LATVIJAS UNIVERSITĀTE
DATORIKAS FAKULTĀTE

ANASTASIJA ŅIKIFOROVA

DATU KVALITĀTES DEFINĒŠANA UN NOVĒRTĒŠANA

Promocijas darbs

RĪGA 2020

LATVIJAS UNIVERSITĀTE
DATORIKAS FAKULTĀTE

ANASTASIJA ŅIKIFOROVA

DATU KVALITĀTES DEFINĒŠANA UN NOVĒRTĒŠANA

Promocijas darbs
datorzinātņu doktora (Dr.sc.comp.) zinātniskā grāda iegūšanai

Nozare: datorzinātnes
Apakšnozare: datu apstrādes sistēmas un datortīkli

Zinātniskais vadītājs:
asoc. profesore Dr. dat. **Zane Bičevska**
profesors Dr. dat. **Jānis Bičevskis**

RĪGA 2020

ANOTĀCIJA

Darbā tiek piedāvāta jauna pieeja datu kvalitātes problēmas risināšanai, kas tika izstrādāta, ņemot vērā esošo pieeju trūkumus. Pieejas pamatā ir trīs komponenti: datu objekts, kvalitātes prasības un process kvalitātes novērtēšanai. To aprakstīšanai tiek piedāvātas trīs grafiskās *DSL*, kuras ir pietiekami vienkāršas, lai tās varētu lietot ne-IT speciālisti. Ir nodrošināta datu kvalitātes pārbaude atkarībā no datu lietojuma, kontekstuālā pārbaude vairāku datu objektu ietvaros, kā arī ir piedāvāta pieejas formalizācija daļēji formālas datu kvalitātes teorijas izveidei.

Piedāvātā pieeja ļauj analizēt "ārējo" datu kopu kvalitāti, nodrošinot iespēju analizēt atvērtus datus, kuri kļūst arvien populārāki visā pasaulē, tajā skaitā arī Latvijā. Tā ir pielietota 30 atvērto datu kopām, atklājot tajās kvalitātes problēmas, tādejādi apliecinot piedāvātās pieejas priekšrocības.

Atslēgvārdi: datu kvalitāte, datu objekts, datu kvalitātes dimensija, datu kvalitātes novērtēšana.

ABSTRACT

DEFINITION AND EVALUATION OF DATA QUALITY

The doctoral thesis proposes a new data object-driven approach to evaluate data quality that was developed taking into account the cons of the existing solutions. The approach is based on 3 main components: data object, data quality requirements and the process of data quality measuring. These components are defined by 3 graphical DSLs, that are easy enough even for non-IT experts. The approach ensures data quality analysis depending on the use-case, contextual analysis within several data objects. The formalization of the approach for partially formalised theory of data quality is proposed as well as well.

Developed approach allows analysing quality of "third-party" data. The proposed solution is applied to 30 open data sets, detecting multiple data quality issues, thus demonstrating the advantages of the proposed approach.

Keywords: data object, data quality, dimension, data quality evaluation.

SATURS

APZĪMĒJUMU SARAKSTS 6

IEVADS 7

1. DATU KVALITĀTES JĒDZIENS 16

1.1. Datu kvalitātes jēdziens 16

1.2. Atvērtie dati 18

1.2.1. Atvērto datu popularitāte 19

1.2.2. Atvērtie medicīnas dati 20

1.2.3. Atvērtie dati Latvijā 22

1.3. Datu kvalitātes problēma un tās aktualitāte 23

2. DATU KVALITĀTES NOVĒRTĒŠANAS RISINĀJUMI 29

2.1. Datu kvalitātes novērtēšana ar dimensiju palīdzību 30

2.2. Atvērto datu portālu kvalitātes novērtēšanas pieejas 40

2.3. Saistīto datu kvalitātes novērtēšanas pieejas 43

2.4. *Data Quality Services* 46

2.5. Apkopojums 49

3. PIEDĀVĀTAIS DATU KVALITĀTES MODELIS 53

3.1. Pieejas vispārīgs apraksts 53

3.2. Datu objekts 61

3.3. Kvalitātes specifikācija 67

3.4. Kvalitātes pārbaudes process 70

3.5. Piedāvātās pieejas implementācija 74

3.6. Kontekstuālā datu kvalitātes pārbaude 77

3.7. Apkopojums 84

4. PIEDĀVĀTĀS PIEEJAS FORMALIZĀCIJA 87

4.1. Datu objekta formalizācija 88

4.2. Datu kvalitātes prasību “pirms-” un “pēc-” nosacījumi 89

5. PIEEJAS PIELIETOŠANAS REZULTĀTI 94
5.1. Uzņēmumu reģistru datu kvalitātes analīzes rezultāti 94
5.2. Datu kopu analīzes rezultāti 103
5.2.1. Datu kopas "Interešu un pieaugušo neformālās izglītības programmu licences" analīzes rezultāti 103
5.2.2. Datu kopas "Statistika par saziņu ar Rīgas pašvaldību" analīzes rezultāti 104
5.2.3. Datu kopas "Valsts informācijas sistēmu reģistrs" analīzes rezultāti 104
5.3. Latvijas atvērto medicīnas datu kvalitātes analīze 105
5.4. Datu kopu datu kvalitātes analīzes rezultātu apkopojums 113
NOBEIGUMS 116
PATEICĪBAS 121
IZMANTOTĀ LITERATŪRA UN AVOTI 122

APZĪMĒJUMU SARAKSTS

Apzīmējums	**Skaidrojums**
DAMA UK Working Group	*Data Management Association International UK Working Group*
DQS	*Data Quality Services*
DSL	*Domain Specific Language*, domēnspecifiskā valoda
ETL	*Extract, Transform, Load*
ISO	*International Organization for Standardization*
IS	Informācijas sistēma
KVS	Kvalitātes Vadības Sistēma
MDA	*Model driven approach*, modeļa virzīta izstrāde
OGD	*Open Government Data*, atvērtie pārvaldes dati
PIM	*Platform independent model*, platformneatkarīgs modelis, PNM
PSM	*Platform specific model*, platformatkarīgs modelis, PAM
SPARQL	*RDF* (*Resource Description Framework*) vaicājumu valoda
SPDQM	*Square-Aligned Portal Data Quality Model*
TDQM	*Total Data Qualtiy Management*

IEVADS

ISO 9000 definē "kvalitāti" kā pakāpi, kurā tiek apmierinātas patērētāja vajadzības, reprezentējot visas klientam pieprasītās produkta vai pakalpojuma pazīmes un īpašības. Savukārt no "kvalitātes" jēdziena atvasināto jēdzienu "datu kvalitāti" visbiežāk definē kā datu piemērotību lietojumam, uzsverot tā relatīvo un dinamisko raksturu, kura kontekstu nosaka datu lietošanas piemērs un no tā atkarīgas prasības, kas laika gaitā var mainīties.

Datu kvalitātes problēma ir aktuāla kopš 60-o gadu beigām, kad tās atsevišķus aspektus sāka pētīt statistikas pētnieki. Datorzinātnieki datu kvalitātes problēmu sāka aktīvi pētīt 90-o gadu sākumā. Taču neskatoties uz datu popularitāti un to apjoma nepārtrauktu pieaugumu, gandrīz 30 gadus vēlāk datu kvalitātes problēma vēl joprojām nav atrisināta un ir aktuāla, kas galvenokārt ir saistīts ar datu un atvērto datu popularitāti. Vairums eksistējošo risinājumu balstās uz datu kvalitātes dimensiju definēšanu, grupēšanu un to pielietošanu datu kopām, ko paši datu pētnieki bieži vien atzīst par problemātisku uzdevumu pat datu kvalitātes speciālistiem. Tādejādi ir pamats apgalvot, ka eksistējošās pieejas nav piemērotas lietotājiem bez padziļinātām zināšanām IT un datu kvalitātes jautājumos, līdz ar to datu kvalitātes specialistu iesaiste kļūst nepieciešama visos datu kvalitātes analīzes posmos. Mūsdienu apstākļos tas nav pieņemami, jo katru dienu lietotāji saskaras ar datiem – tie ir visur, līdz ar ko iespējai pārbaudīt to kvalitāti ir jābūt katram lietotājam neatkarīgi no viņa zināšanām IT un datu kvalitātes jomās.

Darba mērķis ir izstrādāt pieeju, kas ļautu definēt analizējamo datu objektu un tā kvalitātes prasības lietotājiem, kuriem var nepiemist padziļinātas zināšanas IT vai datu kvalitātes jomās, kā arī pielietot to atvērto datu kopām, nodemonstrējot to darbībā, un praktiskā risinājuma formalizācijas rezultātā piedāvāt datu kvalitātes teoriju.

Darba mērķa sasniegšanai tika izvirzīti vairāki **uzdevumi**:

1) izpētīt "datu kvalitātes" un "atvērto datu kvalitātes" jēdzienus, problēmas, to aktualitāti, kā arī to popularitāti zinātniskajos rakstos;
2) izpētīt un novērtēt eksistējošas datu kvalitātes analīzes un novērtēšanas pieejas, nosakot tās priekšrocības un trūkumus;
3) izvirzīt prasības jaunai datu kvalitātes novērtēšanas pieejai, ņemot vērā noteiktus eksistējošo pieeju trūkumus;
4) piedāvāt jaunu datu kvalitātes analīzes un novērtēšanas pieeju, kas atbilst izvirzītājām prasībām, uzsverot tās priekšrocības, salīdzinot ar eksistējošām pieejām;

5) piedāvāt izstrādātās pieejas formalizāciju daļēji formālas datu kvalitātes teorijas izveidei;
6) novērtēt piedāvāto datu kvalitātes novērtēšanas pieeju, pielietojot to vairākām atvērto datu kopām, ar mērķi identificēt tajās datu kvalitātes problēmas;
7) veikt secinājumus par atvērto datu kvalitāti, balstoties uz piedāvātās pieejas pielietošanas rezultātiem;
8) veikt secinājumus par pētījuma rezultātā iegūto pieredzi;
9) darīt iegūtus rezultātus zināmus pasaulei, publicējot zinātniskus rakstus un uzstājoties starptautiskās konferencēs.

Promocijas darbā ir piedāvāta jauna, saukta datu objekta virzīta, lietotājorientēta pieeja datu kvalitātes definēšanai un novērtēšanai. Par pieejas oriģinalitāti liecina darba autores eksistējošo risinājumu analīze, kā arī *Batini* (datu kvalitātes jautājumos vadošā pētnieka) datu kvalitātes problēmas dziļš izpētes darbs un eksistējošo metodoloģiju pārskats ((Batini et al., 2009, 2016)). Taču neskatoties uz pieejas pamatidejas būtisku atšķirību no citiem risinājumiem, atsevišķas piedāvātā risinājuma idejas saskaņojas ar *Batini* uzskatiem.

Piedāvātais kvalitātes modelis sastāv no trim komponentiem: (1) datu objekts, kura kvalitāte tiek vērtēta, (2) datu kvalitātes prasības – nodefinētajam datu objektam definētas kvalitātes prasības, kas ir atkarīgas no konkrēta datu lietojuma, un (3) datu kvalitātes pārbaudes process, kura izpildes rezultātā tiek lemts par dotā datu objekta kvalitāti, analizējot tajā konstatētās datu kvalitātes problēmas. Piedāvātā pieeja būtiski atšķiras no eksistējošām pieejām – tā neizmanto "datu kvalitātes dimensijas" jēdzienu, ļaujot lietotājiem pašiem definēt specifiskās kvalitātes prasības ar viņu noteiktiem datu objektiem atkarībā no datu lietojuma jeb lietošanas piemēra. "Datu kvalitātes dimensijas" jēdziena vietā tiek izmantots plašāks "datu kvalitātes prasības" jēdziens, kas var tikt uzskatīts par uz datu kvalitāti attiecināmu datu kvalitātes dimensiju virskopu. Datu objektu un kvalitātes prasības konkrētam datu objektam definē lietotājs, līdz ar ko lietotājiem ir sniegta iespēja pārbaudīt konkrētas datu kopas datu kvalitāti saviem nolūkiem. Katrs komponents tiek definēts, izmantojot grafiskas blokshēmām līdzīgas diagrammas, kas ļauj atvieglot datu kvalitātes analīzes procesu, kā arī nodrošināt vairāku lietotāju mijiedarbību, veicinot lietotāju savstarpēju saziņu ar diagrammu palīdzību, ko ir iespējams ātri un vienkārši veidot un labot. Tas tiek panākts katram komponentam izstrādājot grafisko domēnspecifisko valodu (*DSL*). Kvalitātes modelis var tikt definēts divos veidos – neformāli, izmantojot dabisko valodu, vai formāli, neformālus tekstus aizstājot ar izpildāmiem, piemēram, *SQL* vaicājumiem. Izstrādātās diagrammas ir attīstāmas līdz izpildāmām, līdz ar ko datu kvalitātes novērtēšanas process kļūst automatizēts. Datu objekta un datu kvalitātes prasību definēšana neprasa no lietotājiem iepriekšējas zināšanas IT vai datu kvalitātes jomā, šīs process

ir intuitīvs, līdz ar ko, atšķirībā no esošo datu kvalitātes risinājumu lielākas daļas, piedāvātā pieeja ir paredzēta plašam lietotāju lokam. IT specialistu iesaiste kļūst nepieciešama tikai beidzamajā posmā - neformālas prasības pārveidojot par izpildāmām.

Dotā pieeja paredz arī konteksta pārbaudes, analizējot datu kopas kvalitāti pret citām datu kopām, kas ir nepieciešamas, veicot padziļinātu datu kvalitātes analīzi. Datu objekts, kura kvalitāte tiek analizēta, kļūst par primāro datu objektu, savukārt pārējie datu kvalitātes analīzē iesaistītie datu objekti, pret kuriem tiek pārbaudīta primārā datu objekta kvalitāte, kļūst par sekundārajiem datu objektiem. Sekundārais datu objekts parasti ir datu kopa, kas tika uzkrāta un apstrādāta ar citu no primārā datu objekta neatkarīgu datu sniedzēju, līdz ar ko kļūst iespējams pārbaudīt primāra datu objekta kvalitāti pret citu neatkarīgu datu objektu. Viena primāra datu objekta datu kvalitātes analīzē iesaistīto sekundāro datu objektu skaits nav ierobežots.

Piedāvātais risinājums nodrošina iespēju veikt "trešo pušu" datu kvalitātes analīzi, t.i. analizēt datus, informācija par kuru uzkrāšanas un apstrādes mehānismiem vai procedūrām var nebūt zināma. Risinājums tiek pielietots atvērtajiem datiem, vienlaicīgi pārliecinoties pieejas efektivitātē un atvērto datu kvalitātē, lielāku uzsvaru liekot uz Latvijas atvērtajiem datiem. Doto risinājumu autore pielietoja arī vienam specifiskam domēnam – medicīnas datiem. Atvērto datu kvalitātes analīze pati par sevi ir izaicinājums, jo, neskatoties uz atvērto datu popularitātes pieaugumu, atvērto datu kvalitātes jautājums tiek pētīts salīdzinoši reti, par ko liecina arī atbilstošās tēmas reprezentējošo pētījumu skaits *Google Scholar*. Statistika rāda, ka laika periodā no 2003. līdz 2014. gadam tika publicēti 4.6 reizes mazāk pētījumu par atvērto datu kvalitāti nekā 2018. gadā, taču, attiecinot ar atvērto datu kvalitāti saistīto pētījumu skaitu pret kopējo ar atvērtajiem datiem saistīto pētījumu skaitu, ir redzams, ka datu kvalitātes jautājums tiek pētīts nepamatoti reti, jo 2018. gadā atvērto datu pētījumu skaits pārsniedz ar atvērto datu kvalitāti saistīto pētījumu skaitu 147 reizēs, t.i. atvērto datu kvalitātes pētījumu īpatsvars pret kopējo ar atvērtajiem datiem saistīto pētījumu skaitu nepārsniedz 0.5%. Pie tam datu kvalitātes pētījumu skaits pārsniedz atvērto datu kvalitātes pētījumu skaitu gandrīz 196 reizēs. Pieaugot atvērto datu apjomam, kļūst nepieciešami risinājumi, kas būtu piemēroti arī lietotājiem bez padziļinātājām zināšanām datu kvalitātes un IT jomā, jo atvērtie dati kļūst par ikdienas parādību, un arī to kvalitātes analīze kļūst par neatņemamu ikdienas darbību. Pētījuma ietvaros, piedāvāto pieeju pielietojot atvērtajiem datiem, autore konstatēja tajos vairākas datu kvalitātes problēmas, kuras, ņemot vērā to raksturu, izdalīja atsevišķās grupās, lai pievērstu uzmanību kopīgajām problēmām, no kurām būtu jāuzmanās datu lietotājiem, un jāņem vērā datu sniedzējiem, izceļot populārākās, kas ir raksturīgas ne tikai Latvijas, bet arī citu valsts datu kopām, kas tika noteiktas gan veiktā pētījumā, gan citu pētījumu izpētes rezultātā.

Datu kvalitātes tēma ir bieži diskutējama arī starptautiskajās zinātniskajās konferencēs, piemērām, *ICEIS*, *QRS*, *RQD*, *MCCSIS*, *DSEA* (*SNAMS)*, *FedCSIS*, *QUATIC*. Uz 2019. gada novembra beigām tika izsludinātas pieteikšanās uz 51 starptautisko konferenci par doto tēmu, kas liecina par tēmas aktualitāti pētniecībā. Tāpat konferencēs arvien biežāk datu kvalitātes problēma tiek uzsvērta kā viena no nozīmīgākajām problēmām, ar kurām sastopas dažādu jomu pētnieki (piemēram, 2019. gadā *SNAMS* konferencē tā tika uzsvērta par vienu no galvenajiem ierobežojumiem ar Lietu Internetu (angl. *Internet of Things*), blokķēdēm, viltus ziņu (angl. *fake news*) atpazīšanu u.c. saistītājos pētījumos), tajā skaitā arī starpdisciplināro pētījumu īstenotāji. Arī Latvijā šī tēma ir aktuāla, par ko vairākkārt tika runāts vairākos pasākumos, tajā skaitā Digitālās nedēļās ietvaros, IKT profesionāļu dienās un ar Ekonomikas ministriju rīkotājā *"Digitālizācijas un Inovāciju foruma DIG-IN"*, iekļaujot datu kvalitātes pētniekus nākotnes profesiju sarakstā.

Darbā izvirzītās **tēzes**:

1) datu kvalitātes problēma ir aktuāla, neskatoties uz savu vecumu;
2) datu kvalitātes un ar to saistītie jēdzieni vēl joprojām nav skaidri un viennozīmīgi nodefinēti;
3) neskatoties uz atvērto datu popularitāti un to strauju pieaugumu, atvērto datu kvalitātes problēma tiek reti pētīta;
4) vairākums eksistējošo risinājumu datu kvalitātes analīzei nav piemērots lietotājiem bez padziļinātām zināšanām IT un datu kvalitātes jomās, prasot to iesaisti visos datu kvalitātes analīzes posmos. Ir iespējams izstrādāt risinājumu, kas ļautu datu kvalitātes analīzes procesā iesaistīties lietotājiem, kuriem var nepiemist padziļinātas zināšanas IT un datu kvalitātes jomās, datu kvalitātes analīzi veicot atbilstoši saviem lietošanas piemēriem;
5) grafisko modeļu izmantošana vienkāršo datu kvalitātes modeļa komponenšu definēšanas procesu un sekmē vairāku lietotāju savstarpējo mijiedarbību;
6) kontekstuālās datu kvalitātes analīzes iespējas ļauj veikt padziļinātu datu kvalitātes analīzi, būtiski uzlabojot tās rezultātus, kā arī nodrošinot veidoto objektu atkalizmantošanas iespējas;
7) atvērtajos datos ir sastopamas datu kvalitātes problēmas, kuru noteikšanai ir piemērota izstrādātā pieeja datu kvalitātes analīzei, nodrošinot "trešo pušu" datu kvalitātes analīzi;
8) izstrādātais risinājums ir pielāgojams izpildlaika datu kvalitātes analīzei;

9) skaidri un viennozīmīgi nodefinētie piedāvātā datu kvalitātes modeļa komponenti un izstrādātā risinājuma specifika ļauj piedāvāt neformālu datu kvalitātes teoriju, kura līdz šīm netika piedāvāta.

Izvirzīto mērķu sasniegšanai pētījuma gaitā tika **izmantotas sekojošas pētniecības metodes**:

- analītiskā metode - dažādu avotu analīze, lai nodefinētu un izpētītu konkrētu tēmu, problēmu un atrisinātu to;
- salīdzinošā metode – salīdzināt vairākus risinājumus definētās problēmas risināšanai, nosakot to priekšrocības un trūkumus, kas būtu ņemami vērā, izstrādājot savu risinājumu;
- eksperimenta metode – praksē pielietot un pārbaudīt literatūras avotos iegūtās zināšanas, izstrādājot savu risinājumu; pielietot izstrādāto pieeju atvērtajiem datiem, ar mērķi novērtēt pieejas efektivitāti un tās pielietojamību "trešo pušu" datiem, un konstatēt tajos datu kvalitātes problēmas;
- aptauja – aptaujas veikšana ar mērķi pārbaudīt veiktos pieņēmumus;
- rezultātu novērtējums – pieejas pielietošanas atvērtajiem datiem to kvalitātes analīzei rezultātu novērtējums, pārbaudot iegūtus datu kvalitātes problēmu protokolus;
- aprakstošā metode – izpētīt problēmu, lai atrisinātu to, apkopojot izlasīto literatūru, aprakstot darba gaitu un iegūtos rezultātus, tos publicējot zinātniskajos rakstos.

Darba rezultāti:

1) ir nodefinēts "datu kvalitātes" jēdziens; ir izpētīta [atvērto] datu kvalitātes problēma un tās aktualitāte;
2) literatūras izpētes rezultātā ir noteiktas populārākas ar datu kvalitāti saistītas tēmas, iezīmējot apgabalus, kas tiek pētīti nepamatoti maz;
3) pētījuma gaitā autore veica vairāk kā 65 eksistējošo risinājumu analīzi, darbā apskatot vairāk kā 25 risinājumus, kas pēta datu kvalitātes problēmu un prezentē risinājumus datu kvalitātes problēmu konstatēšanai un novēršanai, nosakot to priekšrocības un trūkumus;
4) ir nodefinēta un izstrādāta jaunā datu objekta virzīta pieeja datu kvalitātes novērtēšanai, kas ir piemērota lietotājiem bez padziļinātām zināšanām IT un datu kvalitātes jomās, kā arī pielietojama "trešo pušu" datu kopām;
5) ir piedāvāta izstrādātās pieejas formalizācija, piedāvājot datu kvalitātes teoriju;

6) piedāvātā datu objekta virzītā pieeja ir aprobēta, pielietojot to atvērtajiem datiem, rezultātā konstatējot tajos vairākas datu kvalitātes problēmas, kas ir iedalītas grupās pēc sava rakstura;
7) pētījuma rezultāti ir nopublicēti zinātnisko rakstu ciklā un prezentēti starptautiskajās konferencēs.

Promocijas darbs ir autores maģistra darba (Nikiforova, 2019b) turpinājums, kas IT noslēguma darbu konkursā ZIBIT tika atzīts par gada labāko maģistra darbu Latvijā (pēc SIA Tieto Latvija, AKF Accenture, AS Emergn un RTU attīstības fonda viedokļa).

Datu kvalitātes tēmas pētījumus darba autore veic gandrīz trīs gadus, tos uzsākot, piedaloties IT kompetences centra pētījumā Nr. 1.8 "Datu kvalitātes pārvaldība ar izpildāmiem biznesa procesu modeļiem" (ERAF līdzfinansētais projekts). Šobrīd uz darba rezultātiem balstās arī 2019. gadā uzsāktais IT kompetences centra pētījums Nr. 1.7. "Biznesa procesu modeļu lietojums pilnai informācijas sistēmas funkcionalitātes testēšanai".

Par saviem sasniegumiem dotajā tēmā 2019. gadā Beļģijas Universitātes prof. *Marc Nyssen* ir nominējis darba autori uz vienu no pasaules prestižākajiem apbalvojumiem datu zinātnes jomā "*WDS Data Stewardship Award*", kas ikgadēji tiek piešķirts perspektīvākajiem datu zinātniekiem.

Pētījuma rezultāti ir apkopoti zinātniskos rakstos, kas publicēti starptautiski atzītos izdevumos:

1. Nikiforova, A. (2019). Analysis of open health data quality using data object-driven approach to data quality evaluation*:* insights from a Latvian context. In *IADIS International Conference e-Health 2019, Part of the IADIS Multi Conference on Computer Science and Information Systems, MCCSIS 2019, July 16 - 19, 2019* (pp. 119-126). IADIS. Indeksēts *Scopus* (ieguldījums: 100%);
2. Nikiforova, A., Bicevskis, J. (2019). An extended data object-driven approach to data quality evaluation: contextual data quality analysis. *Proceedings of the 21st International Conference on Enterprise Information Systems (ICEIS 2019)*, 274-281. DOI: 10.5220/0007838602740281. Indeksēts *Scopus* (ieguldījums: 70%);
3. Bicevskis, J., Nikiforova, A., Bicevska, Z., Oditis, I., Karnitis, G. (2019). A step towards a data quality theory. In *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)* (pp. 303-308). IEEE. Indeksēts *Scopus* (ieguldījums: 50%);
4. Bicevskis, J., Bicevska, Z., Nikiforova, A., Oditis, I. (2019). Towards data quality runtime verification. In *2019 Federated Conference on Computer Science and Information Systems (FedCSIS)*. IEEE. Indeksēts *Scopus* (ieguldījums: 40%);

5. Nikiforova, A. (2018). Open data quality evaluation: a comparative analysis of open data in Latvia. *Baltic Journal of Modern Computing*, *6*(4), 363-386. Indeksēts *Web of Science* (ieguldījums: 100%);
6. Nikiforova, A. (2018). Open data quality. In *Doctoral Consortium/Forum@ DB&IS* (pp. 151-160). Indeksēts *Scopus* (ieguldījums: 100%);
7. Bicevskis, J., Bicevska, Z., Nikiforova, A., Oditis, I. (2018). An approach to data quality evaluation. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)* (pp. 196-201). IEEE. Indeksēts *Scopus* un *Web of Science* (ieguldījums: 50%);
8. Bicevskis, J., Bicevska, Z., Nikiforova, A., Oditis, I. (2018). Data quality evaluation: a comparative analysis of company registers' open data in four European countries. In *FedCSIS Communication Papers* (pp. 197-204) Indeksēts *CrossRef* (ieguldījums: 35%);
9. Nikiforova, A., Bicevskis, J., Bicevska, Z., Oditis, I., (2020). User-Oriented Approach to Data Quality Evaluation. *Journal of Universal Computer Science, 26*(1), 107-126. Indeksēts *Web of Science*.

Ar citu ar kvalitāti saistītu tēmu saistītais zinātniskais raksts, kurš ir indeksēts *Web of Science*:

10. Nikiforova, A., Bicevska, Z. (2018). Application of LEAN Principles to Improve Business Processes: a Case Study in Latvian IT Company. *Baltic Journal of Modern Computing*, *6*(3), 247-270. Indeksēts *Web of Science* (ieguldījums: 75%).

Darba rezultātus darba autore prezentēja piecās **starptautiskās konferencēs**:

- *"Data Science Engineering and its Application (DSEA 2019)"* (apvienojumā ar *"Sixth International Conference on Social Networks Analysis, Management and Security"(SNAMS 2019)*), Spānija, 2019;
- *"13th Multi Conference on Computer Science and Information Systems" (MCCSIS 2019)* (*"11th International Conference on e-Health"* apakškonference), Portugāle, 2019;
- "*21st International Conference on Enterprise Information Systems (ICEIS 2019)"*, Krēta, Grieķija, 2019;
- *"Data Science Engineering and its Application (DSEA 2018)"* (apvienojumā ar *"Fifth International Conference on Social Networks Analysis, Management and Security"(SNAMS 2018)*), Spānija, 2018;
- "*Baltic DB&IS 2018 Joint Proceedings of the Conference Forum and Doctoral Consortium*"), Lietuva, 2018.

IT Kompetences Centra pētījuma Nr. 1.8 "Datu kvalitātes pārvaldība ar izpildāmiem biznesa procesu modeļiem" rezultātus darba autore prezentēja Ekonomikas ministriju rīkotājā *"Digitālizācijas un Inovāciju foruma DIG-IN" POP*^{up} *Demo* centrā.

Savukārt pētījumu rezultātus, kas attiecas uz atvērto datu kvalitāti, 2020. gadā autore prezentēja LATA (Latvijas Atvērto Tehnoloģiju Asociācijas) konferencē "Datu virzītā nācija".

Tāpat dotais pētījums ir kalpojis par pamatu Specsemināriem "Datu kvalitāte" un "Atvērtie dati un datu kvalitāte", kurus darba autore 2019./ 2020. akadēmiska gadā ir sākusi lasīt Latvijas Universitātes Datorikas fakultātes bakalaura studiju programmas "Datorzinātne" studentiem.

Darbs sastāv no ievada, pamatdaļas ar 5 nodaļām un 21 apakšnodaļām ar vēl 6 punktiem un nobeiguma.

1. nodaļā ir sniegtas pamatjēdzienu (datu kvalitātes, atvērto datu) definīcijas, datu kvalitātes nozīmīguma apraksts un problēmas pamatojums. Tiek aplūkota arī atvērto datu popularitāte Latvijā, ar mērķi novērtēt nepieciešamību to analīzē. Tiek izpētīta arī datu un atvērto datu kvalitātes pētījumu popularitāte, balstoties uz *Google Scholar* datiem, nosakot tēmas, kurām tiek pievērsts mazāk uzmanības, identificējot apgabalus, kuru izpēte varētu sniegt pievienoto vērtību sabiedrībai.

2. nodaļā ir aprakstīti eksistējošie risinājumi un datu kvalitātes problēmas pētījumi, īsi aprakstot to trūkumus un priekšrocības. Ir nodefinētas kopīgas negatīvas raksturiezīmes, kuras būtu vērts ņemt vērā, izstrādājot jaunu pieeju.

3. nodaļā ir prezentēta piedāvātā jaunā pieeja datu kvalitātes definēšanai un novērtēšanai, kura ir izstrādāta, ņemot vērā eksistējošo risinājumu trūkumus. Piedāvātā pieeja saista datu kvalitātes jēdzienu ar datu lietojumu, ļaujot veikt datu kvalitātes analīzi katram lietotāja definētam lietošanas piemēram. Ir pamatota tās komponenšu izvēle, uzsverot tās priekšrocības. Ir aprakstīta datu kvalitātes analīze gan viena datu objekta, gan vairāku datu objektu ietvaros, kas atbilst kontekstuālai datu kvalitātes analīzei, primārā datu objekta kvalitāti pārbaudot pret sekundārajiem no tā neatkarīgiem datu objektiem, ar nolūku veikt tā padziļināto analīzi. Tā kā datu kvalitātes modelis var tikt definēts gan neformāli, gan formāli, kas atbilst modeļa virzītas izstrādes jeb *MDA* pamatidejai, piedāvātais risinājums tiek apskatīts no *MDA* skatupunkta, veicot to atbilstības *MDA* pamatidejām analīzi, uzsverot gan līdzības, gan atšķirības.

4. nodaļā tiek piedāvāta pieejas formalizācija, tādā veidā praktisko risinājumu pārveidojot datu kvalitātes teorijā, kura, neskatoties uz vairākiem mēģinājumiem, līdz šim vēl netika piedāvāta. Tas galvenokārt ir saistīts ar to, ka pētnieki nav spējuši piedāvāt viennozīmīgu definīciju visiem ar datu kvalitāti saistītājiem jēdzieniem – īpaši datu kvalitātes dimensijām, kuru skaits, nosaukumi un semantiskā nozīmē var atšķirties atkarībā no pētījuma. Neskatoties

uz ilggadējiem pētījumiem, datu kvalitātes dimensijām nav tiešas un viennozīmīgas definīcijas (gan terminam tā plašajā nozīmē, gan katrai atsevišķai dimensijai), kā arī to pielietošana konkrētam lietošanas gadījumam un mērīšanas procesi izraisa vairākas diskusijas un nesaprašanas gan datu lietotājiem, gan datu pētniekiem. Pie tam, vairākums risinājumu nav piemērots lietotājiem bez padziļinātājām ne tikai IT, bet arī datu kvalitātes zināšanām, līdz ar ko datu kvalitātes specialistu iesaiste kļūst nepieciešama visos datu kvalitātes analīzes posmos, kas nav pieņemami. Piedāvātās idejas pamatā ir skaidri un viennozīmīgi definētie jēdzieni, taču visu komponenšu definēšanā ir iesaistīts galalietotājs - tiek ņemts vērā datu un datu kvalitātes relatīvs raksturs, kas ir atkarīgi no konkrēta lietotāja un lietošanas piemēra.

5. nodaļā ir apkopoti piedāvātās pieejas pielietošanas rezultāti. Autore pielietoja pieeju vairāk kā 30 datu kopām, 4 no kurām ir četru valsts Uzņēmumu Reģistri, un 15 datu kopas reprezentē konkrētu domēnu – Latvijas medicīnas datus (t.i., atvērtie medicīnas dati). Tādā veidā autore analizē gan dažādu domēnu datu kopu kvalitāti, gan vienam domēnam piederošo datu kopu kvalitāti, veicot rezultātu salīdzinājumu. Rezultātā ir noteiktas kopīgās datu kvalitātes problēmas, kas ir raksturīgas visām datu kopām un konkrētiem domēniem piederošām, grupējot datu kvalitātes problēmas pēc to rakstura.

Promocijas darbā veiktā analīze un pētījums pamatojas uz zinātnieku darbiem, publikācijām zinātnisko rakstu krājumos un periodikā, internetā pieejamiem materiāliem, statistikas datiem. Kopumā darba izstrādes gaitā autore izpētīja vairāk kā 168 avotus, 106 no kuriem ir zinātniskās publikācijas, 17 - grāmatas, 6 – rokasgrāmatas un ziņojumi, kurus darba autore atlasīja, ņemot vērā rakstu citējāmību, 2 – promocijas darbi un 2 - maģistra darbi.

Darbā ir iekļauti 19 attēli, 5 tabulas un 168 literatūras avoti.

1. DATU KVALITĀTES JĒDZIENS

Nodaļā "Datu kvalitātes jēdziens" ir apskatīti kvalitātes, datu kvalitātes un atvērto datu jēdzieni, datu kvalitātes problēma un tās aktualitāte, uzmanību pievēršot atvērtajiem datiem, kas kļūst arvien populārāki visā pasaulē, kas sekmē arī datu kvalitātes problēmas aktualitātes un popularitātes pieaugumu. Tiek aplūkota arī zinātnisko pētījumu, kas fokusējas uz datu kvalitāti, atvērtiem datiem un atvērto datu kvalitāti, popularitāte pēdējos gados, veicot to savstarpējo salīdzinājumu, nosakot apgabalus, kas tiek pētīti nepamatoti reti. Tādejādi ir iezīmēti svarīgāki aspekti, kas tiek apskatīti darbā, uzsverot veiktā pētījuma svarīgumu un nozīmīgumu. Šī nodaļa balstās uz (Nikiforova 2018a, 2018b, 2019b).

1.1. Datu kvalitātes jēdziens

Pētījuma veikšana, kura rezultātā tiek piedāvāta jauna pieeja datu kvalitātes analīzei, paredz pamatjēdzienu definīciju, sākot no pašiem pamatiem, t.i. "kvalitātes" un "datu kvalitātes" jēdzieniem. Abiem jēdzieniem mēdz būt dažādas definīcijas, taču tiek aplūkotas visbiežāk sastopamas un visprecīzāk konkrētu jēdzienu raksturojošas definīcijas.

Atbilstoši *ISO 9000* "kvalitāte" ir pakāpe, kurā tiek apmierinātas patērētāja vajadzības, t.i. kvalitāte reprezentē visas klientam pieprasītās produkta vai pakalpojuma pazīmes un īpašības. Datu kvalitātes jēdzienam nav vienas vispārpieņemtas definīcijas, savukārt visbiežāk to definē kā piemērotību lietojumam (piemēram, (Tayi et al., 1998), (Olson, 2003) utt.). Daži pētnieki papildina to ar prasību, atbilstoši kurai datos nedrīkst būt datu kvalitātes problēmas, dažreiz norādot, kurām nepieciešamām/ "vēlamām" pazīmēm/ īpašībām datiem ir jāpiemīt (Scannapieco et al., 2002a), (Redman, 2001), (Wang et al., 1996). Šīs prasības mēdz atšķirties atkarībā no konkrētā risinājuma. Piemēram, *Lee* ar līdzautoriem uzskata, ka kvalitatīvus datus raksturo tādas īpašības kā datu pilnīgums (angl. *completeness*), nepretrunīgums (angl. *consistency*), ticamība (angl. *believability*), savlaicīgums (angl. *timeliness*), piekļūstamība (angl. *accessibility*), kā arī datiem ir jābūt pieejamiem atbilstoša apjomā (angl. *appropriate amount of data*) un kļūdas nesaturošiem (angl. *free of error*) (Lee at el., 2009). *Juran* piedāvātais īpašību, kurām ir jāpiemīt kvalitatīviem datiem, saraksts daļēji pārklājās ar (Lee et al., 2009). Atbilstoši (Juran, 1995) datiem ir jābūt pieejamiem, precīziem (angl. *accurate*), savlaicīgiem, pilnīgiem, nepretrunīgiem ar citiem avotiem (angl. *consistent with other source*), relevantiem (angl. *relevant*), visaptverošiem (angl. *comprehensive*), atbilstošiem detalizācijas līmenim (angl. *proper level of detail*), viegli lasāmiem (angl. *easy to read*), viegli

interpretējamiem (angl. *easy to interpret*) utt. (Nikiforova, 2019b). Datu kvalitātes jautājumam ir pievērsušies arī ISO standartu izstrādātāji, 2015. gadā izstrādājot *ISO/IEC 25024 Systems and software engineering - Systems and software Quality Requirements and Evaluation (SQuaRE) - Measurement of data quality* standartu, kurā datu kvalitāti raksturo ar tādām 15 raksturiezīmēm kā precizitāte, pilnīgums, nepretrunīgums, ticamība (angl. *credibility*), izplatība (angl. *currentness*), piekļūstamība, atbilstība, konfidencialitāte, efektivitāte, precizitāte, trasējamība, saprotamība, pieejamība, pārnesamība un atkopjamība (ISO 25024:2015, 2015). Ir jāatzīmē, ka īpašību saraksti, kas raksturo datu kvalitāti, mēdz būt ļoti daudzveidīgi (atbilst darba sākumā izvirzītajai 2. tēzei). Šīs jautājums ir detalizētāk apskatīts 2. nodaļā.

Kopumā, atbilstoši (Nikiforova, 2018b, 2019b) datu kvalitāte ir datu kopas un tās īpašību piemērotība konkrētam jautājumam/ uzdevumam jeb lietošanas piemēram, kas ir atkarīgs no datu lietotāja jeb datu patērētāja, kas izmanto tos, piemēram, analītikā, darījumlēmumu pieņemšanā un plānošanā. Vadošie datu kvalitātes pētnieki *Batini* un *Scannapieco* uzsver, ka datu kvalitāte ir ārkārtīgi svarīga gan tā saucamajos lēmumu, gan operatīvajos procesos (Batini et al., 2009).

No "datu kvalitātes" jēdziena definīcijas seko, ka atkarībā no lietošanas piemēra vieni un tie paši dati mēdz būt piemēroti vienam lietošanas piemēram vai lietotājam, bet nelietojami nepieņemami zemas kvalitātes dēļ citam ((Tayi et al., 1998), (Nikiforova, 2018b)). Datu kvalitātes relatīvais raksturs un tās atkarība no lietošanas piemēra (Bertossi et al., 2016) ir datu kvalitātes pamatprincipi, kas ir jāņem vērā, analizējot datu kvalitāti. Tas nozīmē, ka atkarībā no lietojuma vieniem un tiem pašiem datiem var būt nepieciešams definēt atšķirīgas datu kvalitātes prasības.

Datu kvalitātes līmenis, pie kura dati apmierinātu visus iespējamus lietošanas piemērus, tiek uzskatīts par absolūto datu kvalitāti. Tiek uzskatīts, ka absolūtā [datu] kvalitāte nav sasniedzama, taču tas ir mērķis, uz kuru tiekties. Šīs princips ir raksturīgs vairākiem ar kvalitāti saistītājiem jautājumiem, piemēram, *KVS* (Kvalitātes Vadības Sistēmai), kā arī daudzām metodoloģijām, piemēram *LEAN* un tās pēctečiem, kuras viens no galvenajiem mērķiem ir darījumprocesu efektivitātes uzlabošana, kur šī īpašība kļūst par filozofijas pamatu (Nikiforova et al., 2018). Ir jāatzīmē, ka arī projektu nesekmības problēma, kuru parasti saista ar nepareizo projektu pārvaldības metodoloģijas izvēli vai arī nevienas metodoloģijas neizmantošanu, ir saistīta arī ar datu kvalitāti, jo atbilstoši (VismaLatvia, 2015) ap 40% darījuminiciatīvu cieš neveiksmi nepietiekamas datu kvalitātes dēļ.

Taču 21. gadsimtā paradījās jauns no "datu" jēdziena atvasinātais jēdziens – "atvērtie dati", rezultātā radot jaunus izaicinājumus, kas izriet no to rakstura. "Atvērto datu" jēdziens un tā popularitāte ir apskatīti nākamajās apakšnodaļā.

1.2. Atvērtie dati

Mūsdienās datu kvalitātes problēmas popularitāte pieaug arvien straujāk, lielākoties pateicoties atvērto datu popularitātei. Arvien biežāk valstis sniedz lietotājiem iespēju iegūt datus no atvērto datu portāliem, kuros savus datus publicē dažādi datu sniedzēji. Šos datus datu lietotāji var izmantot saviem nolūkiem, sākot ar vienkāršām datu analīzēm saviem nolūkiem, turpinot ar mūsdienās populāru lietotņu, kas balstās uz atvērtajiem datiem, izstrādes tendenci.

Atvērtie dati ir vēlamais veids, kādā vispārpieejamus datus nodot atkalizmantošanai (VARAM, 2016). Tie ir publiski pieejami dati, uz kuriem neattiecas privātuma, drošības vai privilēģiju ierobežojumi, kā arī tie nesatur personīgus, drošību ietekmējošus, kultūrsensitīvus, nepilnīgus vai maldinošus datus, kā arī datus, kas var tikt atzīti par komercnoslēpumu. Atbilstoši (VARAM, 2016) "… atvērto datu pieeju var attiecināt uz visu informāciju, kuru valsts iestāde publicē, piemēram, publiskiem reģistriem un valsts informācijas sistēmu publiskajām daļām, pētījumiem, statistiku, tabulām utt. Pieejas pamatā ir ideja, ka informācija sabiedrībai ir jānodod tādā formā, lai to varētu apstrādāt un brīvi lietot".

Atbilstoši ((Bauer et al., 2011), (Latvijas Atvērto datu portāls, 2018a), (Sunlight Foundation, 2017)), lai dati tiktu atzīti par atvērtajiem datiem, tiem ir jābūt:

- pilnīgiem – visiem datiem ir jābūt pieejamiem un pēc iespējas pilnīgiem (likumu robežās). Ir jābūt pieejamiem arī metadatiem, kas definē un apraksta publicēto datu kopu, skaidrojot arī kā tika iegūtas vai aprēķinātas konkrēto parametru vērtības, sniedzot datu lietotājiem iespēju pētīt datus pēc iespējas dziļākā detalizācijas līmenī;
- primārajiem – publicētājiem datiem ir pilnībā jāatbilst datu avotam, no kura tie tika izgūti, ar augstāko iespējamo granularitātes pakāpi, neveicot to apstrādi vai apkopošanu;
- laicīgiem – datiem ir jābūt pieejamiem pēc iespējas ātrāk, nodrošinot pēc iespējas laicīgāku datu nodošanu galalietotājam;
- pieejamiem – datiem ir jābūt pieejamiem visplašākajam lietotāju lokam visiem iespējamajiem nolūkiem;
- mašīnlasāmiem – datiem ir jābūt strukturētiem, lai tie varētu tikt automatizēti apstrādāti. Piemēram, dati, kas ir sniegti *.pdf* formātā ir grūti apstrādājami un netiek uzskatīti par atvērtajiem datiem. Datiem ir jābūt pieejamiem plaši lietojama mašīnlasāma formātā (piemēram, *.csv, .xls* utt.);

- nediskriminējošiem – datiem ir jābūt pieejamiem visiem bez ierobežojumiem, izslēdzot nepieciešamību reģistrēties to iegūšanai;
- atvērtajā datu formātā (angl. *non-proprietary*) – datiem ir jābūt pieejamiem tādā formātā, par kuru nevienam nav īpašas kontroles;
- bez licences – uz datiem nedrīkst būt attiecinātām nekādām autortiesībām, preču zīmju vai patentu likumiem. Uz tiem var tikt attiecināti tikai un vienīgi pamatoti privātuma, drošības un privilēģiju ierobežojumi, ja to nosaka likumi (Nikiforova, 2018a, 2019b).

Atbilstoši autores rakstā (Nikiforova, 2018b) norādītajam, neviens atvērto datu princips nav saistīts ar datu kvalitāti tās plašajā nozīme, līdz ar ko ir pamats veikt pieņēmumu, ka atvērtajos datos (pat tajos, kas pilnībā ievēro visus iepriekšminētos principus) mēdz būt datu kvalitātes problēmas.

Tā pati tendence ir novērojama *OGD* novērtēšanas gadījumā, jo atbilstoši (Klein et al., 2018), kvalitātes aspekts ieņem tikai 4. pozīciju (no četrām) pēc popularitātes, sekojot aiz politikas/ likumības (angl. *policy*), labuma (angl. *benefit)* un riska, neskatoties uz to, ka kvalitāte var ietekmēt katru aspektu. Taču atbilstoši *European Data Portal* pētījumiem, datu kvalitāte ir problemātiskākais atvērto datu portālu aspekts. Tas atbilst arī vairāku pētījumu rezultātiem.

Ņemot vērā, ka atvērto datu portālos tiek publicēti dažādu datu sniedzēju dati, datu kopu kvalitāte viena portāla ietvaros mēdz variēt, kas atbilst arī ((Ngomo et al., 2014), (Kuk et al., 2011), (Petychakis et al., 2014)). Tas mēdz būt saistīts ar to, ka, publicējot datu kopas, datu portāli reti pārbauda to kvalitāti, kas parasti ir saistīts ar datu kvalitātes pārbaudes procesa sarežģītību. Rezultātā (Kuk et al., 2011) un (Yi, 2019) ir secināts, ka mūsdienās *OGD* datu kvalitāte ir nepietiekoši augsta, un kvalitātes problēmas sākās ar neregulāro datu atjaunošanu un nepareizo formātu izvēli, kas lielākoties attiecās uz datu kopu kvalitāti, turpinot ar datu kvalitātes problēmām, kuras visbiežāk ir saistītas ar datu nepilnīgumu, nosaukumu un identifikatoru pretrunīgumu, zemo granularitātes līmeni.

1.2.1. Atvērto datu popularitāte

Viens no atvērto datu portālu piemēriem ir *European Data Portal* *https://www.europeandataportal.eu*, kurā 2019. gada novembra beigās bija pieejamas vairāk kā 970 tūkstoši datu kopas. Atvērtie dati būtiski ietekmē arī pasaules ekonomiku, un atbilstoši (Tinholt, 2013) tikai lietotņu, kas balstās uz atvērtajiem datiem, ietekme uz *EU27* ekonomiku pārsniedz €140 miljardi gadā. Tajā pašā laikā atbilstoši 2013. gada pētījumam, ko veica

McKinsey Global Institute, atvērtie dati potenciāli ir spējīgi uzlabot globālo ekonomiku par $3 triljoniem gadā (Castro et al., 2015).

G8 Open Data Charter īpaši uzsver atvērto datu un to kvalitātes lomu inovāciju izstrādē un ieviešanā, un pārvalžu "caurspīdīgumā" (atbilst arī (Ubaldi, 2013). (Bullinger et al., 2012)). Atvērtie dati ļauj paaugstināt/ uzlabot valsts ekonomiku, uzlabot pārvalžu pakalpojumu kvalitāti un samazināt vai pat novērst krāpšanas, pārtēriņus un ļaunprātīgu izmantošanu pārvalžu programmās. Paaugstināts caurspīdīgums nodrošina sabiedrības iesaisti un sadarbību ar mērķi veidot inovatīvus pakalpojumus, kas sniedz pievienoto vērtību sabiedrībai. Taču ar jaunām iespējām nāk arī jauni izaicinājumi, un valdēm ir jārēķinās ar to, ka datiem, ko tās publicē, ir jābūt kvalitatīviem, jo tikai tad no tiem ir jēga.

