukums)	4 57/48347 0.18%	0	30 unikāli rezultāti, pēc manuālas pārbaudes 3 izglītības programmu nosaukumi nesakrīt ar atrodamajiem nosaukumiem VIIS reģistrā, kopumā ietekmēti 56 lauki. Tāpat arī 1 kļūda reģistrācijas numurā.	20 unikāli rezultāti, darba autore secina, ka nosaukumu nesakritības starp abām datnēm ir neaktīvu izglītības licenžu dēļ. Kļūdu nav.

11.	Izglītojamo_skaits_pirmsskolā_kopā	Oficiālais statistikas portāls	-	-	Oficiālajā statistikas portālā 100292, bet pārbaudāmajā datu objektā 103738	-
12.	Kopā_1_12_klasē	Oficiālais statistikas portāls	-	-	Oficiālajā statistikas portālā 206923, bet pārbaudāmajā datu objektā 213423	-
13.	Izglītojamo_skaits_kopā	Oficiālais statistikas portāls	-	-	Oficiālajā statistikas portālā 307215, bet pārbaudāmajā datu objektā 317161	-

2. tabula

Kontekstuālās pārbaudes rezultāti datnei "Izglītojamo skaits uz 01.05.2020" no "Izglītojamo skaits profesionālās izglītības programmās" datu kopas

Nr.	Lauka nosaukums	VIIS datu kopa	Kļūdu skaits	Kļūdu skaits VIIS	Komentāri	Pret 2021 failu
1.	Novads_republikas_pilsēta	iestades_export (Atrašanās_vieta)	0	0	-	4 likvidētās iestādes, pēc manuālas pārbaudes Beauty School mainījusi atrašanās vietu no Liepājas uz Rīgu. Kļūdu nav.
2.	Iestādes_reģistrācijas_Nr	iestades_export (Reģistrācijas_numurs)	0	13 72/7056 1.02%	Tika atrasti 14 unikāli rezultāti, bet pēc darba autores manuālas izpētes kļūdas tika atrastas tieši VIIS reģistrā.	5 unikāli rezultāti, no kuriem 4 likvidētas iestādes un 1 iestāde, Beauty school, kura mainījusi atrašanās vietu.
3.	Iestādes_nosaukums	iestades_export (Iestādes_nosaukums)	0	0	3 unikāli rezultāti. Manuāli pārbaudot, tika noskaidrots, ka kļūdu nav.	4 unikāli rezultāti – likvidētas iestādes, kļūdu nav.
4.	Iestādes_veids	iestades_export (Veids)	0	0	1 unikāls rezultāts. Manuāli pārbaudot, tika noskaidrots, ka kļūdu nav.	4 unikāli rezultāti, visas uzrādītās iestādes ir likvidētas. Kļūdu nav.
5.	Iestādes_tips	iestades_export (Tips)	0	0	1 unikāls rezultāts. Manuāli pārbaudot, kļūdu nav.	4 unikāli rezultāti, visas uzrādītās iestādes ir likvidētas. Kļūdu nav.

6.	Pakļautība	iestades_export (Pakļautība)	0	0	1 unikāls rezultāts. Manuāli pārbaudot, tika noskaidrots, ka kļūdu nav.	4 unikāli rezultāti, visas parādītas iestādes ir likvidētas. Kļūdu nav.
7.	Faktiskais_dibinātājs	iestades_export (Dibinātājs)	8 42/7056 0.6%	0	45 unikāli rezultāti, pēc manuālas pārbaudes nesakrīt 7 unikāli rezultāti (piemēram, Valsts policijas koledžas dibinātājs ir norādīts Izglītības un zinātnes ministrija, bet VIIS failā Valsts policijas koledža). 35 rezultāti saturēja vērtību "Latvijas Republikas Ministru kabinets", kas ir vispārinājums, piemēram, Kultūras ministrijai, utt. Pēc darba autores domām tā netiek uzskatīta kā kļūda.	5 unikāli rezultāti, visas ir likvidētas iestādes pēc manuālas pārbaudes kļūdu nav.
8.	Adreses_ATVK_kods	vid.lv informācija	0	0	Tika norādīti 20 unikāli rezultāti, bet pēc darba autores manuālas izpētes kļūdu nav. Datu kopā ATVK kodi tiek vispārināti.	4 likvidētās iestādes, pēc manuālas pārbaudes Beauty School mainījusi atrašanās vietu no Liepājas uz Rīgu. Kļūdu nav.
9.	Izglītības_programmas_otrais_klasifikācijas_līmenis	-	0	0	-	83 unikāli rezultāti, pēc manuālas pārbaudes kļūdu nav.
10.	Izglītības_programmas_kods	Izglitibas_programmas_export (Izglītības_programmas_kods)	3 3/7056 0.043%	0	-	83 unikāli rezultāti, pēc manuālās pārbaudes tika atklātas 3 kļūdas (3364001, 35b64001, 32543041)

11.	Izglītības_programmas_nosaukums	-	0	0	-	3 unikāli rezultāti, programmu nosaukumi, kuri vairs neeksistē 2021. gada failā. Kļūdu nav.
12.	Izglītojamo_skaits_kopā_1_4_kurss	-	-	-	Oficiālajā statistikas portālā 27734, bet datnē 24244	-

3. tabula

Kontekstuālās pārbaudes rezultāti datnei "SAM 8.4.1 dati par izglītības iestādēm un programmām"

Nr.	Lauka nosaukums	Fails salīdzināts ar	Kļūdu skaits	Komentāri
1.	Izglītības_iestāde	VIIS reģistra dati "8.4.1. Demogrāfiskie dati" datne "Uzņēmumu reģistra atvētie dati"	0	31 unikāli rezultāti (4571), pēc autores manuālas izpētes VIIS reģistrā atrodamas 21 iestādes. Iestāžu nosaukumi tika salīdzināti arī ar Uzņēmuma reģistra datiem, kur bija iespējams atrast 10 iestādes, kuras nav atrodamas VIIS reģistrā. Kļūdu nav.
2.	Programmas_veids	"8.4.1. Demogrāfiskie dati" datne	0	-
3.	Programmas_nosaukums	"8.4.1. Demogrāfiskie dati" datne	0	-
4.	Programmas_kvalifikācija	"8.4.1. Demogrāfiskie dati" datne	0	-
5.	Mācību_uzsākšanas_datums	"8.4.1. Demogrāfiskie dati" datne	0	-
6.	Mācību_pabeigšanas_datums	"8.4.1. Demogrāfiskie dati" datne	0	-
7.	Programmas_ilgums	"8.4.1. Demogrāfiskie dati" datne	0	-
8.	Mācību_kārta	"8.4.1. Demogrāfiskie dati" datne	0	-