Iepriekšminēto mērķu sasniegšanai atvērtajiem datiem ir jābūt kvalitatīvajiem (atbilst arī (Zuiderwijk et al., 2012)).

Diskusija par atvērto datu popularitāti zinātniskajos pētījumos, salīdzinājumā ar datu kvalitātes pētījumiem, ir pieejama 2.5. apakšnodaļā.

1.2.2. Atvērtie medicīnas dati

Atvērtus datus mēdz iedalīt dažādās kategorijās (piemēram, transports, kultūra, valsts pārvalde, izglītība, vide, veselība utt.), taču viena no populārākām un svarīgākām atvērto datu kategorijām, ir medicīnas dati. Pēdējo dekāžu laikā medicīnas datu apjoms nepārtraukti pieaug, un ir sagaidāms, ka tuvāko gadu laikā tas pieaugs vēl vairāk (Raghupathi et al., 2014).

Atbilstoši *European Open Data* portālam, Eiropā atvērtie medicīnas dati ieņem 6. pozīciju no 13 kategorijām. Latvijā 2019. gadā tie ieņēma 6. no 14 pozīcijām pēc publicēto datu kopu skaita (salīdzinājumam, 2014. gadā tie ieņēma 7. no 9 pozīcijām (Bojārs et al., 2014)). Mērķi un iespējami atvērto medicīnas datu lietojumi mēdz būt ļoti dažādi, jo medicīnas dati un informācija ir raksturojami ar dažādiem iespējamiem lietojumiem, lietotājiem un lietotnēm (Cabitza et al., 2016). Daži no tiem var tikt salīdzināti ar (Schmidt et al., 2015), jo atvērtie medicīnas dati var (1) veidot pamatu veselības un zāļu pārvaldes slimnīcu statistikai vai veselības ekonomiskajiem aprēķiniem, (2) nodrošināt iestādes ar datiem, kas atbalsta slimnīcu plānošanu, (3) sniegt datus, lai atbalstītu iestādes, kas atbild par slimnīcu auditiem, (4) uzraudzīt dažādu slimību un ārstēšanas biežumu, (5) nodrošināt medicīnisko pētījumu paraugu ņemšanu, (6) veicināt veselības (aprūpes) pakalpojumu kvalitātes nodrošināšanu utt..

Atbilstoši (Andreassen et al., 2007) ASV veikto pētījumu rezultāti rāda, ka no 56% līdz 79% lietotājiem meklē ar medicīnu un veselību saistītu informāciju internetā, ieskaitot arī atvērtos datus. Latvijai šīs rādītājs ir 35%, Polijai – 42%, savukārt zemākais rādītājs ir

dienvidvalstīm, piemēram, Grieķijai – 23%. Augsts medicīnas datu un informācijas meklētāju rādītājs atkārtoti liecina par to, ka atvērtajiem datiem ir jābūt augstas kvalitātes, neskatoties uz to, ka dažas datu kopas apkopo skaitliskus rādītājus, kas, iespējams, parastajam lietotājam nesniedz svarīgu informāciju, dažas datu kopas mēdz būt galalietotājam informatīvākas, piemēram, sniedzot informāciju par medikamentiem, to lietošanas noteikumiem un devām.

Neskatoties uz medicīnas datu kvalitātes svarīgumu un nozīmīgumu, esošais [atvērto] datu kvalitātes stāvoklis nav apmierinošs. To konstatēja dažādu valsts pētnieki, piemēram, 2015. gadā datu kvalitātes problēmu esamība tika konstatēta Dāņu Nacionālajā Pacientu Reģistrā (Schmidt et al., 2015), neskatoties uz to, ka datu kvalitāte tika vērtētā, ņemot vērā tikai divas dimensijas – derīgums un pilnīgums. Ir jāatzīmē, ka, neskatoties uz to, ka analizētie dati bija "slēgtie", identificētās datu kvalitātes problēmas ir raksturīgas arī "atvērtajiem" datiem.

Vēl viens piemērs ir (Kerr et al., 2007a, 2007b) pētījums, kas pēta Jaunzēlandes medicīnas datu kvalitāti, un, neskatoties uz to, ka autori atzīst dažādu semantisko nozīmju piešķiršanu dimensijām ar vienādiem nosaukumiem un citas ar datu kvalitātes dimensiju izmantošanu saistītas problēmas, savā risinājumā autori datu kvalitāti sasaista ar 6 datu kvalitātes dimensijām, piešķirot tām 24 raksturīpašības un 69 kvalitātes kritērijus. Tik augsts kritēriju skaits samazina iespējamību, ka šīs risinājums tiks bieži izmantots, it īpaši ar lietotājiem bez padziļinātām IT un datu kvalitātes zināšanām. Taču, kamēr daži pētījumi izmanto pārāk lielu datu kvalitātes dimensiju skaitu, mēdz būt pētījumi, kuros autori izmanto tikai 2 kvalitātes dimensijas – precizitāte un pilnīgums. Tik zems izmantoto datu kvalitātes dimensiju skaits ir raksturīgi lielākoties ar medicīnas datiem saistītājiem pētījumiem ((Dahbi et al., 2018), (Weiskopf et al., 2013)).

Pētījumu virkne, kuros tiek pētīta medicīnas datu kvalitāte dažādās valstīs – Jaunzēlande (Kerr et al., 2007a, 2007b), (Raghupathi et al., 2014), Dānija (Schmidt et al., 2015), Brazīlija (Oliveira et al., 2016), ASV, Lielbritānija un Japāna (Yi, 2019), Šveice (Wanner et al., 2018), Zviedrija (Tomic et al., 2015), Kolumbija (Prieto Rodríguez, 2018), nonāk pie secinājuma, ka atvērtajos medicīnas datos ir izplatītas datu kvalitātes problēmas (ar vienu izņēmumu Norvēģijas gadījumā (Larsen et al., 2009)).

Ņemot vērā, ka (1) medicīnas datu kvalitāte ir ārkārtīgi svarīga, taču medicīnas datos ir bieži novērojamas datu kvalitātes problēmas, (2) esošie risinājumi nav piemēroti cilvēkiem bez padziļinātām IT zināšanām, taču atvērto medicīnas datu popularitāte strauji pieaug, līdz ar kuru pieaug arī nepieciešamība sniegt lietotājiem iespēju veikt to kvalitātes analīzi, (3) atvērto medicīnas datu saturs ir vienkāršāks salīdzinājuma ar "slēgtajiem" datiem, līdz ar ko arī nepieciešamas datu kvalitātes pārbaudes ir vienkāršākas, Latvijas atvērtie medicīnas dati ir piemēroti darbā piedāvātās pieejas aprobācijai uz tiem (5. nodaļa).

1.2.3. Atvērtie dati Latvijā

Atbilstoši (VARAM, 2016) "Latvijā publiskās pārvaldes datu pieejamība atvērtā veidā ir viens no e-pārvaldes politikas pamatprincipiem 2014. - 2020. gada attīstības plānošanas periodā". Par atvērto datu popularitātes pieaugumu liecina arī tas, ka, lai pastiprinātu to izmantošanu Eiropas Savienībā, 2019. gadā Latvijā piedalījās "*EU Datathon 2019*", kas pulcē vairāku valsts dalībniekus, ar mērķi apmainīties ar idejām par to kā tas varētu tikt panākts.

Latvijā ir pieejami vairāki atvērto datu portāli, taču populārākais no tiem ir *data.gov.lv*. Atbilstoši (VARAM, 2016) tas "tika izveidots Eiropas reģionālās attīstības fonda līdzfinansētā projekta Nr. 2.2.1.1/16/I/001 "Publiskās pārvaldes informācijas un komunikācijas tehnoloģiju arhitektūras pārvaldības sistēma" (PIKTAPS) ietvaros". Tas tika izveidots 2017. gadā ar Vides aizsardzības un reģionālās attīstības ministrijas atbalstu, un tā palaišanas brīdī tajā bija pieejamas 33 datu kopas no 13 dažādiem datu sniedzējiem, savukārt 2018. gada jūlijā pieejamo datu kopu skaits ir pieaudzis līdz 139 datu kopām no 41 datu sniedzēja (Nikiforova, 2018a), 2019. gada augustā to skaits ir sasniedzis 281 datu kopas no 64 datu publicētājiem, savukārt 2019. gada novembra beigās – 322 datu kopas no 68 datu publicētājiem. Pēdējā gadā pieejamo datu kopu skaits svārstās (2019. gada jūnijā – augustā svārstoties no 214 datu kopām līdz 267), t.i. periodiski samazinoties un pieaugot, kas var tikt skaidrots ar datu portāla pārvaldību, kuras laikā tiek pārbaudīta datu kopu atbilstība atvērto datu principiem, kas, iespējams, netiek veikts datu kopu publicēšanas brīdī. Portālā datu kopas ir pieejamas dažādos formātos, taču populārākie formāti ir.*csv* un *.xls* (*.csv* formāta popularitāte atbilst arī *European Data Portal* tendencēm), kopā veidojot 78.4% no visām pieejamām datu kopām. Atsevišķas datu kopas ir pieejamas formātos, kas netiek uzskatīti par mašīnlasāmiem, taču tās ir papildinātas ar mašīnlasāmo datu versiju (Nikiforova, 2019b). Atvērto datu kopu sniegšana nemašīnlasāmā formātā ir raksturīga vairāku valsts atvērtajām datu kopām, jo atbilstoši (Yi, 2019), pat tādās valstīs kā Lielbritānija un ASV vairāk kā 50% atvērto datu kopu nav pieejamas mašīnlasāmajā formātā. Par to liecina arī citi pētījumi (Zuiderwijk et al., 2012), (Sáez Martín et al., 2016) utt.

Ir jāatzīmē, ka pēdējo gadu laikā Latvija strauji uzlabo atvērto datu portālu stāvokli, par ko liecina arī *European Data Portal* pētījuma rezultāti (European Data Portal, 2018), kuros ir apkopots atvērto datu brieduma (angl. *maturity*) novērtējums dažādās Eiropas valstīs. Taču, neskatoties uz salīdzinoši augstiem atvērto datu portālu vērtējumiem, datu kvalitātes rādītājs Latvijas atvērtajiem datiem ir zemāks nekā citām valstīm. *European Data Portal* vērtē katras valsts atvērtos datus pēc 4 kritērijiem: ietekme, likumi (angl. *policy*), portāls un kvalitāte. Latvijas gadījumā kvalitātes rādītājs ir sliktākais (62%), salīdzinot ar vidējo kvalitātes rādītāju citām valstīm (71%). Taču neskatoties uz to, 4 kategoriju starpā: iesācēji, sekotāji, ceļrāži,

tendenču noteicēji (angl. *trend setters*), atbilstoši *European Data Portal*, Latvija tiek pieskaitīta pie ceļrāžiem, savukārt Lietuva, Igaunija – pie sekotājiem. Latvija ieņem augstāku pozīciju Baltijas un Skandināvu valsts starpā (Nikiforova, 2019a). Ir jāatzīmē, ka citu ar kvalitāti nesaistītu augsto rādītāju dēļ, 2018. gadā Latvija ieņēma 12. pozīciju ar vidēja brieduma (angl. *maturity*) līmeni 66.2%, kas ir par 1.2% labāks nekā vidējais *EU28* rezultāts (vislabākie rezultāti ir Īrijai – 87.8%, Spānijai – 87% un Francijai – 83%). Pie tam ir svarīgi atzīmēt, ka Latvija strauji attīstās, jo 2015. gadā tā ieņēma 27. pozīciju, 2016. – 28., 2017. – 18. un 2018. – 12. pozīciju, kas ļauj secināt, ka Latvija cenšas attīstīties un realizēt ieplānoto. Pie galvenajiem šķēršļiem pēc *European Data Portal* viedokļa ir pieskaitāmi (a) likumu trūkums, kas piespiestu organizācijas publicēt datus, kā arī (b) organizācijas esamība, kas sniedz datus par papildu maksu.

Papildus, kopš 2017. gada Latvijā ir viena no 70 *Open Government Partnership* starptautiskās platformas (Open Government Partnership, 2015), dalībniecēm, kuras mērķis ir veikt pārvaldes atvērtākas, labāk pārvaldāmas un atbildīgākas pret saviem iedzīvotājiem.

Par šīs problēmas aktualitāti tiek plaši runāts arī dažādos pasākumos, tajā skaitā Digitālās nedēļās ietvaros, IKT profesionāļu dienās un ar Ekonomikas ministriju rīkotājā *"Digitālizācijas un Inovāciju forumā DIG-IN"*, iekļaujot datu kvalitātes pētnieku profesiju nākotnes profesiju sarakstā.

1.3. Datu kvalitātes problēma un tās aktualitāte

Datu kvalitātēs problēmu uzsāka pētīt vēl 60-o gadu beigās. Sākumā tika pētīti tikai atsevišķi šīs problēmas aspekti, kas izraisīja interesi statistikas pētniekiem, lielākoties pētot dublikātvērtību problēmu, tās risināšanai piedāvājot matemātiskus teorētiskus risinājumus. 80-o gadu sākumā tika uzsākti arī datu kvalitātes problēmas pārvaldības (angl. *management*) pētījumi, galvenokārt fokusējoties uz datu kvalitātes problēmu noteikšanu un novēršanu ražošanas sistēmu kontroles risinājumos. Savukārt 90-o gadu sākumā datu kvalitātes problēmu ir sākuši pētīt arī datorzinātnieki, fokusējoties uz datu, kas glabājās datubāzēs, datu noliktavās un mantotās sistēmās, kvalitātes jēdziena definēšanu, kvalitātes mērīšanu un uzlabošanu (Scannapieco et al., 2002b), kā arī "datu kvalitātes" jēdzienu sasaistīšanu ar "datu kvalitātes dimensiju" jēdzienu, piedāvājot dažādus dimensiju grupējumus (Cai et al., 2015). Taču neskatoties uz datu popularitāti un to apjoma nepārtrauktu pieaugumu ((Hashem et al., 2015), (Kitchin, 2014), (Cai et al, 2015)), gandrīz 30 gadus vēlāk datu kvalitātes problēma vēl joprojām nav atrisināta un vēl joprojām ir aktuāla (Cai et al., 2015).

Datu kvalitātes problēmu iemesli un izraisītājfaktori mēdz būt dažādi ((Cannon, 2016), (Lehmannm et al., 2016)), taču tipiskākie ir:

1) **datubāzē uzkrāto datu pilnība, precizitāte un nepretrunība**, ko saista ar datu pakāpenisko uzkrāšanu datubāzēs, t.i. ievadot to sistēmā dažādos darījumprocesa soļos, dažādos laika posmos un no dažādiem datu avotiem. Šī problēma ir visbiežāk sastopama IS un e-pārvalžu (angl. *e-government*) gadījumos, jo tām visbiežāk ir raksturīga atkārtota datu uzkrāšana no fiziskām un juridiskām personām savām vajadzībām, neskatoties uz to, ka šie dati jau tika uzkrāti iepriekš, taču ar citu datu turētāju. Vairāki datu turētāji savā starpā nesadarbojas un nedalās ar datiem, kam arī mēdz būt dažādi skaidrojumi: (a) nav zināma cita datu sniedzēja/ administratora datu kvalitāte, (b) dažreiz datu izplātīšana ir ierobežota ar likumiem vai datu turētāja noteikumiem, (c) datu izplatīšanas procedūras sarežģītība, tajā skaitā datu formātu dažādība un nesavietojamība (atbilst Scannapieco et al., 2002a, 2002b). Taču e-pārvalžu gadījumā datu kvalitāte ir ārkārtīgi svarīga, jo to mērķi atbilstoši ((Batini et al., 2006), (Ubaldi, 2013)) ir attiecību uzlabošana starp (a) valdību un valsts iedzīvotājiem, (b) aģentūrām un uzņēmumiem, kas tiek panākts ar informācijas un komunikācijas tehnoloģiju palīdzību. Šīs problēmas novēršanai atsevišķos gadījumos ir paredzētas kontroles datu lauku ierobežojumu izpildei, vēlāk ievadītos datus sasaistot ar citiem datiem, ņemot vērā to kontekstu jeb semantiku. Ņemot vērā, ka mainoties kontekstam, var mainīties datu aktualitāte un to kvalitāte, augstas datu kvalitātes nodrošināšanai ir atkārtoti jāpārbauda datu atbilstība kvalitātes prasībām. Taču, lai gan šī problēma ir aktuāla jau daudzus gadus, datu kvalitātes pārbaudes tiek īstenotas katrā konkrētā IS individuāli, ja vispār tiek īstenotas, un universāls risinājums pagaidām nav atrasts;
2) **datu migrācija**, kas rodas darījumprasību izmaiņu dēļ, kuru rezultātā ir jāmodificē ilgi darbojošās sistēmas. Ņemot vērā, ka lietotājam nedrīkst prasīt ievadīt visus iepriekšējā versijā uzkrātos datus atkārtoti, iepriekšējā sistēmas versijā uzkrātie dati, parasti tiek migrēti uz jauno datubāzi. Tā kā dati tiek uzkrāti ar dažādām programmu versijām, un to struktūra var atšķirties no iepriekšējām versijām, IS izstrādātāji sniedz atbalstu migrējamo datu kvalitātes nodrošināšanai, tajā skaitā precīzai datu atribūtu aprakstīšanai un datu migrācijas veikšanai. Tas nozīme, ka piedāvātajam risinājumam jābūt elastīgam, lai tas ļautu katru reizi definēt jaunus datu saderības nosacījumus un to pārbaudes;

3) **datu uzkrāšana datu noliktavās** sastopas ar jau minētām problēmām, t.i. dati tiek izgūti no dažādām datu struktūrām, par dažādiem laika posmiem, kā arī izmantojot dažādu programmu versiju uzkrātus un dažādi strukturētus datus. Lietotājiem ir nepieciešami datu sakarību aprakstīšanas līdzekļi, kas ļautu pārbaudīt datu noliktavā ievadāmo datu kvalitāti. Tradicionāli datu noliktavā ievadāmo datu kvalitātes pārbaudi veic ar *ETL* (*Extract, Transform, Load*) procedūru palīdzību, kuras ir datu noliktavu sistēmu komponents, taču IS darbībā šīs risinājums netiek lietots.

Par datu kvalitātes problēmas "vecumu" liecina viens no agrīnajiem piemēriem - 1992. gada Arnolda pētījums (Arnold, 1992), kurā 500 uzņēmumu analīzes rezultātā tika konstatēts, ka vairāk kā 60% vidēja izmēra uzņēmumiem ar gada ienākumu virs 20 miljoniem ASV dolāru ir raksturīgas datu kvalitātes problēmas.

Mūsdienās katru gadu tiek veikti dažādi pētījumi, kas ietver sevi aprēķinus un aptaujas, kuru mērķi ir noteikt datu kvalitātes ietekmi, t.sk. zudumus, ko izraisa zemas kvalitātes dati. Gūtie rezultāti liek aizdomāties par šīs tēmas aktualitāti un aicina meklēt risinājumus esošās situācijas uzlabošanai. Dažu aptauju un analīžu rezultāti:

- 2018. gadā zemas kvalitātes dati tika atzīti par galveno jauno inovatīvo tehnoloģiju neveiksmes cēloni/ iemeslu, kas izraisa 9.7 miljardu ASV dolāru lielus zaudējumus ASV ekonomikai gadā (Lebied, 2018).
- 2017. gada *Gartner* pētījuma (Moore, 2017) rezultāti liecina, ka katru gadu datu kvalitātes problēmu dēļ uzņēmumi zaudē aptuveni 15 miljonus ASV dolāru. Šī tendence pēdējo gados ir nemainīga (piemēram, (Friedman et al., 2013), (Kelly, 2009), (Moore, 2017));
- *IBM* pētījuma (Singh, 2017) rezultāti liecina, ka darījumlēmumi, kas tiek pieņemti, balstoties uz nekvalitatīviem datiem, maksā ASV ekonomikai 3.1 miljardus ASV dolāru gadā;
- nekorektu adrešu datu dēļ ASV pasta pakalpojumu sniedzēja *USPS* zaudējumi ir 3.4 miljardus ASV dolāru gadā (Karel, 2015);
- atbilstoši *Gartner* pētījumiem, aptuveni 40% no uzņēmumos esošiem datiem ir nekvalitatīvi, savukārt datu kvalitāte ir cieši saistīta ar procesu kvalitāti un kā rezultāts – ar uzņēmējdarbības sekmību (Friedman et al., 2011), un, piemēram, atbilstoši *PwC* 2014. gada pētījumam aptuveni 44% uzņēmumos reizi mēnesī un 35% - reizi četros mēnešos, balstoties uz esošiem datiem, tiek pieņemti nopietni darījumlēmumi (Witchalls, 2014).

Kamēr zemas datu kvalitātes ietekmi visbiežāk izsaka finansiāli, tai mēdz būt arī nopietnākas globāla mēroga sekas. Atbilstoši (Fisher et al., 2001) kosmiskās "atspoles" *Challenger* sprādziena iemesls bija nekvalitatīvi dati, galvenokārt datu pretrunība, nepilnība un neprecizitāte.

Neskatoties uz dažu novērtējumu skaitlisko dažādību, tie visi dod pamatu uzskatīt, ka datu kvalitātes problēma ir svarīga un aktuāla. Tāpat ir jāņem vērā, ka dati tiek bieži izmantoti darījumlēmumu pieņemšanai, taču tie būs pareizi un efektīvi tikai tad, ja izmantotie dati būs kvalitatīvi (atbilst (Ubaldi, 2013)). Dati, uz kuriem nevar paļauties, ir iemesls kļūdainiem lēmumiem un samazinātai produktivitātei.

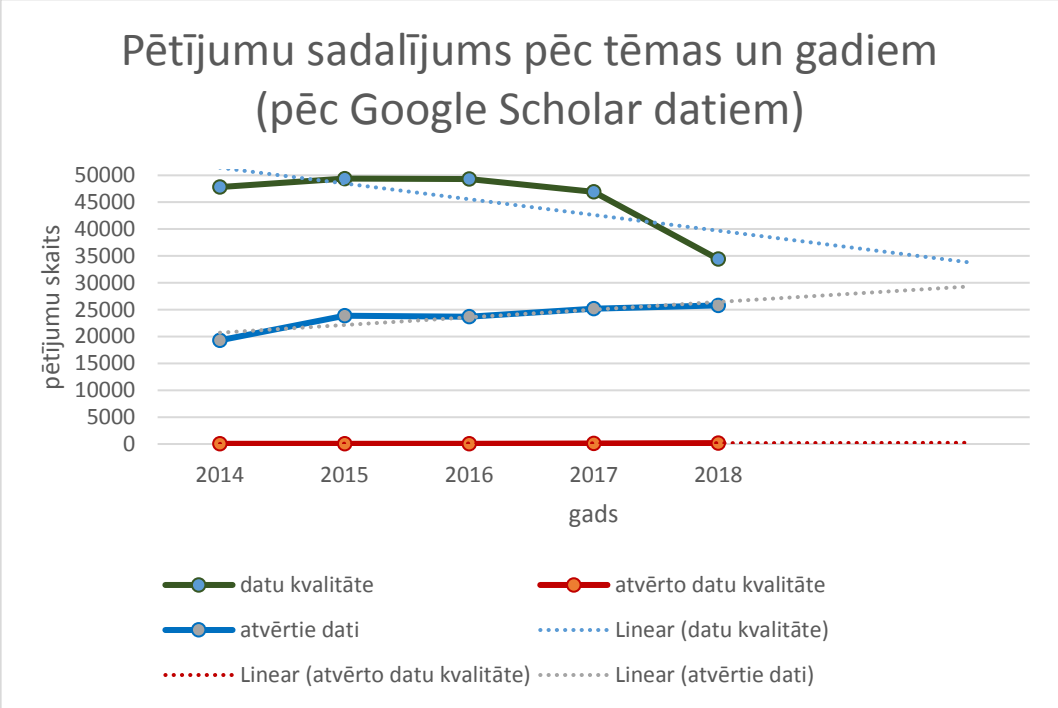
Atbilstoši Loshin (2001), zemas kvalitātes dati pazemina darba efektivitāti, līdz ar ko, strādājot ar datiem, ir jābūt pārliecībai, ka tie ir korekti, pirms tiek tiks papildināti vai apstrādāti – tiem ir jābūt korektiem katrā datu uzkrāšanās posmā. Ja datos ir kļūda, tai ir jābūt izlabotai vai arī ierakstam ir jābūt izdzēstam, pirms tas tiks turpmāk izmantots. Jo vairāk kļūdu uzkrājās, jo vairāk resursu ir nepieciešams, lai izlabotu tās, taču, neizlabojot, to apstrādes rezultāti kļūst nekorekti. Atbilstoši (Ross, 2017) datu kvalitātes problēmai ir pielietojams arī *TDQM* "1-10-100" likums, atbilstoši kuram 1 ASV dolārs, kas tiek tērēts problēmas novēršanai/ nepieļaušanai, ļauj ietaupīt 10 ASV dolārus, kas tiktu tērēti problēmas izvērtēšanai un 100 ASV dolārus, kas būtu jātērē problēmas iestāšanās gadījumā (Nikiforova, 2019b). Zemas kvalitātes dati ietekmē darījumlēmumu pieņemšanu, savukārt, augstas kvalitātes dati uzlabo arī datu noliktavu lietderību, jo parasti datu izguve, tīrīšana un ielāde aizņem 80% laika. Tas atbilst arī (Gabernet et al., 2017), atbilstoši kuram plaši lietotais "80-20 likums" ir piemērojams arī datu kvalitātei, jo tiek uzskatīts, ka 80% no datu pētnieka laika prasa kvalitātes nepilnību meklēšana, datu tīrīšana un organizēšana, 20% atstājot to izmantošanai, tajā skaitā analīzes veikšanai (Nikiforova, 2018a, 2019b).

Atbilstoši ((Jetzek, 2017), (Chen et al., 2016), (Colpaert et al., 2013)), [atvērto] datu kvalitāte ietekmē zināšanu kvalitāti, to ticamību un nozīmīgumu, kas ir iegūstamas, apstrādājot datus. Datu kvalitātes pārbaude kopā ar citām aktivitātēm, kas tiek veiktas datus gatavojot apstrādei un analīzei, tiek uzskatīta par vienu no vislaikietilpīgākajām un sarežģītākajām aktivitātēm (Pyle, 1999).

Visi šie dati liecina par to, ka datu kvalitātes problēma ir aktuāla un bieži pētāma, taču datu kvalitātes rādītāji ir neapmierinoši un diemžēl nemainīgi, jo ar esošiem risinājumiem un pētījumiem nepietiek. Datu kvalitātes problēmas esamību apliecina arī vairāku pētījumu rezultāti ((Nikiforova, 2018, 2018b, 2019a), (Nikiforova et al., 2019), (Yi, 2019), (Bicevskis et al., 2018, 2018b), (Ferney et al., 2017), (Färber et al., 2016), (Vetrò et al., 2016), (Martin, 2014),

(Kontokostas et al., 2014), (Acosta et al., 2013), (Kuk et al., 2011), (Kerr et al., 2007a, 2007b), (Guha-Sapir et al., 2002), utt.).

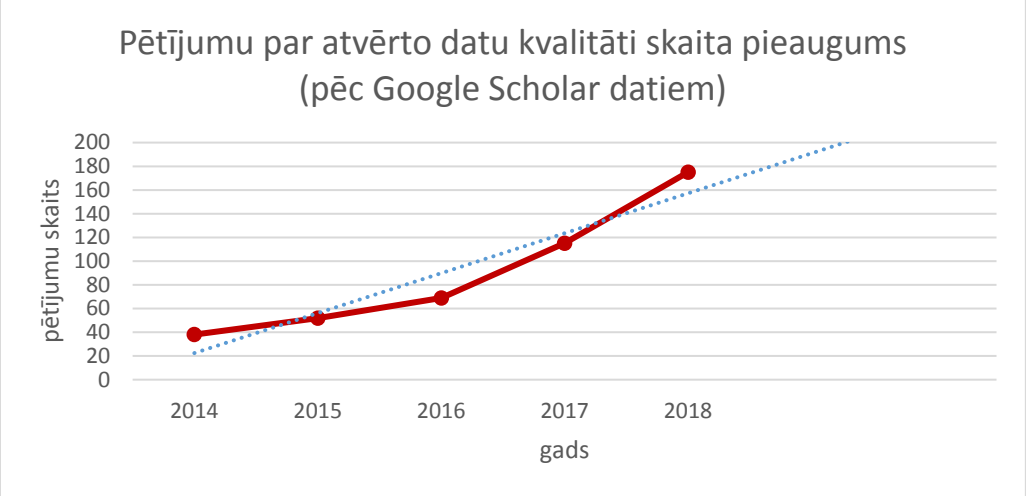
Par [atvērto] datu kvalitātes problēmas aktualitāti un tās popularitātes pieaugumu liecina arī veikto pētījumu skaits. Izpētot *Google Scholar* datus, darba autore konstatēja, ka attiecībā uz atvērtajiem datiem atbilstoši *Google Scholar* laika periodā no 2003. līdz 2014. gadam tika publicēti 4.6 reizes mazāk pētījumu par atvērto datu kvalitāti nekā 2018. gadā. Pēc 1.3.1. att. ir redzams, ka kopš 2017. gada tika novērots atvērto datu kvalitātes popularitātes straujš pieaugums, kad pasaulē ir sācis pieaugt atvērto datu kopu skaits un atvērto datu portālu skaits. Tādējādi, tiek apstiprināta darba sākumā izvirzītā 1. tēze, t.i. datu kvalitātes problēma ir aktuāla, neskatoties uz savu vecumu.



1.3.1. att. **Datu kvalitātes pētījumu popularitāte (2014-2018) [izveidoja autore]**

Taču, attiecinot ar atvērto datu kvalitāti saistīto pētījumu skaitu pret kopējo ar atvērtajiem datiem saistīto pētījumu skaitu, ir redzams, ka datu kvalitātes jautājums tiek pētīts nepamatoti reti, jo 2018. gadā atvērto datu pētījumu skaits pārsniedz ar atvērto datu kvalitāti saistīto pētījumu skaitu 147 reizēs (2019. gadā – 179 reizēs), t.i. atvērto datu kvalitātes pētījumu īpatsvars pret kopējo ar atvērtajiem datiem saistīto pētījumu skaitu nepārsniedz 0.5%. Tas atbilst darba sākumā izvirzītajai 3. tēzei, t.i. neskatoties uz atvērto datu popularitāti un to strauju pieaugumu, atvērto datu kvalitātes problēma tiek reti pētīta.

Pie tam datu kvalitātes pētījumu skaits pārsniedz atvērto datu kvalitātes pētījumu skaitu gandrīz 196 reizēs (t.i. atvērto datu kvalitātes pētījumu īpatsvars pret kopējo ar datu kvalitāti saistīto pētījumu skaitu, ir ~0.2%). Tas liecina par nepieciešamību veikt pētījumus, kas būtu saistīti ar atvērto datu kvalitāti, jo datu kvalitātes pētījumu rezultāti liecina par problēmas esamību, taču veikto pētījumu sadalījums liecina par to, ka mūsdienās esošie risinājumi ir lielākoties domāti tā saucamo "slēgto" datu kvalitātes novērtēšanai un nav piemēroti atvērtajiem datiem un/ vai lietotājiem bez padziļinātājām IT un datu kvalitātes zināšanām. 1.3.2. att. ir attēlota atvērto datu kvalitātes jautājuma popularitāte zinātniskajā literatūrā, kas ļauj saprast, ka, neskatoties uz to, ka šī tēma tiek pētīta retāk nekā citas iepriekšminētās, atvērto datu kvalitātes jautājuma popularitāte tomēr arī pieaug un pat straujāk nekā atvērto datu popularitāte.



1.3.2. att. **Atvērto datu kvalitātes pētījumu popularitāte (2014-2018) [izveidoja autore]**

Pēc 1.3.1. un 1.3.2. att. var redzēt, ka [slēgto] datu kvalitātes tēma zaudē savu popularitāti, taču atvērto datu un atvērto datu kvalitātes pētījumu popularitāte nepārtraukti pieaug. Par to liecina arī 1.3.2. att. tendences līkne, atbilstoši kurai atvērto datu kvalitātes popularitāte turpinās pieaugt arī tuvāko gadu laikā (analīze tika veikta 2019. gada rudenī). Pieaugot atvērto datu apjomam, kļūst nepieciešami risinājumi, kas būtu piemēroti lietotājiem bez padziļinātājām zināšanām datu kvalitātes un IT jomā, jo atvērtie dati kļūst par ikdienas parādību un atbilstoši to kvalitātes analīze kļūst par neatņemamu ikdienas darbību.

Šīs nodaļas rezultāti ir publicēti (Nikiforova, 2019a, 2018b).

Nākamajā nodaļā ir apskatīti eksistējošie risinājumi, nosakot to trūkumus un priekšrocības, kas būtu ņemami vērā, izstrādājot jaunu pieeju datu kvalitātes novērtēšanai risināšanai.

2. DATU KVALITĀTES NOVĒRTĒŠANAS RISINĀJUMI

Nodaļā "Datu kvalitātes novērtēšanas risinājumi" ir apskatīti ar datu kvalitātes problēmu saistītie pētījumi, eksistējošie datu kvalitātes analīzes risinājumi, identificējot to trūkumus un priekšrocības, kas būtu ņemami vērā, izstrādājot alternatīvo risinājumu.

Šī nodaļa balstās uz (Nikiforova, 2019a, 2019b, 2018a, 2018b), (Nikiforova et al., 2019) un (Bicevskis et al., 2018a).

Atbilstoši (Nikiforova, 2018a) eksistējošie pētījumi un piedāvātie datu kvalitātes problēmas risinājumi var tikt iedalīti vairākās grupās:

- uz dimensiju definēšanu, to grupēšanu un datu kvalitātes novērtēšanu, pielietojot definētas dimensijas datu kopām, vērstie pētījumi ((Wang et al., 1996), (Van den Berghe et al., 2017), (Ferney et al., 2017), (Redman, 2001) utt.);
- atvērto datu portālu un/ vai atvērto pārvaldes datu (angl. *Open Government Data (OGD)*) kvalitātes vērtēšana ((Vetro et al., 2016), (Kučera et al., 2013), (Neumaier et al., 2016), (Sáez Martín, 2016), (Sasse et al., 2017) utt.);
- saistīto datu kvalitātes vērtēšana ((Acosta et al., 2013), (Paulheim et al., 2014) utt.).

Daži pētījumi veic industrijai specifisko datu un informācijas kvalitātes analīzi, visbiežāk pielietojot datu kopām konkrētam sektoram specifiskās vai maksimāli pielāgotās tam metodes:

- vēža reģistri ((Bray et al., 2009), (Sigurdardottir et al., 2012), (Larsen et al., 2009), (Parkin et al., 2009), (Tomic et al., 2015) utt.);
- veselības aprūpe ((Dahbi et al., 2018), (Weiskopf et al., 2013), (Van den Berghe et al., 2017), (Schmidt, 2015), (Kerr, 2007a, 2007b) utt.);
- ķīmisko ieroču un risku novērtēšana (Bevan et al., 2012) utt..

Ņemot vērā šo pētījumu pielietošanas jomas ierobežotību, tikai dažus no tiem autore apskatīja darbā (dažu pētījumu analīze ir pieejama (Bicevskis et al., 2018a)).

Atsevišķos gadījumos risinājumiem tiek piedāvāti arī datu kvalitātes novērtēšanas satvari ((Vetrò et al, 2016), (Kučera et al., 2013), (Neumaier, 2015), (Umbrich et al., 2015).

Ir jāatzīmē, ka atsevišķie pētījumi var vienlaicīgi piederēt vairākām grupām, jo šīs grupas ir savstarpēji saistītas, kā arī vairākas grupas varētu tikt iedalītas apakšgrupās, piemēram, atsevišķie pētījumi izstrādā arī vadlīnijas datu kvalitātei (piemēram, (Aarshi et al., 2018), (Kučera et al., 2013), (Vetro et al., 2016), (Sasse et al., 2017) utt.), taču tās parasti izriet no pētījuma rezultātiem.

Nozīmīgākie risinājumi ir apskatīti turpmākajās apakšnodaļās.

2.1. Datu kvalitātes novērtēšana ar dimensiju palīdzību

Datu kvalitāti tradicionāli saista ar datu kvalitātes dimensiju jēdzienu. Parasti datu lietotāji visbiežāk saista to ar vienu konkrētu datu kvalitātes dimensiju - datu precizitāti (angl. *accuracy*), taču datu kvalitātes jēdziens ir daudz plašāks un iekļauj vairākas datu kvalitātes dimensijas.

90-o gadu sākumā *MIT* izstrādātā *Total Data Qualtiy Management* (*TDQM*), kas ir viena no visplašāk zināmām datu kvalitātes "programmām" jeb metodoloģijām, piedāvāja datu kvalitātes dzīvesciklu, kas sastāv no trim savstarpēji saistītām fāzēm (Wang et al., 1993a): (1) datu kvalitātes mērīšanas un definēšanas fāze; (2) datu kvalitātes analīzes fāze; (3) datu kvalitātes uzlabošanas fāze. Citu metodoloģiju pārskats ir pieejams (Batini et al., 2009), (Zaveri et al., 2016). Pilnīgas datu kvalitātes vadības process ir cikls, jo, lai panāktu un saglabātu augstu datu kvalitāti, visas ciklā iekļautās fāzes ir sistemātiski jāatkārto (atbilst arī ISO/IEC 25024:2015). Tas ir nepieciešams, lai (a) pārbaudītu kā notiek datu kvalitātes uzlabošanas mehānisma realizācija, kas parasti ir sasniedzams, atkārtojot mērīšanas un analīzes fāzes, (b) nodrošinātu jaunu vai modificētu datu kvalitātes pārbaudi, jo dati datu krātuvēs pastāvīgi mainās, kas savukārt var izraisīt jaunas datu kvalitātes problēmas vai rast nepieciešamību jaunu datu kvalitātes prasību definēšanai utt. (Nikiforova, 2018a).

Atbilstoši ((Wang et al., 1993a), (Linstedt et al., 2015)) datu kvalitātes definēšanas un mērīšanas fāze var tikt iedalīta divās atsevišķās apakšfāzēs: (1) datu kvalitātes definēšanas fāze, kuras ietvaros tiek (a) formulēti datu kvalitātes raksturojumi jeb datu kvalitātes dimensijas, piemēram, pareizība, pilnība, nepretrunīgums, unikalitāte, ticamība, un (b) uzdotas datu kvalitātes metrikas, katram raksturojumam izvirzot konkrētas datu kvalitātes prasības, kurām ir jāatbilst datiem; (2) datu kvalitātes mērīšanas fāze, kuras ietvaros tiek (a) formulēti izvēlēto datu kvalitātes kritēriju mēri, (b) pārbaudīts, vai reālie dati atbilst definēšanas fāzē formulētajiem nosacījumiem, konstatējot vērtības, kas pārkāpj definētās kvalitātes prasības. Datu kvalitātes analīzes fāzē tiek veikta mērīšanas fāzē saņemto datu kvalitātes pārbaudes rezultātu analīze. Tās mērķis ir datu kvalitātes problēmu konstatēšana, cenšoties noskaidrot identificēto problēmu iemeslus un izstrādājot priekšlikumus datu kvalitātes uzlabošanai. Datu kvalitātes uzlabošanas fāzes ietvaros notiek datu kvalitātes uzlabošanas mehānisma izvēle un tā realizācija.

Tā kā datu kvalitātei atkarībā no lietošanas piemēra mēdz tikt definētas dažādas kvalitātes prasības, arī lietotāju interesējošas datu kvalitātes dimensijas mēdz būt dažādas. Rezultātā, ņemot vērā, ka tradicionāli datu kvalitāti saista ar "datu kvalitātes dimensijas" jēdzienu, vairākums pētījumu fokusējas uz datu kvalitātes dimensiju definēšanu, to grupēšanu un

pielietošanu datu kopām (piemēram, (Bovee et al., 2001), (English, 2009), (Färber et al., 2018), (Ferney et al., 2017), (Jarke, 1999), (Kučera et al., 2013), (Loshin, 2001), (Naumann, 2003), (Redman, 2001), (Vetro et al., 2016), (Wang et al., 1996)).

Atbilstoši *TDQM* literatūrā visbiežāk sastopamas datu kvalitātes dimensijas ir precizitāte, uzticamība (angl. *reliability*), pilnīgums, nepretrunīgums, atbilstība (angl. *relevance*) un citas (Wang et al., 1993b). Datu kvalitātes dimensiju skaits un to klasifikācijas mēdz būt ļoti daudzas un dažādas. Viena no visplašākzināmām dimensiju klasifikācijām, kas bieži tiek uzskatīta par vienu no "tradicionālajam" klasifikācijām ir (Wang et al., 1996) pētījuma rezultāts, atbilstoši kuram datu kvalitāti raksturo 15 datu kvalitātes dimensijas, kas iedalās 4 grupās - iekšējā, kontekstuālā, attēlošanas un pieejamības datu kvalitāte (Nikiforova, 2019b). Pie šāda iedalījuma autori nonāca aptaujas rezultātā, apkopojot lietotāju viedokļus/ uzskatus. *Batini* un *Scannapieco* klasificē datu kvalitātēs dimensiju noteikšanai pielietoto pieeju kā empīrisko (Batini et al., 2016).

Intuitīvas pieejas pielietošanas rezultātā *Redman* (Redman, 2001) ir iedalījis datu kvalitātes dimensijas 9 kategorijās: pieejamība, satura kvalitāte, vērtību kvalitāte, pasniegšanas (prezentācijas) kvalitāte, elastīgums, uzlabošana, privātums, uzticamība, arhitektūra. Katrai dimensijai *Redman* ir sagatavojis kritēriju sarakstu, kas kopā veido 51 kritērijus jeb datu kvalitātes dimensijas, kuru starpā ir precizitāte, pilnīgums, redundance (angl. *redundancy*), lasāmība (angl. *readability*), piekļūstamība, nepretrunīgums, noderīgums (angl. *usefulness*), uzticamība (angl. *trust*) un citas. Atbilstoši (Nikiforova, 2018b) tik liels kritēriju skaits samazina iespējamību, ka šīs risinājums tiks izmantots reālajā pasaulē, it īpaši ar lietotājiem bez padziļinātājām zināšanām datu kvalitātes apakšjomā.

Ņemot vērā, ka dimensiju skaitam risinājuma ietvaros nedrīkst būt pārāk augstam, kā arī to, ka speciālistu vienošanās par dimensiju nozīmi nav panāktas, līdz ar ko tas būtu jāizvēlas ļoti piesardzīgi, 2013. gadā *DAMA UK Working Group* (Ashkham et al., 2013) piedāvāja lietot mazāk detalizētu dimensiju klāstu, kas sastāv no 6 dimensijām:

1) pilnīgums – attiecībā starp pieejamiem un visiem potenciāli iespējamiem datiem. Piemēram, obligāto lauku aizpildījums konkrētā datu objektā pret visiem obligāti aizpildāmiem laukiem;
2) unikalitāte (angl. *uniqueness*) – reālās pasaules objekta identificēšana pēc konkrēta parametra. Piemēram, persona tiek identificēta ar tās personas kodu, nevis ar citiem parametriem – vārdu, uzvārdu, dzimšanas laiku un vietu;
3) savlaicīgums - pakāpe, kurā dati reprezentē realitāti laika dimensijā, raksturojot laika intervālu no notikuma stāšanas reālā pasaulē līdz tā reģistrēšanai datubāzē;

4) derīgums (angl. *validity*) – datu atbilstība definētām sintakses prasībām, piemēram, formāts, vērtību tips, minimālās vai maksimālās vērtības ierobežojumi. Piemēram, datuma tipa vērtībā parādās 31. februāris utt.;
5) precizitāte – pakāpe, kurā dati apraksta reālās pasaules objektus un notikumus. Piemēram, objekta izmēru uzdošanas precizitāte;
6) nepretrunīgums – atšķirības starp dažādām viena reālās pasaules objekta reprezentācijām. Piemēram, divās datubāzēs vienam un tam pašam objektam noteikta parametra vērtības ir dažādas (Ashkham et al., 2013), (Nikiforova, 2019b).

No iepriekšējiem piemēriem ir redzams, ka atsevišķas datu kvalitātes dimensijas vairākām pieejām ir kopīgas, taču ir arī tādas, kuras piemīt tikai atsevišķām pieejām. Ņemot vērā, cik daudz dažādu pārbaužu var būt nepieciešamas datu kvalitātes analīzei, vairāki autori definē savas datu kvalitātes dimensijas. Datu kvalitātes dimensiju izvēle un definēšana mēdz notikt dažādos veidos: (a) izstrādātāju/ studentu/ lietotāju aptaujas rezultātā (piemēram, (Vetro et al., 2016)); (b) literatūras un esošo risinājumu izpētes darba rezultātā; (c) ar pašiem autoriem atbilstoši viņu skatupunktam/ "vīzijai". Rezultātā definētās datu kvalitātes dimensijas visbiežāk kļūst aktuālās tikai konkrētam risinājumam. Autoru vēlmi un nepieciešamību definēt jaunās savam risinājumam domātas datu kvalitātes dimensijas mēdz saistīt ar to, ka šobrīd nav zināmas universālas datu kvalitātes dimensijas, savukārt, atbilstoši *Scannapieco* un *Catarci*, pat ja universāli svarīgas dimensijas ir noteiktas, nav noteiktas to nozīmes, jo dažādos pētījumos viens un tas pats nosaukums tiek bieži izmantots dimensijām ar semantiski dažādām nozīmēm un otrādi, t.i., dimensijām ar vienādu nozīmi dažādi autori mēdz piešķirt dažādus nosaukumus. Šī problēma tiek bieži diskutēta, tajā skaitā ar tādiem plaši pazīstamiem datu kvalitātes pētniekiem kā *Scannapieco, Batini, Price* (Price et al., 2004, 2005), *Jayawardene* (2013, 2015) utt., secinot, ka dimensiju skaidras un viennozīmīgas nozīmes, tāpat kā veids kā tām ir jābūt izmērāmām, nav zināmi - nav zināma universālā klasifikācija, kurai būtu jāseko datu kvalitātes risinājumu izstrādātājiem un to lietotājiem. Dažas datu kvalitātes dimensijas, to definīcijas un grupēšanas ir plaši zināmas un bieži lietotas, taču vairākums eksistējošo risinājumu fokusējas uz jauno datu kvalitātes dimensiju definēšanu un grupēšanu. Rezultātā vairākas dažādas datu kvalitātes dimensijas un to grupējumi tiek izmantoti tikai konkrēta risinājuma ietvaros un nav atkalizmantojami.

Piemēram, savā darbā (Batini et al., 2016) *Batini* ar līdzautoriem veic vienas konkrētas datu kvalitātes dimensijas - savlaicīguma (angl. *timeliness*) analīzi 7 autoru darbos. Autori konstatēja, ka attiecībā uz šo dimensiju 7 darbos ir sastopami 3 dažādi nosaukumi - izplatība (angl. *currency*), svārstīgums (angl. *volatility*) un savlaicīgums (angl. *timeliness*). Vēl

interesantāk ir tas, ka vienādu nosaukumu izmantošana negarantē vienādu dimensiju būtību un otrādi – dažādu nosaukumu izmantošana neliecina par dimensiju būtības dažādību, jo *Wand* (1996) savlaicīguma definīcija atbilst *Redmana* izplatības dimensijas definīcijai, savukārt *Wang* un *Liu* sauc to par savlaicīgumu. Analizējot datu kvalitātes dimensiju "pilnīgums", autori ir nonākuši pie secinājuma, ka, neskatoties uz to, ka šo jēdzienu vairāki autori definē vienādi vai ļoti līdzīgi, dažādos risinājumos šī dimensijai ir raksturīgs dažāds granularitātes līmenis un tā tiek saistīta ar dažādiem datu modeļa elementiem, piemēram, informācijas sistēma, datu noliktava vai entitāte. Savukārt, neskatoties uz to, ka "pilnīgums" ir viena no vienkāršākajām un viennozīmīgi definējamām datu kvalitātes dimensijām, (Chen et al., 2014) pētījuma rezultātā, izpētot 24 dažādus pētījumus, autori ir konstatējuši, ka arī tai dažādos risinājumos mēdz tikt piešķirtas dažādas nozīmes, un pielietoti dažādi mērīšanas mehānismi. Kopumā autori ir izanalizējuši 49 dažādus atribūtus jeb dimensijas, iedalot tās divās grupas, kuras atbilstoši raksturo (a) zemas kvalitātes datus – 11 atribūti un (b) augstas kvalitātes datus – 38 atribūti. Šī iedalījuma rezultātā, autori ir konstatējuši, ka zemas kvalitātes datu atribūtam "trūkstošie dati" (angl. *missing values*) vienā risinājumā atbilst citos risinājumos sastopamais augstas kvalitātes atribūts "pilnīgums", kuram kā jau tika minēts iepriekš dažādos risinājumos arī mēdz piešķirt dažādas nozīmes. Abu pētījumu autori secina, ka dažādi autori definē vienādas pēc nosaukuma dimensijas ļoti dažādi, un otrādi – vienādām pēc nozīmes dimensijām piešķir dažādus nosaukumus.

Līdzīga pētījuma (Scannapieco et al., 2002a) autori veic 6 pētījumu datu kvalitātes dimensiju un to nozīmju salīdzinājumu, secinot, ka no 23 dažādām datu kvalitātes dimensijām tikai viena dimensija – "precizitāte", visos pētījumos tika definēta vienādi. 14 dimensijām ir vienādi dimensiju nosaukumi un to definīcijas no 2 līdz 4 pētījumos, savukārt dažām dimensijām ir raksturīgas dažādas definīcijas, pat gadījumos, ja 4 no 6 pētījumos tiem ir vienādas definīcijas. Tas nozīmē, ka ne tikai datu kvalitātes jēdziens ir sarežģīts jēdziens, kuru mēdz definēt dažādi (Scannapieco, et al., 2005), bet arī datu kvalitātes dimensijas jēdziens pats par sevi un katra individuālā dimensija ir sarežģīti.

Loshin (2001) sava pētījuma rezultātā iedala datu kvalitātes dimensijas 5 grupās: datu modeļa, datu vērtību, datu domēnu/ apgabalu, datu attēlojuma/ reprezentācijas un informācijas noteikumu (angl. *policy*) kvalitātes dimensijas. Ir jāatzīmē, ka dažu dimensiju nosaukumi ir sastopami vairākās grupās, taču atkarībā no grupas, kurai tā pieder, mēdz mainīties dimensijas definīcija. Datu modeļa kvalitāti raksturo 15 dimensijas, datu vērtību – 5, datu domēnu – 3, datu reprezentāciju – 8 un informācijas noteikumus – 6 dimensijas. Arī viņš nonāca pie secinājuma, ka dažādām dimensijām mēdz būt dažādi nozīmības līmeņi atkarībā no datu lietotāja, atzīmējot, ka dimensiju sarakstu var turpināt/ paplašināt un, visticamāk, tas nekad nebūs pilnīgs. Ar to ir

domāts, ka atkarībā no lietošanas piemēra bieži vien būs iespējams izvirzīt kvalitātes prasību, kuru nebūs iespējams attiecināt uz eksistējošo un iepriekšnodefinēto dimensiju.

Pie līdzīga secinājuma ir nonācis arī *Eppler* (2006), kas savā pētījumā ir mēģinājis konsolidēt eksistējošas dimensijas. Izpētot 70 visbiežāk sastopamus kritērijus, savam risinājumam viņš izvēlējas 16 dimensijas, ko iedalīja 4 līmeņos: (1) "kopiena" (angl. *community*), ko saista ar relevanci jeb atbilstību; (2) "produkts", ko saista ar stabilitāti; (3) process; (4) infrastruktūra. Savu izvēli viņš ir veicis, no 70 dimensijām izdalot tās, kuras, viņaprāt, vislabāk raksturo datu kvalitāti, izslēdzot tās dimensijas, kuru definīcijas pārklājās savā starpā, jo datu kvalitātes dimensijai ar vienu semantisko nozīmi varēja atbilst līdz pat 5 dažādiem nosaukumiem. Taču ir jāatzīmē, ka sākotnēja analizējamo datu kvalitātes dimensiju kopa nav pilnīga, līdz ar ko arī šī klasifikācija ir noderīga lielākoties konkrētā risinājuma ietvaros un var būt noderīga tikai atsevišķajos gadījumos, ja pētnieki, izstrādājot savu risinājumu izvēlās izmantot jau eksistējošo klasifikāciju, taču nav pārliecības, ka tiks izmantota tieši šī klasifikācija.

Papildus, vairāki autori, t.sk. *Batini* (Batini et al., 2016) un *Scannapieco* ar saviem līdzautoriem (2005), nonāk pie secinājuma, ka pat ja dažādu risinājumu autori definē datu kvalitātes dimensijas vienādi, katrai dimensijai var tikt nodefinētas dažādas metrikas, taču pie vienāda metriku saraksta, metodes to mērīšanai visbiežāk ir ļoti dažādas, kuru starpā mēdz būt ne tikai metodes un testi, piemēram *Grubbs, Dixon, Walsh* testi (Czechowski et al., 2015), kuru pielietošanas rezultātā tiek iegūti skaidri un viennozīmīgi rezultāti, bet arī tādas subjektīvas metodes kā galalietotāju aptaujas.

Piemēram, datu kvalitātes dimensiju pārskatu un to savstarpēju salīdzinājumu veic arī *Jayawardene* ar līdzautoriem (2013, 2015), savā pētījumā analizējot 16 dažādus pētījumus, kuri kopā izmanto 127 datu kvalitātes dimensijas (vidēji 8 dimensijas uz vienu risinājumu), no kuriem autori izceļ 30 dominējošas datu kvalitātes dimensijas. Šīs 30 dimensijas autori iedala 8 "klasteros": pilnīgums, pieejamība un piekļūstamība, izplatība, precizitāte, derīgums, uzticamība, nepretrunīgums, lietojamība un interpretējamība, kas saskaņojas ar (Zhang, 2014). Katram klasterim autori sniedz dažādu autoru definīcijas, iedalot tās vienā no divām grupām atkarībā no pieejas – deklaratīvā (angl. *declarative perspective*) vai lietojuma (angl. *usage perspective*), kas ir atkarīgi no dimensijas kontekstuālās nozīmes, kuru tai piešķir konkrētā risinājuma autors. Rezultātā viens klasteris pieder dažādām grupām, un no piederības grupai mainās arī tās definīcija un tās mērījumi.

Kā seko no iepriekšrakstītā, dažādi pētnieki savus risinājumus balsta uz dažādām dimensijām, kuru nosaukumi un nozīmes mēdz atšķirties, taču piedāvāto risinājumu atšķirības

mēdz pastāvēt arī datu kvalitātes mērīšanas mehānismos vai implementācijās. Tālāk tiek apskatīti pēc autores domām interesantākie ar datu kvalitāti saistīti pētījumi.

Pretstatā pētījumiem, kuros pētnieki izmanto pārāk un dažreiz nepamatoti augstu datu kvalitātes dimensiju skaitu, kas apgrūtina šo pieeju izmantošanu lietotājiem bez atbilstošām zināšanām un iemaņām, citos pētījumos to skaits ir samazināts līdz pat dažām datu kvalitātes dimensijām. Šīs lēmums atsevišķos gadījumos arī mēdz būt riskants, jo zems izmantojamo datu kvalitātes dimensiju skaits mēdz neļaut veikt pietiekoši padziļināto datu kvalitātes analīzi, ierobežojot datu kvalitātes analīzi. Viens no piemēriem ir 1.2.2. punktā minētais (Schmidt et al., 2015) pētījums, kurā Dāņu Nacionālā Pacientu Reģistrā datu kvalitātes analīze tika veikta, pielietojot tam tikai divas dimensijas – derīgums un pilnīgums.

Cits piemērs ir (Ferney et al., 2017) pētījums, kura autori atvērto datu kvalitātes analīzi veic datu kopām pielietojot tikai trīs datu kvalitātes dimensijas - trasējamība (angl. *traceability*), pilnīgums un atbilstība (angl. *compliance*). Šo datu kvalitātes dimensiju mērīšanai autori ir izstrādājuši programmatūru *RapidMiner*, kas atbalsta visas "klasiskās" datu kvalitātes analīzes darbības - datu iegūšanu, apstrādi, glabāšanu un novērtēšanu. Rezultātā datu kvalitātes analīzes process ir relatīvi intuitīvs: (1) datu izgūšana no datu avota, pētījumā ietvaros izanalizējot 6 datu kopas no *www.datos.gov.co* repozitorija, saglabājot datus *MongoDB* datubāzē; (2) *MongoDB* datubāzes satura lasīšana, sasaistot metadatus, ja tie ir pieejami, ar datu kopu, pielietojot trasējamības metriku; (3) datu transformēšana, apstrādājot datus, kas pēc 1. un 2. soļu izpildes tiek glabāti *.json* formātā, sadalot tos rindās un kolonnās; (4) atribūtu ģenerēšana, pārbaudot, (a) izveides datuma un avota pieejamību, (b) visu lauku aizpildījumu, pārliecinoties, ka neviena vērtība nav tukša vai nav vienāda ar "?", kas *RapidMiner* gadījumā ir pēc noklusēšanas tukšiem laukiem piešķirtā vērtība, (c) pēdējās izmaiņas datumu un izmaiņu veikšanas datumu saraksta pieejamību; (5) metriku izvēle, veicot atbilstošās darbības (piemēram, trasējamības metrikas paredz ierakstu skaitīšanu, novirzes aprēķinus utt.); (6) atribūti ģenerēšana, veicot sākotnējo un iepriekšējā solī iegūto rezultātu salīdzinājumu; (7) atribūtu izvēle. Analīzes rezultāti sniedz lietotājiem informāciju par datu kopas kvalitāti grafisko diagrammu veidā. Risinājuma autori fokusējas uz (1) metadatu esamības pārbaudi, (2) informācijas par datu kopas izcelsmi esamības pārbaudi, (3) informācijas esamības par izmaiņu biežumu pārbaudi, kā arī (4) *null* vērtību esamības pārbaudi. Neskatoties uz pārbaužu un izstrādātā risinājuma ierobežotību, un nespēju definēt datu kvalitātes prasības konkrētiem laukiem, t.sk. vienkāršas sintaktiskās pārbaudes, datu kopu analīzes rezultātā autori atklāja vairākas datu kopās esošas datu kvalitātes problēmas.

Cits pētījums (Zaveri et al., 2016) fokusējas uz saistīto datu kvalitātes novērtēšanas pieejām, veicot 30 dažādu pieeju un 12 rīku analīzi. Avotu izvēli pētījuma ietvaros veiktai

analīzei, autori balstīja uz sekojošiem kritērijiem: (1) rakstiem ir jābūt publicētiem starp 2002. un 2014. gadiem, darbā apskatot (2.1) saistīto datu kvalitātes jautājumus, (2.2) saistīto datu uzticamības novērtēšanu, (2.3) piedāvājot vai implementējot pieejas saistīto datu kvalitātes novērtēšanai, (2.4) vērtējot saistīto datu vai informācijas sistēmu kvalitāti, balstoties uz saistīto datu principiem, rezultātā apkopojot informāciju par atklātām kvalitātes problēmām. Pirmkārt, ņemot vērā vairāku nosaukumu piešķiršanu vienādām pēc savās semantiskās nozīmes datu kvalitātes dimensijām, autori formalizē rakstos izmantoto terminoloģiju, rezultātā iegūstot 18 datu kvalitātes dimensiju un 69 metriku sarakstu. Datu kvalitātes dimensijas autori sadalīja 4 grupās - pieejamības, kontekstuālās, attēlošanas un iepriekšdefinētas dimensijas, katrā grupā esošajām dimensijām sniedzot to definīciju, metriku sarakstu, nosakot mērījumu raksturu - kvantitatīvs vai kvalitatīvs. Analizētās pieejas tika salīdzinātas pēc dimensijām, kas tajos tiek izmantotas, tām piekārtotām metrikām, datu tipiem, kurus ir iespējams analizēt, un rīka esamības. Rīkus autori salīdzināja pēc to pieejamības, licencēšanas, pielāgojamības, lietojamības un citiem kritērijiem. Savas analīzes rezultātā autori atzīmē, ka mūsdienās pētījumos visbiežāk tiek piedāvātas teorētiskās metodoloģijas, ļoti reti implementējot rīkus datu kvalitātes novērtēšanai, iezīmējot nepieciešamību pēc jauniem pētījumiem, kuru rezultātā tiktu piedāvāti praktiskie risinājumi datu kvalitātes novērtēšanai, iezīmējot arī specifiskus aspektus, kas būtu jāaplūko turpmākos pētījumos. Galvenais pienesums ir literatūras pārskats, kuru pēc autoru domām citi pētnieki var izmantot, uzsākot savus pētījumus un pieņemot lēmumu par savu pētījumu virzienu, kā arī aicinājums iesaistīties datu kvalitātes pētījumos. Veiktais datu kvalitātes dimensiju apkopojums, kārtējo reizi liek pārdomāt datu kvalitātes dimensiju izmantošanas lietderīgumu un piemērotību datu kvalitātes problēmas risināšanai.

Viens no risinājumiem, kas pēc savas pamatdomas būtiski atšķiras no pārējiem, ir *Zhang* un līdzautoru (2014) pētījums. Ņemot vērā no datu kvalitātes dimensijas jēdziena izrietošas problēmas, precīzāk, dažādas datu kvalitātes dimensiju klasifikācijas un vienotās klasifikācijas trūkumu, kā arī to faktu, ka dimensijas un prasības mēdz būt uztvertas dažādi, *Zhang* ar līdzautoriem (2014) piedāvā datu virzītu pieeju datu kvalitātes prasību noteikšanai. Neskatoties uz to, ka autori paši apzinās datu kvalitātes dimensiju jēdziena problēmas, autori savu risinājumu tomēr saista un pat balsta uz dimensiju jēdzienu, taču cenšas risināt vismaz dažas ar to saistītas problēmas. Izanalizējot eksistējošās datu kvalitātes dimensijas, darba autori ir atlasījuši 8 dimensijas, kas, viņuprāt, vislabāk raksturo datu kvalitāti. Šīs dimensijas ir pilnīgums, pieejamība un piekļūstamība, izplatība, precizitāte, derīgums, uzticamība, nepretrunīgums, lietojamība un interpretējamība (angl. *interpretability*). Piedāvātais risinājums paredz analizējamo datu ielādi datubāzē un to apstrādi ar *SQL* vaicājumiem. Ir jāatzīmē, ka šie soļi daļēji pārklājās ar piedāvāto risinājumu (sk. 3. nodaļu). Arī metadatu analīze tiek veikta

manuāli vai ar *SQL* vaicājumu palīdzību, analizējot kolonnu saturu pret datu sniedzēja dokumentāciju. Datu kopa tiek apstrādāta, ar mērķi atrast un novērst dublikātvērtības, un noteikt prasības datu kopu kvalitātei. Datu tipa vai raksta noteikšanai tiek veikta apakškopas analīze, taču, ja apakškopā, kurā autoru piemērā ir 200 ierakstu liela, bet kopējais ierakstu skaits pārsniedz 60 miljonus, netika identificēta datu kvalitātes kļūda, visas datu kopas atbilstība konkrētai prasībai netiek pārbaudīta. Neskatoties uz to, ka autoru mērķis ir identificēt datu kvalitātes prasības konkrētai datu kopai, šīs paņēmiens nav piemērots datu kvalitātes analīzei, jo datu kvalitātes problēmai ir jābūt identificētai arī vienas vērtības gadījumā, kā arī 200 ierakstu no vairāk kā 60 miljonu liela apakškopa ir pārāk maza, lai noteiktu konkrētas prasības novērtēšanas nepieciešamību, un tas var tikt izmantots tikai kā sagatavošanas darbs.

Datu kvalitātes dimensijas mēdz tikt izmantotas arī specifiskākajos risinājumos, viens no kuriem ir (Färber et al., 2018) risinājums, kas fokusējas uz zināšanu grafu kvalitātes novērtēšanu. Pētījuma autori veic piecu brīvi pieejamu zināšanu grafu - *DBpedia, Freebase, openCyc, Wikidata* un *YAGO* kvalitātes analīzi un to savstarpēju salīdzinājumu. Šīm nolūkam autori izvēlās 11 dimensijas, katrai no tām piekārtojot vairākus kritērijus, ko viņi ir aizguvuši no citu autoru darbiem, kuru starpā ir iepriekšminētie (Wang et al., 1996), (Bizer, 2007) un (Zaveri et al., 2016). Kopumā autori izvēlēja 34 kritērijus, savukārt 11 dimensijas sagrupēja 4 kategorijās: (1) iekšējā (angl. *intrinsic*), kas ietver sevī precizitāti, ticamību (angl. *trustworthiness*) un nepretrunīgumu; (2) kontekstuālā, kas ietver sevī relevanci (angl. *relevancy*), pilnīgumu un savlaicīgumu; (3) reprezentācijas/ pieraksta (angl. *representation*), kas ietver sevī saprotamību (angl. *ease of understanding*), sadarbspēju (angl. *interoperability*); (4) piekļūstamības, kas ietver sevī piekļūstamību, licenci un savstarpēju saistību (angl. *interlinking*) (Nikiforova, 2018a, 2019b). Lietotāju ērtībām autori piedāvā satvaru, izmantojot kuru, izvēlēto dimensiju pielietošanas katram no pieciem zināšanu grafiem rezultātā, autori identificēja vairākas zināšanu grafos sastopamas datu kvalitātes problēmas. Viena no visbiežāk sastopamam kvalitātes problēmām ir literāļu sintaktiskas datu kvalitātes problēmas, kas ir raksturīgas arī relāciju datubāžu datu kopām. Piedāvātais satvars ir daļēji pielāgots lietotāja vajadzībām, nodrošinot iespēju veikt zināšanu grafu datu kvalitātes analīzi, iepriekš izvēloties konkrētam gadījumam piemērotākas datu kvalitātes dimensijas no iepriekšsagatavota dimensiju saraksta, katram kritērijam norādot tā svaru konkrētas pārbaudes ietvaros, t.i., cik svarīgs ir konkrēts kritērijs konkrētas analīzes gadījumā. Šī ideja līdzās darba 3. nodaļā piedāvātai pieejai, atbilstoši kurai katram lietotājiem ir jābūt iespējai izteikt kvalitātes prasības atbilstoši konkrētam lietošanas piemēram. Taču (Färber et al., 2018) pieeja ir ierobežotāka, jo dimensiju saraksts ir sagatavots iepriekš, taču: (1) dimensiju sarakstā var nebūt konkrētam lietotājam un lietošanas piemēram nepieciešamas dimensijas; (2) atbilstoši iepriekšminētajam, dimensijas

bieži vien mēdz būt neviennozīmīgas un var tikt interpretētas dažādi atkarībā no lietotāja un lietošanas piemēra. Ņemot vērā, ka šī pētījuma galvenā ideja ir zināšanu grafu kvalitātes analīze, neskatoties uz risinājuma pamatidejas daļēju atbilstību darba ietvaros piedāvātas pieejas idejai, tā nav paredzēta atsevišķu datu kopu, piemēram, atvērto datu kopu kvalitātes analīzei. Lietotājiem ir sniegta iespēja kontrolēt zināšanu grafam pielietotus kritērijus un to svarus noteiktās robežās, neļaujot vērtēt to izmantošanas iespējas noteiktam nolūkam. Pieejas specifikas dēļ, šī pieeja paredz padziļinātu zināšanu grafu izpratni un IT cilvēku ar atbilstošām zināšanām obligāto iesaisti. Pēc autores domām dotais risinājums ir perspektīvs, un, iespējams, viens no labākajiem savā kategorijā, t.i. zināšanu grafu datu kvalitātes analīzei domāto risinājumu starpā, jo, piesaistot lietotājus ar atbilstošām zināšanām, kuru iesaiste risinājuma specifikas dēļ ir obligāta, risinājuma izmantošanas rezultāts var būt noderīgs galalietotājiem.

Viens no spilgtākajiem dotās kategorijas pētījumiem, kas 2015. gadā rezultējas ar standartu izstrādi ir *ISO/IEC 25024 Systems and software engineering - Systems and software Quality Requirements and Evaluation (SQuaRE) - Measurement of data quality*, kas pieder *ISO/IEC 250xx* kvalitātes standartu saimei (ISO/IEC 25024, 2015). *ISO/IEC 250xx* saimei ir iedalīta 4 daļās – kvalitātes pārvaldība (*ISO/IEC 2500n*), kvalitātes modelis (*ISO/IEC 2510n*), kvalitātes mērīšana (*ISO/IEC 2502n*), kvalitāte prasības (*ISO/IEC 2503n*), kvalitātes novērtēšana (*ISO/IEC 2504n*), kuras ir iedalītas vēl smalkāk atsevišķos standartos. To savos darbos izmanto vairāki pētnieki (arī daži turpmāk apskatītie pētījumi balsta savus risinājumus tieši uz *SQuaRE* standartu (piemēram, (Moraga et al., 2009), (Vetro et al., 2016)). Šis standarts satur katras datu kvalitātes raksturiezīmes mērījumu aprakstu, entitāšu, kurām ir piemērojamas konkrētas raksturiezīmes datu kvalitātes dzīvescikla laikā, skaidrojumu par datu kvalitātes mērījumu piemērošanu, vadlīnijas organizācijām datu kvalitātes prasību un novērtēšanai paredzētu mērījumu noteikšanai. Šī standarta pamatā ir 15 datu kvalitātes dimensijas, kas ir apskatīti standarta dokumentācijā. Šīs standarts ir paredzēts tā pielietošanai jebkuriem strukturētiem datiem. Tā pozitīva iezīme ir katra datu kvalitātes aspekta definīcija un datu kvalitātes raksturiezīmju jeb dimensiju mērīšanas darbību apraksts, kā arī līdzība programmatūras kvalitātes aspektiem, kas var tikt pamatots ar šī standarta piederību *ISO/IEC 250xx* kvalitātes standartu saimei un visciešāko saistību ar programminženierijas standartu *ISO/IEC 25000* un *25012:2008.* Taču ir jāatzīmē, ka ISO standartu ieviešana un to turpmāka ievērošana ir ārkārtīgi resursietilpīgs process, kas parasti pat lieliem uzņēmumiem prasa daudz resursu un nepārtrauktu to izpildes uzraudzību, līdz ar ko tajā aprakstītie ar datu kvalitāti saistīties aspekti var kalpot par pamatu (a) praktiskiem risinājumiem, ko ir pierādījuši arī vairāki pētnieki, kas savus risinājumu balsta uz datu kvalitātes dimensijām, (b) datu turētājiem, kas

uzkrāj un apstrādā savus datus, taču datu kvalitātes dimensiju neviennozīmības problēma ir spēkā.

Basciani ar līdzautoriem savā pētījumā (Basciani et al., 2016) piedāvā izmantot modeļa virzītu inženierijas pieeju, kas ļauj sniegt precīzu kvalitātes modeļa definīciju, identificējot kvalitātes atribūtus, kas interesē konkrētu ieinteresēto personu. Galvenais eksistējošo risinājumu trūkums pēc autoru domām ir to ierobežota paplašināmība, artefaktu specifika un manuālā novērtēšana. Rezultātā autori piedāvā pieeju, kas atbalsta kvalitātes modeļu definīciju, kas sastāv no hierarhiski organizētiem kvalitātes atribūtiem, kuru novērtēšana ir atkarīga no metrikām, kas ir izvēlētas un pielietotas modelēšanas artefaktam, kas ir jāanalizē. Tiek piedāvāta domēnspecifiska valoda, kura ļauj definēt kvalitātes atribūtus un metrikas. Autori piedāvā arī izpildes vidi, lai modelēšanas artefaktiem pielietotu definētus kvalitātes modeļus un nodrošināt to kvalitātes automātisko novērtējumu. Metamodeļa īpašību definēšanai un pārbaudei, uz kuru savā darbā autori pamatā fokusējas, priekšroka ir dota valodai *mmSpec*, savukārt metamodeļa atribūtu definēšanai tiek izmantota *OCL* valoda. Šīs risinājums, kā jau paši autori atzīst, pamatā ir domāts izstrādātajiem, kam ir iespēja un nepieciešamība atkalizmantot modelēšanas artefaktus, kas vēlās veikt uz pierādījumiem balstītu lēmumu. Šī pieeja ir no artefakta atkarīga, tā var tikt pielietota dažāda veida modelēšanas artefaktam. Atkarībā no konkrēta uzdevuma un lietotāja vēlmes, ir iespējams pielāgot, paplašināt jau eksistējošo vai arī izveidot savu kvalitātes modeli, paplašinot piedāvāto metriku sarakstu, kas ir nepieciešamas konkrēta kvalitātes atribūtu mērīšanai. Tas nozīmē, ka dotā pieeja ir domāta datu kvalitātes speciālistiem, kas ikdienā nodarbojas ar dažādu datu kvalitātes analīzi, taču papildus padziļinātājām zināšanām, viņiem ir jāpiemīt arī atbilstošajām zināšanu inženierijas zināšanām. Taču, neskatoties uz risinājuma piemērotību tikai ierobežotam lietotāju lokam, tas paredz arī galalietotāju iesaisti, sniedzot iespēju piedalīties kvalitātes atribūtu metriku definēšanā datu kvalitātes speciālistu vadībā. Papildus ir jāatzīmē, ka šī risinājuma atsevišķas pazīmes ir raksturīgas arī piedāvātajam risinājumam.

Apskatīto pētījumu viens no galvenajiem trūkumiem ir nepieciešamība datu kvalitātes dimensiju būtības izpratnē, kas izraisa problēmas gan IT un datu kvalitātes speciālistiem, gan tā saucamiem sfēras speciālistiem jeb ne-IT speciālistiem. Savukārt vēl grūtāks uzdevums abu lietotāju grupām ir to pielietošana, piemēram, attiecinot uz konkrētu datu kvalitātes problēmu risinājumā paredzēto datu kvalitātes dimensiju. Pie tam, vairākums pētījumu piedāvā savu dimensiju sarakstu, kas kļūst aktuāls tikai viena risinājuma ietvaros, jo dimensiju būtība dažādos pētījumos var atšķirties pēc savas nozīmes, pat ja nosaukumi ir vienādi un otrādi. Rezultātā, dimensiju definīciju un pielietošanas apgabala izpēte kļūst nepamatoti resursietilpīgs, kā arī atbilstošo zināšanu un pieredzes prasīgs process, it īpaši ne-IT un ne-datu

kvalitātes speciālistiem. Tas būtiski ierobežo atbilstošo risinājumu potenciālo lietotāju loku. Papildus, vairākas pieejas neļauj galalietotājiem definēt specifiskās viņu interesējošas pārbaudes specifiskiem laukiem/ parametriem (Nikiforova, 2018a).

2.2. Atvērto datu portālu kvalitātes novērtēšanas pieejas

Atvērto datu popularitāti autore apskatīja 1.2. apakšnodaļā. Atvērto datu popularitātes pieaugumu rezultātā parādās arvien vairāk atvērto datu portālu, un kā rezultāts tiek izstrādāti atvērto portālu kvalitātes analīzes risinājumi, kas ir populārāki, t.i. biežāk sastopami nekā atvērto datu kopu kvalitātes analīzes risinājumi. Eksistējošie risinājumi arvien biežāk sniedz datu portālā publicēto visu vai atsevišķo datu kopu kvalitātes novērtējumu apkopotā veidā, neļaujot noteikt konkrētas datu kvalitātes problēmas, kuras būtu jāņem vērā galalietotājiem un jālabo datu sniedzējiem.

Piemēram, (Neumaier, 2015) un (Umbrich et al., 2015) pētījumos ir sniegti automatizēto kvalitātes novērtējuma satvaru, kas nosaka un mēra atvērto datu portālu kvalitātes un viendabīguma (angl. *heterogeneity*) problēmas, apraksti.

Atbilstoši (Neumaier, 2015) ņemot vērā, ka mūsdienās nav zināmas visaptverošas un objektīvas atvērto datu portālu esoša stāvokļa un kvalitātes atskaites, kā arī satvaru, kas ļautu nepārtraukti pārraudzīt portālu izmaiņas, ko autori sauc par "evolūciju", savā pētījumā autori cenšas atrisināt šīs problēmas. Datu kvalitāti autori saista ar sešām kvalitātes metrikām: izgūstamību (angl. *retrievability*), lietojumu (angl. *usage*), pilnīgumu, precizitāti, atklātību (angl. *openness*) un sasniedzamību, t.i. iespēju sazināties ar atbildīgo par datiem (angl. *contactability*). Vairākums izvēlēto dimensiju ir autoru definētās datu kvalitātes dimensijas, kuras citos pētījumos neparādās. Nodefinētas datu kvalitātes dimensijas tiek izmantotas, periodiski pārraugot portālu saturu un izskaitļojot kvalitātes metriku kopu, lai iegūtu ieskatu par datu kopu (un metadatu) "evolūciju", t.i., cik lielā mērā dati un to kvalitāte ir mainījušies, salīdzinot esošus rezultātus ar iepriekšiegūtājiem, kas tiek īstenots, pateicoties izstrādātājam satvaram. Izstrādāto risinājumu autori pielietoja trim atvērto datu portāliem, atklājot tajos dažādas datu kvalitātes problēmas. Dotais risinājums ir noderīgs, veicot datu portālu kvalitātes analīzi, taču atsevišķu datu kopu kvalitātes novērtēšanai un analīzei tas nav paredzēts. Savukārt autoru lēmums ieviest jaunas datu kvalitātes dimensijas ir apšaubāms, jo veicot datu kvalitātes analīzi, kur datu kvalitāte ir saistīta ar datu kvalitātes jēdzienu, katras izmantotas datu kvalitātes dimensijas būtība ir jāizprot pietiekoši labi, lai spētu korekti izvēlēties nepieciešamu dimensiju, no kuras būs atkarīgi datu analīzes rezultāti, taču šīs zināšanas ir noderīgas tikai šī risinājuma

ietvaros, jo citas pieejas šīs dimensijas neizmanto vai vismaz dēvē tās savādāk. Iespējams, tādos gadījumos būtu ieteicams izmantot iepriekšdefinētas un vispārpieņemtākas dimensijas, kuras, nepieciešamības gadījumā, tiktu pielāgotas atbilstoši autoru vīzijai. Ir jāatzīmē, ka šīs risinājums pieder arī iepriekšējai pētījumu kategorijai.

Caro un viņa kolēģu projekts (Caro et al., 2007) *PoDQA* piedāvā datu kvalitātes modeli tīmekļu portāliem. Piedāvātā modeļa pamatā ir vairāk kā 30 datu kvalitātes atribūti jeb dimensijas. Risinājuma mērķis ir noteikt konkrētā datu portāla kvalitātes līmeni, kas tiek noteikts, veicot datu lietotāju/ patērētāju izvēlēto kvalitātes atribūtu novērtējumu. Autoru pamatideja piedāvāt risinājumu, kas būtu piemērots gan portālu lietotājiem, gan izstrādātājiem, atbilst 3. nodaļā prezentētās pieejas pamatidejai. Taču 33 datu kvalitātes atribūtu, kas ir iepriekšdefinēti ar autoriem, nevis ar lietotāju, izmantošana ļauj veikt pieņēmumu, ka dotais risinājums nevar tikt uzskatīts par galalietotājam piemērotu, jo pastāv augsta varbūtība, ka galalietotāju loks, kuriem šīs risinājums ar visiem tajā piedāvātajiem atribūtiem būs saprotams, būs ierobežots, prasot no lietotājiem datu kvalitātes jomas zināšanas (saskaņojas ar (Eppler, 2006)). Ir jāatzīmē, ka arī autoru pieeja datu kvalitātes dimensiju izvēlei ir apšaubāma, jo to izvēle balstās uz datorzinātņu fakultātes maģistrantūras studentu aptauju, no piedāvātajiem 34 atribūtiem izvēloties svarīgākos. Risinājumā autori izmanto 33 no tiem, taču, ņemot vērā datu kvalitātes dimensiju daudzveidību un augstu skaitu, ir pamats apgalvot, ka sākotnējais saraksts varēja nebūt pietiekoši pilnīgs. Papildus, ņemot vērā, ka aptaujas dalībnieku, kas ir tīmekļa portālu lietotāji nevis datu kvalitātes analītiķi, viedoklis nevar tikt uzskatīts par pietiekoši autoritatīvu, jo tie nav datu kvalitātes speciālisti, kas spētu objektīvi un pamatoti novērtēt piedāvāto datu kvalitātes dimensiju svarīgumu, kā arī papildināt to ar citām, lai gan atbilstoši (Scannapieco et al., 2002b), (Batini et al., 2016), (Price et al., 2004, 2005), (Jayawardene, 2013, 2015)) arī datu kvalitātes speciālistiem šāda veida uzdevums mēdz izraisīt problēmas. Piedāvātais risinājums sniedz priekštatu par visa datu portāla kvalitāti kopumā, neļaujot noteikt ne tikai konkrētas datu kopas kvalitāti, bet arī datu portālā esošo datu konkrētas kvalitātes problēmas. Savukārt pozitīva risinājuma iezīme ir iespēja veikt (a) viena portāla datu kvalitātes evolūcijas novērtējumu, t.i. datu kvalitātes līmeņa izmaiņas laika gaitā, kā arī (b) vairāku portālu savstarpēju salīdzinājumu, sniedzot lietotājam iespēju izvēlēties piemērotāko.

Vairākums apskatīto risinājumu nav pielietojams vienas konkrētas datu kopas kvalitātes analīzei, sniedzot datu kvalitātes novērtējumu apkopotā veidā, piemēram, visam portālam (Caro et al., 2007). Tas ļauj iegūt ieskatu par "kopējo ainu", taču neļauj noteikt konkrētas datu kvalitātes problēmas, kuras datu galalietotājam un datu sniedzējiem/ turētājiem vajadzētu ņemt vērā, izmantojot atbilstošās datu kopas. No tā seko, ka arī efektīvai datu kvalitātes uzlabošanai šie risinājumi nav piemēroti. Papildus ir jāatzīmē, ka, neskatoties uz autoru tieksmi pēc garāka

un pēc iespējas pilnīgāka datu kvalitātes dimensiju saraksta, prakse rāda, ka, jo augstāks ir izmantoto datu kvalitātes dimensiju skaits, jo sarežģītāks ne-IT un ne-datu kvalitātes speciālistam kļūst konkrēts risinājums un datu kvalitātes analīzes uzdevums, rezultējoties ar augstāku iespējamību datu kvalitātes speciālistu piesaistei datu kvalitātes analīzes veikšanai, kas, kā jau tika minēts mūsdienīgajos apstākļos nav pieļaujams.

Atbilstoši (Kučera et al., 2013), (Vetro et al., 2016) mūsdienās ir pieejami vairāki atvērto datu portāli un atvērtie publiskie dati (*OGD*), taču bieži vien lietotājiem ir grūti vai pat vispār nav iespējams atrast viņus interesējošus datus, kā arī esošu datu kvalitāte parasti ir apšaubama. (Vetro et al., 2016) ir uzsverts, ka mūsdienās pētījumos arvien biežāk uzmanība tiek pievērsta atvērto datu platformu nevis atsevišķu datu kopu kvalitātei. Atbilstoši autoru literatūras pārskata darbam, šobrīd pasaulē ir novērojams visaptverošo teorētisko satvaru, kas varētu nodrošināt atvērto datu portālu datu kopu augstu kvalitāti, trūkums. Tāpēc autori izstrāda satvaru, kas ļauj veikt *OGD* datu kvalitātes analīzi, pielietojot datu kopām 7 datu kvalitātes dimensijas, kas tika noteiktas 15 izstrādātāju aptaujas rezultātā. Papildu dimensiju vai pārbaužu definēšana, kā arī esošo pārbaužu pielāgošana lietotāja vajadzībām nav iespējama. Aptaujā iekļauto datu kvalitātes dimensiju sarakstu autori ir aizguvuši no izvēlētās metodoloģijas *SPDQM* (*Square-Aligned Portal Data Quality Model*) (Moraga et al., 2009) izstrādātājiem, kas satur 42 datu kvalitātes raksturīpašības, kas savukārt tika aizgūtas no šīs metodoloģijas priekštečiem – *SquaRE - Software product Quality Requirements and Evaluation* (ISO, 2008) un *PDQM* (Moraga et al., 2009). Priekšroka *SPDQM* tika dota, veicot *SPDQM, TDQM, DWQ, TIQM, AIMQ, AMEQ* metodoloģiju atbilstības autoru prasībām pārbaudi, atbilstoši kurām metodoloģijai būtu jāatbalsta 3 savstarpēji saistītas, secīgas aktivitātes/ darbības: (1) stāvokļa noteikšana, (2) novērtēšana un mērīšana, (3) uzlabošana. *SPDQM* tika atzīta par vispiemērotāko metodoloģiju, jo pēc autoru domām tā satur gan pamatīpašības, kas ir raksturīgas lielākai metodoloģiju daļai, gan tādas īpatnējas datu kvalitātes dimensijas kā trasējamība (angl. *traceability*), atbilstība (angl. *compliance*) un saprotamība (angl. *understability*). Katrai datu kvalitātes dimensijai autori ir nodefinējuši vairākas metrikas, kas var tikt pielietotas tikai konkrētai datu kopas šūnai vai visai datu kopai, taču automātiska metriku skaitļošana ir iespējama tikai divām datu kvalitātes dimensijām - pilnībai un precizitātei, uz parējām dimensijām attiecināmas metrikas skaitļojot manuāli. Visi mērījumu rezultāti ir normalizēti, t.i. atrodas intervālā [0,1]. Ir jāatzīmē, ka dotā risinājuma pamataspektu izvēle, t.i. metodoloģijas un izmantoto rādītāju izvēle balstās uz lietotāju aptaujām, ko paši autori ir atzinuši par problemātisko paņēmienu, jo vairākkārt ir saskarušies ar lietotāju atbilžu dažādu iespējamo interpretāciju. Viens no piemēriem, kas vislabāk apraksta datu kvalitātes raksturu un sarežģītības, ko izraisa datu kvalitātes jēdziena sasaiste ar datu kvalitātes dimensijām, ir autoru

pieredze, kas tika iegūta, lietotāju definētajām/ novērtētajām kvalitātes problēmām, piekārtojot datu kvalitātes dimensijas. Autori ir konstatējuši, ka vienai problēmai atkarībā no tās interpretācijas var tikt piekārtotas vairākas dažādas dimensijas, taču noskaidrot lietotāja sākotnējo viedokli, t.i. datu kvalitātes problēmas skaidru un viennozīmīgu formulējumu, parasti ir grūti. Ir paredzēts, ka kvalitātes īpašības izvēlas lietotāji, taču nav skaidrs, vai tiem būtu (a) jāveic arī atbilstošo dimensiju, kurām ir jābūt piekārtotām izvēlētājām īpašībām, izvēli, vai arī (b) jā nosaka problēmas, kuras būtu jāizpēta un tad, dimensiju pārzinis izvēlas atbilstošās dimensijas. Autoru pētījums parādīja, ka dimensiju izvēle atbilstoši noteiktām problēmām mēdz būt problemātiska pat tiem cilvēkiem, kuri izprot katras kvalitātes dimensijas būtību. Risinājuma iespējas tika nodemonstrētas, veicot *OGD* datu kvalitātes analīzi, rezultātā izstrādājot arī vadlīnijas datu turētājiem/ sniedzējiem, kas būtu jāievēro, gatavojot datu kopas to publicēšanai. Salīdzinājumā ar citiem risinājumiem, ir jāatzīmē tā pozitīva raksturiezīme, atbilstoši kurai risinājums var tikt pielietots konkrētu datu kopu analīzei. Taču datu kvalitātes analīzei ir nepieciešama datu kvalitātes ekspertu iesaiste, kas korekti izvēlētos konkrētām pārbaudēm nepieciešamas datu kvalitātes dimensijas. Ir jāatzīmē arī risinājuma ierobežojumi, viens no kuriem ir nespēja veikt formāta parezības pārbaudes. Risinājuma autori ir norādījuši, ka analīzes rezultāti lietotājiem tiek attēloti grafiskā veidā, taču rezultātu detalizācijas līmenis nav zināms, t.i. nav zināms, vai lietotājiem tiek izveidoti datu kvalitātes analīzes protokoli, kuros būtu pieejami visi ieraksti, kuros tika konstatētas datu kvalitātes problēmas.

Tāpat kā (Neumaier, 2015) un (Umbrich et al., 2015) šie risinājumi ir noderīgi, veicot datu kvalitātes uzlabošanu, taču tie galvenokārt ir noderīgi portālu turētājiem un pārbaudītājiem, gatavojot datu kopas to publicēšanai nevis datu lietotājiem, analīzējot to izmantošanas iespējas atbilstoši noteiktiem nolūkiem un lietošanas piemēriem.

2.3. Saistīto datu kvalitātes novērtēšanas pieejas

Vairāki pētījumi analīzē saistīto [atvērto] datu kvalitāti, galvenokārt fokusējoties uz *DBpedia* analīzi ((Acosta et al., 2013), (Färber et al., 2018), (Kontokostas et al., 2014), (Redman, 2001), (Zaveri et al., 2016)). Neskatoties uz to noderīgumu saistīto datu analīzei, tie nav piemēroti atsevišķu datu kopu kvalitātes analīzei, papildus prasot lietotājiem padziļinātas zināšanas atbilstošājā apgabalā (Nikiforova, 2018b).

Ņemot vērā, ka vairāki risinājumi mēdz piederēt dažādām grupām, daži no šī grupai piederošajiem risinājumiem - zinātniskas literatūras apskati par saistīto datu kvalitātes novērtēšanas pieejām (Zaveri et al., 2016) un (Färber et al., 2018), ir apskatīti 2.1. apakšnodaļā.

Cits piemērs ir (Kontokostas et al., 2014) piedāvātā testu virzītā metodoloģija saistīto datu kvalitātes novērtēšanai. Piedāvātās pieejas mērķis ir vārdnīcu, ontoloģiju un zināšanu bāzu testpiemēru izveide, kas nodrošinātu kvalitātes "pamatlīmeni". Piedāvātā metodoloģija ir paredzēta saistīto datu resursu kvalitātes novērtēšanai, balstoties uz datu kvalitātes problēmu formalizāciju. Formalizācijā tiek izmantoti *SPARQL* vaicājumu paraugi, uz kuriem balstās specifiskie kvalitātes testa vaicājumi. Izmantojot *RDF* datu modeli, autori ir izstrādājuši modeļos balstītu pieeju *RDF* zināšanu bāzes datu kvalitātes testu izveidei. Datu kvalitātes testa paraugs *DQTP* ir kortežs (*V, S*), kur *V* ir tipizēto paraugu mainīgo kopa, kuru starpā var būt *IRI*, literāļi, operatori, datu tipu vērtības un regulāras izteiksmes, savukārt *S – SPARQL* vaicājuma paraugs ar vietturi (angl. *placeholder*) mainīgajam no kopas *V*. Izveidoto *RDF* testpiemēru pārklājuma notācija balstās uz sešām pārklājuma metrikām, četras no kurām ir paredzētas īpašībām un divas – klasēm, datu kvalitātes testu paraugu bibliotēku veidojot lietotāju aptaujas rezultātā. Testi tiek definēti vienā no sekojošajiem veidiem: (1) ieinteresēto pušu atgriezeniskā saite no datu kopas izmantošanā iesaistītas personas, un (2) eksistējoša *RDFS/ OWL* datu kopas shēma. Tie tiek veidoti, (a) izmantojot *RDFS/ OWL* ierobežojumus tiešajā veidā, (b) bagātinot *RDFS/ OWL* ierobežojumus, (c) atkalizmantojot testus, balstoties uz kopīgām vārdnīcām, (d) instancējot eksistējošas *DQTP*, (e) ar pašrakstītājiem *DQTP*. Testpiemēri tiek attiecināti uz zināšanu bāzes konkrētām īpašībām vai klasēm, t.i. piedāvātā metodoloģija ļauj veikt kvalitātes novērtējumu atbilstoši konkrētam lietošanas piemēram, atbalstot arī datu uzlabošanu. Veidojamā vārdnīca ir atkalizmantojama, t.i. tā var tikt pielietota citai datu kopai, pielāgojot to konkrētam lietošanas piemēram. Piedāvātais risinājums ir paredzēts automātiskās testu instancēšanai no paraugiem, atbalstot arī automātisku atvasināšanu no *OWL* shēmu aksiomām. Tā kā pēc autoru viedokļa, kuram piekrīt arī darba autore un daudzi citi pētnieki, daudzas saistīto atvērto datu shēmas nav pietiekoši izteiksmīgas, piedāvātā metodoloģija nodrošina arī pusautomātisku shēmu bagātināšanu. Rezultātā izstrādātais risinājums ļauj veikt testu instancēšanu (a) automātiski, balstoties uz shēmu ierobežojumiem, (b) pusautomātiski, sniedzot lietotājiem iespēju ģenerēt specifiskas testu instancēšanas, kas turpmāk tiek pielietotas shēmai vai datu kopai. Atbilstoši risinājuma autoriem viena no piedāvātās pieejas priekšrocībām ir iespēja domēnspecifiskas semantikas iekodēt datu kvalitātes testpiemēros, tādejādi nodrošinot iespēju pētīt datu kvalitātes problēmas ārpus parastās kvalitātes heiristikas. Savukārt galvenais uz doto brīdi esošais risinājuma trūkums, ko atzīmē paši risinājuma autori, ir tam raksturīgais zems testu pārklājums, kas ir saistīts ar shēmu lielu izmēru un pārāk zemu izteiksmību. Taču neskatoties uz to, izstrādāto risinājumu aprobējot uz 297 shēmām, kurām tika izveidoti 32 293 testpiemēri, autori secināja, ka saistītājiem datiem ir raksturīgs ļoti augsts datu kvalitātes problēmu skaits.

Saistīto datu kvalitātes problēmu noteikšanai un risināšanai paredzētais risinājums, kas izmanto tā saucāmu kolektīvu pakalpojumu izmantošanu (angl. *crowdsourcing*) ir (Acosta et al., 2013). Kolektīvu pakalpojumu izmantošanas dēļ piedāvātais risinājums būtiski atšķiras no citiem. Metodoloģijas implementācijai tika izmantotas tādas kolektīvo pakalpojumu izmantošanas pieejas kā: (1) saistīto datu ekspertu iesaistīšana kvalitātes problēmu identificēšanai un klasificēšanai, kas ietver sevī *RDF* resursu izpēti un nepareizu trijnieku identificēšanu, kas tiek panākts, rīkojot iesaistīto dalībnieku "sacensības"; (2) izejas datu publicēšana *Amazon Mechanical Turk* (*MTurk*) platformā kā mikrouzdevumus (angl. *paid microtasks)* datu kvalitātes problēmu verificēšanai, kas paredz cilvēkam lasāmas informācijas par dotiem *RDF* trijniekiem, t.i. 1. solī iegūto trijnieku, analīzi, tādejādi verificējot rezultātus. Risinājuma pamatā ir tā saucama "atrast-salabot-verificēt" (angl. *find-fix-verify)* pieeja. Eksperimenta veikšanai un iegūto rezultātu apkopošanai autori ir izstrādājuši rīku, kurā tiek reģistrētas visas identificētas datu kvalitātes problēmas. Pētījumā tika iesaistīti datu pētnieki un entuziasti, kas pētīja un klasificēja kvalitātes problēmas *DBpedia*. Savā risinājumā autori priekšroku deva mikrouzdevumiem, jo pēc viņu viedokļa tas ir ātrs un izmaksu ziņā efektīvs ar ekspertiem konstatēto kļūdu pārbaudes veids. Dotās pieejas izmantošanas rezultātā tiek pārklāti sekojošie kvalitātes apgabali: (1) objekta vērtības izvilkšanas pareizība vai pilnība; (2) datu tipu izvilkšanas pareizība; (3) saišu pareizība starp *DBpedia* entitātēm un saistītiem tīmekļa resursiem. Lielu uzmanību pievēršot eksperimenta projektējuma izveidei, t.sk. pārdomājot atlīdzības piešķiršanas procedūru eksperimentu dalībniekiem, autori ir veltījuši pārāk maz uzmanības tehniskai daļai. Lai gan arī pašas pieejas izvēle ir diezgan apšaubāma, par ko liecina arī risinājuma autoru atziņas, atbilstoši kurām veikto eksperimentu gaitā saistīto datu eksperti pieļāva salīdzinoši daudz kļūdu, t.sk. tika novērotas kļūdainas atbilsmes (angl. *false-positive*) un kļūdainas neatbilsmes (angl. *false-negative*), it īpaši objektu vērtību un datu tipu analīzes gadījumā, kopēja 1. posma jeb datu ekspertu veiktās analīzes rezultāta precizitāte atkarībā no analizējamas datu kvalitātes problēmas veida svārstās no 0.1525 līdz 0.8270. 2. posmā iesaistīto *MTurk* dalībnieku, kas verificē 1. posmā iegūtus rezultātus, analīžu rezultātu precizitāte svārstās no 0.4752 līdz 0.9412. Tā kā visa analīze pamatā balstās uz analīzē iesaistītājiem cilvēkiem, t.i. viņu zināšanām, pieredzi, uzmanību un citiem faktoriem, rezultātu pareizība un izmantotas pieejas piemērotība šāda veida uzdevumiem ir apšaubāma.

Apskatītās pieejas ir paredzētas IT-speciālistiem ar padziļinātām zināšanām saistīto datu un citās zināšanu inženierijas apakšjomās, t.sk. resursi, *RDF*, *OWL*, ontoloģijas utt.. Tā kā saistīto datu kvalitātes analīze nav tiešā veidā saistīta ar promocijas darba tēmu, detalizēti šīs pieejas netiks aplūkotas.

2.4. *Data Quality Services*

SQL Server Data Quality Services jeb *DQS* ir viens no *Microsoft SQL Server* komponentiem, kas ir paredzēts *EIM* (*Enterprise Information Management*) risinājumu, kuru starpā ir arī *SQL Server Integration Services* jeb *SSIS* un *SQL Server Master Data Services* jeb *MDS*, izstrādei. *DQS* ir zināšanu virzīts risinājums, kas sniedz interaktīvo un datorizēto iespēju datu avotu kvalitātes un integritātes pārvaldīšanai.

Atbilstoši *DQS* dokumentācijai (Laudenschlager et al., 2012a, 2012b) tas piedāvā veikt datu kvalitātes uzlabošanu, nodrošinot datu piemērošanu konkrēta lietotāja vajadzībām, tādejādi veicot to uzlabošanu. Atbilstoši rīka izstrādātājiem tā izmantošanas rezultātā dati kļūst vairākkārt izmantojami, uzticami, pieejami, kā arī tiek nodrošināta iespēja uzlabot datu pilnīgumu, precizitāti, nepretrunīgumu un atbilstību.

Tā kā *DQS* ir zināšanu virzīts risinājums, datu kvalitātes analīzes pamatā ir zināšanu bāzes izveide, kas turpmāk tiek izmantota tādu ar datu kvalitāti saistīto uzdevumu izpildei kā datu labošana, bagātināšana, standartizācija un dublikātu izslēgšana. Tā pamatkomponenti ir:

1) zināšanu bāze - tā saucamais "zināšanu virzīts risinājums", kurš analīzē datus, balstoties uz zināšanām, kas tiek definētas *DQS*. Tas ļauj veidot datu kvalitātes procesus, kas nepārtraukti uzlabo un papildina zināšanas par datiem, tādā veidā uzlabojot arī datu kvalitāti. Zināšanu pārvaldības galvenie procesi ir (1) datu izvēršana, (2) izlūkošana, (3) zināšanu pārvaldība, (4) datu labošana un (5) standartizēšana, (6) pārbaude uz saskanību un (7) labošana. Tā ietver sevī saskanības politiku (angl. *matching policy*), domēnus, t.sk. saliktus jeb jauktus domēnus (Laudenschlager et al., 2012a). Rīka izstrādātāji apgalvo, ka, neskatoties uz to, ka zināšanu izveide ir sarežģīts process, kas prasa rūpīgu pieeju un augstu detalizācijas līmeni, *DQS* ir nodrošināta automātiskā datu izvilkšana (angl. *extracting*) no datu izlases jeb parauga (angl. *sample*), kas būtiski paātrina procesu, samazinot arī iespējamo kļūdu skaitu un kļūdu rašanas iespējamību ((Laudenschlager et al., 2012a), (Prakash, 2016)). Ir nodrošināta *DQS* zināšanu bāzes uzlabošanas iespēja, veicot papildinājumus vai izmaiņas tajā tiklīdz rodas jaunas kvalitātes prasības. Vienreiz izveidota zināšanu bāze var tikt atkalizmantota, balstot jaunu zināšanu bāzi uz jebkuru iepriekšēju, nepieciešamības gadījumā pielāgojot to konkrētas analīzes vajadzībām;
2) datu saskanības noteikšana (angl. *data matching*) - aktivitāte, kuras ietvaros likumos balstītie procesi identificē dublētus ierakstus, nosakot pilnu vai daļēju

saskanību, lietotājam ļaujot veikt dublikātu izslēgšanu (Laudenschlager et al., 2012a);

3) datu tīrīšana - aktivitāte, kas ļauj rediģēt, dzēst vai standartizēt datus, kas nav patiesi vai pilnīgi, ko īsteno datorizētie un interaktīvie procesi, parasti izpildot tos secīgi, balstoties uz iepriekšējos posmos izveidoto zināšanu bāzi (Masson, 2011). Parasti datu tīrīšana notiek 2 soļos:
    1) datorizētais datu tīrīšanas process: dati tiek analizēti un pārbaudīti pret zināšanu bāzes zināšanām, galalietotājam sniedzot ieteikumus par ieteicamajām izmaiņām. Apstrādātie dati tiek iedalīti 5 grupās: (1) ieteiktie (angl. *suggested*), ja vērtību atbilstība nebija precīza, un *DQS* iesaka vērtības labošanas iespējas (ieslēdzot "Uzticības" iespēju, ir iespējams ierakstus zem konkrēta uzticības līmeņa automātiski apstiprināt, parējos pārvietojot atsevišķajā tabulā to turpmākai analīzei), (2) jaunie (angl. *new*), (3) nederīgie (angl. *invalid*), ja *DQS* uzskata, ka konkrēta vērtība neatbilst jeb nepieder konkrētam domēnam, (4) izlabotie (angl. *corrected*), ja vērtība tiek uzskatīta par nepareizu un tika izlabota, atbilstošu vērtību saglabājot kolonnā "Izlabotie" (angl. *corrected*), (5) pareizie (angl. *correct*);
    2) interaktīvais datu tīrīšanas process: iepriekšēja solī iegūtu rezultātu apstiprināšana, noraidīšana vai manuālā labošana (Masson, 2011).
4) datu profilēšana, kas ļauj analizēt datu integritāti, kas tiek panākts, analizējot datu avotus, kā arī sniedzot ieskatu par datiem katrā datu izpētes, domēna pārvaldības, saskanības noteikšanas un datu tīrīšanas procesu posmā. Tā sniedz informāciju par datu kvalitātes uzlabošanas rezultātiem pēc datu tīrīšanas un saskanību noteikšanas procesu veikšanas, sniedzot informāciju par divām datu kvalitātes dimensijām: pilnīgums un precizitāte (Laudenschlager et al., 2017).

DQS risinājuma iespēju un piemērotības datu kvalitātes analīzes uzdevumam novērtēšanai autore izmēģināja uz testa datubāzes piemēra, kas saturēja gan kvalitatīvus, gan nekvalitatīvus datus. Lai gan *DQS* izmantošanas rezultātā autore konstatēja atsevišķas datubāzē esošās datu kvalitātes problēmas, tā izmantošanas gaitā autore identificēja sekojošus risinājuma trūkumus:

1) neskatoties uz to, ka *DQS* nenosaka stingru datu analīzes darbību secību, t.i. ļaujot veikt darbības brīvā secībā, šī iespēja nav realizēta pietiekoši augstā līmenī. Piemēram, neskatoties uz to, ka domēnu izveide ir iespējama arī zināšanu bāzes izveides procesā, domēna likumu izveide tajā brīdī vairs nav iespējama, jo ir iespējama tikai daļēja domēna definēšana;

2) noklusētais saskanības noteikšanas minimālais saskanības slieksnis ir 80%, līdz ar ko šī saskanības noteikšanas iespēja nevar tikt izmantota, ja galalietotājs vēlās noteikt ierakstus, kuru saskanība ir, piemēram, 70 vai 79%;
3) domēnu iespējamo datu formātu saraksts ir nepilns un to papildināšana nav iespējama. Rezultātā, neskatoties uz to, ka tiek piedāvāti 17 dažādi datuma pieraksta varianti, nav neviena varianta, kur datuma pieraksts sāktos ar gadu, kuram sekotu mēnesis vai diena, piemēram, "*yyyy-mm-dd*" un "*yyyy-dd-mm*". Tas nozīmē, ka lietotāji, kuriem ir jāpārbauda datuma vērtību pieraksta atbilstība vienam no iepriekšminētajiem paraugiem, pārbaudes *DQS* rīkā nav paredzētas;
4) pārbaudīt datuma/ laika laukus (*datetime* vai *smalldatetime*) pēc noklusēšanas nav iespējams, jo tie ir pieejami datu ierakstu tabulā, taču nevar tikt izvēlēti kartēšanas sadaļā (*angl. mapping*). Vienīgais datuma tips, kas *DQS* tiek atbalstīts, ir *date*. Tas nozīmē, ka atbilstošo vērtību kvalitātes pārbaudei, lietotājiem ir jāparūpējas pašiem par atbilstošo parametru vērtību konvertāciju *date* formātā, kas nav lietotājam draudzīgi;
5) skaitļu formātu izmaiņu veikšana nav iespējama. Tas attiecās arī uz to vērtībām datu analīzes rezultātu sadaļā, kas virkņu gadījumā strādā;
6) datu kvalitātes analīze var tikt veikta tikai vienai tabulai. Vairāku tabulu apvienošana analīzes veikšanai vai kontekstuālā datu kvalitātes pārbaude vairāku tabulu kontekstā nav iespējama. Vienīgais iespējamais vairāku tabulu vienlaicīgas apstrādes variants ir iepriekšēja tabulu skatu izveide, veicot vairāku tabulu apvienošanu. Taču šīs variants ir aktuāls (a) lietotājiem ar atbilstošām zināšanām, (b) lietotājiem, kuriem pie tam ir piekļuve datubāzei un tabulām, kuru skati ir jāveido. Citiem lietotājiem vairāku tabulu vienlaicīga izmantošana datu kvalitātes analīzē nav iespējama. Tas nozīmē arī to, ka nav iespējama kodifikatoru pārbaude, kas nav pieņemami. Ir jāatzīmē, ka šī problēma promocijas darba piedāvātā risinājumā ir atrisināta ar datu objekta jēdziena ieviešanu.
7) *DQS* ir iespējams analizēt lielus datu apjomus, taču, pieaugot ierakstu skaita apjomam, būtiski pieaug arī resursu patēriņš. Piemēram, 1 000 000 ierakstu lielas tabulas apstrāde *DQS* prasa vismaz 60 minūtes, 40-50 minūtes no kurām aizņem datu kvalitātes analīzes process, savukārt 20 minūtes - datu ielāde no datu avota. Rezultējošo datu eksports prasa 7 – 10 minūtes. Papildus, brīžiem rodas nepieciešamība atkārtot procesu, jo *DQS* paziņo, ka nespēj ielādēt un apstrādāt tik lielus datu apjomus. Ir jāatzīmē, ka šī procesa izpildei *DQS* ir nepieciešami visi datora resursus, t.i. *CPU*, diska un atmiņa noslodze palielinās līdz 100%. No tā

seko, ka *DQS* nav rekomendējams lielu datu apjomu analīzei uz parastajām mašīnām. Ir jāatzīmē, ka šī problēma ir raksturīga arī citiem risinājumiem, piemēram *Talend Data Quality* rīkam, it īpaši, ja datu kvalitātes analīzes ir tikai viena no rīkā nodrošinātājām iespējām.

DQS risinājums tiek pozicionēts kā *Microsoft SQL Server* glabājamo datu kvalitātes analīzes un uzlabošanas rīks, kas sniedz daudzas iespējas datu kvalitātes analīzei un uzlabošanai, taču tam ir raksturīgi vairāki ierobežojumi, kā arī pieejamās iespējas neļauj veikt padziļinātu un visaptverošu datu kvalitātes analīzi. *DQS* var tikt uzskatīts par pirmo soli datu kvalitātes analīzē, ko nosaka gan tehnoloģiski ierobežojumi, piemēram, ar datu tipiem saistītie ierobežojumi, nespēja veikt kontekstuālo datu kvalitātes analīzi vai kodifikatoru pārbaudi, gan vispārēji principiālu jautājumu risinājumi, kas atbilst arī (Leonard et al., 2014) un (McGilvray, 2013).

2.5. Apkopojums

Datu kvalitātes dimensiju definēšana un metodes to kvantitatīvai novērtēšanai ir viens no svarīgākajiem soļiem, kas līdz šīm tika veikts datu kvalitātes jomā (Nikiforova, 2018b). Tulkojot (Bicevskis et al., 2018a) rakstīto, "... mūsdienās esošie datu kvalitātes analīzes risinājumi lielākoties ir vērsti uz datu kvalitātes neformālo definēšanu un iegūto vērtību mērīšanu, taču mehānismi datu kvalitātes īpašību noteikšanai formalizētājās valodās nav zināmas" (vai pietiekoši populāras). Tāpat nav zināmi risinājumi, kas ļautu lietotājiem vienkārši pārbaudīt konkrētu datu kopu kvalitāti, definējot konkrētas datu kvalitātes prasības atsevišķiem lietotājus interesējošiem parametriem (Nikiforova, 2018b, 2019b).

Apkopojot iepriekšējās apakšnodaļās teikto, ir jāuzsver, ka "datu kvalitāte" ir sarežģīts jēdziens, kas ir atkarīgs no konkrētā datu lietojuma (Nikiforova, 2018a). To apstiprina arī pētījuma ietvaros veiktā datorzinātņu studentu aptauja, kurā piedalījās 55 respondenti. 55 respondentu liela grupa var tikt uzskatīta par reprezentatīvo respondentu kopu, jo tradicionāli tiek uzskatīts, ka aptaujās un eksperimentos ir jāpiedalās vismaz 30 respondentiem (Roscoe, 1975). Pētījuma ietvaros darba autore prasīja respondentiem nodefinēt "datu kvalitātes" jēdzienu. Rezultātā netika konstatētās vairākas pilnībā vienādas definīcijas, jo pat pie līdzīgas jēdziena vispārīgās definīcijas, tās tika papildinātas ar dažādām raksturīpašībām, kas pēc respondentu viedokļa vislabāk raksturo datu kvalitāti. Attiecībā uz datu kvalitātes dimensiju dažādību un daudzveidību, ir jāatzīmē, ka 55 respondentiem prasot nosaukt trīs svarīgākās datu kvalitātes dimensijas, autore konstatēja tikai 9 pēc nosaukuma vienādas dimensijas –

“pareizība”, “atbilstība”, “integritāte”, “dublēšanās”, “precizitāte”, “pieejamība”, “viennozīmība”, “uzticamība”, “pilnība/ pilnīgums”. Ir jāatzīmē arī tas, ka 10% respondentu ir atzīmējuši, ka nevar nosaukt nevienu datu kvalitātes dimensiju, jo par “dimensijas” jēdzienu datu kvalitātes kontekstā nav dzirdējuši, savukārt 12% respondentu atbildes liecina par to, ka arī viņi nav iepriekš saskārušies ar datu kvalitātes dimensijas jēdzienu. 78% respondentu starpā visbiežāk nosauktā dimensija ir “pilnīgums” - to ir nosaukuši 5 respondenti, “precizitāte” – 4 respondenti, savukārt parējās dimensijas ir nosaukuši no 2 līdz 3 respondentiem. Taču ir jāatzīmē, ka neskatoties uz nosaukto dimensiju vienādu nosaukumu, nevar apgalvot, ka respondenti tās izprot vienādi. Tad tika veikta ne-IT lietotāju aptauja ar mērķi identificēt, vai viņiem ir zināms “datu kvalitātes dimensijas” jēdziens, kā, viņuprāt, tas varētu tikt definēts, un kādas varētu būt datu kvalitātes dimensijas, tika konstatēts: (1) 96.4% nav iepriekš dzirdējuši “datu kvalitātes dimensijas” jēdzienu; (2) tikai 7.3% respondentu pieņēmums par to kā varētu tikt definēts “datu kvalitātes” jēdziens ir korekts un 17.1% definīcijas ir daļēji korektas; (3) 16.4% spēja nosaukt vismaz vienu eksistējošo datu kvalitātes dimensiju. Kopumā ir jāatzīmē, ka pat bakalaura studiju programmas “Datorzinātnes” 3. kursa studentu zināšanas par datu kvalitāti un it īpaši ar datu kvalitātes dimensijām saistītājos jautājumos ir ļoti ierobežotas, kas ļauj vēlreiz pārliecināties par “datu kvalitātes” jēdziena sasaisti ar “datu kvalitātes dimensijas” jēdzienu nepiemērotību galalietotājiem bez padziļinātām zināšanām datu kvalitātes jomā.

Vairākums eksistējošo risinājumu nav piemērots ne-IT un ne-datu kvalitātes speciālistiem, prasot no lietotājiem specifiskās zināšanas ne tikai IT jomā, bet arī datu kvalitātes jomā, it īpaši, ja tiek pielietots viens no ((Caro et al., 2007), (Ferney et al., 2017), (Neumaier, 2015), (Umbrich et al., 2015), (Vetro et al., 2016)) iepriekšējās nodaļās aprakstītājiem risinājumiem. Šie risinājumi ir paredzēti lietotājiem ar atbilstošām zināšanām, prasmēm un pieredzi datu kvalitātes jomā vai arī iesaistot tādus cilvēkus visos datu kvalitātes analīzes posmos, jo (a) izmanto lielu dimensiju skaitu, kas būtiski un bieži vien pārlieku apgrūtina uzdevumu, it īpaši praktiķu gadījumā, (b) prasa datu kvalitātes prasību definēšanu, (c) prasa attiecināt definētas vai jau eksistējošas datu kvalitātes prasības uz atbilstošām dimensijām, kas vēlāk tiek pielietotas datu kopām (Nikiforova, 2019b). Speciālistu iesaiste visos datu kvalitātes analīzes posmos nav pieņemama, jo neatbilst datu kvalitātes specifikai – datu atbilstībai lietošanas piemēram, kuram ir jābūt definētam tikai un vienīgi ar pašu galalietotāju, kas analīzē datu kopas kvalitāti savos nolūkos. Viņu iesaiste ir pieļaujama tikai atsevišķos datu kvalitātes analīze vēlākājos posmos, savukārt gan datu, kuriem ir jābūt analizētiem, gan kvalitātes prasībām, atbilstoši kurām būtu jāpārbauda datu kvalitāte, ir jābūt pašu galalietotāju definētiem. IT-speciālisti drīkst tikai atbalstīt datu kvalitātes analīzi, sniedzot nepieciešamo palīdzību jautājumos, kuros ir nepieciešamas atbilstošās zināšanas un prasmes.

Datu kvalitātes dimensiju izmantošana datu kvalitātes analīzē ir problemātiska, jo, neskatoties uz datu kvalitātes analīzes pētījumu, kuros datu kvalitāte tiek saistīta ar datu kvalitātes dimensijām, vecumu, vēl joprojām nav skaidrs, kā un kuru konkrētu datu kvalitātes dimensiju piesaistīt noteiktam lietošanas piemēram. Uz šo esošo risinājumu trūkumu norāda arī vairāki datu pētnieki, t.sk. *Batini* - datu kvalitātes jautājumos pasaules vadošais pētnieks, datu kvalitātes problēmas dziļa izpētes darba un eksistējošo metodoloģiju pārskata autors (Batini et al., 2009, 2016). To apstiprina arī pētījuma ietvaros veiktā datorzinātņu studentu aptauja (kopā 55 respondenti), konstatējot, ka 64,4% respondentu uzskata, ka datu kvalitātes prasībām ir jābūt galalietotāja vai potenciālo lietotāju definētām, savukārt 35,6% pieļauj, ka vismaz daļa no tām varētu būt iepriekšdefinēta ar risinājuma izstrādātājiem vai datu kvalitātes speciālistiem. Datu kvalitātes dimensiju definēšanas gadījumā vairāk kā 92% respondentu uzskata, ka dimensijām ir jābūt risinājumu izstrādātāju (40,4%) vai datu kvalitātes speciālistu (51,9%) definētām, jo galalietotājam, kuram var nebūt datu kvalitātes zināšanas, datu kvalitātes dimensiju definēšanas uzdevums ir pārāk sarežģīts. Daži risinājumi prasa specifisko metodoloģiju pārzināšanu vai arī padziļinātas zināšanas citās jomās, piemēram, saistīto datu, zināšanu inženierijas jomās.

Atbilstoši Tartu Universitātes prof. *Marlon Dumas* ieteikumam, datu kvalitātes analīzes lietotāji, kuriem piemīt specifiskās šīm uzdevumam nepieciešamas zināšanas, var tikt iedalīti divās grupās: (1) IT-speciālisti un (2) datu kvalitātes speciālisti, pieļaujot arī abu kopu pārklāšanās. IT un datu kvalitātes speciālistu nošķirums ir nepieciešams, jo (1) IT-speciālistam var nebūt datu kvalitātes zināšanas, t.i. zināšanas, kas ir nepieciešamas datu kvalitātes analīzei, (2) lietotājs var tikt uzskatīts par datu kvalitātes speciālistu, ja viņa kvalifikācija un/ vai pieredze ir pietiekoša, lai veiktu datu kvalitātes analīzi, savukārt šīs zināšanas var būt lietotājam, kuram nav padziļinātu zināšanu IT jomā. 2. grupai piederošie cilvēki parasti ir bankās vai citās sfērās strādājošie datu analītiķi, kuru zināšanas līmenis datu kvalitātes jomā ir pietiekošas datu kvalitātes analīzes veikšanai, neskatoties uz IT izglītības trūkumu. Tas nozīmē, ka lietotājs var tikt uzskatīts par datu kvalitātes speciālistu, ja tam piemīt padziļinātas ar datu kvalitāti saistīto jēdzienu zināšanas, un tā ir spējīga veikt iepriekšminētos uzdevumus, kas attiecas uz datu kvalitātes analīzi (Nikiforova, 2018a, 2019b). Atbilstoši (Nikiforova, 2018a), ja datu kvalitātes analīzes risinājums prasa padziļinātas zināšanas IT jomā (piemēram, (Acosta et al., 2013), (Färber et al., 2018), (Gandhi, 2016), (Kontokostas et al., 2014), (Redman, 2001), (Zaveri et al., 2016)), ir spēkā IT-speciālista jēdziens. Lietotājs tiek uzskatīts par IT-speciālistu, ja viņam ir izglītība un/ vai pieredze IT jomā, kas pārklāj atbilstošas tēmas (t.i. specifiskās tehnoloģijas, pieejas, zināšanu inženierija utt.).

Ir jāatzīmē, ka neskatoties uz eksistējošo risinājumu datu kvalitātes analīzei, tajā iesaistīto pušu, nepieciešamu zināšanu vai tehnoloģiju dažādību, kopīga ar datu kvalitāti saistīto pētījumu

pazīme ir datu kvalitātes problēmu konstatēšana datos, kuriem piedāvātie risinājumi tiek pielietoti. Tas nozīmē, ka datu kvalitātes problēma vēl joprojām nav atrisināta, un ir (a) jāturpina pētījumu veikšana šajā jomā, piedāvājot jaunus risinājumus, (b) jāpārbauda eksistējošas [atvērto] datu kopas, sniedzot informāciju par tajās konstatētajām nepilnībām, tādejādi uzlabojot to kvalitāti. Šīm nolūkam tika izstrādāta jauna pieeja datu kvalitātes analīzei, kas ir apskatīta nākamajā nodaļā.

3. PIEDĀVĀTAIS DATU KVALITĀTES MODELIS

Nodaļā "Piedāvātais datu kvalitātes modelis" ir nodefinētas prasības datu objekta virzītai pieejai datu kvalitātes analīzei, aprakstīts piedāvātais datu kvalitātes modelis, sniedzot pārskatu par katru tā komponentu. Ir aprakstīta kontekstuālās datu kvalitātes analīzes iespēja. Ir novērtētas piedāvātās pieejas priekšrocības. Ir sniegts pieejas vispārīgs apraksts, tās apraksts modeļa virzītas izstrādes (*MDA*) kontekstā, pamatojot komponenšu izvēli un to apvienošanu kopējā risinājumā. Nodaļas nobeigumā ir sniegts piedāvātās pieejas iespēju un ierobežojumu uzskaitījums.

Šī nodaļa pamatā balstās uz (Nikiforova, 2018a, 2018b, 2019b) un (Nikiforova et al., 2019, 2020).

3.1. Pieejas vispārīgs apraksts

Ņemot vērā datu kvalitātes relatīvo un dinamisko raksturu, atbilstoši kuram prasības datu kvalitātei nosaka datu lietojums, katram konkrētam lietojumam var būt nepieciešamas specifiskas datu kvalitātes pārbaudes. Tas atbilst arī (Batini et al., 2009), kurā eksistējošo datu kvalitātes risinājumu analīzes rezultātā *Batini* ar līdzautoriem secina, ka datu kvalitātes risinājuma pamatā ir trīs aspekti: (1) datu un procesu analīze, (2) datu kvalitātes prasību analīze un (3) datu kvalitātes analīze. Datu un procesa analīze autoru izpratnē ir datu shēmas izpēte, kas paredz lietotāju aptauju, lai panāktu vienošanās ar datu lietotājiem par datiem, uz tiem attiecināmajiem ierobežojumiem un likumiem, procesiem, kas izmanto datus, vai kuru rezultātā tiek iegūti jauni dati. Datu kvalitātes prasību analīze bieži vien ietver sevī datu lietotāju un administratoru aptaujas, kuru mērķi ir identificēt kvalitātes problēmas, kuru rezultātā kļūst iespējams noteikt "kritiskas" datu kopas, nodefinēt datu kvalitātes metrikas un kvalitātes mērķus. Datu kvalitātes analīze ir aktivitāšu, kas ir saistītas ar datu kopas izpētes, novērtēšanu un profilēšanu pret definētajam datu kvalitātes metrikām, kopums. Ņemot vērā arī to faktu, ka dati parasti tiek uzkrāti pakāpeniski, tiek izvirzītas sekojošās pamatprasības datu kvalitātes vadības sistēmai:

- ņemot vērā datu kvalitātes jēdziena atkarību no lietošanas piemēra jeb datu lietojuma, datu kvalitātes prasības ir jāformulē platformas neatkarīgos jēdzienos, t.i. neiekļaujot pārbaudes informācijas sistēmas programmas kodā;

- datu kvalitātes prasības ir jāformulē vairākos līmeņos, t.i. atsevišķam datu objektam, datu objektam tā atribūtu kontekstā, datu objektam datubāzes kontekstā, datu objektam daudzu informācijas sistēmu kontekstā;
- datu kvalitātes modeļa, t.sk. datu kvalitātes prasību, komponenšu definēšanas valodai ir jābūt pietiekoši vienkāršai, nodrošinot iespēju definēt datu objektus un tām izvirzāmās datu kvalitātes prasības arī nozaru speciālistiem ar minimālu IT speciālistu piesaisti. Atbilstoši (Zhao et al., 2003) tas var tikt panākts, datu kvalitātes modeļa komponenšu izveidei izmantojot grafisko *DSL*, kuras sintaksi un semantiku ir viegli piemērot katrai jaunai informācijas sistēmai;
- datu kvalitātei ir jābūt pārbaudītai vairākos datu apstrādes posmos, katru reizi izmantojot savu individuālu datu kvalitātes prasību aprakstu. Ir ieteicams necensties iekļaut datu kvalitātes prasības vienā visaptverošā prasību specifikācijā, ar kuru datu kvalitāte tiktu pārbaudīta tikai datu uzkrāšanas gala posmā, t.i. tad, kad dati jau uzkrāti datubāzē.

Attiecībā uz iepriekšapskatīto *Batini* un viņa līdzautoru (Batini et al., 2009) skatījumu, piedāvātā pieeja paredz lietotāja maksimālo iesaisti nevis tā aptaujāšanu, kur galalietotājs ir tikai viens no datu kvalitātes analīzē iesaistītājām personām, ļaujot lietotājam pašam definēt ar datu kvalitātes analīzi saistītas detaļas.

Atbilstoši (Nikiforova, 2019b) no *TDQM* viedokļa, atbilstoši kuram datu kvalitātes dzīvescikls sastāv no 4 savstarpēji saistītām fāzēm (datu kvalitātes definēšana, datu kvalitātes mērīšana, datu kvalitātes analīze un datu kvalitātes uzlabošana), piedāvātā pieeja var tikt īsumā aprakstīta sekojoši:

- 1. fāze – **datu kvalitātes definēšana** notiek (1) datu objekta izvēle, kuru kvalitāte tiek analizēta, (2) datu kvalitātes prasību definēšana datu objektu klasei. Tā ir izsakāma ar nosacījumu kopumu, kuru izpilde tiek pārbaudīta. Datu kvalitātes prasības tiek fiksētas ar grafisku diagrammu palīdzību, kur grafa virsotnes attēlo izpildāmās operācijas un pārbaudes, bet loki - to izpildīšanas secību. Datu kvalitātes prasības var tikt formulētas dažādos abstrakcijas līmeņos, sākot ar neformālu tekstu (piemēram, dabiskajā valodā) un beidzot ar precīziem, automātiski izpildāmiem programmu moduļiem, aizvietojot neformālos tekstus ar izpildāmu programmkodu vai *SQL* vaicājumiem.

  Attiecībā uz datu nolasīšanu no datu avotiem, ir jāatzīmē, ka "datu objekta" jēdziena izmantošana paredz, ka tiek atlasīti tikai tie dati, kas ir nepieciešami konkrētai analīzei, tādā veidā samazinot apstrādājamo datu apjomus, taupot laiku

un citus resursus (to mēdz saistīt ar denormalizāciju). Dati var tikt atlasīti, izmantojot programmēšanā labi zināmos un plaši izmantotos *SQL* vaicājumus (precīzāk, *SELECT* operatoru), taču darbiniekiem, kuriem nepiemīt padziļinātas datubāzes zināšanas, tie var likties pārāk sarežģīti un mazsaprotami, līdz ar ko praksē bieži tiek izmantotas individuālas datu atlasīšanas un transformācijas operācijas, līdzīgi kā to piedāvā *Microsoft SQL Server 2017* komponents *SQL Server Integration Services* ((Laudenschlager et al., 2012a), (Laudenschlager et al., 2012b));

- 2. fāzē - **datu kvalitātes mērīšanas** fāzē (1) analizējamie dati tiek atlasīti no datu avotiem (ekrānformu laukiem, datubāzēm, failiem, datu noliktavām u.c.) un (2) tiek veikti datu kvalitātes mērījumi, katram datu objektam pārbaudot iepriekš formulēto prasību izpildi. Vienā kvalitātes mērīšanas procesā var iekļaut vairāku datu objektu nolasīšanu un prasību pārbaudi ar vairākām kvalitātes specifikācijām. Kvalitātes novērtēšanas procesa izpildes rezultātā var tikt sagatavots protokols, kas satur pārbaudes procesā atrastās neatbilstības kvalitātes specifikācijai;
- 3. fāzē - **datu kvalitātes analīzes fāzē** tiek veikta mērīšanas fāzē saņemto datu kvalitātes pārbaudes rezultātu analīze. Tās mērķis ir datu kvalitātes problēmu konstatēšana un šo problēmu galveno cēloņu noskaidrošana;
- 4. fāzē - **datu kvalitātes uzlabošana** fāzē tiek veikta kvalitātes uzlabošanas mehānisma izvēle un tā realizācija. To var veikt gan ar individuāli izstrādātiem programmu moduļiem, gan, izmantojot datu kvalitātes analīzes un uzlabošanas rīka *DQS* piedāvātus līdzekļus, kura priekšrocības, trūkumus un potenciālo pielietojumu autore apskatīja (Nikiforova, 2018a) un (Bicevskis et al., 2018a), līdz ar ko piedāvātais risinājums šo fāzi neapskata.

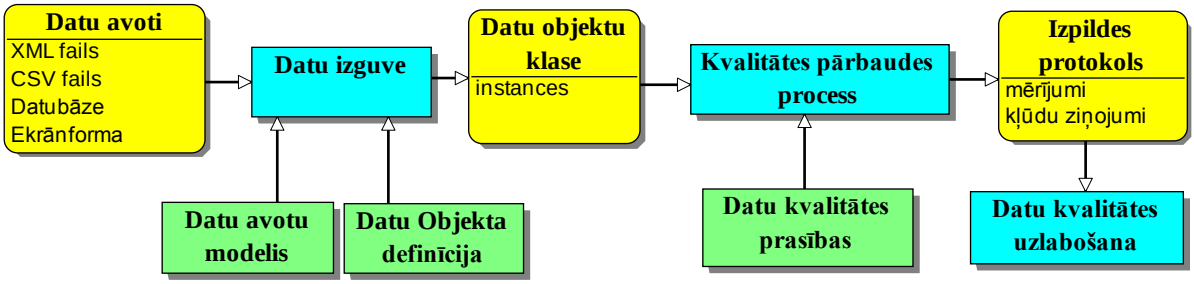
Tā kā atbilstoši *TDQM* datu kvalitāte var tikt nodrošināta, sistemātiski atkārtojot datu kvalitātes cikla fāzes, kas ir nepieciešams, jo dati nepārtraukti mainās, kā arī jauni vai modificētie dati var izraisīt jaunas datu kvalitātes problēmas vai datu kvalitātes prasību maiņu, piedāvātā pieeja nodrošina datu kvalitātes novērtēšanas kritēriju maiņu un jaunu datu kvalitātes prasību definēšanu katrā jaunā ciklā iterācijā, tādējādi nodrošinot datu augstu kvalitāti.

Datu kvalitātes vadības sistēmas arhitektūra ir attēlota 3.1.1. attēlā. Piedāvātās pieejas un atbilstoši piedāvātā datu kvalitātes modeļa galvenie komponenti ir:

1) datu objekts, kas definē tos datus, kuru kvalitāte tiks analizēta;

2) datu kvalitātes prasības – nosacījumi, kuriem jāizpildās, lai dati tiktu atzīti par kvalitatīviem;
3) datu kvalitātes pārbaudes process – visas aktivitātes, kas ir jāveic, lai novērtētu datu objekta kvalitāti.

Visi trīs datu kvalitātes modeļa komponenti tiek aprakstīti ar valodu metamodeļiem. Papildus sintaksei tie uzdod arī modeļa grafisku reprezentāciju, savukārt modeļa semantika tiek aprakstīta, uzdodot grafisko diagrammu izpildes likumus. Šī nostādne atbilst arī (Zhao et al., 2003). Tādā veidā datu kvalitātes modeļi kļūst izpildāmi un praktiski lietojami.



3.1.1. att. **Datu kvalitātes vadības arhitektūra [pēc (Bicevskis et al., 2019a) izveidoja autore]**

Saskaņā ar izplatīto datu kvalitātes iedalījumu sintaktiskajā un semantiskajā precizitātē, un daudzām problēmām, kas izriet no datu kvalitātes dimensiju jēdziena lietojuma datu kvalitātes risinājumos, piedāvātā datu objektu virzītā pieeja atsakās no datu kvalitātes dimensijas jēdziena, aizstājot to ar universālāku jēdzienu "datu kvalitātes prasība" – datu kvalitātes jēdziens netiek saistīts ar datu kvalitātes dimensijas jēdzienu. Kā seko no 2. nodaļas, datu kvalitātes dimensijas jēdziena izpratne, tāpat kā konkrētu dimensiju definēšana ar pieejas izstrādātājiem un izpēte ar konkrētas pieejas lietotājiem, prasa nepamatoti daudz resursu, ko vairākkārt ir atzinuši pat spilgtākie datu kvalitātes tēmas pētnieki (*Batini, Scannapieco, Price, Shanks* utt.). Pie tam kā atbilstoši 2. nodaļā minētajam, daži pētījumi izmanto pārāk lielu datu kvalitātes dimensiju skaitu, savukārt citi – ierobežo to skaitu līdz pat divām dimensijām. Piedāvātā pieeja neuzstāda tādus ierobežojumus, sniedzot lietotājiem iespēju definēt datu kvalitātes prasības, kas ir atkarīgas no lietošanas piemēra. Atbilstoši (Nikiforova, 2019b), "datu kvalitātes prasības" jēdziena izmantošana "datu kvalitātes dimensiju" jēdziena vietā dod ievērojamus ieguvumus, jo, pirmkārt, tas sniedz iespēju datu kvalitātes analīzes procesā iesaistīties lietotājiem, kuriem var nepiemīt padziļinātās zināšanas IT un datu kvalitātes jomās, otrkārt, sekmē vairāku lietotāju savstarpējo sadarbību, treškārt, neierobežo izvirzāmo kvalitātes prasību raksturu (salīdzinājumā ar datu kvalitātes dimensijām, kur katrai konkrētai dimensijai atkarībā no tās specifikas un risinājuma implementācijas var būt nodefinēts ierobežots

iespējamo datu kvalitātes pārbaužu saraksts). Papildus, tas ļauj ietaupīt pieeju datu kvalitātes analīzei izstrādātāju un lietotāju laiku, atvieglojot gan to izstrādes, gan izmantošanas procesu, neprasot vairāku ar datu kvalitātes dimensijām saistītu resursietilpīgu darbību veikšanu, t.i. no izstrādātāju puses – datu kvalitātes dimensiju definēšanu ar visām no tā izrietošām sekām, t.sk. to grupēšana, atribūtu, metriku izvēle, mērīšanas mehānismu izvēle un nodrošināšana utt., no lietotāja puses – visu risinājumā esošo dimensiju un to komponentu apgūšana, kas parasti ir noderīgs tikai viena konkrēta risinājuma ietvaros. Taču, neskatoties uz atteikšanos no datu kvalitātes dimensijas jēdziena, aizstājot to ar universālāko, piedāvātā pieeja ievēro vispārpieņemtās datu kvalitātes jēdziena definīcijas.

Attiecībā uz datu kvalitātes nodrošināšanas apgabaliem, t.i. atsevišķam datu objektam, datu objektam savu atribūtu kontekstā un datu objektam datubāzes kontekstā, kas attiecas uz dažādiem IS komponentiem, piedāvātā risinājuma stratēģija ļauj veidot vienotu risinājumu. Tas tiek panākts, piedāvājot *DSL* platformu, kas nodrošina plašu datu kvalitātes prasību valodu definēšanas spektru, ļaujot definēt datu kvalitātes prasības modulāri, un pārbaudot prasību izpildi dažādos datu apstrādes posmos, t.i. neiekļaujot datu kvalitātes prasības vienā visaptverošā prasību specifikācijā, datu kvalitāti pārbaudot tikai datu uzkrāšanas gala posmā. Tādejādi tiek nodrošināta iespēja datu kvalitāti pārbaudīt daudzos datu apstrādes posmos, katru reizi izmantojot savu individuālu datu kvalitātes prasību aprakstu.

Cita piedāvātās datu objekta virzītas pieejas ideja ir datu lietotāju iesaiste datu kvalitātes analīzē. Tā ir svarīga ne tikai pašiem lietotājiem, kuriem ir jābūt iespējai analizēt datu kvalitāti saviem nolūkiem, bet arī datu sniedzējiem, jo atbilstoši (Tinholt, 2013), ņemot vērā pieaugošo publicēto atvērto datu apjomu (datu kopu skaitu), un lielu iespējamo lietošanas piemēru skaitu, ko paši datu publicētāji nav spējīgi izvirzīt un pārbaudīt, lietotāju iesaiste datu kvalitātes analīze un to atgriezeniskā saite būtiski paaugstina potenciālo datu kvalitāti, un paaugstina varbūtību, ka pēc iespējas lielāks datu kvalitātes defektu skaits tiks noteikts un rezultātā arī novērsts, tāda veidā būtiski paaugstinot kopējo datu kvalitāti ((Ruijer et al., 2019), (Attard et al., 2015)). Piedāvātais risinājums ļauj vērtēt datu kvalitāti atbilstoši objektīvām un subjektīvām prasībām. Ar objektīvām prasībām tiek saprastas no lietotāja neatkarīgas prasības, kas ļauj novērtēt datu atbilstību iepriekšdefinētājām prasībām, integritātes likumiem vai ārējiem avotiem (Price et al., 2004, 2005). Savukārt ar subjektīvām prasībām tiek saprastas no datu lietojuma atkarīgas prasības, kas mēdz mainīties atkarībā no uzdevuma, kuram ir nepieciešami dati konkrētajā brīdī. Tās ir lielā mērā atkarīgas no datu lietotāja "datu kvalitātes" jēdziena uztveres. Objektīvās prasības ir universālākas, t.i. tās prasības, kuru izpilde ir jāpārbauda neatkarīgi no datu lietojuma, piemēram, primāro datu pilnīgums. Ir jāatzīmē, ka atbilstoši (Jayawardene et al., 2015) dotā pieeja apskata datu kvalitāti gan no (a) deklaratīvas perspektīvas (D), kas fokusējas

uz no lietotāja neatkarīgajām datu raksturīpašībām, kas raksturo datus pašus par sevi, t.i., izmantojot metadatus, shēmas standartus un darījumlikumus, kas izriet no datu darbības organizācijas, gan no (b) lietojamības perspektīvas (U), kas fokusējas uz no lietotāja atkarīgajam datu raksturīpašībām, kas attiecas uz datu izveides un lietošanas efektivitāti, kas ir atkarīga no lietotāja definētās datu kvalitātes jēdziena definīcijas. Lielāks uzsvars tiek likts uz lietojamības (U) perspektīvu, ko citi risinājumi bieži vien ignorē.

Piedāvātais datu kvalitātes modelis var tikt formulēts un izmantots divos veidos/ līmeņos: (a) neformāli (līdzīgi *PIM*), kur nepieciešamas pārbaužu darbības tiek aprakstītas dabiskā valodā – diagrammu simboli satur darbību tekstuālus aprakstus; (b) izpildāmā veidā (līdzīgi *PSM*), kas var tikt panākts, pārveidojot neformālo modeli, neformālus tekstus/ aprakstus aizstājot ar programmkodu, *SQL* vaicājumiem vai cita veida izpildāmiem objektiem. *PIM* var tikt uztverts kā informācijas sistēmas neformālais apraksts, kas tiek veidots ar industrijas speciālistiem, t.i. lietotājiem, kuriem var nebūt padziļināto IT zināšanu. Tāda veida tiek nodrošināta modeļa pakāpeniskā detalizācija.

Tas nozīmē, ka piedāvātais datu kvalitātes modelis var tikt aprakstīts no *MDA* skatupunkta. Iespējams, tas neatbilst *MDA* tā tradicionālajā izpratnē, taču piedāvātās pieejas pamatprincipi atbilst tai. Atbilstoši (Kleppe et al., 2003), *MDA* "… balstās uz plaši izplatītājiem industrijā [programmatūras projektējumu un] modeļu vizuālizācijas, glabāšanas un apmaiņas standartiem". Šī ideja ir piedāvātās pieejas pamatā.

Tāpat kā *Object Management Group* (*OMG*) gadījumā (Soley, 2000), piedāvātājā risinājumā "modeļi ir palīglīdzekļi nevis "izdevumi"", kā arī viens no galvenajiem diagrammu izmantošanas mērķiem ir "modelēšanas tehnoloģijas pielietošana pilnas ainas apzināšanai".

Atbilstoši šī nostādnei visi piedāvātās pieejas komponenti ir definēti, izmantojot grafiskās domēnspecifiskās valodas (*DSL*). Izstrādāto *DSL* sintakse un semantika ir (a) vienkārši pielietojama jaunajām informācijas sistēmām, (b) pietiekoši vienkārša, lai datu objektu un kvalitātes specifikāciju definēšana varētu tikt veikta ar lietotājiem bez IT zināšanām, t.i. ar minimālo IT-ekspertu iesaisti. Priekšroka grafisko modeļu lietošanai tika dota arī tādēļ, ka modeļi tiek uzskatīti par komunikācijas līdzekli (Mellor et al., 2004), kas uzlabo informācijas lasāmību, jo grafiski attēlota informācija parasti tiek uztverta labāk nekā tekstuālā – vizuālā informācijā ir vienkāršāk un ātrāk lasāma un maināma. Modeļu lietošana samazina lietotāju savstarpējas nesaprašanas risku, kā arī atbilstoši (Mellor et al., 2004) modeļu izveide ir lētāka nekā reāla produkta izveide. *Mellor* uzsver divus pamataspektus, kuri ir jāievēro modeļu izveidei: abstrakcija un klasifikācija. Ar abstrakciju *Mellor* saprot visas tās informācijas ignorēšanu, kura ir ārpus konteksta konkrētā situācijā. Piedāvātās pieejas gadījumā šīs aspekts ir ievērots, definējot datu objektu, kurš satur tos un tikai tos parametrus, kas raksturo reālo

objektu, kuru kvalitātes interesē konkrētu lietotāju konkrētā lietošanas piemēra gadījumā. Parametri, kuri nav nepieciešami un nav svarīgi konkrētajam lietotājam, var tikt ignorēti – nav iekļauti konkrētajā datu objektā. Ar klasifikāciju *Mellor* saprot elementu grupēšanu, balstoties uz to kopīgām pazīmēm. Šīs princips tiek daļēji ievērots, grupējot katram parametram definētās datu kvalitātes prasības. Atbilstoši piedāvātajam risinājumam, ir nodrošināta arī veidojamo modeļu pievienotā vērtība – tie ir mašīnlasāmi nevis papīra formāta (Kleppe et al. 2003), nodrošinot arī to glabāšanu standartizētājos repozitorijos, lai samazinātu liekas un visbiežāk resursietilpīgas darbības. Piedāvātās pieejas ietvaros, šīm nolūkam tiek izmantots grafiskais *DSL* redaktors *DIMOD*. Tāpat kā (Kleppe et al., 2003) risinājumā modeļi ir ātri un vienkārši veidojami, rediģējami un atkalizmantojami. Tie var tikt vairākkārtēji mainīti atkarībā no lietotāja vajadzībām, katru reizi mainoties lietošanas piemēram vai parādoties jaunām detaļām, izmaiņas var tikt neaizkavējoties veiktas arī atbilstošajos modeļos. Jebkura rakstura modeļu izmaiņas var tikt veiktas ar dažādiem lietotājiem, jo datu kvalitātes analīzes procesā, t.i. modeļu izveidē, var tikt iesaistīti vairāki lietotāji. Papildus, iesaistītie lietotāji var pārstāvēt dažādas strukturālās vienības vai arī nevienu no tām. Veidotās diagrammas demonstrē katru datu kvalitātes posmu nevis galarezultātu, kas arī atbilst (Kleppe et al., 2003). Ir jāatzīmē, ka līdzīga nostādne, t.i. grafisko modeļu izmantošana datu kvalitātes analīzes uzdevumos, ir raksturīga arī vienas no vadošas datu kvalitātes pētnieces *Scannapieco* pētījumam (Scannapieco et al., 2002b), kurā autori norada uz datu kvalitātes uzlabošanas fāzei, kas ir dotā pētījuma centrālais objekts, domāto modelēšanas valodu trūkumu. Autori norāda, ka datu kvalitātes uzlabošanas uzdevumiem piemērotai modelēšanas valodai ir jābūt pietiekoši formālai, lai nodrošinātu valodas konstrukciju unikālu un viennozīmīgu interpretāciju. Šajā pētījumā autori norāda uz nepieciešamību pēc valodas, kura būtu pietiekoši vienkārša, lai to varētu izmantot arī lietotāji bez padziļinātam zināšanām IT jomā, uzsverot, ka mijiedarbība ar galalietotāju ir primārais uzdevums datu kvalitātes analīzes jautājumā. Atšķirībā no autoriem, kuri priekšroku ir devuši *Unified Modelling Language* (*UML*), kas atbilst arī *OMG* redzējumam, piedāvātajā risinājumā priekšroka ir dota blokshēmām līdzīgām grafiskām *DSL*.

UML diagrammas ir *MDA* visbiežāk izmantotais līdzeklis (Kleppe et al., 2003), ko savā risinājumā izmanto arī *Scannapieco* (Scannapieco et al., 2002b), taču, ņemot vērā, ka *UML* prasa lietotājiem specifiskās zināšanas un iepriekšējo pieredzi, tā ir viena no labākām izvēlēm inženieriem, jo ļauj dokumentēt idejas un apmainīties ar tām (atbilst (Kleppe et al. 2003)), taču lietotājiem bez padziļinātājām zināšanām IT un datu kvalitātes jomā, kuriem ir jābūt piemērotai piedāvātai pieejai, *UML* nav paredzēts (atbilst (Sproģis, 2014) un citiem). Papildus tradicionāla *UML*, t.i. bez paplašinājumiem, mēdz tikt uzskatīta par pārāk virspusīgu un vispārīgu (Scannapieco et al., 2002b). *UML* trūkumu dēļ *MDA* (Haubold et al., 2010) *UML* tiek izmantots

kombinācijā ar *DSL*, kuru kombinācija pēc autoru viedokļa dod labākus rezultātus, novēršot katras tehnikas trūkumus, pie *DSL* trūkumiem attiecinot tikai to pielietojamības ierobežojumus. Savukārt *UML* kombinēšana ar *DSL* būtiski atvieglo metamodeļu izveidi (Haubold et al., 2010).

Tajā pašā laikā blokshēmu struktūra ir vienkārša, kā arī vismaz atsevišķas Latvijas vidusskolas tās ir iekļautas mācību programmā, līdz ar ko iespējamība, ka uz tām balstītas diagrammas būs labi saprotamas visiem cilvēkiem neatkarīgi no zināšanām un darba jomas, ir augstāka, salīdzinot ar alternatīvajam iespējām. Tas ļauj veikt pieņēmumu, ka šādas diagrammas bez grūtībām varēs veidot un rediģēt arī ne IT-speciālisti, kā arī tās veicinās vairāku datu kvalitātes analīzes procesā iesaistīto personu savstarpējo komunikāciju. Šo iemeslu dēļ priekšroka tika dota blokshēmām līdzīgām diagrammām.

Ņemot vērā visu iepriekšminēto, tika izveidots datu kvalitātes modelis, kas sastāv no grafiskajiem modeļiem, kur katra diagramma apraksta konkrētu datu kvalitātes pārbaudes posmu. Visas viena darījumprocesa pārbaudes tiek apvienotas pakotnēs, savukārt visas pakotnes kopā veido kvalitātes modeli. Katra diagramma sastāv no virsotnēm un lokiem, kur (a) virsotnes, kas tiek apzīmētas ar mnemoniskiem grafiskiem simboliem, attēlo elementārās datu kvalitātes vadības darbības, (b) loki savieno virsotnes, norādot uz veicamo darbību secību (Nikiforova, 2018a). Diagrammās ir iespējams iekļaut arī citas darbības, piemēram, kļūdu ziņojumu sagatavošanu, kas ir paredzēti datu kvalitātes problēmu reģistrēšanai, t.i. izpildes protokola izveidei, kurā tiek reģistrēti datu kvalitātes prasībām neatbilstošie dati. Iegūtie izpildes protokoli turpmāk tiek izmantoti datu labošanai. Diagrammu izmantošana ļauj lietotājiem definēt datu objektu un atbilstoši datus, kuru kvalitāte tiks analizēta, datu kvalitātes prasības, kuras būtu jāapmierina datiem, lai būtu iespējams veikt secinājumus par to kvalitātes līmeni un atbilstību noteiktam uzdevumam, neatkarīgi no savu zināšanu līmeņa. Prasību aprakstīšana šādā veidā izslēdz nepieciešamību aprakstīt prasības tekstuālā formā, kas savukārt var tikt dažādi interpretētas, tādejādi atvieglojot arī trešā datu kvalitātes analīzes posma realizāciju, t.i. datu kvalitātes procesu, jo ir izslēgta vai vismaz būtiski samazināta nesaprašanas starp datu galalietotāju un datu analītiķi iespējamība. Tas atbilst darba 5. tēzei.

Ņemot vērā programmēšanas valodu un platformu dažādību to semantikā, *PIM* transformācija *PSM* dotā risinājuma ietvaros tiek veikta manuāli. Neskatoties uz dažādiem risinājumiem *PIM* modeļa automātiskai vai pusautomātiskai transformācijai *PSM* modelī, nodrošināt lietotāja definētā *PIM* modeļa transformāciju konkrētajā *PSM* modelī ir nepamatoti resursietilpīgs process. Papildus, vairāki pētnieki apgalvo, ka *PIM* automātiskā transformācija *PSM* modelī nespēj radīt no lietojamības un uzticamības viedokļa kvalitatīvu produktu. Savukārt manuālā transformācija ir relatīvi vienkāršs uzdevums, īpaši lietotājiem ar programmēšanas pamatzināšanām, kuru iesaiste kļūst nepieciešama tikai kvalitātes analīzes

beidzamajos posmos. Pie šī secinājuma savos pētījumos ir nonākuši vairāki autori - (Lano, 2005), (Miller et al., 2003), (Ostadzadeh et al., 2008), (Pauker et al., 2016), (Carroll et al., 2006), (Chungoora et al., 2013) utt..

Nākošās apakšnodaļās ir veltītas piedāvātā kvalitātes modeļa komponentiem un kontekstuālās kvalitātes pārbaudes iespējai.

3.2. Datu objekts

Datu objekts ir viens no piedāvātas pieejas pamatjēdzieniem. Par datu objektu tiek uzskatīta konkrēta reālās pasaules objekta raksturojošo parametru vērtību kopa (Nikiforova, 2019b).

Piemēram, par datu objektu var tikt uzskatīta (1) universitāte un to raksturojošie parametri – nosaukums, reģistrācijas numurs, dibināšanas datums, interneta vietnes adrese, kontakttālrunis, fakultāšu saraksts un to nosaukums, adrese utt., (2) valsts un tās nosaukums, galvaspilsēta, valsts valoda, likumdevējs, platība, valūta, *ISO* kods, robežvalsts (robežvalsts vai vairāku robežvalstis saraksts) utt.. Viens no mūsdienās plaši sastopamiem datu objekta piemēriem ir infokastes *Wikipedia* un tajās esoša informācija par katru meklējamo vienību. Tāpat par konkrēta datu objekta reprezentācijas piemēru var kalpot dokuments ar aizpildītām lauku vērtībām, piemēram, anketa, rēķins utt.. Tos visus vieno īpašība, ka datu objekta parametru vērtības ir atklāti redzamas, vērtības nekodējot, kas savukārt ir raksturīgs datu glabāšanai datubāzēs. Ir jāatzīmē, ka datu objekta raksturs ļauj definēt to kā procesa izpildes rezultātu, piemēram, no navigatora iegūtais izbraukto maršrutu saraksts.

Datu objekts var tikt uzdots dažādos veidos, viens no kuriem ir datu objekta uzdošana ar atslēgas vārdiem, atbilstoši kuram datu objekts "Universitāte" var tikt uzdots sekojošajā veidā:

<nosaukums: Latvijas Universitāte, reģistrācijas numurs: 90000076669, dibināšanas datums: 1919, interneta vietne: www.lu.lv, kontakttālrunis: +371 67034444, fakultāte: nosaukums: Datorikas fakultāte, adrese: Raiņa bulvāris 19; nosaukums: Fizikas, matemātikas un optometrijas fakultāte, iela: Zeļļu iela 25; nosaukums: Juridiskā fakultāte, adrese: Raiņa bulvāris 19>.

Atbilstoši (Nikiforova, 2019b) dotais risinājums paredz datu kvalitātes pārbaudes aprakstīšanu ar grafiskām diagrammām, aizstājot tradicionālo ievadāmās informācijas pārbaudes implementāciju ar programmu palīdzību, pārbaudes iekodējot programmkodā, ar izpildāmām grafiskām diagrammām. Kvalitātes pārbaudes tradicionālajā implementācijā ir aizstātas ar universālu risinājumu, "atdalot" tās no programmakoda, jo prasības, kas nav atklāti

definētas, ir grūti pārvaldāmas un testējamas. Šī pieeja nodrošina "ārēja" datu kvalitātes analīzes mehānisma izveidi, neprasot informāciju par datu uzkrāšanas, apstrādes, apkopošanas principiem un mehānismiem utt., jo pēc datu objekta izveides, t.i., izgūstot datu objektu no IS, datu kvalitātes analīze tiek veikta neatkarīgi no sākotnējas datu kopas un, piemēram, sistēmas, no kuras tie tika izgūti.

Datu objekta un datu kvalitātes specifikācijas pamatā ir lietošanas piemērs, t.i. nolūks, kuram konkrēts datu objekts tiek izmantots – lietotāja vajadzības, vēlmes utt.. Tas nozīmē, ka datu kvalitātes analīzei ir nepieciešami tikai tie reālu objektu raksturojošie lauki, kuri lietotājam ir svarīgi, un dažādos gadījumos tie būs dažādi. Rezultātā vienu reālu objektu reprezentējošs datu objekts atkarība no lietošanas piemēra var izskatīties dažādi gan to raksturojošo parametru skaita, gan struktūras ziņā (Nikiforova, 208a, 2019b).

Pētījuma ietvaros autore veica vairāk kā 30 atvērto datu kopu datu kvalitātes analīze, kuras rezultāti ir pieejami arī atbilstošajos zinātniskajos rakstos - (Nikiforova, 2018a, 2018b, 2019a), (Nikiforova et al., 2019) un (Bicevskis et al., 2018b). Lielāka uzmanība šajā darbā tiek veltīta (a) viena konkrēta domēna analīzei – Latvijas atvērto medicīnas datu kvalitātes analīzei, (b) četru Eiropas valsts (Latvijas (Latvijas Republikas Uzņēmumu reģistrs, 2018), Norvēģijas (Brønnøysundregistrene, 2018), Lielbritānijas (GOV.UK, 2018), Igaunijas (RIK, 2018)) Uzņēmumu reģistru datu kvalitātes analīzei. Uzņēmumu reģistru gadījumā no datu avotu nosaukumiem seko, ka datu objekts visos četros gadījumos ir "Uzņēmums". Medicīnas datu gadījumā datu objekts dažādām datu kopām būs dažāds, jo ir atkarīgs no datu kopas rakstura. Lai sniegtu ieskatu par doto pieeju, visi datu kvalitātes analīzes posmi tiek apskatīti uz konkrēta piemēra, sniedzot atbilstošo datu kvalitātes analīzes posmu diagrammas datu kopai "Lielbritānijas Uzņēmumu reģistrs". Autore dod priekšroku Lielbritānijas Uzņēmumu Reģistram, jo tas ļauj aprakstīt katru ar datu objektu saistītu jēdzienu. Lielbritānijas Uzņēmumu reģistrā katru uzņēmumu raksturo 55 parametri. Pateicoties datu objekta jēdziena raksturam dotās pieejas ietvaros, atbilstoši kurai datu objekts ietver tikai tos parametrus, kuri interesē konkrēto lietotāju atkarībā no lietošanas piemēra, parametru skaits ir atkarīgs no lietošanas piemēra, kas atbilst abstrakcijas principam (Mellor et al., 2004) skatījumā. Vienam datu objektam var tikt definēti dažādi lietošanas piemēri. Piemēram, veiktās aptaujas rezultāti liecina, ka Uzņēmumu reģistram var tikt nodefinēti vismaz 19 dažādi lietošanas piemēri. Pētījuma ietvaros Uzņēmumu reģistru analīzei tika izvēlēti divi vienkāršie un intuitīvie lietošanas piemēri:

1) atrast/ identificēt uzņēmumu pēc tā nosaukuma, reģistrācijas numura un dibināšanas datuma;

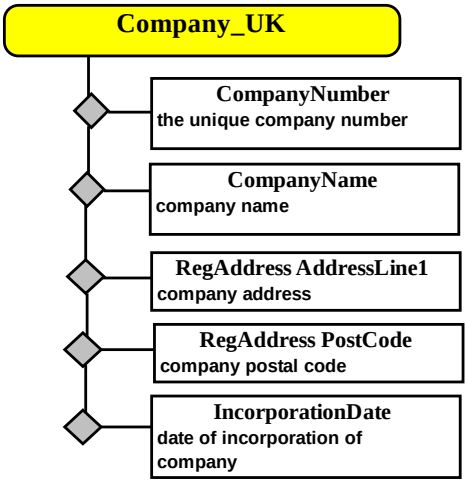
2) sazināties ar uzņēmumu pa pastu, izmantojot tā adresi un pasta indeksu (Nikiforova, 2018a, 2019b).

Vienādu lietošanas piemēru definēšana visiem analizētājiem Uzņēmumu reģistriem ļauj veikt dažādu valsts Uzņēmumu reģistru kvalitātes savstarpēju salīdzinājumu. Tas nozīmē, ka Lielbritānijas Uzņēmumu reģistra gadījumā datu objekts "*Company_UK*" satur tikai 5 parametrus: *[CompanyNumber]* – uzņēmuma reģistrācijas numurs, *[CompanyName]* – uzņēmuma nosaukums, *[IncorporationDate]* – dibināšanas datums, *[RegAddress_AddressLine1]* – uzņēmuma adrese, *[RegAddress_PostCode]* – uzņēmuma pasta indekss. Savukārt pārējie 50 parametri var tikt ignorēti.

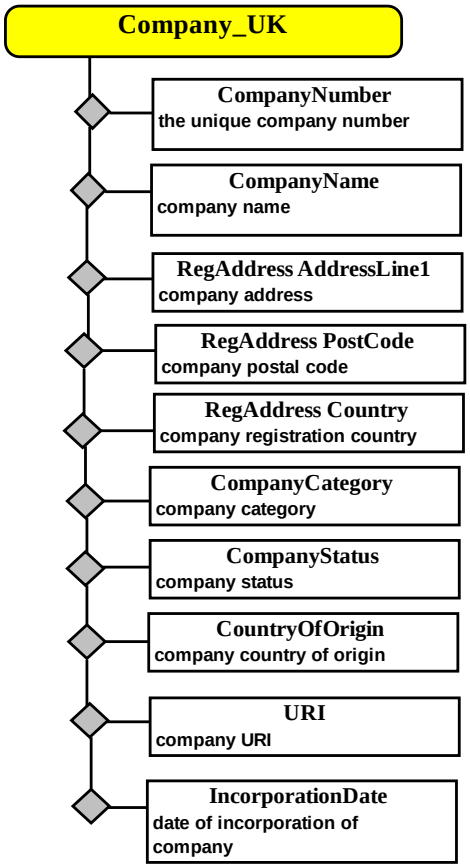
PIM modeļa gadījumā katram parametram tiek definēts glabājamo vērtību neformāls apraksts (dabiskajā valodā). Apraksts ir neformāls, jo nekādi sintakses nosacījumi tā atribūtiem netiek definēti.

Datu apraksts var tikt iegūts dažādos veidos: (a) no dokumentācijas, ko ir sniedzis datu sniedzējs (atbilst arī (Zhang et al., 2014)); (b) no parametru nosaukumiem; (c) izpētot datu kopas saturu. Pirmais veids ir laikietilpīgs un lietotājam draudzīgs, jo nav nepieciešamības lieko soļu veikšanai, taču šī informācija tiek sniegta reti. Otrais veids atbilst vadlīnijām par datu kopu izveidi un publicēšanu. Tas ir plašāk izplatīts. Lielbritānijas Uzņēmumu Reģistra gadījumā datu sniedzēji sniedz dokumentāciju, kurā ir apkopota informācija par datiem, to formātiem, pieļaujamo vērtību sarakstiem utt.. Papildus, arī parametru nosaukumi ir pietiekoši izteiksmīgi, līdz ar ko šīs solis neprasa papildu darbību veikšanu, taču ir jāatzīmē, ka Lielbritānijas Uzņēmumu Reģistrs ir vienīgā tik "lietotājam draudzīga" datu kopa analizējamo uzņēmumu reģistru starpā.

Atbilstoši (Kleppe et al., 2003) *PIM* modelis nav saistīts ar galaplatformu – tas ir neatkarīgs no tās specifikas un detaļām, līdz ar ko tehniskās detaļas tajā netiek iekļautas. Rezultātā, katram datu objekta parametram tiek piešķirts parametra nosaukums un tajā glabājamās vērtības neformāls apraksts. Tā notācija ir ļoti vienkārša un atbilstošais datu objekts ar tā 5 parametriem ir attēlots 3.2.1. att.. Taču šī pētījuma ietvaros katrai datu kopai autore veica padziļināto datu kvalitātes analīzi, analizējot katru atribūtu. Ar papildus parametriem paplašinātais datu objekts ir attēlots 3.2.2. att.. Atbilstoši (Nikiforova, 2019b) padāvātā pieeja paredz datu objekta modificēšanas iespēju tiklīdz rodas tāda nepieciešamība, piemēram, mainās lietošanas piemērs. Tas var tikt veikts ar jebkuru datu kvalitātes analīzē iesaistīto personu.



3.2.1. att. **Datu objekta "Company_UK" *PIM* modelis [izveidoja autore]**



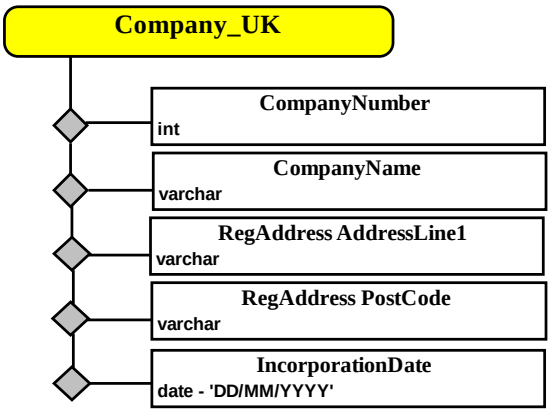
3.2.2. att. **Paplašināta datu objekta "Company_UK" *PIM* modelis [izveidoja autore]**

Salīdzinājumā ar *PIM* modeli, *PSM* modelī ir jābūt iekļautām tehniskām detaļām. Datu objekta parametru apraksts šajā posmā ir daļēji formāls, jo tā atribūtiem tiek definēti sintakses likumi. Tie var tikt definēti vismaz divos dažādos veidos: (1) neformāli, piemēram, dabiskā valodā; (2) formāli, iekļaujot mainīgos, kas ir definēti, piemēram, *C#* programmēšanas valodā. 2. gadījumā datu objekta modelis ir tuvs modeļa implementācijas videi un *PIM* pārvēršas *PSM* modelī. Neformālās prasības ir aizstātas ar formālajām, nosakot laukiem atbilstošākus datu tipus atkarībā no vērtībām, kas glabājās tajos. Arī šī informācija var tikt iegūta (a) no dokumentācijas, ja tāda ir pieejama, (b) veicot datu kopas vai apakškopas sākotnējo apstrādi, analizējot vairākumu datu vērtības, kas glabājās konkrētajā laukā. Ir jāatzīmē, ka parametru formāts ir atkarīgs arī no tehnikas, kas tiks pielietota, aizstājot neformālus aprakstus ar izpildāmiem. No 3.2.1. att. *PIM* iegūtais *PSM* modelis ir attēlots 3.2.3. att., savukārt paplašināta datu objekta *PSM* modelis ir attēlots 3.2.4. att..

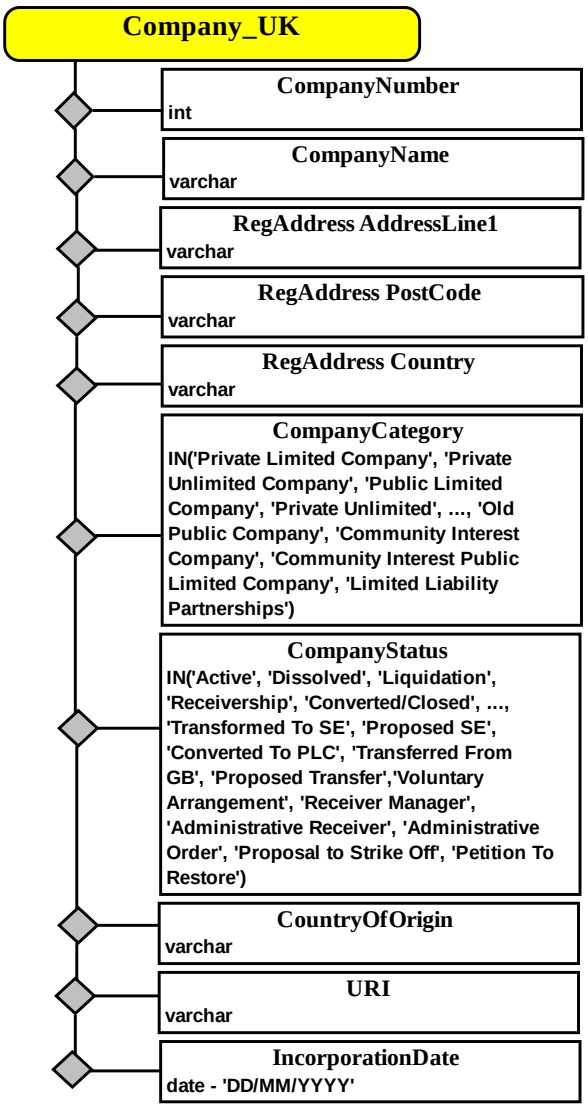
Atbilstoši (Nikiforova et al., 2020) *PSM* modeļa gadījumā katram datu objekta parametram tiek norādīts atbilstošākais datu tips, nepieciešamības gadījumā norādot uz citām parametru vērtību īpatnībām, balstoties uz datiem, kas glabājās konkrētajā laukā. Sākotnēji visu

lauku datu tips ir virkne (*varchar*). Piemēram, tiek norādīts, ka nosaukums ("*CompanyName*") ir patvaļīga garuma virkne, datumu saturoša lauka vērtības formāts ir "*DD/MM/YYYY"*, uzņēmējdarbības forma ("*CompanyCategory*") tāpat kā statuss ("*CompanyStatus*") ir viena no pieļaujamām vērtībām.

Parametru vērtību datu tips ir atkarīgs no datu kvalitātes pārbaudes procesā izmantotas tehnoloģijas. Konkrētā gadījumā darba autore par piemērotāko atzīst *SQL* iespēju, taču atkarībā no iesaistīto pušu zināšanām, pieredzi un vēlmi, tās vietā varētu tikt izmantota arī cita iespēja, piemēram, *C#* (atbilst valodai, kurā uz doto brīdi ir uzrakstīts kompilators). Tādā gadījumā arī definētie datu tipi būtu citi, piemēram, uzņēmuma nosaukumam ("*CompanyName*") *varchar* vietā tiktu piešķirts *string* datu tips (Nikiforova, 2019b).



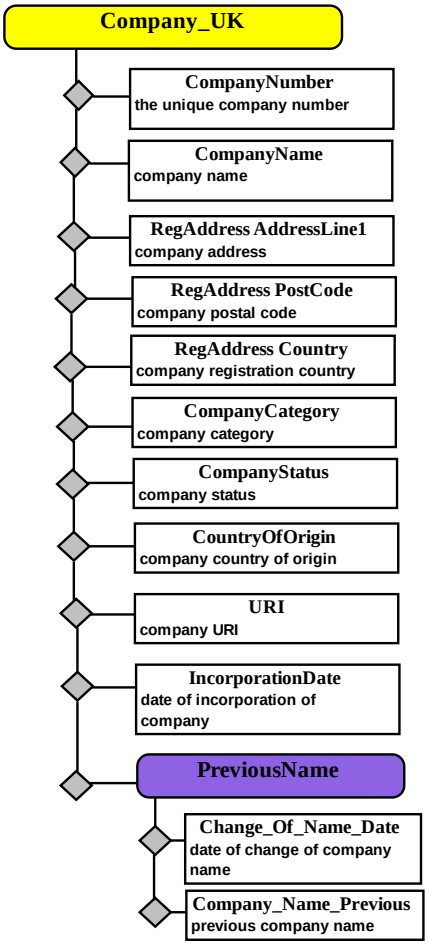
3.2.3. att. **Datu objekta "Company_UK" *PSM* modelis [izveidoja autore]**



3.2.4. att. **Paplašināta datu objekta "Company_UK" *PSM* modelis [izveidoja autore]**

Vēl viens jēdziens, kurš ir nepieciešams dotā risinājuma ietvaros ir "datu objektu klase". Datu objektu klase ir vienādas struktūras objektu kolekcija, ko veido daudzu vienveidīgu datu objektu kopa. Datu objektu klase satur vairākus datu objektus, kas tiek uzskatīti par tās instancēm. Datu objektu klase satur galīgu patvaļīgu parametru un citu datu objektu skaitu. Atbilstoši (Nikiforova, 2019b) katrs konkrēts datu objekts var saturēt visu vai tikai daļu parametru vērtības, tādejādi veidojot kokveida struktūras datu objektu klasi. Iepriekšapskatīta datu objekta "Universitāte" gadījumā, "Universitāte" var tikt uzskatīta par datu objektu klasi, kas satur četrus to raksturojošus parametrus – "*nosaukums*", "*reģistrācijas numurs*", "*interneta vietne*", "*kontakttālrunis*", un vienu datu objektu "*fakultāte*", kas savukārt satur vēl 2 parametrus – "*nosaukums*" un "*iela*".

Piemēram, Lielbritānijas Uzņēmuma reģistra gadījumā uzņēmumu raksturo ne tikai tā esošais nosaukums, bet arī līdz pat 10 iepriekšējiem nosaukumiem. Katru iepriekšējo uzņēmuma nosaukumu raksturo divi parametri: (a) "*Company_Name_Previous*" – iepriekšējais uzņēmuma nosaukums, (b) "*Change_of_Name_Date*" – nosaukuma maiņas datums. Šajā gadījumā "*Company*" ir datu objektu klase, kura ir attēlota att. 3.2.5.. Tā satur 11 parametrus un datu objektu "*PreviousName*", kas satur vēl divus parametrus.



3.2.5. att. **Datu objektu klase "Company_UK" [izveidoja autore]**

Atbilstoši (Nikiforova, 2018a) datu objekta klases jēdziena ieviešana ļauj definēt datu kvalitātes prasības datu objektu kolekcijām. Tas ļauj noteikt datu kvalitāti, ieviešot slieksni, kuram nedrīkst būt pārsniegtam. Piemēram, ja kopējais datu klases "*Company_UK*" kļūdu rādītājs nepārsniedz 5%, tā var tikt uzskatīta par kvalitatīvu, citādi tās kvalitātei ir jābūt uzlabotai nekavējoties. Šīs rādītājs tiek aprēķināts nekvalitatīvo ierakstu skaitu attiecinot pret kopējo ierakstu skaitu (atbilst arī (Batini et al., 2016)). Atbilstoši piedāvātās pieejas pamatidejai, arī šo slieksni nosaka galalietotājs.

Atbilstoši (Nikiforova, 2019b) datu objekta un datu objektu klases jēdzienu izmantošana ļauj vērtēt datu kopas kvalitāti konkrētam nolūkam, ko definē galalietotājs. Tas nozīmē, ka, pat ja konkrētai datu kopai ir raksturīgas vairākas datu kvalitātes problēmas, tā var tikt atzīta par pietiekoši kvalitatīvu konkrētam nolūkam, ja konkrētu lauku vērtības, kas interesē konkrētu lietotāju, nesatur datu kvalitātes problēmas (piemēri ir aplūkoti 5. nodaļā).

3.3. Kvalitātes specifikācija

Otrais datu kvalitātes analīzes posms ir kvalitātes specifikācijas definēšana iepriekšējā solī nodefinētajam datu objektam. Atbilstoši (Bicevskis et al., 2018a) konkrēta datu objekta datu kvalitātes specifikācija sastāv no nosacījumiem, kuriem ir jāatbilst datiem, lai tie tiktu uzskatīti/ atzīti par kvalitatīviem.

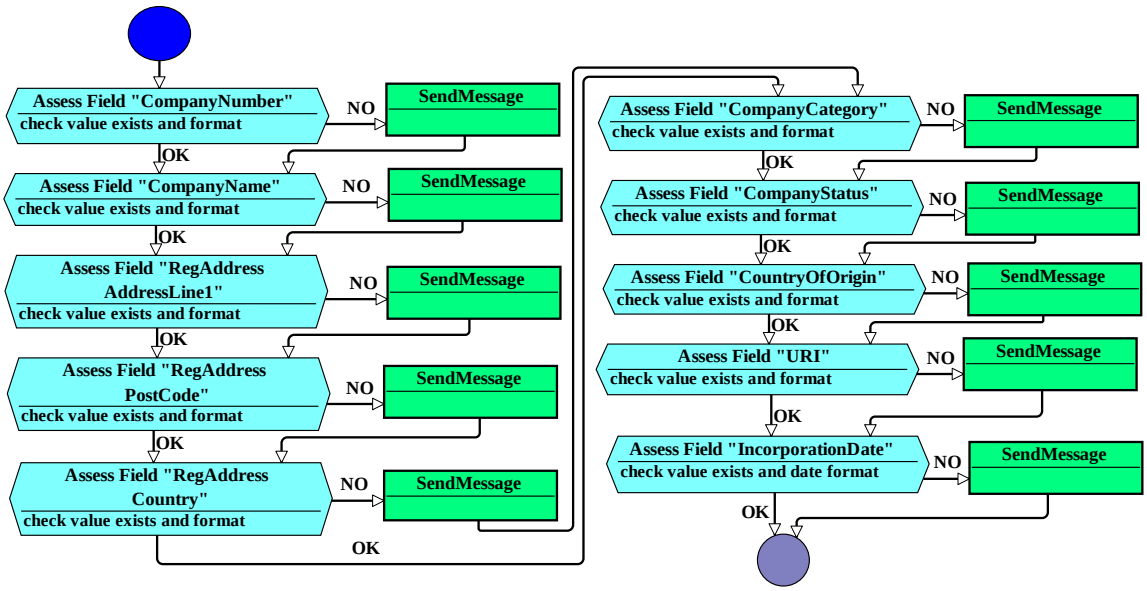
Datu objekta parametru vērtību datu kvalitātes kontrole reducējas uz atsevišķo vērtību pārbaudi. Piemēram, attiecībā uz datu objektu "Universitāte": (1) vai simbolu virkne, kas ir pirmā parametra vērtība, var kalpot par universitātes nosaukumu, (2) vai reģistrācijas numura vērtība atbilst reģistrācijas numura paraugam (angl. *pattern*), tāpat kā kontakttālrunis un interneta vietnes vērtības, (3) vai fakultātes nosaukuma vērtība atbilst vienai no pieļaujamām vērtībām, (4) vai dibināšanas datums ir ticams, (5) vai norādītā adrese eksistē utt.. Proti, datu objekta kvalitātes pārbaudes var saturēt arī sarežģītāku nosacījumu pārbaudes, piemēram, kontrolcipara pareizības pārbaudi paškontrolējošos kodos utt., ko nosaka datu specifika un galalietotāja definētais lietošanas piemērs, datu objekts un datu kvalitātes prasības.

Datu kvalitātes prasības izriet no datiem, kas glabājās konkrētajā datu objekta parametrā. Parasti kvalitātes prasības var tikt definētas: (a) balstoties uz dokumentāciju, ja datu sniedzējs nodrošina to; (b) veicot datu kopas vai apakškopas sākotnējo apstrādi (līdzīgi (Zhang et al., 2014)); (c) tā kā datu kvalitāte ir atkarīga no datu lietotāja un lietošanas piemēra, prasības tiek definētas ar lietotāju, t.i. tās balstās uz lietotāja prasībām pret konkrētu datu kopu un datu objektu. Ir paredzēts, ka pirmie divi varianti ir tikai palīglīdzekļi, savukārt, prasības galvenokārt

tiek definētas ar lietotāju atkarībā no definētā lietošanas piemēra. Otrais variants var būt lietotājam noderīgs gadījumos, kad, piemēram, ir jānosaka, vai konkrētajā laukā ir pieļaujamas tukšas vērtības, nosakot slieksni, piemēram, 3%, tukšam vērtībām pārsniedzot kuru, tiek uzskatīts, ka konkrētajā parametrā vērtība mēdz būt nenorādīta. Lielbritānijas Uzņēmumu reģistra gadījumā tā tiek daļēji iegūta no datu sniedzēju dokumentācijas, no kuras ir izgūstama informācija par datiem, to garumu, pieļaujamām vērtībām utt., papildinot to ar prasībām, ko definē galalietotājs. Vairākas kvalitātes prasības var tikt definētas neatkarīgi no lietošanas piemēra, kas atbilst objektīvām (Price et al., 2004, 2005) vai deklaratīvām (Jayawardene et al., 2015) prasībām. Piemēram, ir acīmredzami, ka uzņēmuma reģistrācijas numuram ir jābūt netukšai vērtībai, kurai ir jāatbilst noteiktam formātam, t.i. paraugam.

PIM modeļa gadījumā kvalitātes prasības tiek definētas neformāli, piemēram, formulējot tās dabiskā valodā vai kā formalizētus aprakstus, kas nav atkarīgi no implementācijas. Tiem ir jābūt saprotamiem lietotājiem, kuriem var nebūt padziļināto zināšanu IT jomā. Šī posma mērķis ir skaidri un saprotami izteikt datu galalietotāja prasības, kas turpmāk tiks pielietotas galalietotāja nodefinētam datu objektam. Katram parametram definētās kvalitātes prasības tiek grupētas kopā, kas atbilst *Mellor* "klasifikācijas" principam (Mellor et al., 2004).

Kvalitātes prasību *PIM* modelis datu objekta "*Company_UK*" paplašinātajai versijai ir attēlots att. 3.3.1., katru kvalitātes nosacījumu aprakstot dabiskā valodā. 1. – 4. un 10. elementos attēlotas kvalitātes prasības atbilst sākotnēja datu objekta versijai.



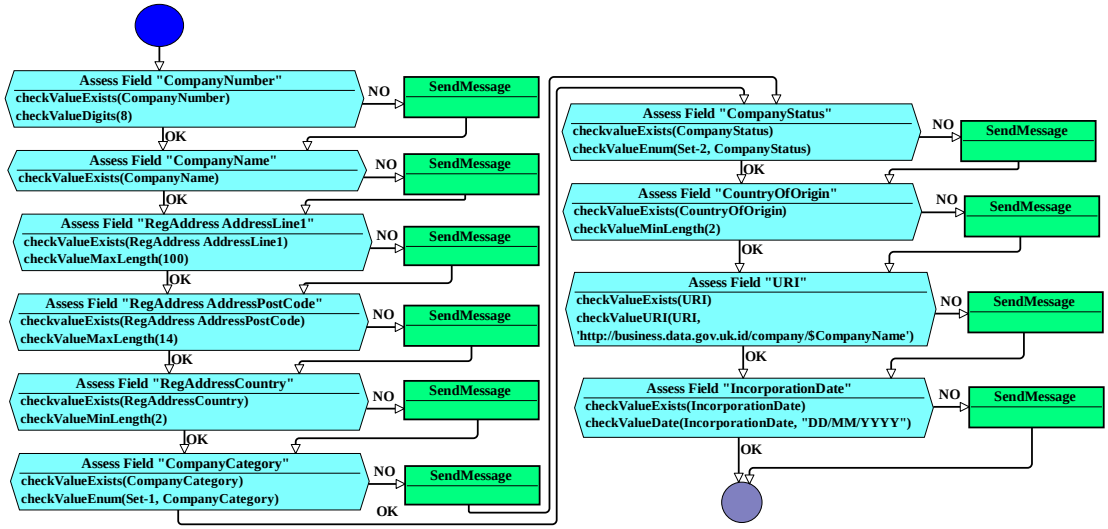
3.3.1. att. **Datu kvalitātes specifikācija datu objektam "Company_UK" (PIM modelis) [izveidoja autore]**

Att. 3.3.1. ir redzams, ka, katram parametram pārbaudot kvalitātes nosacījumu izpildi, ir paredzēta kļūdas ziņojumu sagatavošana, ja parametra vērtība neatbilst izvirzītai prasībai. Kļūdas ziņojumi tiek ierakstīti protokolā tā turpmākai apstrādei datu kvalitātes uzlabošanas posmā. Neatkarīgi no iepriekšējas pārbaudes rezultāta, t.i. datu atbilstības vai neatbilstības datu kvalitātes nosacījumiem, notiek pāreja pie nākama parametra pārbaudes līdz pēdējās pārbaudes izpildei.

Atbilstoši (Nikiforova et al., 2020) nākamais solis ir *PIM* modeļa transformācija *PSM*. *PSM* gadījumā datu kvalitātes prasības tiek aizstātas ar formālām prasībām, kas ir definētas, izmantojot loģiskās izteiksmes. Diagrammu struktūra paliek nemainīga, mainoties tikai kvalitātes prasību definīcijām neformālās aizstājot ar loģiskām izteiksmēm. Atbilstoši (Nikiforova, 2019b) loģisko izteiksmju operandiem kalpo datu objekta atribūtu/ parametru nosaukumi, savukārt operācijām var izmantot gan tradicionālos programmēšanas valodās lietotos līdzekļus - loģisko izteiksmju operācijas, gan datu kvalitātei specifiskas operācijas. Loģiskās izteiksmes ir vienlaicīgi gan pietiekoši izteiksmīgas, gan vienkārši saprotamas, kas paaugstina iespējamību procesā iesaistīties lietotājiem bez padziļinātām IT un datu kvalitātes zināšanām.

Loģisko izteiksmju raksturs ir atkarīgs galvenokārt no datu objekta *PSM* modeļa un iepriekšējā posmā nodefinētā kvalitātes specifikācijas *PSM* modeļa. Lielbritānijas Uzņēmumu reģistra gadījumā tās ir (a) vai parametra "*IncorporationDate*" vērtības formāts atbilst definētājām – "*DD/MM/YYYY*", (b) vai parametra "*CompanyCategory*" vērtība eksistē un atbilst kādai no pieļaujamo vērtību sarakstā esošajām, (c) vai "*URI*" parametra vērtība atbilst paraugam, atbilstoši kuram vērtībām ir jāsatur "*http://business.data.gov.uk/id/company/*" virkne, un tai ir jāseko uzņēmuma nosaukumam, kas atbilst "*CompanyName*" parametra vērtībai (att. 3.3.2.).

Tāpat kā iepriekšējos darba autores pētījumos ((Nikiforova, 2018a, 2018b, 2019a), (Nikiforova et al., 2019), (Bicevskis et al., 2018b)) visbiežāk datu objektu parametru definētās prasības fokusējas uz tādām pārbaudēm kā (a) vērtību eksistence, (b) atbilstība noteiktam datu tipam, (c) formāta, piemēram, virknes garuma pārbaude, (d) atbilstība noteiktam paraugam, (e) vērtības atbilstības pieļaujamo vērtību sarakstam, (f) vērtību derīguma, piemēram, vai datums ir ticams, un citām, kas ir atkarīgas no datu rakstura un lietošanas piemēriem. Šīs prasības galvenokārt atbilst deklaratīvajam jeb objektīvām prasībām.



3.3.2. att. **Datu kvalitātes specifikācija datu objektam "Company_UK" (*PSM* modelis) [izveidoja autore]**

Atbilstoši (Nikiforova et al., 2020) datu kvalitātes prasības datu objektam tiek definētas, izmantojot pseidokodu, kas tiek ierakstīts diagrammas elementos. Neskatoties uz to, ka dažreiz pseidokods tiek saistīts ar *PIM* (piemēram, (Ruiz, 2018)), piedāvātā risinājuma ietvaros tas var tikt uzskatīts par *PSM*, jo pseidokods ir cieši saistīts ar tā implementāciju, piemēram, *C#* valodu (atbilst (Coutinho et al., 2012), (Kessler et al., 2010), (Shi et al., 2015) utt.).

Attēlotās prasības lielākoties atbilst sintaktiskajam pārbaudēm, savukārt kontekstuālās pārbaudes, kas paredz papildu datu objektu iesaisti datu kvalitātes analīzē, tiek apskatītas 3.6. apakšnodaļā.

Datu objekta un datu kvalitātes prasību definēšanā var tikt veikta gan ar vienu lietotāju, gan ar vairākiem lietotājiem, kuru skaits nav ierobežots, jo dotā pieeja paredz un atbalsta lietotāju savstarpēju mijiedarbību gan vairāku līmeņu/ struktūru (piemēram, tehniskās un tirgvedības (angl. *marketing*) nodaļu), gan dažādu organizāciju līmenī. Pie tam, to var veikt arī nevienai organizācijai nepiederoša persona (t.i. savām personiskām vajadzībām).

3.4. Kvalitātes pārbaudes process

Datu kvalitātes pārbaudes process paredz visu aktivitāšu, kas būtu jāveic, lai novērtētu datu objekta kvalitāti, aprakstu. Pirmkārt, datu objekta vērtības tiek nolasītas no datu avota un ierakstītas datubāzē. Atbilstoši (Nikiforova, 2018a, 2019b) šī soļa sarežģītība ir atkarīga no datu formāta, jo strukturēto datu ielāde datubāzē parasti neizraisa nekādas problēmas, kas ir skaidrojams ar to struktūras līdzību datubāzes tabulai, savukārt daļēji strukturētu datu atlasei

un ielādei var būt nepieciešamas papilddarbības, kas mēdz būt atkarīgas no dokumenta struktūras, tajā esošajām datu hierarhijām, atbilstoši kurām atsevišķos gadījumos katram hierarhijas līmenim jāparedz atsevišķa tabula, sasaistot tās savā starpā ar primāro un ārējo atslēgu palīdzību, kā arī citām īpatnībām, kas ir atkarīgas no datu sniedzēja, piemēram, datu sniedzēju nepareiza jeb vienādu vērtību un parametru atdalītāju izvēle.

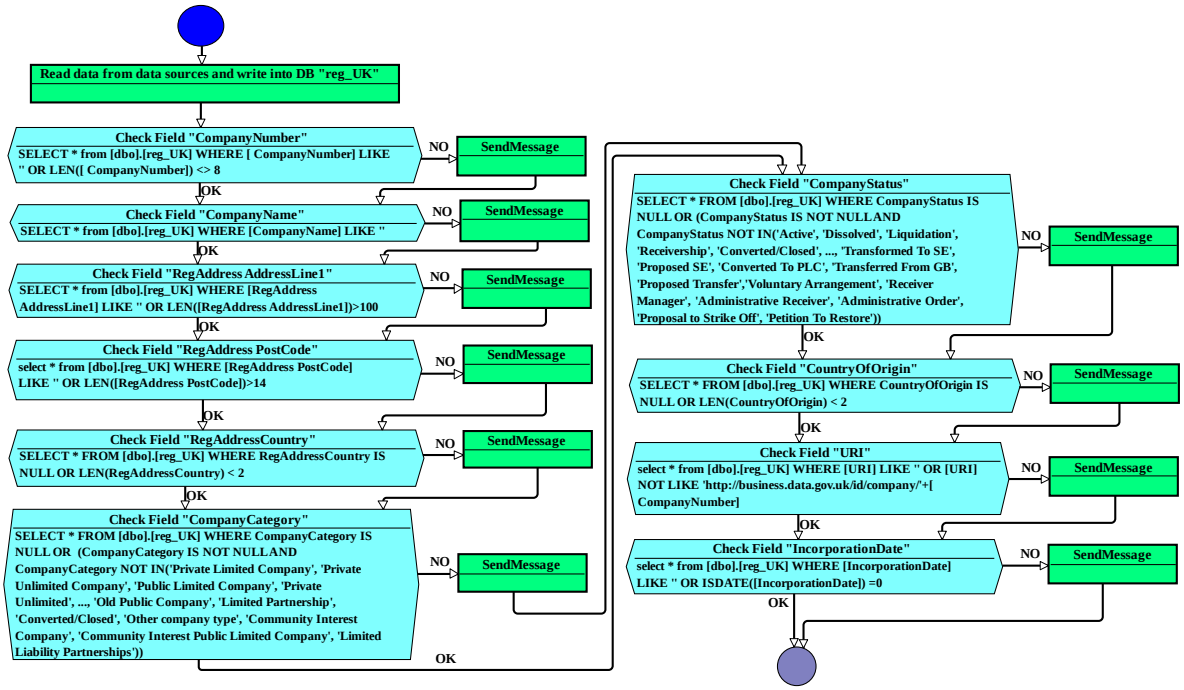
Šīm solim seko viens vai vairāki soli, kas būtu jāveic, lai novērtētu vērtību kvalitāti. Katrs solis apraksta vienu vai vairākus soļus, kas būtu veicami, pārbaudot datu objekta atbilstību nodefinētajām datu kvalitātes prasībām. Datu kvalitātes pārbaudes procesā tiek apstrādātas datu objektu klases. Datu objekta instances tiek atlasītas no datu avota un ierakstītas kolekcijā. Cikliski apstrādājot visas atlasītās instances, katrai instancei tiek pārbaudīta datu kvalitātes nosacījumu izpilde - tāpat kā tas tiek veikts vienam datu objektam. Šī procesa izpildes rezultāts ir katrai individuālai instancei noteiktās datu kvalitātes problēmas. Tādejādi atbilstoši att. 3.4.1., ja konkrēta vērtība neatbilst nodefinētai kvalitātes prasībai, atbilstošais ziņojums tiek sagatavots un nosūtīts galalietotājam. Visas "netukšas" elementu "*SendMessage*" vērtības veido datu kvalitātes protokolu, kas tiek noglabāts datubāzē tā turpmākai apstrādei. Tas var tikt izmantots datu kvalitātes uzlabošanai, automātiski vai manuāli veicot izmaiņas datu avotā.

Atbilstoši (Nikiforova, 2019b) "valoda, kas apraksta datu kvalitātes novērtēšanas procesu, ietver sevī konkrēta datu objekta verifikācijas darbības, kas arī var tikt formulēti (a) neformāli – dabiskā valodā, (b) izmantojot *UML* aktivitāšu diagrammas, vai (c) veidojot savu domēnspecifisku valodu jeb *DSL*".

Datu kvalitātes novērtēšanas *PSM* modelis ir izpildāms. Datu kvalitātes prasību specifikācijas izpildāmība ļauj iekļaut kvalitātes prasību izpildes pārbaudes dažādos datu apstrādes posmos. Tas atrisina datu kvalitātes pārbaudes problēmas situācijās, kad dati tiek uzkrāti pakāpeniski, pieļaujot tādu datu ievadīšanas secību datubāzē, kas atšķiras no notikumu iestāšanās "reālā pasaulē" vai to reģistrēšanas secības (Nikiforova, 2019b).

Datu objekts vai datu objekta klase ir datu kvalitātes novērtēšanas procesa ievads. Nolasot datus no datu avota vai vairākiem avotiem un saglabājot tos datubāzē (pētījuma ietvaros autore izmantoja *Microsoft SQL Server 17*), tiek veikta katras instances atbilstības kvalitātes nosacījumiem pārbaude, iepriekšējā posmā neformāli nodefinētas datu kvalitātes prasības, aizstājot ar izpildāmiem *SQL* vaicājumiem.

Att. 3.4.1. ir attēlots datu objekta "*Company_UK*" datu kvalitātes pārbaudes procesa *PSM* modelis. Pirmais elements atbilst datu nolasīšanas un ierakstīšanas datubāzē darbībai, kurai seko parametru vērtību kvalitātes pārbaude, automātiski izpildoties *SQL* vaicājumiem.

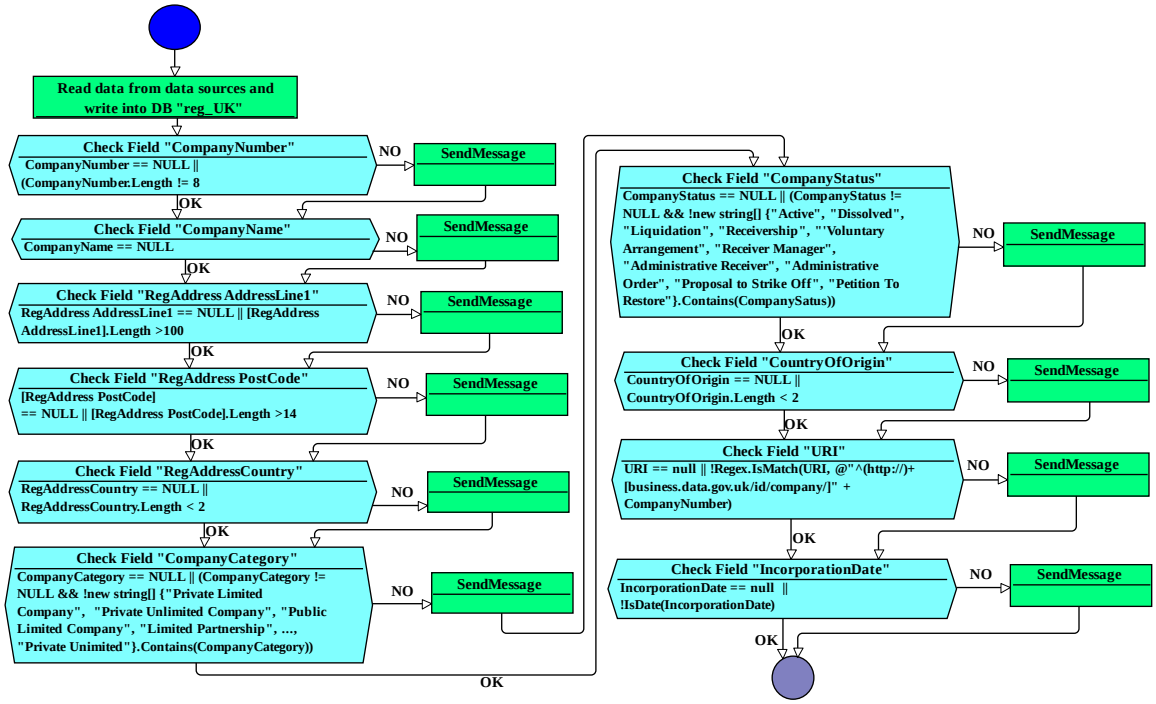


3.4.1. att. **Datu kvalitātes pārbaudes process datu objektam "Company_UK" [izveidoja autore]**

Darba autore priekšroku dod *SQL*, jo *SQL* vaicājumu konstrukcija ir tuva datu un datu kvalitātes raksturam, jo *SQL* vaicājuma *SELECT* operators ir piemērots datu nolasīšanas darbībai, savukārt *WHERE* nosacījums tiek izmantots datu kvalitātes prasību definēšanai, grupējot vienam parametram definētās kvalitātes prasības, izmantojot *OR* operatoru. Šo variantu tiek atzīts par vispiemērotāko dotā uzdevuma izpildei arī respondentu vidē, jo absolūtais dalībnieku vairākums priekšroku, definējot izpildāmas prasības nodefinētajam datu objektam, ir devis *SQL* valodai.

Savukārt izpildāmu objektu raksturs ir atkarīgs no šajā posmā iesaistītās personas zināšanām un pieredzes, jo, lai panāktu atbilstošo prasību izpildi, ir jānodrošina to sintaktiskā pareizība, līdz ar ko šī fāze paredz IT-speciālista iesaisti. Tas nozīmē, ka izpildāmiem tekstiem nav obligāti jābūt pierakstītiem ar *SQL* vaicājumu palīdzību – tie var tikt aizstāti ar kādā programmēšanas valodā rakstītiem tekstiem. Viena no paredzētājām un uzdevumam atbilstoša alternatīvām ir kvalitātes prasību pieraksts, izmantojot programmkodu C# valodā (3.4.2. att.).

Atbilstoši (Nikiforova, 2019b) šāda veida datu kvalitātes prasību implementācija ir raksturīga datu apstrādei relāciju datubāzēs, un dotais modelis ir platformatkarīgs, jo ir cieši saistīts ar izpildes vidi.



3.4.2. att. **Datu kvalitātes pārbaudes process datu objektam "Company_UK" [izveidoja autore]**

Datu kvalitātes pārbaudes procesa apraksta valodas konkrēta datu objekta galvenā atšķirība no datu kvalitātes prasību apraksta valodas ir datu nolasīšanas no datu avota vai vairākiem avotiem un to ierakstīšanas datubāzē operāciju veikšana. Ņemot vērā datu kvalitātes pārbaužu dažādību atkarībā no lietošanas piemēra, ko definē galalietotājs, *DSL* ir nodefinēta, ievērojot šo principu, t.i. katru reizi, veicot datu kvalitātes analīzi, definējot konkrētam lietošanas piemēram atbilstošas datu kvalitātes pārbaudes.

Šīs posms ir pirmais datu kvalitātes analīzes posms, kas paredz IT cilvēku iesaisti, kas var būt nepieciešams iepriekšējā posmā nodefinēto loģisko izteiksmju prasību pārrakstīšanai. Taču, atbilstoši sākotnējai idejai (Nikiforova, 2019b), nākotnē šīs solis varētu tikt pielāgots lietotāju vajadzībām, bieži definētājām kvalitātes prasībām, izveidojot atbilstošās pārbaudes metodes, ko lietotāji spētu patstāvīgi pielāgot savām vajadzībām bez IT-speciālistu piesaistes. Potenciālās pārbaudes varētu būt datu esamības, virkņu garuma, vērtību iekļaušanos diapazonā un atbilstības noteiktam formātam pārbaudes.

Tā kā iepriekšaprakstītais galvenokārt atbilst sintaktiskai kontrolei, pārbaudot datu objektu vērtības viena datu objekta ietvaros, t.i. ievaddatu atbilstību to sintaksei, darba autore konstatēja nepieciešamību pieejas paplašināšanai, nodrošinot iespēju veikt arī semantisko jeb kontekstuālo pārbaudi vairāku datu objektu ietvaros. Dotais pieejas paplašinājuma apraksts un tā uzlabošana salīdzinājumā ar (Nikiforova et al., 2019) un (Nikiforova, 2019a, 2019b) prezentēto ir apskatīts 3.6. apakšnodaļā.

Pēdējais solis ir *PSM* modeļa izpilde. Atbilstoši (Nikiforova, 2019b) un (Nikiforova et al., 2020) tas var tikt implementēts vairākos veidos, divi no kuriem ir: (a) kvalitātes prasību specifikācija kā programmēšanas darbs, kas ļauj precīzi formulēt prasības, bet tiek implementēts ar tradicionālās programmēšanas metodēm; (b) vispārīgāks implementācijas variants ir interpretators vai kompilators, kas spēj izpildīt kvalitātes pārbaudes nosacījumus, kas noglabāti repozitorijā. Pirmais variants parasti ir sastopams informācijas sistēmās, kurās ievaddati tiek ievadīti ar ekrānformu palīdzību. Piedāvātās kvalitātes prasību specifikācijas priekšrocība ir tas, ka, atdalot prasību specifikāciju no pirmkoda, nepieciešamības gadījumā ir iespējams mainīt programmas kodu atbilstoši specifikācijai. Objektorientētā programmēšanā kvalitātes prasību pārbaudi ir iespējams veidot kā atsevišķu metodi, kura ir pielietojama konkrētam datu objektam. Neskatoties uz to, ka otrais variants ir sarežģītāks, priekšroka tika iedota tam. Šīm nolūkam šī risinājuma ietvaros tika izmantots *DIMOD* rīks. Tas nodrošina repozitoriju izveidi kvalitātes prasību modelēšanas laikā, ļaujot veikt izmaiņas izveidotājos modeļos, neskarot IS programmas, tādejādi ieintegrējot kvalitātes pārbaudes procesu IS procesos. Datu kvalitātes pārbaudei tiek izsaukts kompilators, izvēloties konkrētai datu kvalitātes analīzei atbilstošo kvalitātes prasību pierakstam izmantoto sintaksi – *C#* vai *SQL*, kas tiek noteikts ar procesā iesaistīto lietotāju, balstoties uz viņa zināšanām un iemaņām katrā tehnoloģijā. Tad kompilatoram tiek nodots datu objekts un no repozitorija tiek izsaukts kvalitātes prasību apraksts. Atbilstoši (Nikiforova, 2019b) kompilatora struktūra balstās uz saistītājiem sarakstiem, kur katram masīva elementam var tikt ierakstīta papildus informācija, t.i. tiek glabāta norāde uz saistīto sarakstu. *DIMOD* rīkā nodefinēta diagramma tiek pārvērsta par grafu, apstaigājot kuru, tas tiek pārvērsts par izpildāmo kodu. Atbilstoši iepriekšminētajam datu kvalitātes pārbaudes procesa posms paredz IT-speciālistu iesaisti, kas ir saistīts ar galveno uz doto brīdi esošo ierobežojumu - visu datu kvalitātes pārbaudes procesu diagrammā definēto formālo izpildāmo tekstu korektums, jo *DIMOD* rīks un kompilators neveic ierakstīta teksta derīguma pārbaudes, t.i. tās ir jānodrošina "gudrajiem lietotājiem" (Bicevskis et al., 2018a).

Īss piedāvātās pieejas implementācijas apraksts, apskatot iepriekšējās apakšnodaļās neapskatītās risinājuma detaļas, ir sniegts 3.5. apakšnodaļā.

3.5. Piedāvātās pieejas implementācija

Kā darba autore minēja iepriekš, piedāvātā pieeja balstās uz grafiskām domēnspecifiskām valodām (*DSL*). Katram datu kvalitātes modeļa komponentam – datu objektam, datu kvalitātes specifikācijai un datu kvalitātes novērtēšanas jeb pārbaudes procesam, ir nodefinēta sava

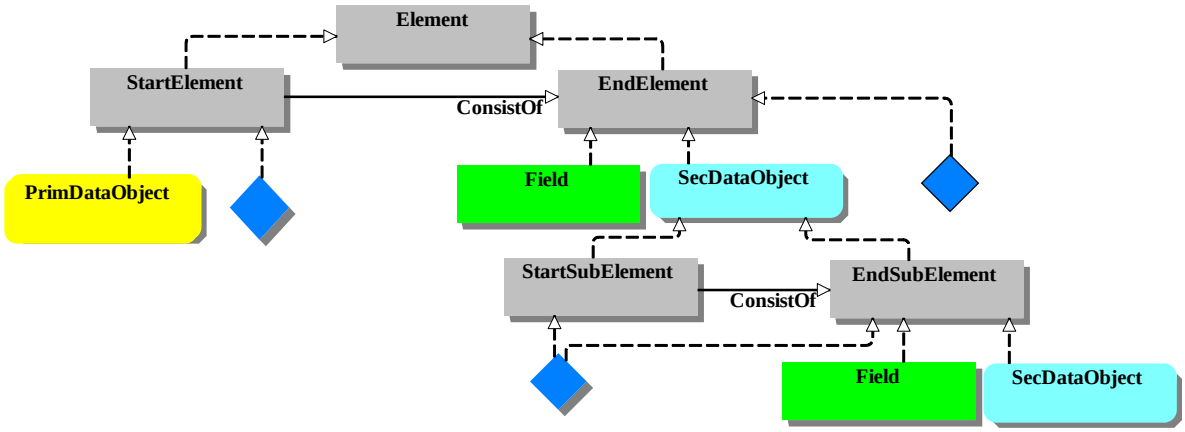
grafiskā *DSL*. Atbilstoši (Nikiforova et al., 2020) ir ieteicams nedefinēt vienu visaptverošu valodu, kas kļūstu pārāk sarežģīta.

Šīm nolūkam pētījuma ietvaros tika izmantota rīku būves platforma *DIMOD*, kas ļauj definēt daudzas dažādas *DSL* ar dažādām datu objektu struktūrām. *DIMOD* ir LUMII izstrādātās grafiskās rīku izveides platformas *GrTp* (Bārzdiņš et al., 2007) atvasinājums, ko izstrādāja SIA "DIVI grupa". Ir jāatzīmē, ka arī *GrTp* risinājums, no kura tika atvasināts izmantotais *DIMOD* rīks, balstās uz *MDA* principiem. *DIMOD* darbības pamati var tikt salīdzināti ar *Monterey Phoenix* ((Auguston et al., 2015), (Zhao et al., 2003)). Pētījuma ietvaros ar *DIMOD* palīdzību tika izveidota trīs valodu saime, kuru definēšanu un reprezentāciju nodrošina *DIMOD*. Katrai *DSL* valodai ir sava struktūra, kura atbilst iepriekšējās apakšnodaļās apskatītājiem piemēriem. Pateicoties tā *DSL* konfigurēšanas iespējām (Sprogis et al., 2013), tiek uzdots *DSL* metamodelis, pēc kura ievadīšanas repozitorijā, *DIMOD* pārvēršas par definētās *DSL* grafisku redaktoru, kas, interpretējot *DSL* metamodeli, piedāvā visas grafiskam redaktoram nepieciešamās iespējas - zīmēt grafiskas diagrammas, rediģēt tās, veidot kokveida struktūras modeļus un citas darbības. Atbilstoši rīka izstrādātājiem (Sproģis, 2014) risinājuma priekšrocība ir tas, ka modeļi tiek automātiski saglabāti kā rīku definēšanas metamodeļa instances, pateicoties rīkam *Configurator*. Tas ir *DSML* (domēnspecifiskās modelēšanas valodas) rīks, kas, salīdzinājumā ar *UML* klašu diagrammām ļauj veikt rīku specificēšanu augstākā abstrakcijas līmenī. Tā specifikāciju interpretēšanai jeb apstrādei ar universālo interpretatoru atbilstošās specifikācijas tiek automātiski transformētas uz universālo metamodeli. Rezultātā risinājuma lietotājiem jeb rīku izstrādātājiem nav nepieciešamības pārzināt rīku definēšanas metamodeli, tā noklusētās vērtības, kā arī pārliecināties veidojamo modeļu pareizībā. Tas ievērojami samazina rīku definēšanas metamodeļa instances izstrādes laiku, izslēdzot iespējamību uzbūvēt interpretatoram neatbilstošu interpretāciju. Cita dotā risinājuma priekšrocība ir tas, ka gan *GrTp*, gan *DIMOD* tika izstrādāti Latvijā, līdz ar ko to izmantošana sekmēs Latvijas izgudrojumu popularizēšanu, kā arī nepieciešamības gadījumā atbilstošu risinājumu autori var sniegt nepieciešamu atbalstu vai arī rīku paplašināšanu vai pielāgošanu jauniem uzdevumiem.

Atbilstoši (Kleppe, 2008) *DSL* definīcijas centrālais komponents ir abstraktā sintakse, kurai valodas specifikācijā ir primārā loma – tā definē valodas jēdzienus un to savstarpējas attiecības, ietverot arī modeļu izveides ierobežojumus. Abstraktās sintakses definēšana tika veikta, izmantojot metamodelēšanas tehniku, kura atbilstoši (Akehurst et al., 2002) tiek uzskatīta par labāku [grafisko projektējumu] normalizācijas veidu, kura garantē datu integritāti modelī, izmantojot formālās tehnikas, samazinot un pat novēršot redundanci un dažādu veidu anomāliju iespējamību. Šīm viedoklim piekrīt arī iepriekšminētais *Basciani* un viņa līdzautori

(Basciani et al., 2016). Metamodeļi ir pietiekoši izteiksmīgi un viegli saprotami, it īpaši salīdzinot ar tekstbāzētām sintaksēm, pateicoties grafisku gramatiku izmantošanai. Tieši šī iemesla dēļ metamodeļiem tiek dota priekšroka, ja starp valodu jēdzieniem pastāv sarežģītas attiecības, kas nevar tikt vienkārši aprakstīti ar tekstuālas sintakses palīdzību. Papildus metamodelēšana nodrošina vairāku ierobežojumu apvienošanu, kas bezkonteksta gramatiku gadījumā tiktu glabāti atsevišķi (Selic, 2009). Kopumā, metamodeļu izmantošana būtiski atvieglo valodas definēšanu. Konkrētajā gadījumā, metamodeļi ļauj vienkārši attēlot (a) katra elementa grafisko reprezentāciju, t.sk. aprakstot elementu formu, krāsu utt., (b) attiecību attēlošanu, "no" un "uz" attiecību attēlošanai izmantojot bultas, kas tekstbazētās gramatikas gadījumā būtu daudz sarežģītāk, prasot vairākas pārlieku sarežģītas darbības. Priekšrokas došana metamodelēšanai kā atbilstošākai tehnikai abstraktas sintakses definēšanai salīdzinājumā ar bezkonteksta gramatiku atbilst arī ((Selic, 2009), (Sproģis, 2014) utt.).

Metamodeļa izveide ir viens no sarežģītākajiem datu kvalitātes risinājuma izveides soļiem, kuram ir nepieciešamas atbilstošās modelēšanas zināšanas (viens no iespējamiem datu objekta definēšanas piemēriem ir nodemonstrēts att. 3.5.1.).



3.5.1. att. **Datu objekta metamodelis (Nikiforova et al., 2020)**

Savukārt, kad uzkonfigurētais grafiskais redaktors ir sagatavots tā turpmākai izmantošanai, publicējot to tiešsaistē, lietotāji var izmantot sniegtās iespējas, neaizdomājoties par metamodeļu izveidi, veidojot atbilstošas grafiskās diagrammas, kuru struktūra ir intuitīva un tuva datu un datu kvalitātes raksturam. Tas nozīmē, ka vienu reizi uzkonfigurējot redaktoru, tas kļūst atkalizmantojams (atbilst arī (Basciani et al., 2016) idejai). Ir jāatzīmē, ka tiešsaistē var tikt publicēti ne tikai grafiskie redaktori, bet arī jau izveidotas diagrammas, kas ļauj galalietotājiem izpētīt iepriekšsagatavotas diagrammas pirms tie veidos savas diagrammas. Šī iespēja nodrošināšana var kalpot par sava veida pamācību galalietotājiem.

3.6. Kontekstuālā datu kvalitātes pārbaude

Praksē bieži vien nepietiek ar datu kvalitātes pārbaudi viena datu objekta ietvaros, radot nepieciešamību veikt arī kontekstuālo datu kvalitātes analīzi vairāku datu objektu ietvaros. Kontekstuālai jeb semantiskai kontrolei ir raksturīga pārbaude, vai datu objekts ir atbilstoši saistīts ar citiem datu objektiem, vai tas ir saderīgs ar citiem jau iepriekš datu avotā ievadītām datu objektu vērtībām, nosakot, vai dati nav pretrunīgi. Semantiskas kontroles raksturs prasa atkārtot semantisko kontroli katru reizi, kad mainās kāda no savstarpēji saistītu datu objekta atribūtu vērtībām (Nikiforova, 2019b).

Tradicionāli semantisko jeb kontekstuālo pārbaudi veic divos posmos:

1) atbilstošā ieraksta atrašana "ārējā" datu kopā,
2) sākotnējas datu kopas ieraksta lauku validēšana pret atrasto ierakstu (Scannapieco et al., 2005).

(Batini et al., 2016) pirmais solis tiek dēvēts "ieraksta identifikācija", savukārt otrais – "lēmuma stratēģija". Tas nozīmē, ka sākumā tiek atrasti visi saskaņoti ieraksti abās datu kopās, sasaistot datu kopas pēc konkrētiem parametriem, kam seko katra atbilstošā pāra vērtību savstarpējās atbilstības pārbaude. Šajā gadījumā primārā datu objekta visas atbilstoša parametra vērtības veido sekundārā datu objekta atbilstoša parametra visu vērtību apakškopu. Abas kopas var būt vienādas, taču primārā datu objekta atbilstoša parametra vērtību kopa nedrīkst saturēt elementus, kas nav pieejami sekundārajā datu objektā. Protams, tas nozīmē, ka konkrēta parametra vērtību pārbaudei ir nepieciešama sekundārā datu objekta augsta kvalitāte un pilnīgums. (Batini et al., 2016) paredz, ka "lēmuma stratēģijas" posmā tiek pieņemts lēmums, vai gadījumā, ja vērtības sakrīt, ir iespējams apgalvot, ka abas vērtības reprezentē vienu un to pašu reālās pasaules objektu. Citos pētījumos pietiek ar vērtību sakrišanu, lai apgalvotu, ka vienādas vērtības norāda uz vienu un to pašu objektu. Ņemot vērā piedāvātās pieejas mērķi kvalitātes analīzi kontroli nodot galalietotāja rokās, arī šajā gadījumā lietotājs patstāvīgi pieņem lēmumu, vai pietiek ar vērtību sakrišanu, vai vienlaicīgi ir jāizpildās arī citiem nosacījumiem, piemēram, veicot arī citu parametru vērtību savstarpēju salīdzinājumu.

Atbilstoši (Bertossi et al., 2016) pirms datu objekts tiek iesaistīts primārā datu objekta kvalitātes analīzē kā sekundārais, ir svarīgi pārliecināties, ka tie ir savietojami. Savā darbā (Bertossi et al., 2016) autori skaidro kontekstuālo kvalitātes analīzi slēgtajiem datiem, kas atbilst arī piedāvātas pieejas idejai. Autori pārbauda datubāzes D kvalitāti pret ārējo no datubāzes D neatkarīgu datu avotu C, kurš veido datubāzes D kontekstu, veidojot datubāzes D kartējumu pret ārējās datubāzes kontekstuālo shēmu pie nosacījuma, ka tās ir savietojamas un to savstarpējais salīdzinājums ir iespējams. Tas nozīmē, ka datu kvalitātes novērtējums balstās

uz primārās datu kopas salīdzinājumu ar ārēju, no kā seko, ka novērtējuma rezultāti ir atkarīgi no ārējas datu kopas, t.i. tās kvalitātes, detalizācijas līmeņa utt.. Kontekstuālās analīzes gadījumā primārajam datu objektam atbilstošo sekundārā objekta meklēšana un tā kvalitātes novērtējums ir problemātiskākie un resursietilpigākie soļi.

Piemēram, analizējot datu objektu "Latvijas Universitāte", kontekstuālās pārbaudes var būt nepieciešamas, veicot lauku "fakultāte", "juridiskā adrese" un "iela" kvalitātes pārbaudi, jo ir nepieciešams pārliecināties, ka šo lauku vērtības atbilst reālas pasaules situācijai, t.i. šie objekti eksistē, piemērām, citās datubāzēs, tie ir ticami un atbilst konkrētam ierakstam.

Cits piemērs, kas izriet datu objekta definēšanas procesa izpildes rezultātā, var būt *Google Maps* piedāvāta maršruta kvalitātes pārbaude pret ar sabiedrisko transportu izbraukto maršrutu, ar mērķi noteikt, vai *Google Maps* ir korekti attēlojis visas pieturas. Piemēram, plānojot braucienu no punkta *A* līdz punktam *B, Google Maps* piedāvā to realizēt, braucot ar transportu *X*, uzskaitot pieturas, kas atrodas starp attiecīgajiem maršruta punktiem. Viens no galalietotāja lietošanas piemēriem (ko ir nodefinējis prof. Ambainis) varētu būt: pārbaudīt ar *Google Maps* sagatavota maršruta kvalitāti un tajā iekļauto pieturu skaitu, ko var papildināt gan ar pieturu nosaukumiem, gan ar atbraukšanas laikiem vai attālumu starp tiem, pret reālo situāciju, dodoties no punkta *A* uz punktu *B* ar piedāvāto transportu *X*. Brauciena jeb procesa rezultāta iegūtie dati, kas tiks izmantoti datu kvalitātes pārbaudē, veido sekundāro datu objektu, t.i. datu objektu, pret kuru tiks pārbaudīta primārā datu objekta jeb *Google Maps* maršruta kvalitāte. Sekundārā datu objekta iegūšanas veids ir atkarīgs no datu uzkrāšanas veida, sākot ar triviālāko – novērojums, ko veic galalietotājs, fiksējot datus, vai citādi, piemēram, automatizēti, t.i. ar navigatora palīdzību, vēlāk izgūstot datus, kas, visticamāk būs pieejami daļēji strukturētā veidā. Tad atkarībā no lietošanas piemēra un no tā izrietošām prasībām, atbilstoši kurām ir jāpārbauda primāra datu objekta kvalitāte, primārā datu objekta kvalitāte tiek pārbaudīta pret sekundāro, pārbaudot *Google Maps* maršruta kvalitāti pret reālās pasaules situāciju.

Kontekstuālās datu kvalitātes pārbaudes nepieciešamība ir novērojama arī Lielbritānijas Uzņēmumu reģistra gadījumā, kurā atbilstoši (Nikiforova et al, 2019) un (Nikiforova, 2019b) ir nepieciešama valsts nosaukumus saturējušu parametru [*CountryOfOrigin*] un [*RegAddress Country*] vērtību kontekstuālā pārbaude. Datu kvalitātes analīze viena datu objekta ietvaros neļauj pieņemt viennozīmīgu lēmumu par konkrēto vērtību kvalitāti, rezultējot tikai potenciāli nekvalitatīvus ierakstus un vērtības. Lai pieņemtu lēmumu par to kvalitāti ir nepieciešama vērtību salīdzināšana pret valsts nosaukumiem, kas atbilst standartiem, t.i. datu objektiem, kuros esošajiem datiem pēc noklusēšanas ir jābūt kvalitatīviem. Šīm nolūkam autore izveidoja datu objektu "Country", kura parametri *ISO, ISO2, ISO3, UNI, UNDP* atbilst noteiktam valsts nosaukumu standartam (FAO, 2019).

Papildus datu objektu ieviešana piedāvātājā risinājumā paredz jaunu definīciju ieviešanu, iedalot datu objektu primārajā un sekundārajā. Par primāro datu objektu tiek uzskatīts datu objekts, kura kvalitāte tiek analizēta – tas ir centrālais datu kvalitātes analīzes objekts. Datu objekts tiek uzskatīts par sekundāro datu objektu, ja tas veido analizējama jeb primārā datu objekta kontekstu.

Gan primāro, gan sekundāro datu objektu definē galalietotājs, līdz ar ko visi primārā datu objekta izveides principi un tā raksturīpašības ir attiecināmi arī uz sekundāro datu objektu. Atbilstoši (Nikiforova et al., 2019) un (Nikiforova, 2019b) arī datu kvalitātes analīzē iesaistīto sekundāro datu objektu skaits ir atkarīgs no galalietotāja un ar viņu definētā lietošanas piemēra rakstura. Tāpat to skaitu nosaka primārā datu objekta un to parametru raksturs – cik parametriem var tikt nodefinēti un piekārtoti sekundārie objekti, vai tiem vispār ir iespējams tos piekārtot utt.. Primārais datu objekts tipiski ir viens – datu kvalitātes analīzes centrālais objekts, kura kvalitāte interesē galalietotāju, kas var tikt saistīts ar neierobežotu, bet galīgu sekundāro datu objektu skaitu.

Sekundāro datu objektu var veidot (a) cita no primārā datu objekta neatkarīga datu kopa, (b) no primārā datu objekta izgūtais datu objekts, kas var tikt izmantots pieļaujamo vērtību pārbaudei. Otrais variants ir paredzēts atbilstošās datu kvalitātes prasības vienkāršošanai, izslēdzot iespēju iekļaut visas pieļaujamas vērtības *SQL* vaicājumā, būtiski paplašinot to. Sekundārā objekta, kas tiek aizpildīts ar visām pieļaujamām vērtībām, definēšana nodrošina arī to atkalizmantošanu citās datu kvalitātes analīzēs.

Apskatītā piemēra ietvaros, primārā datu objekta analīzei pret sekundāro, tāpat kā viena datu objekta ietvaros, var tikt definētas vairākas kvalitātes prasības, piemēram, atbilstoši (Nikiforova, 2019b):

1) reģistra esošajam valsts nosaukumam ir jāatbilst vismaz vienam valsts nosaukumu standartam. Tas ļauj pārliecināties, ka Lielbritānijas Uzņēmumu reģistra esošās vērtības ir derīgas - atbilst reālai pasaulei;
2) visiem reģistra esošajiem valsts nosaukumiem vienas datu kopas ietvaros ir jāatbilst vienam standartam. Tas ļauj pārliecināties, ka datu kopā noteiktajos parametros ir nodrošināts datu viendabīgums. Atkarībā no galalietotāja arī šajā gadījumā ir iespējami divi varianti:
    2.1) atbilst vienam no vispārpieņemtajiem standartiem;
    2.2) atbilst ar lietotāju noteiktam vispārpieņemtam standartam.

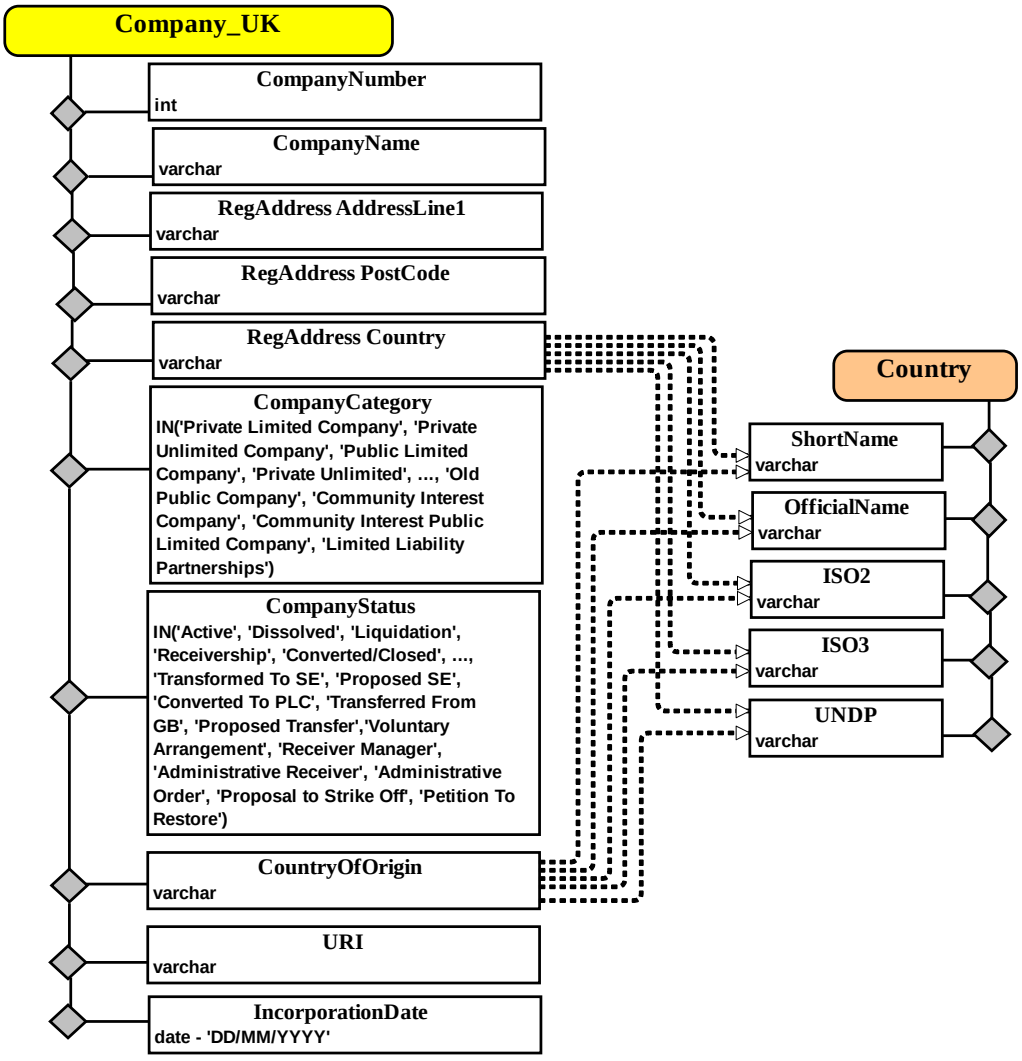
Kontekstuālā datu kvalitātes pārbaudi autore veica atbilstoši 1. un 2.1. prasībām, lai gan datu kvalitātes jēdzienam tā tradicionālajā izpratnē 1. variants, t.i. vērtību ticamības pārbaude, neatbilst.

Dotā kvalitātes kontekstuālā pārbaude paredz iepriekšējās apakšnodaļās pieejamo diagrammu modifikācijas atbilstoši dotajam uzdevumam, ievērojot tradicionālus kontekstuālās datu kvalitātes pārbaudes principus ((Scannapieco et al., 2005), (Bertossi et al., 2016)). Datu objekta definēšanas posmā, pārbaudot analizētas datu kopas kvalitāti pret citu datu kopu, papildus primārajam datu objektam tiek definēts(-i) arī sekundārais(-ie) datu objekts(-i). Sniegtajā piemērā ir tikai viens sekundārais datu objekts, taču (Nikiforova, 2019a) ir parādīta primārā datu objekta kvalitātes analīze pret 3 sekundārajiem datu objektiem. Grafiskas attēlošanas ziņā sekundārais objekta definēšana pilnībā atbilst primārā datu objekta, piešķirot sekundārajām datu objektam citu krāsu. Primārā un sekundārā(-o) datu objekta(-u) savstarpējas attiecības tiek attēlotas ar bultām. Tiek atzīmēta vairāku datu objektu savstarpēja saistība, norādot, ar kuru sekundāra datu objekta parametru ir saistīts konkrēts primārā datu objekta parametrs. Detalizētāka datu objektu parametru saistība tiek definēta ar kvalitātes prasību palīdzību nākamajā, t.i. datu kvalitātes specifikācijas definēšanas posmā, kurā tiek definētas prasības pret primāra datu objekta lauku kvalitāti sekundārā datu objekta kontekstā. Ir jāatzīmē, ka tas ir tikai vienas no iespējamajiem datu objektu attēlošanas veidiem, kas ir piedāvāts dotā risinājuma ietvaros, kas atšķiras no sākotnējas risinājuma versijas, ko darba autore piedāvāja (Nikiforova et al., 2019), kur tika iezīmēta datu objektu savstarpēja saistība, nenorādot uz sekundārā datu objekta parametriem, ar kuriem ir saistīts atbilstošs primārā datu objekta parametrs, savstarpējo divu datu objektu saistību un attiecību raksturu attēlojot tikai 2. posmā. Savukārt piedāvātais attēlošanas veids nodrošina labāku detalizācijas līmeni, ļaujot lietotājiem "izteikt" vairāku datu objektu savstarpējas attiecības jau 1. posmā.

Visi turpmākajos soļos definētie nosacījumi attiecās galvenokārt uz primāro datu objektu - piedāvātais risinājums neparedz sekundārā datu objekta kvalitātes pārbaudi, jo tas ir palīglīdzeklis primārā datu objekta datu kvalitātes analīzei. Ņemot vērā sekundārā datu objekta datu kvalitātes nozīmīgumu primārā datu objekta analīzei, ir paredzēts, ka tā kvalitāte tika pārbaudīta iepriekš, definējot to kā primāro datu objektu, vai arī uzskatot to par pietiekoši kvalitatīvu pārbaužu veikšanai pret to.

3.6.1. att. ir nodefinēti primārais datu objekts "Company_UK" un sekundārais datu objekts "Country". Atbilstošie primārā datu objekta parametri [*RegAddress Country*] un [*CountryOfOrigin*] ar bultu palīdzību ir sasaistīti ar sekundārā datu objekta "Valsts" ("Country") attiecīgajiem parametriem. Atbilstoši iepriekšnodefinētājam 1. lietošanas piemēram – primārā parametra vērtībai ir jāatbilst vismaz kādai no sekundārā datu objekta parametra vērtībai, abi analizētie primārā datu objekta parametri ir saistīti ar katru sekundārā datu objekta parametru. Atkarībā no lietošanas piemēra katrs primārā datu objekta parametrs varētu tikt saistīts ar dažādiem sekundārā datu objekta parametriem vai arī nevienu no tiem. 2.

lietošanas piemēra gadījumā – primārā datu objekta parametra vērtībai ir jāatbilst ar lietotāju noteiktam vispārpieņemtam standartam, katrs primārā datu objekta parametrs tiktu saistīts ar vienu sekundārā datu objekta parametru. Ir jāatzīmē, ka, mainoties lietošanas piemēram, attiecīgā diagramma varētu tikt pielāgota tam jebkurā datu kvalitātes analīzes posmā. Savukārt rodoties nepieciešamībai veikt datu kvalitātes analīzi atbilstoši vairākiem lietošanas piemēriem, vienreiz izveidota un saglabāta diagramma var tikt atkalizmantota, veicot tājā atbilstošās izmaiņas, neprasot lietotājam tās atkārtotu definēšanu.

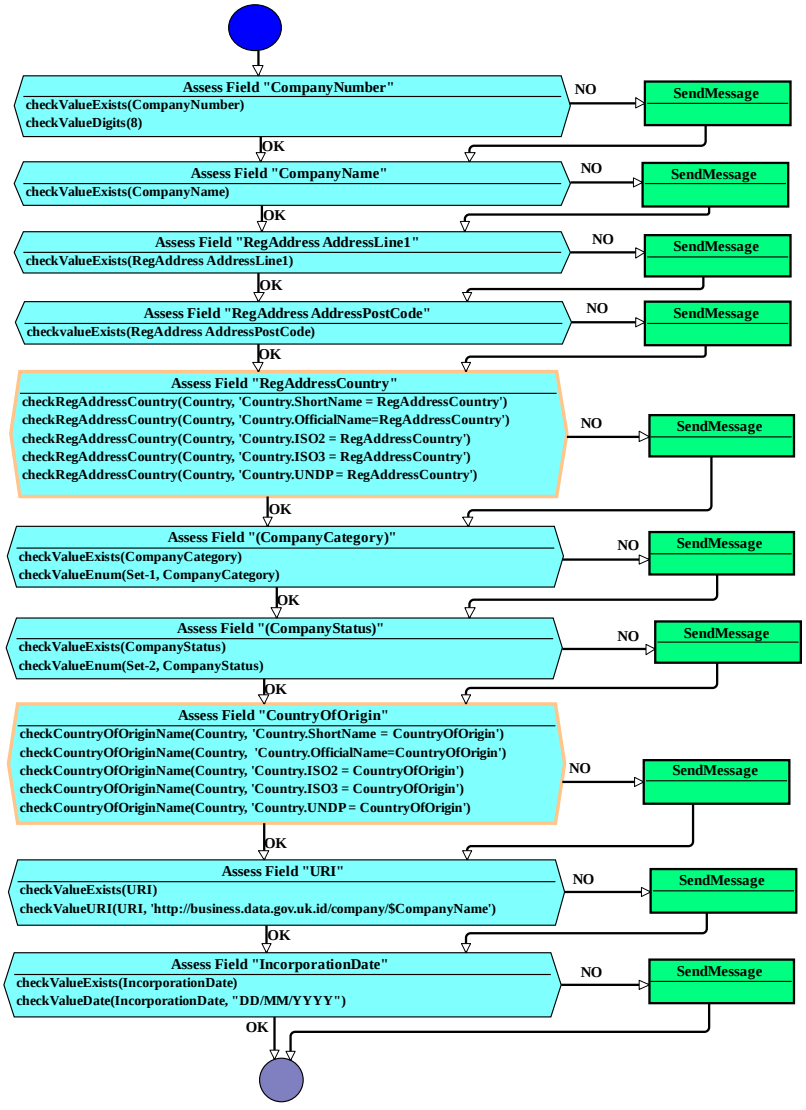


3.6.1. att. **Datu objekts "Lielbritānijas Uzņēmums" [izveidoja autore]**

Datu kvalitātes prasību specifikācijas diagramma kontekstuālās pārbaudes gadījumā 3.3. apakšnodaļā aprakstītā diagramma tiek papildināta ar datu kvalitātes prasībām primārā datu objekta attiecīgo parametru pārbaudi pret sekundāro datu objektu parametriem. Prasību formulējuma specifika atbilst viena datu objekta kvalitātes analīzēs pārbaudes ietvaros formulētājām prasībām. Definējot datu kvalitātes prasības primārā datu objekta parametru

vērtībām sekundārā datu objekta(-u) kontekstā, ir jāatceras, ka dažas viena datu objekta ietvaros veikta datu kvalitātes prasības var tikt ne tikai papildinātas, bet arī aizstātas ar citām. Piemēram, analizēto parametru gadījumā, iepriekšdefinētas vispārīgas prasības "*checkValueExists(RegAddressCountry)*" un "*checkValueMinLength(2)*" tiek aizstātas ar pārbaudēm pret sekundāro datu objekta parametru vērtībām, līdz ar ko sākotnējas pārbaudes kļūst liekas – konkrēta lietošanas piemēra ietvaros nav nepieciešamības pārbaudīt vērtības esamību un garumu, ja tiek pārbaudīta vērtību pareizība pret standartiem atbilstošām vērtībām.

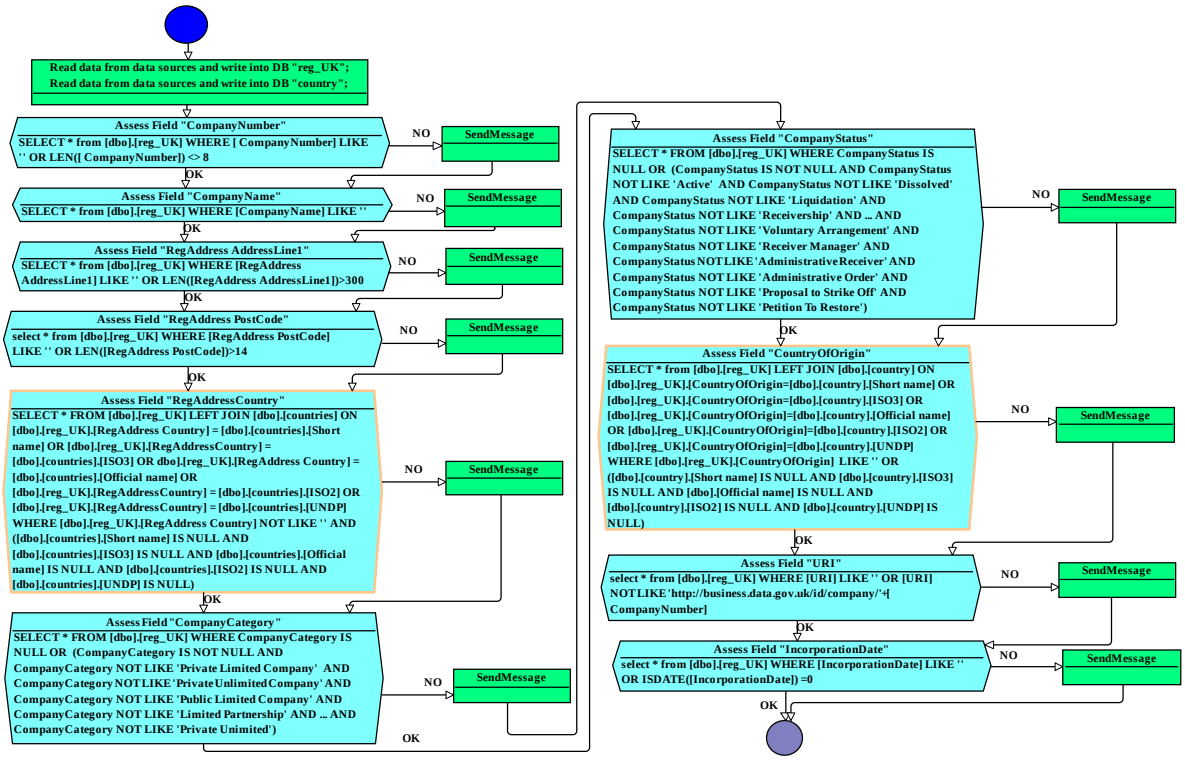
Lai grafiski izceltu parametrus, kuriem tiek veida konteksta pārbaudes, atbilstošu elementu apmalēm ir piešķirta attiecīga sekundāra datu objekta krāsa, kas tika piešķirta tam datu objekta definēšanas posmā. 3.6.2. att. attēlotais modelis atbilst *PSM* modelim.



3.6.2. att. **Datu kvalitātes prasības datu objektam "Lielbritānijas uzņēmums" [izveidoja autore]**

Datu kvalitātes pārbaudes posmā ar loģisko izteiksmju palīdzību definētās datu kvalitātes prasības tiek aizstātas ar izpildāmām (att. 3.6.3.). Atbilstošās izmaiņas diagrammā ir: (1)

diagrammas papildinājums ar datu nolasīšanas no sekundārā datu avota operāciju, (2) datu kvalitātes pārbaudes definēšana atbilstošajiem parametriem.



3.6.3. att. **Datu kvalitātes prasību izpildes pārbaude datu objektam "Uzņēmums" [izveidoja autore]**

Tā kā visas kvalitātes pārbaudes tiek veiktas primārajam datu objektam, arī izpildāmās prasības ir definētas primārā datu objekta elementos. Datu kvalitātes pārbaude pret sekundāro datu objektu paredz vairāku datu objektu savstarpēju saistību aprakstu. Šīm nolūkam tāpat kā 3.4. apakšnodaļā tiek izmantoti *SQL* vaicājumi, kuros vairāku datu objektu savstarpējai saistībai pēc noteiktiem parametriem ir paredzēts *JOIN* operators, kura piemērotība šīm uzdevumam atkārtoti parāda *SQL* atbilstību dotā uzdevuma nostādnei.

Tādējādi piedāvātais modelis ir paplašināts ar kontekstuālās datu kvalitātes pārbaudes iespēju, primārā datu objekta datu kvalitāti pārbaudot pret neierobežotu sekundāro datu objektu skaitu. Tas ļauj veikt dziļāku datu kvalitātes analīzi, ieguldot tajā mazāk resursu. Šīs iespējas priekšrocības, izsakot tās kvantitatīvi, ir pieejami 5.1. apakšnodaļā un rakstā (Nikiforova et al., 2019).

Kopumā kontekstuālo datu kvalitātes analīzi darba autore veica 18 datu kopām, kvalitātes problēmas konstatējot 17 no tām (94.4%), 18. datu kopā kvalitātes problēmas nekonstatējot pēc datu paraugu saskaņošanas to turpmākai analīzei, kas arī mēdz tikt uzskatīts par datu kvalitātes problēmu. Parametru skaita ziņā kontekstuālo analīzi darba autore veica 61

parametram, 83.6% parametros konstatējot vismaz dažas datu kvalitātes problēmas. Tā kā identificētās problēmas netika konstatētas datu kvalitātes analīzes viena datu objekta ietvaros rezultātā - tikai atsevišķu pārbaužu ietvaros, konstatējot potenciālās datu kvalitātes problēmas, nespējot pieņemt lēmumu par to neatbilstību reālai pasaulei, ir pamats apgalvot, ka kontekstuālās pārbaudes iespēja ļauj būtiski uzlabot datu kvalitātes rezultātus, nodrošinot iespēju veikt padziļinātu un visaptverošāku datu kvalitātes analīzi. Tas nozīmē, ka 6. darba sākumā izvirzītā tēze ir apstiprināta. Detalizētāk konstatētas datu kvalitātes problēmas un to raksturs dažādiem datu avotiem ir apskatīti 5. nodaļā.

3.7. Apkopojums

Piedāvātais datu objekta virzītais risinājums kardināli atšķiras no eksistējošiem datu kvalitātes risinājumiem. Tā ideja nav sastopama citos darbos, par ko liecina gan eksistējošo risinājumu patstāvīgā analīze, gan *Batini* - datu kvalitātes jautājumos pasaules vadošā pētnieka, datu kvalitātes problēmas dziļš izpētes darbs un eksistējošo metodoloģiju pārskats, kas ir publicēts vairākās grāmatas un zinātniskajos rakstos (Batini et al., 2006, 2009, 2016). Tajā pašā laikā jāatzīmē, ka piedāvātais risinājums ir vienkāršs un pat intuitīvs, jo ir tuvs datu un datu kvalitātes raksturam.

Piedāvātais risinājums ir "ārējais" mehānisms, kas ļauj veikt datu kvalitātes analīzi datu lietotājiem, nezinot kā dati tika uzkrāti un apstrādāti pie datu sniedzējiem. Tas nozīmē, ka tā: (a) var tikt pielietota "trešo pušu" datiem, kas var būt gan "slēgtie", gan "atvērtie" dati, kas mūsdienās ir populāri un kļūst arvien plašāk izplatīti visā pasaulē, tajā skaitā arī Latvijā; (b) ir domāta gan datu sniedzējiem, gan datu lietotājiem. Ir jāatzīmē, ka dotā pieeja ir paredzēta gan strukturēto, gan daļēji strukturēto datu kvalitātes analīzei.

Ņemot vērā pieejas pamatjēdzienu vienkāršību, to viennozīmīgu un skaidru definīciju, grafisko *DSL* izmantošanu un divu modeļu iesaisti, ir pamats apgalvot, ka dotā pieeja ir piemērota lietotājiem bez padziļinātājām zināšanām IT un datu kvalitātes jomā, jo IT-speciālistu iesaiste var būt nepieciešama tikai beidzamajos posmos, neformālus aprakstus, aizstājot ar izpildāmiem. Piedāvātā pieeja atbalsta un pat sekmē lietotāju savstarpēju sadarbību, ļaujot jebkurā posmā nepieciešamības gadījumā iesaistīties arī vairākām personām. Pie tam, jebkura veida izmaiņas var tikt iniciētas tiklīdz rodas nepieciešamība tajās, t.i. jebkurā datu kvalitātes analīzes posmā. Tādejādi ir panākts risinājuma "elastīgums". Tādējādi darba tiek apstiprināta darba sākumā izvirzītā 4. tēze.

Dotā pieeja paredz datu objekta un datu kvalitātes specifikāciju definēšanu, kas ir pilnībā atkarīgas no konkrētā datu lietotāja un lietošanas piemēra, kas nodrošina datu kvalitātes analīzes precīzu atbilstību datu lietotāja vēlmēm un vajadzībām.

Ir jāatzīmē arī tas, ka piedāvātā risinājuma potenciāls neierobežojas ar atsevišķo datu kopu kvalitātes analīzi, jo (Bicevskis et al., 2019b) ir aprakstīta datu kvalitātes izpildlaika (angl. *runtime*) verifikācija, kas balstās uz piedāvātā kvalitātes modeļa. Tas ļauj pārliecināties datu kvalitātē gandrīz reālā laikā, veicot datu kvalitātes analīzi darījumprocesa izpildes laikā. Tas nozīmē, ka kļūst iespējams pārbaudīt, ka konkrēts process nesabojāja sistēmā esošos datus, bet gadījumā, ja kāda darījumprocesa izpildes gaitā tika pārkāptas kādas datu kvalitātes prasības, šīs darījumprocess tiek identificēts tāpat ka nekvalitatīvi dati datu kopas kvalitātes analīzes gadījumā. Tas ļauj nodrošināt nepārtrauktu datu kvalitātes analīzi. Tas apstiprina darba sākumā izvirzīto 8. tēzi, t.i. izstrādātais risinājums ir pielāgojams izpildlaika datu kvalitātes analīzei.

Piedāvātajam risinājumam ir arī savi ierobežojumi. Pirmkārt, dotais risinājums neparedz modeļu elementos ierakstīta teksta derīguma pārbaudes, kas atbilstoši iepriekšrakstītajam būtu jānodrošina "gudrajiem lietotājiem" (Bicevskis et al., 2018a). Šīs ierobežojums tiek pamatots ar *DIMOD* rīka un pētījumā Nr. 1.8 "Datu kvalitātes pārvaldība ar izpildāmiem biznesa procesu modeļiem" ietvaros izstrādāta kompilatora ierobežojumiem.

Otrkārt, dotais risinājums ir paredzēts strukturēto un daļēji strukturēto datu kvalitātes analīzei, no kā seko, ka tas nav paredzēts nestrukturēto datu kvalitātes analīzei. Tā kā dotais risinājums ir domāts lietotājiem bez padziļinātām zināšanām IT un datu kvalitātes jomās, arī risinājumam bija jābūt pēc iespējas vienkāršam un galalietotājam pielāgotam, savukārt risinājuma pielāgošana arī nestrukturēto datu analīzei pārlieku sarežģītu piedāvāto risinājumu. Taču ir jāatzīmē, ka atsevišķos gadījumos tas var tikt pielietots arī nestrukturētajiem datiem. Piemēram, ja, apstrādājot tekstu, galalietotājs vēlās pārliecināties, ka minētie fakti ir semantiski pareizi, ir iespējams nodefinēt datu objektu, kas saturēs tos atribūtus, kuru vērtības tiks analizētas, atbilstošās vērtības saglabājot tajā, un turpmāk veicot to analīzi atbilstoši iepriekšaprakstītajai procedūrai. Tas būtu īpaši pamatoti gadījumā, ja tiek apstrādāts teksts, kurā parādās vairāki vienveidīgie objekti un to apraksts – piemēram, dažādu valsts apraksti, ieskaitot to nosaukumu, galvaspilsētu, platību, valsts valodu utt.. Nestrukturētu datu saglabāšanu datu objektā var automatizēt, izmantojot kādu ārēju risinājumu (viens no tādiem risinājumiem ir tapis Latvijā, LU MII sadarbojoties ar SIA "Tildes").

Treškārt, dotais risinājums lielākoties ir piemērots datu kvalitātes analīzei, nefokusējoties uz datu kopu atbilstības atvērto datu principiem analīzi, piemēram, metadatu analīzi. Taču, balstoties uz diskusiju ar Beļģijas Universitātes prof. *Marc Nyssen* attiecībā uz autores rakstu (Nikiforova, 2019a), neskatoties uz to, ka dotais risinājums sākotnēji nebija paredzēts metadatu

kvalitātes analīzei, ir jāatzīmē, ka, uzskatot konkrētu datu kopu aprakstošus metadatus par datu objektu, kuram tiek izvirzītas lietotāju interesējošās datu kvalitātes prasības, metadatu kvalitāte var tikt novērtēta atbilstoši iepriekšaprakstītai procedūrai.

Vēl viens ierobežojums ir risinājuma praktiskais raksturs, taču nākamajā nodaļā tiek piedāvātā šī risinājuma formalizācija, ar mērķi pārveidot to datu kvalitātes teorijā.

4. PIEDĀVĀTĀS PIEEJAS FORMALIZĀCIJA

Neskatoties uz lielu datu kvalitātes risinājumu skaitu, kas tika izstrādāti pēdējo dekāžu laikā, datu kvalitātes problēmas pētnieki nebija spējīgi piedāvāt datu kvalitātes teoriju.

Viens un, iespējams, galvenais iemesls, kāpēc datu kvalitātes teorija līdz šīm netika piedāvāta, neskatoties uz lielu mēģinājumu skaitu, ir tas, ka jebkuras teorijas pamatā ir jābūt skaidri un viennozīmīgi definētiem jēdzieniem. Atbilstoši Nacionālajai Enciklopēdijai jēdzienu "teorija" lieto, lai "aplūkotu noteiktu definīciju, aksiomu, teorēmu, piemēru vai modeļu kopumu, kas var būt gan plašāks, gan šaurāks jēdziens par apakšnozari ..." (Bula, 2019). Ņemot vērā, ka tradicionāli datu kvalitāti saista ar datu kvalitātes dimensijas jēdzienu, kuram ir raksturīgs universālās definīcijas, klasifikācijas un mērīšanas mehānismu trūkums (kritika ir pieejama arī 2. nodaļā), šī pamatprasība nevar tikt izpildīta. Dotā pieeja neizmanto datu kvalitātes dimensiju jēdzienu, tā vietā izmantojot vispārīgāku jēdzienu "datu kvalitātes prasība", citādi ievērojot visus vispārpieņemtās datu kvalitātes un ar to saistīto jēdzienu definīcijas, kas ļauj veikt pieņēmumu, ka šīs pieejas formalizācija varēs kalpot par neformālu datu kvalitātes teoriju.

Kā seko no izmantotās "teorijas" jēdziena definīcijas, vispirms, ir jānodefinē piedāvātā risinājuma komponenti. No iepriekšējas nodaļas seko, piedāvātais kvalitātes modelis sastāv no trim pamatkomponentiem:

1) **datu objekts** - reālās pasaules objektu raksturojošo parametru vērtību kopa, kas definē tos un tikai tos datus, kuru kvalitāte tiks analizēta konkrētā lietošanas piemēra ietvaros. Datu objekts mēdz būt (a) **primārais**, (b) **sekundārais**. Gan primārajam, gan sekundārajam objektam mēdz būt **apakšobjekti**. Kā tika minēts iepriekšējā nodaļā vienādas struktūras objektu kolekcija veido **datu objektu klasi**;
2) **datu kvalitātes specifikācija** – visas tās prasības, kurām ir jāatbilst datiem, lai tie tiktu atzīti par kvalitatīvajiem;
3) **datu kvalitātes pārbaudes process** – visu to darbību kopums, kas ir jāveic, lai novērtētu datu atbilstību izvirzītājām kvalitātes prasībām, secinot par to kvalitāti.

Visi komponenti tiek reprezentēti ar grafiskām *DSL*.

4.1. Datu objekta formalizācija

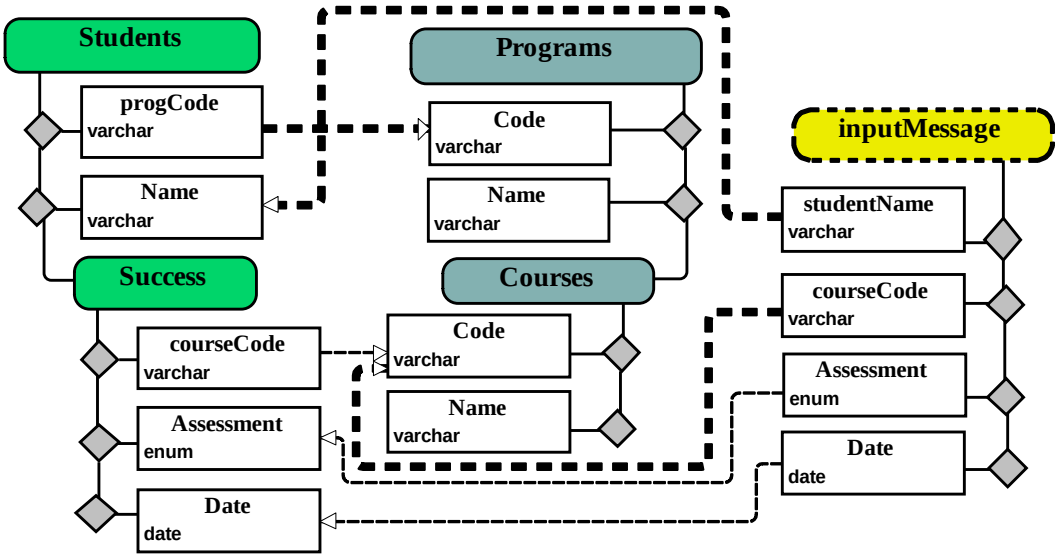
Datu objekts ir raksturots ar tā atribūtiem, kuru apraksts ir neformāls, jo noteikumi to vērtību sintaksei netiek definēti. Par individuāla datu objekta atribūta vērtības adresi kalpo **<*dataObjectName.attributeName*>**. Datu objekta atribūtu adreses ir izmantotas datu kvalitātes nosacījumos.

Viena datu objekta atribūta vērtības mēdz būt saistītas ar cita datu objekta atribūtu vērtībām, kas ir ievadītas datubāzē jau iepriekš. Piemēram, pievienojot datubāzei datus par studenta nokārtotu eksāmenu, datubāzē ir jābūt ierakstam par studentu, kārtoto kursu utt.. Datu objekts, kura kvalitāte tiks analizēta, tiek uzskatīts par **primāro datu objektu**, bet pārējie, kas nosaka kontekstu, t.i. tie, pret kuriem tiek pārbaudīta primārā datu objekta kvalitāte - par **sekundāriem datu objektiem**. Parasti primārais datu objekts ir viens, jo tas ir centrālais datu kvalitātes analīzes objekts, taču sekundāro datu objektu skaitu nosaka primārā datu objekta raksturs un konkrēts lietošanas piemērs. Abu datu objektu veidu gadījumā **datu apakšobjektu** skaits nav ierobežots.

Visi piedāvātā risinājuma komponenti tiek aplūkoti uz konkrētā piemēra, par pamatu ņemot "tradicionālo" studiju piemēru. Piemērā 4.1.1. attēlā ir attēloti datu objekti *inputMessage*, *Students* un *Programs*. Datu objektiem *Students* un *Programs* ir datu apakšobjekti.

Datu objekts *inputMessage* ir individuāls datu objekts, bet pārējie – datu objektu klases:

- datu objekts *inputMessage* satur informāciju par konkrēta studenta nokārtotu eksāmenu;
- datu objekts *Students* ir datu objektu klase, kura katra instance satur informāciju par vienu konkrētu studentu, raksturojot to ar parametru *progCode*, kas norāda apmācības programmu, kurā students ir reģistrējies apmācībai, un *Name* - studenta identificējošu vārdu, vērtībām. Papildus tam katra instance satur datu apakšobjektu *Success* – zināšanu novērtējums, kur katra instance satur datus par viena konkrēta studenta nokārtotiem eksāmeniem. Atbilstošie atribūti ir *courseCode* – kursa kods, kurš ir jākārto studentam, *Assessment* – vērtējums, *Date* – datums;
- datu objekts *Programs* ir datu objektu klase, kura katra instance satur datus par vienu apmācības programmu un apakšobjektu *Courses*, kas satur programmā ietilpstošos apmācības kursus, kurus studentam studiju laikā ir jānokārto.



4.1.1. att. **Datu objektu definēšana [pēc (Bicevskis et al., 2019a) modificēja autore]**

Relācija starp datu objekta *inputMessage* atribūtu *studentName* un datu objekta *Students* atribūtu *Name* ir attēlota ar raustītu līniju, un nozīmē, ka datu objekta *inputMessage* atribūta *studentName* vērtībai ir jāatbilst kādai no datu objekta Students atribūta *Name* vērtībām.

4.2. Datu kvalitātes prasību "pirms-" un "pēc-" nosacījumi

Datu kvalitātes prasības individuālam datu objektam tiek uzdotas ar loģisku izteiksmju palīdzību, kur par loģisko izteiksmju operandiem kalpo individuāla datu objekta atribūtu/ lauku vārdi, bet operācijām var lietot gan tradicionālos programmēšanas valodās lietotos līdzekļus - loģisko izteiksmju operācijas, gan datu kvalitātei specifiskas operācijas (4.2.1. attēls).

Apstrādājot datu objektu klases, datu objektu klases instances tiek atlasītas, katrai instancei pārbaudot datu kvalitātes prasību izpildi, kas savukārt atbilst individuālā datu objekta apstrādei. Cikls instanču apstrādei definē lietotājs. Populāri ir 2 paņēmieni:

1) caurskatot visas klases instances, mainot adresi <***dataObjectName(instanceIdent).attributeName***>, kura tiek aprēķināta vispirms, izvēloties pirmo instanci ar metodi ***instanceIdent = getFirst(dataObjectName)***, kam seko pāreja pie nākošās instances ar metodi <***instanceIdent = getNext(dataObjectName)***>. Šis paņēmiens tiek lietots, ja datu kvalitāte jāpārbauda visām datu objekta instancēm;

2) izmantojot dinamiski izrēķināmu adresi <***instanceIdent = seekInst(dataObject, expression)***>, kur ***expression*** ir loģiska izteiksme ar operandiem, t.i. atribūtu vārdiem, un tradicionālām loģiskām operācijām. Ja

izpildes rezultātā tiek atrasta datu objekta instance, tad (1) norāde uz datu objektu tiek piešķirta mainīgajam *instanceIdent*, (2) videi tiek atgriezta vērtība *TRUE*. Citādi mainīgajam tiek piešķirta vērtība *NULL*, videi atgriežot vērtību *FALSE*. Šis paņēmiens tiek lietots, ja datu kvalitāte jāpārbauda vienai atsevišķai datu objekta instancei.

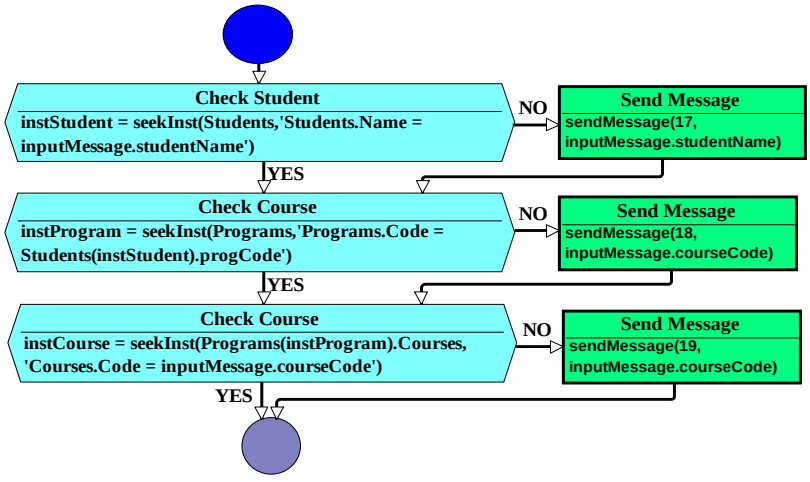
Kontekstuālo datu kvalitātes prasību iekļaušanai kvalitātes prasībās, nosacījumos tiek izmantotas sekundāro datu objekta parametru adreses ***<secondaryDataObjectName(instanceIdent).attributeName>***. Ja sekundārais datu objekts tiek meklēts pēc tā atribūta vērtības, tiek izmantota primārā datu objekta meklēšanai analoģiska sekundārā datu objekta meklēšanas komanda ***<instanceIdent = seekInst (secondaryObjectName, expression)>***.

Kā tika minēts iepriekš, apstrādājamo ziņojumu *inputMessage* ievadu datubāzē veic process, piemēram, *INSERT_data_into_DB*, kurš ir uzskatāms par "melno kasti", jo tā struktūra un funkcionēšana nav zināmi. Procesa izpildes rezultātā datubāzei jābūt korekti pievienotiem *inputMessage* datiem.

Šīm nolūkam tiek definēti divi izpildāmi procesi - *Pre-condition* un *Post-condition* – tā saucamie "pirms-" un "pēc-" nosacījumi, kas ļauj datu kvalitāti pārbaudīt gan pirms izmaiņas tika veiktas datu objektā, gan pēc tām. Piedāvātā piemēra ietvaros:

1) *Pre-condition* pārbauda (treknas raustītas bultas 4.1.1. attēlā):
   - vai eksistē students, uz kuru attiecas *inputMessage*;
   - vai students ir reģistrēts kādā apmācības programmā;
   - vai *inputMessage* norādītais kurss ir apmācības programmas kurss.

   Pre-condition process lasa datubāzē iepriekš uzkrātus datus, tos nemainot un tādejādi izpildāms pirms *INSERT_data_into_DB*.

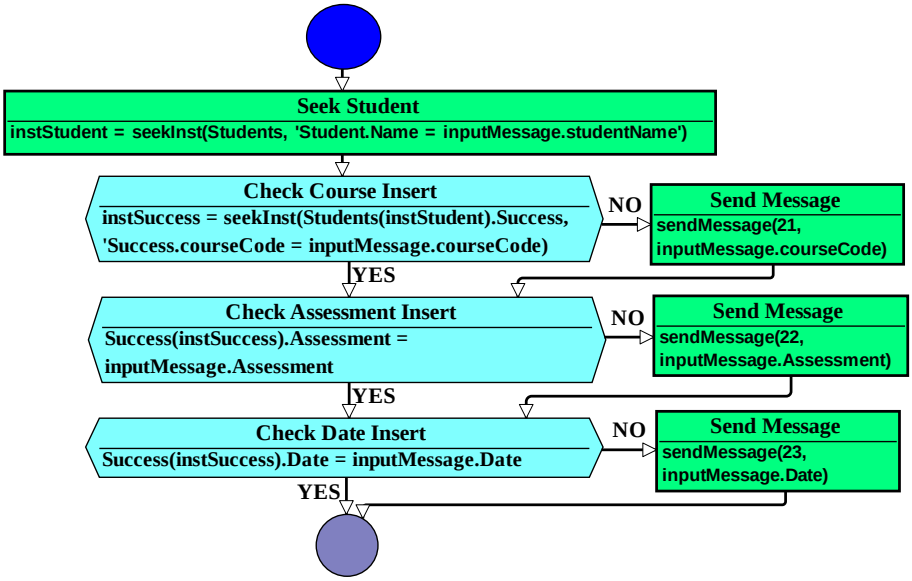


4.2.1. att. ***Pre-condition* kvalitātes definēšana [pēc (Bicevskis et al., 2019a) modificēja autore]**

2) *Post-condition* tiek izpildīts pēc *INSERT_data_into_DB* un pārbauda (tievas raustītas bultas 4.1.1. attēlā):

- vai datu objekta *Students* apakšobjektam *Success* ir pievienota jauna instance;
- vai datu objekta *Students* apakšobjektam *Success* ir pievienota jauna instance ar atbilstošā kursa vērtējumu;
- vai datu objekta *Students* apakšobjektam *Success* ir pievienota jauna instance ar atbilstošā eksāmena datumu.

  Arī *Post-condition* process tikai lasa datubāzē iepriekš uzkrātu informāciju, to nemainot.



4.2.2. att. ***Post-condition*** **kvalitātes definēšana [pēc (Bicevskis et al., 2019a) modificēja autore]**

Aplūkotajā piemērā tiek veikta datubāzes datu pārbaude gan pirms, gan pēc datu ievada datubāzē, pirms datu ievada pārbaudot datu ievada iespējamību, datus neievadot datubāzē - tikai lasot tos. Pēc datu ievada no datubāzes tiek nolasīts atbilstošs datu objekts, kas tiek salīdzināts ar ievadāmajiem datiem. Atšķirību gadījumā tiek identificēta nekorekta datu ievada programmas darbība.

Tādejādi datu ievads datubāzē tiek kontrolēts ar ārēju procedūru, kura, saņemot ievadāmo ziņojumu, pirms datu ievada datubāzē pārbauda izpildes priekšnosacījumus un pēc datu ievada datubāzē pārbauda, vai tas ir noticis korekti.

"Pirms-" un "pēc-" nosacījumu izmantošana ir populārs paņēmiens, kas ir sastopams gan modeļu pārbaudes risinājumos, gan testēšanā (Khurshid et al., 2003), pārbaudot specifikācijas pareizību. Piedāvātās idejas kontekstā un 3.7. apakšnodaļā minēta izpildlaika datu kvalitātes verifikācijas gadījumā par "pirms-" jeb priekšnosacījumiem kalpo katras iepriekšējas pārbaudes

rezultāti, kas tiek izmantoti kā ievaddati, savukārt "pēc-" nosacījumu pārbaude tiek veikta pēc pārbaudes izpildes. Tas nozīmē, ka 4 stāvokļu, kas ir raksturoti ar dažādu nosacījumu kopu, piemēram, "*Start*", "*Pause*", "*Resume*" un "*Finish*", kvalitātes pārbaudi sākot ar "*Start*", "*Pause*" gadījumā "*Start*" dati kļūst par priekšnosacījumiem, pret kuriem tiek pārbaudīta "*Pause*" datu kvalitāte.

Ir skaidrs, ka priekšnosacījumu izmantošana datu kvalitātes analīzē paredz to pareizību, precīzāk - augsto kvalitāti. Datu kvalitātes analīzē izmantotie priekšnosacījumi tiek uzskatīti par kvalitatīviem, jo izpildes laika datu kvalitātes verifikācijas gadījumā to kvalitāte tiek pārbaudīta katrā iepriekšējā solī. Savukārt pieņēmums, ka (a) formalizētā risinājuma gadījumā priekšnosacījuma izmantotie un (b) sākotnējie dati datu kvalitātes verifikācijas gadījumā (sauksim tos q_0) ir kvalitatīvi, tiek veikts, pieņemot, ka statiski glabājošos datu kvalitāte tiek pastāvīgi pārbaudīta un kā rezultāts, atbilstošie dati ir kvalitatīvi un var tikt izmantoti kvalitātes analīzē, veicot datu kvalitātes analīzi pret tiem.

Datu kvalitātes pārbaudes process notiek atbilstoši 3.5. apakšnodaļā aprakstītai procedūrai, atbilstoši kurai datu objekts vai datu objektu klase kalpo par datu kvalitātes pārbaudes procesa ievaddatiem, secīgi pārbaudot tam datu kvalitātes prasību izpildi. Ja dati neatbilst izvirzītājām kvalitātes prasībām, tiek izsaukta *SendMessage* metode. Kvalitātes pārbaudes procesa rezultātā tiek izveidots protokols, kurā ir apkopoti identificēti datu kvalitātes problēmas ziņojumi.

Kopumā datu kvalitātes pārbaudes process var tikt vienkārši izskaidrots matemātiskās loģikas jēdzienos:

- katra datu kvalitātes prasība var tikt apskatīta/ uztverta kā izteikums jeb propozīcija (angl. *proposition*), kurai var tikt piešķirta viena no divām patiesuma vērtībām, t.i., patiess jeb 1 vai aplams jeb 0;
- vairāk kā viena datu kvalitātes prasība, kas ir nodefinētas vienam konkrētam datu objektam, tiek grupētas, veidojot kvalitātes prasību kopu. Katra propozīcija tiek reprezentēta ar propozīcijas mainīgo $\mathbf{r_k}$, kur k ir konkrēta parametra kvalitātes prasības kārtas numurs;
- pārbaudot kvalitātes prasību izpildi, konkrēta parametra kvalitātes prasības tiek kombinētas, *SQL* vaicājumu gadījumā izmantojot "*OR*" operatoru. Tādā veidā nodefinētie viena parametra propozīcijas veido saliktu propozīciju, pielietojot bināro operatoru – loģisko konjunkciju (^);
- sešstūra elements, kas satur konkrēta datu objekta parametra kvalitātes prasības, kopā ar vienu no bultām, kas no tā iziet, var tikt uzskatītas par divnosacījuma (angl. *biconditional*) operatoru "tad un tikai tad" (angl. "*if and only if*") – "*IFF*".

Kvalitātes prasības, kas ir attēlotas vienā sešstūra elementā, veido salikto propozīciju, kas var tikt izskatīta par hipotēzi vai priekšnoteikumu (angl. *premise*), savukārt "*OK*" bulta, kas savieno divus sešstūra elementus, var tikt uztverta kā propozīcija "*datu objekta parametram kvalitātes problēmu nav*" (q) (formula 4.2.1). Šī propozīcija ir secinājums. Savukārt "*NO*" bulta ir šīs propozīcijas negācija – "*datu objekta parametram ir kvalitātes problēma(-s)*" ($\neg q$) (formula 4.2.2).

$$r_1 \wedge r_2 \wedge \ldots \wedge r_n \leftrightarrow q \quad (4.2.1)$$

$$r_1 \wedge \neg r_2 \wedge \ldots \wedge r_n \leftrightarrow \neg q \quad (4.2.2)$$

Tas nozīmē, ka atbilstoši patiesuma tabulām, visai izteiksmei ir piešķirta vērtība "patiess", ja katra prepozīcija ir patiesa. Ja vismaz viena propozīcija ir aplama, visa izteiksme ir aplama, un atbilstošs datu objekta parametrs tiek uzskatīts par nekvalitatīvu jeb vismaz vienu datu kvalitātes problēmu saturošu. Citos vārdos, ja $r_1..r_n$ ir patiesi, tad q arī ir patiess.

Ir jāatzīmē, ka šīs skaidrojums ir spēkā tikai kvalitātes prasību grafiskās reprezentācijas gadījumā - tas neattiecās uz visu datu objekta kvalitāti, jo atbilstoši piedāvātās pieejas pamatiem, konkrēta datu objekta datu kvalitātes novērtējums ir atkarīgs no lietotāja un lietošanas piemēra, pie tam, katrs lietotājs savas datu kvalitātes analīzes ietvaros var uzstādīt slieksni, kuru nepārsniedzot, datu objekts var tikt uzskatīts par kvalitatīvu un otrādi. Lēmumu par datu objekta kvalitāti galalietotājs pieņem, balstoties uz datu kvalitātes pārbaudes rezultātā iegūto protokolu (daļēji atbilst (Batini et al., 2016) "lēmumu stratēģijas" koncepcijai).

Dotā risinājuma formalizācijas ideja ir publicēta rakstā (Bicevskis et al., 2019a). Šī ideja apstiprina darba sākumā izvirzīto 9. tēzi, t.i. skaidri un viennozīmīgi nodefinētie piedāvātā datu kvalitātes modeļa komponenti un izstrādātā risinājuma specifika ļauj piedāvāt neformālu datu kvalitātes teoriju.

Datu kvalitātes modeļa implementāciju autore aplūkoja iepriekšējā nodaļā, savukārt nākamajā nodaļā ir apskatīti piedāvātā risinājuma pielietošanas rezultāti, analizējot "trešās puses" datus.

5. PIEEJAS PIELIETOŠANAS REZULTĀTI

Piedāvāto datu kvalitātes novērtēšanas pieeju darba autore pielietoja vairākām datu kopām, darbā apkopojot 26 datu kopu analīzes rezultātus (daži no tiem ir publicēti (Nikiforova, 2019a, 2018a, 2018b), (Nikiforova et al., 2019), (Bicevskis et al., 2018a, 2018b)). Tā kā 11 atvērto datu kopu analīzes rezultāti darba autore detalizēti apskatīja maģistra darbā (Nikiforova, 2019b), šī diskusija netiks atkārtota, sniedzot (1) Uzņēmumu reģistru analīzes datu kopu kvalitātes analīzes rezultātus apkopotā veidā, uzmanību pievēršot (a) iepriekšējā nodaļā apskatītam Lielbritānijas Uzņēmumu reģistram, (b) Latvijas Uzņēmuma reģistram, (2) īsi aprakstot pieredzi, kas gūta, analizējot citas atvērto datu kopas, izceļot izplatītākās atvērto datu kvalitātes problēmas, (3) uzmanību pievēršot vienu konkrētu domēnu reprezentējošām datu kopām un to kvalitātes analīzei – Latvijas atvērtie medicīnas dati. Ir jāatzīmē, ka analizētās datu kopas ir atvērtās datu kopas, ko sniedz dažādi datu sniedzēji, līdz ar ko analīzes rezultāti ļauj spriest par vispārīgo atvērto datu kvalitāti, kuras pakāpe nevar tikt saistīta ar datu sniedzēju. Datu kvalitātes līmeņa likumsakarības ar datu sniedzēju un datu uzkrāšanas veida jautājumu, t.i. centralizēti vai decentralizēti uzkrātie dati, darba autore apskatīja (Nikiforova, 2018a).

Ir jāatzīmē, ka dotā pieeja var tikt pielietota ne tikai atvērtajiem datiem, tā var tikt pielietota dažāda veida strukturētiem un daļēji strukturētiem datiem, taču atvērto datu analīze ļauj (a) pārbaudīt datu, kas ir brīvi pieejami lietotājiem, kvalitāti, novērtējot to lietošanas iespējamību un piemērotību lietotāja nolūkiem, (b) pielietot piedāvāto pieeju datiem, nepārkāpjot privātuma, drošības un privilēģiju ierobežojumus, vienlaicīgi parādot, ka dotā pieeja var tikt pielietota "svešiem" jeb "trešo pušu" datiem, nezinot kā tie tika uzkrāti un apstrādāti.

5.1. Uzņēmumu reģistru datu kvalitātes analīzes rezultāti

Veiktā četru Eiropas Uzņēmumu reģistru (Latvijas, Igaunijas, Norvēģijas un Lielbritānijas) salīdzinoša datu kvalitātes analīze ir publicēta (Bicevskis et al., 2018b) un (Nikiforova, 2018a), savukārt kontekstuālās analīzes rezultātu demonstrēšana uz Lielbritānijas Uzņēmumu reģistra piemēra ir publicēta (Nikiforova et al., 2019). Uzņēmumu reģistru pirmā posma datu kvalitātes analīzi darba autore veica, balstoties uz diviem visiem Uzņēmumu reģistriem nodefinētajiem lietošanas piemēriem, veicot to savstarpēju salīdzinājumu, otrajā analīzes posmā katra Uzņēmuma reģistra datu kvalitāti pārbaudot tikai viena Uzņēmuma

reģistra ietvaros, analizējot katru uzņēmumu raksturojošo parametru, tādejādi veicot reģistru padziļināto analīzi.

Ņemot vērā nodefinēto lietošanas piemēru vienkāršību, datu kvalitātes analīzē iesaistot tikai primāro atribūtu kvalitātes analīzi, darba autore izvirzīja pieņēmumu, ka datiem ir jābūt (a) pilnīgiem, (b) aizdomīgas vērtības nesaturošiem, (c) korektiem. Taču datu kvalitātes analīzes rezultāti pierādīja, ka šīs pieņēmums ir aplams.

Atbilstoši pirmajam lietošanas piemēram autore pārbaudīja, vai konkrētais Uzņēmumu reģistrs ļauj viennozīmīgi identificēt/ atrast jebkuru uzņēmumu pēc tā nosaukuma, reģistrācijas numura un dibināšanas datuma. Analīzes rezultāti, kas ir apkopoti tabulā 5.1.1. rāda, ka Lielbritānijas un Latvijas Uzņēmumu reģistros ir reģistrēti uzņēmumi, kuriem nav norādīts to nosaukums, kā arī ir konstatētas datu kvalitātes problēmas dibināšanas datuma parametros. Igaunijas un Norvēģijas Uzņēmumu reģistros problēmas netika konstatētas, taču ir jāatzīmē, ka Igaunijas Uzņēmumu reģistrs nesniedz datus par uzņēmumu dibināšanas datumu, līdz ar ko tas neatbilst lietošanas piemēram pilnā mērā. Atbilstoši (Nikiforova, 2019b) Norvēģijas un Lielbritānijas Uzņēmumu reģistros uzņēmuma dibināšanas datumu laukā 12 datumi ir aizdomīgi, jo atbilstoši tiem viens Norvēģijas uzņēmums tika reģistrēts "1277-09-13", savukārt Lielbritānijas – "25/04/1552", kas ir maziespējams.

5.1.1. tabula

Uzņēmumu reģistru datu kvalitātes analīzes rezultātu apkopojums (Nikiforova, 2019b)

Uzņēmumu reģistrs (valsts)	Nosaukums	Reģistrācijas numurs	Dibināšanas datums	Adrese	Pasta indekss
Lielbritānija	1 (0.0001%)	0	3 (0.0004%)	7 518 (0.997%)	12 151 (1.6%)
Latvija	10 (0.0025%)	0	94 (0.02%)	366 (0.09%)	20 498 (5.16%)
Igaunija	0	0	-	29 918 (11.24%)	22 621 (8.5%)
Norvēģija	0	0	9 (0.0008%)	68 128 (6.2%)	14 683 (1.3%)

Pārbaudot datu kvalitāti atbilstoši otrajam lietošanas piemēram, tika veiktas vienkāršas adreses un pasta indeksa parametru vērtību kvalitātes pārbaudes. Tika pārbaudītas (a) adreses vērtības esamība, (b) pasta indeksa vērtības esamība, un (c) pasta indeksa vērtības atbilstība noteiktam paraugam, katrai valstij nodefinējot tai atbilstošo formātu. Visos Uzņēmumu reģistros autore konstatēja vairākas datu kvalitātes problēmas. Atbilstoši (Nikiforova, 2018a)

attiecībā uz adreses lauku, labākus rezultātus uzrādīja Latvijas Uzņēmumu reģistrs (0.09% kvalitātes defektu), kuram seko Lielbritānija (0.997%), Norvēģija (6.2%) un sliktākais rezultāts - Igaunijai (11.24%). Pasta indeksu gadījumos Norvēģijas un Lielbritānijas Uzņēmumu reģistros ir identificētas vismazāk datu kvalitātes problēmu (1.3 un 1.6%), kuriem seko Latvijas Uzņēmumu reģistrs (5.16%), savukārt visaugstākais datu kvalitātes problēmu rādītājs ir Igaunijas Uzņēmumu reģistrā (8.5%).

Tas nozīmē, ka neviens Uzņēmumu reģistrs neapmierināja nevienu no ļoti vienkāršajiem/ intuitīvajiem un pat acīmredzamajiem lietošanas piemēriem, kuros tika analizētas primāro parametru vērtības. Taču Igaunijas un Norvēģijas Uzņēmumu reģistri var tikt izmantoti, lai viennozīmīgi identificētu jebkuru uzņēmumu pēc tā nosaukuma un reģistrācijas numura, jo tikai tie izturēja atbilstošo lauku kvalitātes pārbaudi. Savukārt arī citos laukos konstatēto datu kvalitātes problēmu esamība neliecina par to, ka atbilstošās datu kopas ir zemas kvalitātes un nevar tikt izmantotas lietotāju vajadzībām, jo konstatēto datu kvalitātes problēmu skaits nav pārāk liels, un to kvalitāte varētu tikt ātri uzlabota, piemēram, izmantojot doto pieeju. Šie uzlabojumi neprasa daudz resursu, taču ļautu būtiski paaugstināt kopējo datu kvalitāti.

Ir svarīgi atzīmēt, ka datu kvalitātes problēmu esamība datu kopās ir bīstama ar to, ka datu sniedzēji, visticamāk, pat nenojauš par tām. Par to liecina arī *Global Open Data Index* (Global Open Data Index, 2018), kas veic 15 valsts sektora atvērto datu kopu novērtējumu, tajā skaitā arī Uzņēmumu reģistrus, kā primārus uzņēmumu raksturojošus parametrus norādot uzņēmuma nosaukumu, unikālo identifikatoru jeb reģistrācijas numuru un uzņēmuma adresi, kas atbilst lietošanas piemēros iekļautajiem parametriem. Sava analīzes rezultātā *Global Open Data Index* ir ievietojis Norvēģijas un Lielbritānijas Uzņēmumu reģistrus 1. pozīcijā, savukārt Latvijas Uzņēmumu reģistru – 18. no 94 Uzņēmumu reģistriem. Atbilstoši šīm novērtējumam, pēdējos gados Latvija uzlabo savus rādītājus, 2014. gadā ieņemot 24., 2015. gadā – 31., 2018. gadā – 18. pozīciju. Tik augsti Uzņēmumu reģistru rezultāti ir skaidrojami ar to, ka *Global Open Data Index* vērtē konkrētu datu kopu atbilstību atvērto datu kopu principiem, nevērtējot datu kvalitāti, kas atbilst autores apgalvojumam par datu kvalitātes aspekta ignorēšanu atvērto datu principu sarakstā. Tas nozīmē, ka galalietotājiem ir jābūt piesardzīgiem un nedrīkst paļauties uz tādiem novērtējumiem, jo augstie rādītāji neobligāti liecina par datu kopu augstu datu kvalitāti.

Analizēto datu kopu kvalitāti darba autore pārbauda arī attiecībā uz citiem datu objektu (t.i. "Uzņēmumu") raksturojošiem parametriem. Rezultātā katrā Uzņēmumu reģistrā autore konstatēja vairākas datu kvalitātes problēmas gan datu sintaksē, gan semantikā. (Nikiforova, 2019b) un rakstos (Nikiforova, 2018a) un (Bicevskis et al., 2018b) ir apkopoti katra Uzņēmuma reģistra datu kvalitātes analīzes rezultāti, sniedzot diskusiju par konstatētājām datu kvalitātes

problēmām, to raksturu un rašanas iespējām, līdz ar ko šī diskusija netiks atkārtota, sniedzot atsevišķus rezultātus un secinājumus apkopotā veidā, detalizētāk apskatot (a) Lielbritānijas Uzņēmuma reģistra kvalitātes analīzes rezultātus, jo iepriekšējā nodaļā visi kvalitātes modeļa komponenti tika demonstrēti uz tā piemēra, un (b) Latvijas Uzņēmumu reģistra kvalitātes analīzes rezultātus.

Tabulā 5.1.2. ir pieejams Lielbritānijas Uzņēmumu reģistra datu kvalitātes analīzes rezultātu kopsavilkums, norādot atbilstošo lauku nosaukumu, lauka formātu un obligātumu, kļūdu skaitu un komentāru attiecībā uz konstatētām kvalitātēs problēmām.

5.1.2. tabula

Lielbritānijas Uzņēmumu reģistrs (pēc (Nikiforova, 2019b, 2018a))

#	Lauka nosaukums	Lauka formāts	Kļūdu skaits	Kļūdu komentāri
1.	CompanyNumber	int, NOT NULL	0	-
2.	CompanyName	Varchar(100) NOT NULL	1 0.0001%	Vērtība (nosaukums) nav norādīta
3.	RegAddress. AddressLine1	Varchar(100) NOT NULL	7 518 1%	7 514 ierakstiem vērtība (adrese) nav norādīta; 4 ierakstu vērtības: “XXXXXXX”, “XXXXXX”, “XXXXXXXXXX”, “XXX XXX”
4.	RegAddress.Post Code	Varchar(20) NOT NULL	12 155 1.6%	Vērtība (pasta indekss) nav norādīta; 4 ierakstu vērtības: “XXXXXXX”, “XXXXXX”, “XXXXXXXXXX”, “XXX XXX”
5.	Company Category	Varchar(50) IN('Public Limited Company', …) NOT NULL	7 202 - 21 524 0.9% - 2.9%	6 kategoriju nosaukumi tikai daļēji atbilst pieļaujamām vērtībām; 4 vērtības ir nederīgas
6.	CompanyStatus	Varchar(50) IN(‘Active’, ‘Liquidation’, …) NOT NULL	191 - 26 538 0.03% - 3.5%	4 vērtības tikai daļēji atbilst pieļaujamām vērtībām; 1 vērtība ir nederīga

#	Lauka nosaukums	Lauka formāts	Kļūdu skaits	Kļūdu komentāri
7.	URI	Varchar(50) , paraugs ‘http://business.data.gov.uk/id/company/X’, kur X – CompanyName, NOT NULL	0	-
8.	IncorporationDate	Date (‘DD/MM/YYYY’) NOT NULL	3 0.0004%	Nederīgas vērtības “16/06/1701”, “09/08/1638”, “25/04/1552”

Lielbritānijas Uzņēmuma reģistram veicot divu parametru vērtību kvalitātes analīzi pret sekundāra datu objekta vērtībām atbilstoši 3.6. apakšnodaļa aprakstītai procedūrai, autore konstatē, ka 208 039 ierakstos esošās vērtības (jeb 48 vērtības) neatbilst nevienam valsts nosaukumu un kodu standartam. Veicot datu kvalitātes problēmu protokola analīzi, (Nikiforova, 2019b) autore konstatē, ka standartiem neatbilstošas datu kvalitātes problēmas var tikt iedalītas sekojošās grupās:

1) vairāk kā viena nosaukuma izmantošana vienas valsts apzīmēšanai (piemēram, (1) “*United Kingdom*” un “*UK*”, (2) “*United States*”, “*United States of America*” un “*USA*”, (3) “*Ireland*”, “*Republic of Ireland*”, (4) “*Nigeria*”, “*Republic of Nigeria*” utt.). To darba autore konstatēja 9 valstīm (6 valsts nosaukumiem “*RegAddress Country*” lauka un 4 – “*CountryOfOrigin*” lauka gadījumā, 1 valsts nosaukums abu lauku gadījumā);
2) zemes vai teritoriju nosaukumu glabāšana laukos, kas ir paredzēti valsts nosaukumu glabāšanai, neskatoties uz to, ka reģistrā vienlaikus kopā ar šīm vērtībām ir sastopamas arī standartiem atbilstošās vērtības (“*Wales*”, “*Scotland*”, “*England & Wales*”, “*England*”, “*Ireland*”, kā arī “*Northern Ireland*” un “*Southern Ireland*” utt., lai gan atbilstoši *Companies House* Apvienotās Karalistes teritorija iedalās Īrijā, Anglijā un Velsā (“*England & Wales*”) un Skotijā);
3) 4 vērtības neatbilst valsts vai zemes nosaukumiem: “*SW7*” - Dienvidkensingtona pasta indekss, Anglijas grāfiste Austrumsaseksa (“*East Sussex*”), “*BWI*” -

Baltimoras-Vašingtonas starptautiskās lidostas kods un "*DE 19901*" - Doveras pasta indekss;

4) vairs neeksistējošas valstis (Čehoslovākija, Dienvidslāvija, PSRS), neskatoties uz to, ka atbilstoši to reģistrēšanas datumam, tie tika reģistrēti pēc šo valsts sabrukuma.

Datu kvalitātes kontekstuālā analīze būtiski uzlabo datu kvalitātes analīzes rezultātus, jo datu kvalitātes analīze viena datu objekta ietvaros norādīja uz datu kvalitātes problēmu esamību, dažreiz norādot tikai uz potenciāli nekvalitatīvām vērtībām. Piemēram, sniegtā piemēra ietvaros darba autore konstatēja vairākas vērtības, kas norāda uz vienu valsti, taču kura no vērtībām ir kvalitatīva, nebija zināms, jo atsevišķās vērtības mēdz būt ticamas, bet tajā paša laikā nekvalitatīvas. Papildus, šo vērtību identificēšana un analīze prasīja daudz resursu, izstrādājot specifiskus vaicājumus, ko autore izstrādāja, apstrādājot visās atbilstošās kolonnās esošas vērtības, veicot to apkopojumu, statistisko rādītāju iegūšanu, analizējot vaicājuma izpildes rezultātus, t.i. katru rezultējošo ierakstu. Otrs paņēmiens, kas atbilst daļēji kontekstuālajai kvalitātes analīzei, ir standartiem atbilstošo vērtību iekļaušana datu kvalitātes nosacījumos. Tas ļauj izteikt secinājumus par vērtību kvalitāti, balstoties uz reālajiem datiem nevis galalietotāja pieņēmumiem un uzskatiem par atbilstošo vērtību pareizību, taču tas pārlieku sarežģi un paplašina kvalitātes nosacījumus, it īpaši gadījumos, kad pieļaujamo vērtību saraksts satur vairākas vērtības, savukārt apskatīta piemēra gadījumā būtu jāizmanto 4 saraksti, kuros ir apkopoti visi valsts nosaukumi. Tas savukārt pārkāptu uzreiz divus piedāvātas pieejas principus: (a) veidojamas diagrammas vairs nebūtu uzskatāmas un viegli saprotamas; (b) nosacījumu struktūra kļūstu sarežģītāka, samazinoties iespējamībai, ka tie varētu būt definējami ar ne-IT cilvēkiem. Arī šīs paņēmiens nebūtu "lietotājam draudzīgs". Savukārt pieejas paplašinājums ar kontekstuālās kvalitātes analīzes iespēju vairāku datu objektu kontekstā (atbilstoši 3.6. apakšnodaļai) risina šo problēmu, nodrošinot iespēju veikt šāda rakstura padziļināto analīzi, saglabājot (a) gan viena datu objekta ietvaros veicamas datu kvalitātes analīzes principus, kas tiek pielāgoti arī paplašinātājai analīzei, (b) diagrammu lasāmību, saprotamību, (c) piemērotību arī ne-IT cilvēkiem. Dotā piemēra ietvaros darba autore to nodemonstrēja, analizējot valsts nosaukumus saturoša parametra vērtības ([*CountryOfOrigin*] un [*RegAddress Country*]) pret citu datu kopu, kas satur oficiālus valsts nosaukumus. Atbilstoši (Nikiforova, 2019b) autore konstatē, ka valsts nosaukumus, kurā tika dibināts/ izveidots uzņēmums, saturošā parametrā [*CountryOfOrigin*] 1 045 ierakstos (0.14%) ir novērojamas datu kvalitātes problēmas, jo 33 no 109 valsts nosaukumiem jeb 30% no visām Lielbritānijas Uzņēmumu reģistrā atbilstošas kolonnas esošajām vērtībām, neatbilst nevienam standartam. Kolonnā ([*RegAddress Country*], kas satur valsts nosaukumu, kurā ir reģistrēts uzņēmums, 206

994 ieraksti (27.44%) satur nekvalitatīvus datus, t.i. 40 no 114 valsts nosaukumi jeb 35% no visiem nosaukumiem neatbilst nevienam standartam.

Atbilstoši 3.4. apakšnodaļā definētājiem lietošanas piemēriem, atbilstoši kuriem var būt nepieciešamība pārbaudīt datu kvalitāti pret konkrēto standartu, autore konstatē, ka parametra [*CountryOfOrigin*] gadījumā visi valsts nosaukumi atbilst īso nosaukumu standartam, savukārt [*RegAddress Country*] gadījumā - 73 no 74 standartam atbilstošajām vērtībām atbilst īso nosaukumu standartam, taču 1 ieraksts atbilst *ISO3/ UNDP* standartam ("*USA*" – ASV). Taču neskatoties uz to, ka kvalitātes problēmas autore konstatēja 208 039 ierakstos (27.6%), tā tiktu novērsta, veicot 48 vērtību labojumu - 33 vērtību labojumus 1. parametrā un 40 - 2. parametrā. Tas nozīmē, ka 208 039 kvalitātes problēmu novēršana prasa samēra maz resursu, savukārt to labošana būtiski uzlabotu datu kopas kopējo kvalitāti ((Nikiforova et al., 2019), (Nikiforova, 2019b)).

Tas nozīmē, ka kontekstuālā datu kvalitātes analīze būtiski atvieglo datu kvalitātes analīzi, ļaujot veikt vairāku datu objektu savstarpējo salīdzinājumu, lēmumus pieņemot, balstoties uz iegūtajiem datu kvalitātes analīzes rezultātiem, neprasot papildus manuālās pārbaudes, vienlaicīgi uzlabojot iegūtos rezultātus, norādot uz viennozīmīgi eksistējošām nevis potenciālajām problēmām. Atbilstoši (Nikiforova et al., 2019), datu kvalitātes analīzes viena datu objekta ietvaros darba autore analizēja 128 dažādas potenciāli nekvalitatīvas vērtības, savukārt kontekstuālās pārbaudes rezultātā noteica 48 nekvalitatīvas vērtības, novēršot nepieciešamību veikt to turpmāku analīzi gala lēmuma pieņemšanai. Veicot abu datu protokolu savstarpējo salīdzinājumu autore konstatēja, ka tikai 13 no 128 vērtībām, kas tika noteiktas datu kvalitātes analīzes viena datu objekta ietvaros, bija nekvalitatīvas, savukārt 115 rezultāti bija kļūdaina atbilsme (angl. *false-positive*). Konkrētajā gadījumā iegūto rezultātu darba autore uzlaboja par 72.9%.

Tabulā 5.1.3. ir apkopoti Latvijas Uzņēmumu reģistra datu kvalitātes analīzes rezultāti (Nikiforova, 2019b).

5.1.3. tabula

Latvijas Uzņēmumu reģistrs (pēc (Nikiforova, 2018a))

#	Lauka nosaukums	Lauka formāts	Kļūdu skaits	Kļūdu komentāri
1.	Reg_number	Int, 9 vai 11 cipari NOT NULL	0	-
2.	Name	Varchar(100), NOT NULL	10 0.0025%	Vērtība (nosaukums) nav norādīta
3.	Type	Varchar(3), NOT NULL	0	-

#	Lauka nosaukums	Lauka formāts	Kļūdu skaits	Kļūdu komentāri
		IN(SIA, AS, ...)		
4.	Type_text	Varchar(50), IN(‘Sabiedrība ar ierobežoto atbildību’, ‘Akciju sabiedrība’, …)	1 403 – 1 578 0.35% - 0.39%	6 uzņēmuma tipi (lauka “type” vērtības atšifrējums) nav norādīti; 2 atbilst pieļaujāmam vērtībām tikai daļēji
5.	Regtype	Char(1), IN(‘B’, ‘K’, ‘U’, ‘M’, ...)	0	-
6.	Regtype_text	Varchar(50), IN(‘Sabiedrisko organizāciju reģistrs’’, ‘Pārstāvniecību reģistrs’…)	71 0.02%	1 vērtība neatbilst reģistrētājiem reģistru nosaukumiem
7.	Address	Varchar(100), NOT NULL	366 0.09%	Vērtība (adrese) nav norādīta
8.	Adress_id	Int, 9 cipari, NOT NULL	4 523 1.14%	Vērtība (adreses kods) nav norādīta
11.	Post_code	Int, 4 cipari, NOT NULL	20 498 5.16%	2 vērtības neatbilst pēc formāta – 3 cipari; 20 496 ierakstos vērtība (pasta indekss) nav norādīta
12.	ATV-code	Int, 7 cipari, NOT NULL	5 521 1.39%	4 574 ierakstos vērtība (reģiona kods) nav norādīta; 947 ierakstos koda garums ir mazāks par 7 cipariem
13.	Date	Date (‘YYYY-MM-DD’) NOT NULL	94 0.024%	Vērtība (dibināšanas datums) nav norādīta
14.	Closed	Char(1), IN(‘L’, ‘R’), NULL	-	-
15.	Terminated	Date (‘YYYY-MM-DD’) NULL	-	-

Atbilstoši (Nikiforova, 2019b) veicot datu semantiskās pārbaudes viena datu objekta ietvaros, t.i. analizējot savstarpēji saistītus parametrus, autore konstatēja pretrunības savstarpēji saistītos parametros "*closed*", kas ir paredzēts neaktīva uzņēmuma statusa norādīšanai – "R" - reorganizēts vai "L" – likvidēts, un "*terminated*", kas ir paredzēts datuma norādīšanai, kad uzņēmumam tika piešķirts atbilstošais statuss. Visiem uzņēmumiem, kuriem parametrā "*closed*" glabājās ne-nulles vērtība, t.i. tie ir reorganizēti (147 764) vai likvidēti (5 809), ir norādīts arī atbilstošais "*terminated*" datums, taču 646 uzņēmumiem, kuriem nav norādīta parametra "*closed*" vērtība, no kā seko, ka tie ir aktīvi, ir norādīta parametra "*terminated*" vērtība. Tāda veidā ir pārkāpti iekšējie ierobežojumi, jo "*terminated*" parametra vērtībai ir jābūt norādītai tad un tikai tad, ja ir norādīta parametra "*closed*" vērtība. Līdzīga veida problēmas autore konstatēja arī citu savstarpēju saistītu parametru gadījumā, piemēram, "*type*" un "*type_text*", kur "*type_text*" parametrā ir jāglabājas parametra "*type*" vērtības atšifrējumam.

Atbilstoši (Nikiforova, 2019b, 2018b) visos četru Uzņēmumu reģistros analīzes rezultātā autore konstatēja datu kvalitātes problēmas. Viszemākais datu kvalitātes problēmu rādītājs ir Norvēģijas uzņēmumu reģistram, jo kvalitātes problēmas tika konstatētas tikai 8 no 42 parametros (19%), kuram seko Lielbritānijas Uzņēmumu reģistrs - 17 no 55 parametros (31%), savukārt sliktākus rezultātus uzrādīja Igaunijas – 7 no 14 (50%) un Latvijas Uzņēmumu reģistrs - 11 no 22 (50%). Visbiežāk sastopamas datu kvalitātes problēmas:

1) *NULL* vērtības primārajos uzņēmumu raksturojošos laukos, piemēram, adrešu laukā;
2) nederīgie un/ vai apšaubāmie datumi (piemēram, Norvēģijā ir uzņēmums, kuru dibināšanas gads ir 1277. gads);
3) *NULL* vērtības laukos, kas satur cita laukā esoša saīsinājuma atšifrējumu (t.i. pilno nosaukumu) un otrādi (atšifrējums, nesniedzot saīsinājumu). Piemēram, Latvijas Uzņēmumu reģistra gadījumā šī problēma ir konstatēta lauku "*type*" un "*type_text*", "*Adressid*" un "*Adress*" gadījumā, Igaunijas Uzņēmumu reģistra gadījumā - "*asukoha_ehak_kood*" un "*asukoha_ehak_tekstina*" utt..

Dažas izplatītas, bet raksturīgas tikai atsevišķiem reģistriem datu kvalitātes problēmas:

1) laukos, kuri ir paredzēti valsts nosaukuma glabāšanai, satur pasta indeksu, grāfistes vai ciemata nosaukumu, neskatoties uz to, kā arī šo vērtību glabāšanai ir paredzēti atsevišķie lauki;
2) vairāku vērtību izmantošana viena reālā objekta apzīmēšanai. Viens no apakšgadījumiem, ir vairāki nosaukumi vienas valsts apzīmēšanai vienas datu kopas ietvaros, vai arī vairs neeksistējošo valsts nosaukumi, neskatoties uz to, ka uzņēmumi tika reģistrēti pēc atbilstošas valsts sabrukuma;

3) kodu neatbilstība paraugam vai garumam;
4) atsevišķu lauku vai ierakstu aizpildīšana ar '*x*' vai '*0*' zīmi, ja *NULL* vērtības nav paredzētas (Nikiforova, 2019b).

Taču atbilstoši datu kvalitātes jēdziena relatīvajam raksturam ir jāatzīmē, ka, neskatoties uz augstu datu kvalitātes problēmu skaitu visos uzņēmumu reģistros, atsevišķos gadījumos atkarībā no definētā lietošanas piemēra, tie var būt kvalitatīvi un to analīzes rezultāti būs precīzi.

5.2. Datu kopu analīzes rezultāti

Šajā nodaļā tiek aplūkoti 7 atvērto datu kopu datu kvalitātes analīzes rezultāti. Tā kā to analīzi autore detalizēti aplūkoja (Nikiforova, 2018b, 2019b), šī diskusija netiks atkārtota, sniedzot īsu katras datu kopas kvalitātes analīzes rezultātu aprakstu, iezīmējot visbiežāk sastopamas datu kvalitātes problēmas.

5.2.1. Datu kopas "Interešu un pieaugušo neformālās izglītības programmu licences" analīzes rezultāti

Datu kopu "Interešu un pieaugušo neformālās izglītības programmu licences" par 2013. – 2015., 2017. un 2018. gadiem (RD IKSD, 2019), ko sniedz Rīgas Dome, kvalitātes analīze liecina par to, ka visās datu kopās ir sastopamas vismaz atsevišķas datu kvalitātes problēmas. Datu kopas par 2013. – 2015. gadiem ir augstas kvalitātes, taču tām ir raksturīgas atsevišķas datu kvalitātes problēmas un anomālijas, kuras autore konstatēja 2 no 9 parametros. Datu kopās par 2017. un 2018. gadiem datu kvalitātes problēmu skaits ir lielāks, datu kvalitātes problēmas konstatējot 3 un 4 no 9 parametros. Vairākums datu kvalitātes problēmu autore konstatēja kontekstuālās analīzes rezultātā, doto datu objektu kvalitāti pārbaudot pret Latvijas Uzņēmumu reģistru. Tas nozīmē, ka no lielākas datu kvalitātes problēmu daļas būtu iespējams izvairīties, ja datu sniedzēji (a) piekļūtu Latvijas Uzņēmumu reģistram un izmantotu tajā esošās vērtības, (b) periodiski veiktu uzkrāto datu kvalitātes analīzi, izmantojot piedāvāto risinājumu vai līdzīgu risinājumu.

Izplatītākās datu kvalitātes problēmas analizētajos datu objektos ir (a) reģistrācijas numuru nederīgums, veicot to salīdzinājumu pret Latvijas Uzņēmumu reģistru, vai uzņēmuma nosaukuma neatbilstība Latvijas Uzņēmumu reģistrā norādītājam, pie vienādiem reģistrācijas numuriem, kas var izpausties kā (1) vienādi nosaukumi, bet dažāda uzņēmējdarbības forma, (2) dažāda notācija jeb pieraksts vienam un tam pašam uzņēmumam, kas atkarībā no lietošanas

piemēra nevar tikt uzskatīts par datu kvalitātes problēmu, jo vērtība ir ticama, (3) pilnībā dažādi nosaukumi, (b) pretrunīgie dati, (c) reģistrācijas numura, licences pieprasītāja vārda vai nosaukuma nenorādīšana.

Tāpat kā Latvijas Uzņēmumu reģistrā, analizētajās datu kopas vienam un tam pašam uzņēmumam var atbilst dažādi tā nosaukuma pieraksti. Atbilstoši (Nikiforova, 2019b) tas attiecas gan uz uzņēmējdarbības formas un uzņēmuma nosaukuma pieraksta secību, gan uzņēmējdarbības formas pierakstu (saīsinātā vai pilnā formā), līdz ar ko automatizēta ierakstu apstrāde ir apgrūtinātā pat vienas datu kopas ietvaros, jo viens un tas pats nosaukums var tikt pierakstīts dažādi, kā arī vienas datu kopas ietvaros (pat Latvijas Uzņēmumu reģistrā) uzņēmumu nosaukumu pieraksts var atbilst dažādiem paraugiem (angl. *pattern*).

5.2.2. Datu kopas "Statistika par saziņu ar Rīgas pašvaldību" analīzes rezultāti

Datu kopas "Statistika par saziņu ar Rīgas pašvaldību" (Latvijas Atvērto datu portāls, 2018b) kvalitātes analīzes rezultātā, kura ietvaros autore veica visu 8 parametru datu kvalitātes analīze, datu kvalitātes problēmas tika identificētas tikai 2 parametros (25%).

Vairākos ierakstos konstatētās problēmas ir saistītas ar datu neatbilstību dokumentācijai, ko ir sniedzis datu sniedzējs, 6 no 9 datu kopā sastopamām vērtībām (sastopamas 25.9% ierakstos), nav uzskaitītas konkrēta parametra pieļaujamo vērtību sarakstā, neskatoties uz to, ka tās ir ticamas. Visticamāk šī problēma varētu tikt novērsta, papildinot dokumentāciju ar 6 vērtībām, ja datu sniedzējs (Rīgas Dome) pieļauj šīs vērtības, citādi labojot tās.

5.2.3. Datu kopas "Valsts informācijas sistēmu reģistrs" analīzes rezultāti

Datu kopā "Valsts informācijas sistēmu reģistrs" datu objektu "Informācijas sistēma" raksturo 36 parametri, 25 no kuriem (69.4%) autore konstatēja datu kvalitātes problēmas. Atbilstoši (Nikiforova, 2019b, 2018a) 7 parametru gadījumā datu objektā konstatēto datu kvalitātes problēmu īpatsvars nepārsniedz 1%, pie tam 20 parametros autore konstatēja nepilnīguma problēmas, kuru novēršana samazinātu kvalitātes problēmas saturošo parametru skaitu līdz 5 (13.9%).

Visizplatītākā no konstatējām datu kvalitātes problēmām ir dažādu vērtību izmantošana viena objekta apzīmēšanai, kas visbiežāk izpaužas, norādot uz vērtības neesamību. Norādot uz vērtības neesamību viena parametra ietvaros *NULL* vērtības vietā vai arī kopā ar to, tiek izmantotas vērtības "-" un "*nav*". Kā autore minēja (Nikiforova, 2019b, 2018a), to varētu pamatot ar datubāzes projektējumu, atbilstoši kuram noteiktiem parametriem ir jābūt

aizpildītiem, taču 12 no 20 parametriem (33.3%), kuros autore konstatēja šī tipa problēma, dažiem ierakstiem ir norādīta *NULL* vērtība. No vienas puses, to nevar uzskatīt par būtisku datu kvalitātes problēmu, jo visi 3 izmantotie apzīmējumi var norādīt uz vērtības neesamību, taču atkarībā no lietošanas piemēra šīs datu neviendabīgums var ietekmēt datu analīzes rezultātus, jo (1) veicot datu analīzi, it īpaši, agregācijas operācijas, kuras rezultāti, visticamāk, būs neprecīzi, jo datu lietotājiem nav zināms, ka uz vērtības neesamību var norādīt vairākas vērtības un otrādi – vērtības esamība neliecina par datu esamību, jo tai ir jābūt pielīdzināmai *NULL* vērtībai, (2) nav pārliecības, ka visām vērtībām ir vienāda nozīme, jo *NULL* var norādīt gan uz to, ka vērtības neeksistē, kas atbilst reālai pasaulei, gan uz to, ka tā vienkārši nav zināma, kas neraksturo reālās pasaules situāciju, kas atbilst arī (Batini et al., 2016) viedoklim. Šī problēma ir raksturīga arī Informācijas Sistēmas tīmekļvietnēm, kur uz tīmekļvietnes neesamību vai datu trūkumu norāda 5 dažādas vērtības.

Papildus autore konstatēja (a) datu pretrunību, (b) datumu formātu dažādību viena parametra ietvaros, (c) nederīgas vērtības, kas neatbilst paraugam, kas tika konstatēts 5 parametriem.

Taču, atbilstoši (Nikiforova, 2018a) neskatoties uz samērā augstu datu kvalitātes problēmu skaitu, nedrīkst apgalvot, ka šī datu kopa ir zemas kvalitātes, un nedrīkst būt izmantota, jo atkarībā no lietošanas piemēra šī datu kopa var izrādīties arī augstas kvalitātes, tās analīzes iegūstot korektus rezultātus. Piemēram, pēc analoģijas ar Uzņēmumu reģistriem definētājiem lietošanas piemēriem, identificējot Informācijas Sistēmas pēc to nosaukuma, reģistrācijas numura, statusa un atbildīgas personas vārda, uzvārda un reģistrācijas numura, datu kopa ir augstas kvalitātes, jo nevienā no šiem parametriem datu kvalitātes problēmas darba autore nekonstatēja. Taču pie citiem nosacījumiem, tās kvalitāte nav apmierinoša un tās analīze var novest pie nekorektiem un neprecīziem rezultātiem.

5.3. Latvijas atvērto medicīnas datu kvalitātes analīze

Iepriekšējās apakšnodaļās darba autore aprakstīja atsevišķu datu kopu kvalitātes analīzi, taču, ņemot vērā medicīnas datu svarīgumu un nozīmīgumu, darba ietvaros tika veikta arī Latvijas atvērto medicīnas datu kvalitātes analīze. Šī domēna analīze ļauj ne tikai pārbaudīt piedāvātās pieejas efektivitāti, bet arī veikt secinājumus par atvērto medicīnas datu kvalitāti. Ņemot vērā, ka datu kvalitātes problēmu esamība Latvijas medicīnas datos darba autore konstatēja arī Latvijas "slēgtajos" medicīnas datos (Konstante, 2016), tajā skaitā Nacionālajā

veselības dienesta (NVD) sistēmā, galvenokārt, norādot uz datu pretrunību, ir pamats uzskatīt, ka datu kvalitātes problēmas tiks noteiktas arī atsevišķās atvērto datu kopās.

Atvērtie medicīnas dati Latvijā pirmo reizi tika publicēti 2018. gada janvāra beigās un 2019. gada 3. ceturksnī tika reprezentēti ar 15 datu kopām, ko publicēja 7 dažādi datu sniedzēji (tabulas 5.3.1. 1. kolonna). Neskatoties uz to, ka dažām datu kopām ir pieejami metadati, kas ir nodefinēti atbilstoši *CKAN (Comprehensive Kerbal Archive Network)*, vairākums datu sniedzēju nesniedz pilnīgus metadatus un neizmanto standarta vērtības, jo tikai 8 no 15 datu kopām ir sniegts visu parametru īss paskaidrojums/ apraksts. Šī problēma ir raksturīga arī citu valsts atvērtajiem datiem (piemēram, Brazīlijai (Oliveira et al., 2016), Vācijai, Francijai un Lielbritānijai (Martin et al., 2013), (Zuiderwijk et al., 2014), (Beno et al., 2017)).

Tikai 6 no 15 datu kopām tiek atjaunotas tik bieži, cik to solas datu sniedzēji. Viens šīs problēmas iespējams skaidrojums ir tas, ka ar datu atjaunošanu datu sniedzēji saprot to atjaunošanu savās sistēmās nevis datu portālos. Arī šī problēma ir raksturīga citu valsts datu kopām, jo atbilstoši (Tinholt, 2013) tas ir raksturīgs 22 no 23 analizētājām valstīm, kā arī Lielbritānijai (Kuk et al., 2011) un citām valstīm (Beno et al., 2017).

Medicīnas datu gadījumā populārākais atvērto datu formāts Latvijā ir *.xlsx*, kurā ir pieejami 53.3% datu kopu, savukārt vēl 26.7% ir pieejami *.zip* arhīvos, kuros ir pieejami datu kopas *.csv* un *.xlsx* formātos un 1 datu kopa, kas ir pieejama *.html* formātā, kas nevar tikt uzskatīta par atvērto datu piemēru, jo tā ir saite uz statistiku (datu kopa "Oficiālās statistikas datubāze"). Ir jāatzīmē, ka vairākums datu kopu ir pieejams mašīnlasāmā formātā.

Tā kā pētījuma centrālais objekts ir datu kvalitāte, neskatoties uz iespēju analizēt tikai atsevišķu parametru kvalitāti atkarībā no lietošanas piemēra ierobežotības, darba autore analizēja katras datu kopas katra parametra kvalitāti, ar mērķi veikt padziļinātu datu kvalitātes analīzi. Medicīnas datu kvalitātes analīzē darba autore izmantoja 15 primārās un 11 sekundārās datu kopas, pret kuru parametriem analizēja 35 primāro datu objektu parametrus. Visplašāk izplatītās datu kvalitātes problēmas analizētājās kopās ir: (a) kontekstuālās datu kvalitātes problēmas; (b) datu nepilnīgums; (c) dažāda viena objekta notācija viena datu objekta un pat viena parametra ietvaros; (d) datu kvalitātes problēmas savstarpēji saistīto parametru gadījumā. Tabulā 5.3.1. ir sniegts populārāko datu kvalitātes problēmu apkopojums pēc parametru skaita, kuros tās tika konstatētas, savukārt tabulā 5.3.2. – pēc ierakstu skaitā, kuros konkrētā datu kvalitātes problēma tika konstatēta.

5.3.1. tabula

Datu kvalitātes problēmu iedalījums pēc parametru skaita, kuros tika identificētas datu kvalitātes problēmas (pēc (Nikiforova, 2019a))

Datu kopa	Kontekstuālās problēmas/ pārbaužu skaits (kopā)	Tukšas / kopā	Dažāda notācija/ kopā	Defekti saistītājos laukos (jā/ nē)	Kvalitatīvi / kopā
Saslimstība ar 2. tipa cukura diabētu Latvijā	0/0	0/6	0/6	nē	6/6 100%
Personu, kuras saņēmušas TPL, skaita sadalījums pēc administratīvās teritorijas	2/2 100%	3/7 43%	0/7	nē	2/7 29%
Sociālo pakalpojumu sniedzēju skaits	2/2 100%	22/27 82%	10/27 37%	nē	4/27 15%
VDEĀVK uzskaitē esošās pilngadīgās personas ar invaliditāti pēc invaliditātes smaguma pakāpes un administratīvās teritorijas	2/2 100%	0/23 0	0/23 0	nē	20/23 87%
VDEĀVK uzskaitē esošo bērnu ar invaliditāti skaits sadalījumā pēc administratīvās teritorijas	2/2 100%	0/10 0	0/10 0	nē	8/10 80%
Nelaimes gadījumi darbā	0-1/1 0 - 100%	1/10 10%	0/10 0	nē	8/10 80%
Apstiprinātas arodslimības	4/5 80%	2/11 18%	1/11 0.09%	nē	9/11 82%
Valsts asinsdonoru centra statistika	0/0	0/4 0	0/4 0	nē	4/4 100%
Farmaceitiskās darbības uzņēmumu reģistrs	½ 50%	17/38 45%	0/38 0	nē	19/38 50%
Zāļu patēriņa statistika	3/3 100%	5/8 63%	2/8 25%	nē	0/8 0
Latvijas Zāļu reģistrs	4/9 44%	21/41 51%	1/41 2%	jā	14/41 34%
Uztura bagātinātāju reģistrs	2/2 100%	30/35 86%	4/35 11%	jā	5/35 14%

Datu kopa	Kontekstuālās problēmas/ pārbaužu skaits (kopā)	Tukšas / kopā	Dažāda notācija/ kopā	Defekti saistītājos laukos (jā/ nē)	Kvalitatīvi / kopā
Diētiskās pārtikas reģistrs	2/2 100%	19/22 87%	4/22 18%	jā	3/22 14%
Veterināro zāļu reģistrs	1/3 33%	16/26 62%	0/26 0	jā	8/26 31%

Tikai vienā analizētājā datu kopā autore neidentificēja nevienu datu kvalitātes problēmu. Ir jāatzīmē, ka tajā ir apkopoti skaitliskie dati, līdz ar ko arī kvalitātes pārbaudes bija vienkāršākas, galvenokārt, pārbaudot datu pilnīgumu un veicot vienkāršākus matemātiskus aprēķinus, kas ir saistīti ar datu agregāciju, veicot datu kvalitātes analīzi vienas datu kopas ietvaros.

Viena no visplašāk izplatītājām datu kvalitātes problēmām ir tukšo vērtību esamība (tabulu 5.3.1. un 5.3.2. 3. kolonna), ko darba autore konstatēja 136 no 137 (81.4%) parametriem. Tukšo vērtību skaits datu kopās un to parametros svārstās no 1 līdz visām konkrētās datu kopas vērtībām. Taču kopējais tukšo vērtību skaits analizētājās datu kopās ir 15%. Atsevišķos gadījumos tukšas vērtības autore konstatēja pat primārajos datu kopas laukos, piemēram, datu kopā "Diētiskās pārtikas reģistrs" 4 ierakstiem nav norādītas parametru [*Nosaukums*] un [*RazotajaNosaukums*] vērtības, neskatoties uz to, ka šī datu kopai šie parametri ir primārie. Ir jāatzīmē, ka atsevišķos gadījumos tukšas vērtības ir pieļaujamas, piemēram, datu kopas "*Uztura bagātinātāju reģistrs*" gadījumā parametram [*AnulesanasIemesls*] vērtības var nebūt, ja parametra [*Statuss*] vērtība ir "*anulēts*". Tas nozīmē, ka, ja vērtībai ir jābūt norādītai tikai tad, kad noteiktas nosacījums ir spēkā vai tā nav obligāta, atkarībā no lietošanas piemēra vērtības neesamība var netikt uzskatīta par datu kvalitātes problēmu. 28 no 136 gadījumos datu neesamība netiek uzskatīta par datu kvalitātes problēmu, taču, kamēr datu sniedzējs nesniedz skaidrojumu vai piezīmi par to, ka ir atļauta tukša vērtība (un kā tai ir jābūt interpretētai), nav pārliecības, ka tā nav datu kvalitātes problēma, jo vērtības neesamība var tikt skaidrota dažādi. Pilnīguma problēma ir viena no visbiežāk sastopamam datu kvalitātes problēmām gan atvērto datu ((Yi, 2019), (Oliveira et al., 2016), (Martin et al., 2013), (Zuiderwijk et al., 2012, 2014), (Janssen et al., 2012),), gan "slēgto" datu ((Schmidt et al., 2015), (Tomic et al., 2015), (Zhang et al., 2014)) gadījumos. Tā ir raksturīga arī dažādu domēnu un valsts datu kopām (Wanner et al., 2018). Taču mēdz būt arī piemēri, kas demonstrē augsts datu pilnīguma rezultātu – Islandes un Norvēģijas Vēža Reģistrs ((Sigurdardottir et al., 2012), (Larsen et al., 2009)), lai gan ir jāatzīmē, ka abi piemēri ir "slēgto" datu piemēri.

5.3.2. tabula

Datu kvalitātes problēmu iedalījums pēc vērtību skaita, kuros tika identificētas datu kvalitātes problēmas

Datu kopa	***Kontekstuālās problēmas/ kontekstuālās pārbaudes (kopā)***	***Tukšas / kopā***	***Kvalitatīvās/ kopā***
Saslimstība ar 2. tipa cukura diabētu Latvijā	0	0	100%
Personu, kuras saņēmušas TPL, skaita sadalījums pēc administratīvās teritorijas	2.4%	2.6%	97%
Sociālo pakalpojumu sniedzēju skaits	2.5%	51.3%	60.7%
VDEĀVK uzskaitē esošās pilngadīgās personas ar invaliditāti pēc invaliditātes smaguma pakāpes un administratīvās teritorijas	2.4%	0	99.8%
VDEĀVK uzskaitē esošo bērnu ar invaliditāti skaits sadalījumā pēc administratīvās teritorijas	2.5%	0	99.5%
Nelaimes gadījumi darbā	0 - 100%	0.01%	88.9% - 99.99%
Apstiprinātas arodslimības	27.8%	2%	85.4%
Valsts asinsdonoru centra statistika	0	0	100%
Farmaceitiskās darbības uzņēmumu reģistrs	45.8%	20.6%	79.3%
Zāļu patēriņa statistika	0.3%	4.7%	95.2%
Latvijas Zāļu reģistrs	6.4%	35.3%	63.5%
Uztura bagātinātāju reģistrs	15.1%	41.2%	58%
Diētiskās pārtikas reģistrs	8.2%	23.6%	75%
Veterināro zāļu reģistrs	16 - 21.1%	28%	70.2%

Cita plaši izplatīta datu kvalitātes problēma ir dažāda viena datu objekta notācija pat vienas datu kopas un pat parametra ietvaros (4. kolonna 4.4.1. tabulā). Šī problēma parādās 6 no 15 datu kopās (40%) 22 no 167 parametros (13.2%). Tā mēdz izpausties dažādos veidos, piemēram kā dažādi nosaukumi (a) vienas valsts nosaukumam, (b) pagatavošanas veidam, sastāvdaļai, vienības izmēram, (c) dažādi paraugi vienas vērtības pierakstam, piemēram, kontakttālrunim vai reģistrācijas numuram, t.i. ar vai bez (1) koda, (2) atdalītāja, vai dažādu atdalītāja tipa izmantošana utt. Divu datu kopu gadījumā tas tika izteikts divu dažādu notāciju veidā, norādot uz vērtības neesamību: *NULL* un ‘*0*’ (bieži vien tiek dēvēts par “neviendabīgumu”). No vienas puses šī problēma var netikt uzskatīta par svarīgu, jo abas

vērtības var norādīt uz vērtības neesamību vai neeksistenci. Tāpat kā "Valsts Informācijas Sistēmu reģistrā" nav pārliecības, ka abām vērtībām ir vienāda nozīme, jo '0' var norādīt uz vērtības vienādību ar nulli, savukārt *NULL* var nozīmēt, ka vērtība nav zināma (atbilst arī (Batini et al., 2016)). Tādā gadījumā, pat ja lietotājs ir pamanījis abu vērtību esamību datu kopā, ir grūti nolemt kā datu kopai ir jābūt apstrādātai. Cits ar eksistējošo vērtību saistīts piemērs, kas attiecās uz "pagatavošanas veidu, sastāvdaļu un vienību izmēru" pierakstu, ir parametru [*Sastavdala*] un [*Mervieniba*] vērtības, kas var tikt pierakstītas dažādos veidos: (a) vienskaitlī, (b) daudzskaitlī, (c) saīsinātajā formā vai kā abreviatūra, (d) vienā no iepriekšējiem veidiem, bet ar pareizrakstības kļūdu. Lai gan šī problēma ir sastopama 8 reizes (2 reizes ir saistīta ar vērtībām, kas norāda uz vērtību neesamību), 5 gadījumos tā varētu tikt atrisināta, ieviešot mehānismu, kas kontrolē pieļaujamo vērtību sarakstu. Uz doto brīdi izskatās, ka datu ievade notiek, lietotājiem pašiem izvēloties parametram piešķiramo vērtību. Arī šī problēma ir sastopama citās valstīs un sektoros, piemēram, (Zhang et al., 2014), Lielbritānijas *OGD* gadījumā (Kuk et al., 2011), ASV, Lielbritānijas un Japānas datos (Yi, 2019) utt..

Tabulas 5.3.1. 2. kolonnā ir apkopoti datu kvalitātes analīzes rezultāti, primārā datu objekta kvalitāti pārbaudot pret sekundāro datu objektu, atbilstoši kuriem tikai 1 no 12 datu kopai (8.3%) – "Nelaimes gadījumu darbā", kurām autore veica atbilstošo analīzi, netika konstatētas šīs kategorijas kvalitātes problēmas. Lai gan ir jāatzīmē, ka, lai konstatētu datu kvalitātes problēmas neesamību šī datu kopā, bija nepieciešamas papildus manipulācijas. Datu kopu "Nelaimes gadījumi darbā" autore analizēja pret sekundāro datu objektu "Profesiju klasifikators", ar mērķi pārliecināties darba kodu pareizībā, veicot primārā datu objektā esošo kodu pareizības analīzi pret sekundāro, kurā ir apkopoti darba kodi, kas atbilst standartiem. Šo analīzi autore veica atbilstoši kontekstuālās datu kvalitātes analīzes procedūrai (3. nodaļa). Pirmo reizi veicot pārbaudi, neviens primārā datu objektā esošais darba kods (parametra [*Cietušās personas profesija, kurā pieņemts darbā (kods)*] vērtība) netika atrasts sekundārajā datu objektā. Veicot atbilstošo parametru vērtību paraugu salīdzinājumu, autore noteica, ka primārā datu objekta atbilstoša parametra vērtība atbilst divu parametru vērtībām sekundārajā datu objektā. Citos vārdos, vērtība "*88.3332-03*" primārajā datu objektā atbilst divu parametru vērtībām "*8332*" un "*03*" sekundārajā datu objektā. Rezultātā bija nepieciešama datu pielāgošana, modificējot primārā datu objekta vērtību paraugu, atbilstoši sekundārā datu objekta atbilstoša parametra vērtības paraugam. Nepieciešamās modifikācijas ietvaros bija nepieciešams: (1) nogriezt 3 zīmes, sākot ar 2. pozīciju; (2) tā kā 2 pēdējās zīmes reprezentē sekundārā datu objekta otrā parametra vērtību, nogriezt domuzīmi, kas atrodas 3. pozīcijā no beigām, rezultātā iegūstot garāko vērtību, kas tiek salīdzināta ar (3) sekundārā datu objekta abu parametru vērtību konkatenāciju. Pēc atbilstošo modifikāciju veikšanas, atkārtotās datu

kvalitātes analīzes rezultātā datu kvalitātes problēmas primārajā datu objektā netika konstatētas. Taču, neskatoties uz iegūto rezultātu, ņemot vērā šo pārbaužu sarežģītību un papildu darbību nepieciešamību, no datu kvalitātes skatu punkta šī datu kopa nevar tikt uzskatīta par pietiekoši kvalitatīvu, jo ir ieteicams ievērot datu viendabīgumu, samazinot vai pat novēršot nepieciešamību papildu darbību veikšanā. Cits dotās kategorijas piemērs attiecās uz četrām datu kopām, ko publicēja Labklājības Ministrija, kurās 3 ierakstiem autore konstatēja datu kvalitātes problēmas parametros [*ATVK kods*] un [*Pilsēta, novads*]. Šie parametri ir paredzēti administratīvo teritoriju koda un pilsētas nosaukuma glabāšanai, kuriem ir jāatbilst datu objekta "Administratīvo teritoriju un teritoriālo vienību klasifikators" atbilstošo parametru vērtībām, taču, veicot primārā datu objekta kontekstuālo kvalitātes analīzi pret sekundāro datu objektu, tika konstatēts, ka 3 vērtības sekundārajā datu objektā nav sastopamas. Veicot sekundārajam datu objektam neatbilstošo vērtību manuālo pārbaudi, autore konstatēja, ka, iespējams, datu sniedzējs ir informēts par tām, jo atbilstošās vērtības ir ticamas - "*Kopā*", "*Ārvalstis*", "*Adrese nav norādīta*". Taču piezīme par atbilstošo vērtību pieļaujamību un datu sniedzēja zināšanu par to esamību neparādās, līdz ar ko, pat ja datu sniedzēji ņem vērā šo vērtību esamību, apstrādājot datus, datu lietotājiem tas nav zināms, līdz ar ko datu lietotāju analīzes rezultāti atkarībā no lietošanas piemēra var būt nekorekti. Tādās situācijas kopā ar datu kopu ir nepieciešams datu sniedzēja komentārs, kas atbilst arī atvērto datu principiem, atbilstoši kuriem datu sniedzējam ir jāsniedz lauku vērtību aprakstus. Pie šīs kategorijas attiecās arī cits novērojums, atbilstoši kuram, pat ja kodu vērtības primārajos un sekundārajos datu objektos mēdz būt vienādas, objektu nosaukumi, kas atbilst kodiem mēdz būt dažādi. Visbiežāk atšķiras vērtību notācijas nevis to semantiskā nozīme, piemēram, "Apstiprinātas arodslimības" datu objektā, ko darba autore analizēja pret 3 sekundārajiem datu objektiem, tika konstatētas 5 datu kvalitātes problēmas parametra [*Arodslimības grupu klasifikācija (kods)*] vērtībās un 871 datu kvalitātes problēma parametra [*Arodslimības grupu klasifikācija*] vērtībās. Konkrētajā piemērā tikai 5 primārie ieraksti un to abu parametru vērtības ir nekvalitatīvas, taču 866 no 871 otra parametra vērtībām var tikt atzītas par nekvalitatīvajām tikai, ja precīza nosaukumu sakrišana ir nepieciešama, tātad atkarībā no lietošanas piemēra tie var tikt uzskatīti par kvalitatīvajiem. Tā pati problēma ir novērojama arī [*Arodslimības izraisītājfaktors (kods)*] un [*Arodslimības izraisītājfaktors*] vērtību gadījumā, kur 22 vērtības ir nederīgas abu parametru gadījumos, taču 1052 atšķiras no sekundārā objekta vērtībām tikai pēc pieraksta. Taču 10 no 12 datu kopās ir novērojamas nopietnākas datu kvalitātes problēmas, kad primārā datu objektu parametru vērtības neatbilst sekundāro datu objektu atbilstošo parametru vērtībām ne tikai pēc sava pieraksta, bet arī pēc savas nozīmes. Kopā, 25 no 35 parametros (71.4%) autore konstatēja vismaz dažas datu kvalitātes problēmas.

Cita kvalitātes problēma (vai, iespējams, tikai anomālija) tika konstatēta datu kopas "Sociālo pakalpojumu sniedzēju skaits" 3 parametru – [*Pakalpojums ar izmitināšanu*], [*Pakalpojums bez izmitināšanas*] un [*Pakalpojums ar izmitināšanu un bez izmitināšanas*] gadījumā. No lietotāju skatupunkta ir gandrīz pašsaprotams, ka katra konkrēta ieraksta parametru [*Pakalpojums ar izmitināšanu*] un [*Pakalpojums bez izmitināšanas*] vērtību kopsummai ir jābūt vienādai ar šī ieraksta parametra [*Pakalpojums ar izmitināšanu un bez izmitināšanas*] vērtību, taču 95 ierakstiem šīs pieņēmums nav spēkā. Tam var būt vismaz divi skaidrojumi: (1) 95 ierakstos ir datu kvalitātes problēma; (2) šie lauki nav savstarpēji saistīti un pirmo divu parametru vērtību summai nav jābūt vienādai ar 3. parametra vērtību. Taču kamēr datu sniedzējs nesniedz skaidrojumu par to kā šīs vērtības tiek iegūtas, kas atbilst atvērto datu principam, nav iespējams noteikt, kurš no iemesliem ir īstais. Līdzīga datu kvalitātes "anomālija" ir novērojama arī datu kopas "VDEĀVK uzskaitē esošo bērnu ar invaliditāti skaits sadalījumā pēc administratīvās teritorijas" gadījumā, kur visu vecuma grupu parametru vērtību kopsumma nav vienāda ar parametra [*Bērni kopā*] vērtību.

Cita problēma, kura ir raksturīga 4 no 15 datu kopām (26.7%) ir dažāds savstarpēji saistīto vērtību skaits. Tas mēdz izpausties dažādos veidos: (a) vērtības dažādās valodās; (b) ID numurs un nosaukums; (c) nosaukums un paskaidrojošie vai papildu dati, piemēram, tips, valsts, kontakttālrunis vai pārstāvis.

Datu kopā "Diētiskās pārtikas reģistrs" ir identificēts visaugstākais datu kvalitātes problēmu skaits, gan nekvalitatīvo ierakstu, gan kvalitātes problēmu tipa ziņā. Viena no izplatītākajām šīs datu kopas datu kvalitātes problēmām ir dažādu datu formātu izmantošana – "*YYYY-MM-DD HH:MM:SS*" un "*DD.MM.YYYY*". Neskatoties uz to, ka vairākums parametru ir obligāts, ir izplatītas tukšas vērtības. Parametrs [*RazotajaVRN*] paredz kontekstuālas datu kvalitātes pārbaudes iespēju pret Latvijas Uzņēmumu Reģistru, veicot primārajā datu kopā esošo vērtību pārbaudi pret sekundāro, kuras rezultātā datu kvalitātes problēmas netika konstatētas, taču, veicot tāda paša rakstura kvalitātes pārbaudi parametra [*IzplatitajaVRN*] vērtībām, autore konstatēja 147 datu kvalitātes problēmas. Padziļinātā analīzes rezultātā autore konstatēja, ka datu kvalitātes problēmu galvenais iemesls ir nepareizs datu formāts, jo reģistrācijas numuru pieraksta formāts variē no pareiza (11-ciparu skaitlis) līdz "*LV-x*", "*LVx*", "*LV x*", kur '*x*' ir 11-ciparu skaitlis. Līdzīga rakstura problēma ir sastopama arī kontakttālruņu numuru gadījumā, jo 47 vērtības parametra [*IzplatitajaKontakttalrunis*] un 100 vērtības parametra [*RazotajaKontakttalrunis*] vērtības neatbilst nevienam no 15 nodefinētajiem un vispārpieņemtajiem kontakttālruņu pieraksta veidiem, t.i. paraugam, kuru starpā ir 3 Latvijas kontakttālruņu paraugi: (a) bez valsts koda; (b) ar valsts kodu, kas ir atdalīts no kontakttālruņa ar (b-1) domuzīmi ('-'), (b-2) atstarpi. Ņemot vērā, ka definētās kvalitātes prasības ir atkarīgas

no lietotāja un lietošanas piemēra, definēto prasību, un šajā gadījumā vērtību paraugu skaits un raksturs ietekmē identificēto prasībai neatbilstošo ierakstu skaitu. Piemērām, definējot tikai vienu paraugu Latvijas kontakttālruņa vērtībām, kas tradicionāli sastāv no valsts koda pieraksta "*+371* ", kuram seko 8 cipari, identificēto ierakstu skaits būtu lielāks. Kopā, 19 no 22 dotās datu kopas parametros autore identificēja vismaz dažas datu kvalitātes problēmas, no kurām 14 parametros tās netika saistītas ar tukšām vērtībām.

Papildus ir jāatzīmē problēma, kas nav saistīta ar darba tematiku, taču pietiekoši izplatīta, lai tiktu pieminēta, ir 3 datu kopām raksturīga nepareiza atdalītāju izvēle, dažādas vērtības viena parametra ietvaros atdalot ar atdalītāju, kas tiek izmantots parametru vērtību atdalīšanai, kas būtiski apgrūtina datu kopu apstrādes procesu, t.i. to ielādi datubāzē. Šī problēma ir raksturīga arī citu valsts datu kopām (Kuk et al., 2011).

Ir jāatzīmē, ka šīs apakšnodaļās minēta pētījumā (Konstante, 2016) NVD sistēmā noteiktās datu kvalitātes problēmas tiktu ātri atrastas un novērstas, it īpaši ņemot to raksturu, kas atbilst izplatītākajām datu kvalitātes problēmām, līdz ar ko to konstatēšanai būtu nepieciešamas visbiežāk definētas datu kvalitātes prasības.

Šīs apakšnodaļas rezultāti ir publicēti (Nikiforova, 2019a).

5.4. Datu kopu datu kvalitātes analīzes rezultātu apkopojums

Veicot atvērto datu kopu datu kvalitātes analīzi, autore secina, ka 83.3% datu kopās ir novērojami vismaz daži datu kvalitātes defekti. Atbilstoši (Nikiforova, 2018a, 2019a, 2019b) par to esamību nezin ne datu lietotāji, kuri var brīvi izmantot šos datus savām vajadzībām, veicot to apstrādi, analīzi un balstot darījumlēmumus uz iegūtajiem rezultātiem, ne [visticamāk] datu sniedzēji, kas ir publicējuši šos datus un izmanto tos savās informācijas sistēmās.

Visplašāk izplatītās datu kvalitātes problēmas ir:

- datu nepilnīgums pat primārajos datos (ir raksturīgs 77% analizētājām datu kopām);
- kontekstuālās datu kvalitātes problēmas;
- dažāda viena objekta notācija viena datu objekta un pat viena parametra ietvaros jeb vērtību pretrunīgums un dažādas vērtības viena reālā datu objekta apzīmēšanai;
- datu kvalitātes problēmas savstarpēji saistīto parametru gadījumā.

Katru kategoriju autore apskatīja iepriekšējas apakšnodāļās, līdz ar ko to skaidrojums netiks atkārtots.

Darba ietvaros datu kvalitātes problēmas datu kvalitātes analīzes vienas datu kopas ietvaros autore konstatēja 83.3% analizētājām datu kopām. Kontekstuālās datu kvalitātes analīzes rezultātā vairāku datu kopu ietvaros, datu kvalitātes problēmas autore konstatēja 94.4% analizētājās datu kopās. Tas nozīmē, ka atvērtajiem datiem ir raksturīgas datu kvalitātes problēmas.

Datu kvalitātes problēmas atvērtajos datos ir raksturīgas ne tikai Latvijas atvērtajiem datiem, bet arī citu valstu atvērtajiem datiem, jo visās analizētajās ārzemju datu kopās autore konstatēja datu kvalitātes problēmas, izvēloties pat vienu no vienkāršākajiem lietošanas piemēriem. Tas saskaņojas arī ar citu autoru pētījumu rezultātiem, ko autore apskatīja 2. nodaļā.

Veiktās analīzes rezultāti rāda, ka vairākas konstatētas datu kvalitātes problēmas ir sistemātiskas un ir novērojamas gan konkrētu domēnu reprezentējošos datos, gan valsts, gan arī dažādu valsts gadījumos. Vairākums datu kvalitātes problēmu var tikt atrisināts, veicot nelielus labojumus, kas neprasa daudz resursu (gan laika, gan cilvēkresursu ziņā), ja šīs problēmas ir sistemātiskas vai ir saistītas ar konkrētu vērtību, kas ir sastopama vairākos ierakstos, t.i. pat vienas vērtības labošana varētu būtiski uzlabot kopēju datu kopas kvalitāti, vienlaicīgi atrisinot problēmas vairākos ierakstos (Nikiforova, 2019b).

Atbilstoši (Nikiforova, 2019a, 2019b) un citiem autoriem, viena no galvenajam atvērto datu izmantošanas problēmām ir tas, ka datu kvalitātes datu galalietotājiem nav zināma, jo vairāk, nav zināms pie kādiem nosacījumiem jeb lietošanas piemēriem konkrētas datu kopas būs pietiekami kvalitatīvas, analīžu veikšanai un darījumlēmumu pieņemšanai, un, kad tās ir nelietojamas, to izmantošanas rezultātā iegūstot neprecīzus vai pat nederīgus rezultātus. Nosakot konkrētas datu kopas atbilstību galalietotāja vajadzībām, t.i. viņa definētajam lietošanas piemēram, tās kvalitātei ir jābūt iepriekš pārbaudītai, pārliecinoties, ka tā apmierina konkrēta lietošanas piemēra nosacījumus, kas ir panākams, pielietojot iepriekšaprakstīto procedūru. Tādējādi tiek apstiprināta darba sākumā izvirzītā 7. tēze, t.i. atvērtajos datos ir sastopamas datu kvalitātes problēmas, kuru noteikšanai ir piemērota izstrādātā pieeja datu kvalitātes analīzei, nodrošinot "trešo pušu" datu kvalitātes analīzi.

Ņemot vērā, ka darbā aprakstītie piemēri atbilst visu datu kopu padziļinātajai kvalitātes analīzei, kas tuvojas "absolūtas" datu kvalitātei, pārbaudot datu kopu kvalitāti atbilstoši vairākiem (bet, protams, ne visiem) iespējamiem lietošanas piemēriem, izveidotas diagrammas un veiktās pārbaudes ir apjomīgākas un sarežģītākas, salīdzinājumā ar tām, kas tiks definētas ar galalietotājiem, analizējot datu kvalitāti saviem nolūkiem. Ņemot vērā, ka visi piedāvātās pieejas komponenti ir vienkārši, viennozīmīgi definēti un intuitīvi tāpat kā viss kvalitātes modelis, ir pamats uzskatīt, ka to izmantos plašs lietotāju loks, ieskaitot lietotājus bez padziļinātām zināšanām IT un datu kvalitātes jomās. Tās nodrošinātu ne tikai datu kvalitātes

pārbaudi savos nolūkos, bet sekmētu sadarbību ar datu sniedzējiem, piemēram, izmantojot šīm nolūkam Latvijas Atvērto datu portālā paredzēto atgriezenisko saiti, ziņojot par atklātām datu kvalitātes problēmām, tādejādi veicinot datu kvalitātes uzlabošanu, tuvojoties absolūtai datu kvalitātei.

NOBEIGUMS

Pētījuma gaitā autore izpētīja literatūru par datu kvalitātes problēmu, tās aktualitāti, eksistējošām pieejām datu kvalitātes analīzei un novērtēšanai. Iegūtās zināšanas tika pielietotas, izstrādājot alternatīvo datu kvalitātes novērtēšanas pieeju.

Apkopojot informāciju, iegūtu no literatūras avotiem, patstāvīgas tēmas izpētes, iegūto zināšanu pielietošanas rezultātā, autore secina:

1) "datu kvalitāte" ir sarežģīts relatīva rakstura jēdziens, atbilstoši kuram datu kvalitāte ir datu piemērotība konkrēta lietotāja lietošanas piemēram. Datu kvalitāte ir atkarīga no konteksta, un tāpat kā paši dati IS mainās laika gaitā, uzkrājoties pakāpeniski, arī datu kvalitātes prasības laika gaitā var mainīties;
2) datu kvalitātes problēma ir sen zināma problēma, par ko liecina arī agrīnie pētījumi, kas tika uzsākti jau 60-o gadu beigās. Pieaugot datu popularitātei, strauji pieaug arī datu kvalitātes problēmas popularitāte un aktualitāte. Mūsdienās, parādoties un kļūstot populāriem atvērtajiem datiem, tā kļūst arvien populārāka, taču atvērto datu kvalitāte tiek nepamatoti maz pētīta, neskatoties uz jauniem izaicinājumiem, kas izriet no atvērto datu rakstura;
3) zinātnisko pētījumu skaits, kas pēta atvērto datu kvalitātes problēmu, ir nepamatoti zems – to īpatsvars pret kopējo ar atvērtajiem datiem saistīto pētījumu skaitu nepārsniedz 0.5%. Datu kvalitātes pētījumu skaits pārsniedz atvērto datu kvalitātes pētījumu skaitu gandrīz 196 reizēs (t.i. atvērto datu kvalitātes pētījumu īpatsvars pret kopējo ar datu kvalitāti saistīto pētījumu skaitu, ir ~0.2%);
4) pētījuma gaitā apskatot vairāk kā 65 pētījumus, kas pēta datu kvalitātes problēmu, darbā apskatot vairāk kā 25 risinājumus, autore secina, ka eksistējošie pētījumi un piedāvātie datu kvalitātes problēmas risinājumi var tikt iedalīti vairākās grupās: (a) uz dimensiju definēšanu, to grupēšanu un datu kvalitātes novērtēšanu vērstie pētījumi; (b) atvērto datu portālu vai atvērto pārvaldes datu kvalitātes vērtēšana; (c) saistīto datu kvalitātes vērtēšana. Eksistējošo risinājumu analīze rada, ka vairākums pētījumu nav piemērots lietotājiem bez padziļinātām zināšanām IT un datu kvalitātes jomās, kas mūsdienās apstākļos nav pieņemami, jo vairākums lietotāju ikdienā saskarās ar datiem, par kuru kvalitāti ir jābūt iespējai pārliecināties. Šīs iespējas nepieciešamība ir saistīta arī ar atvērto datu popularitāti.

   Vairāki risinājumi izmanto lielu dimensiju skaitu, paredz datu kvalitātes dimensiju un prasību definēšanu, prasa attiecināt definētas vai jau eksistējošas datu kvalitātes prasības uz atbilstošām dimensijām, kas mēdz izraisīt grūtības pat datu kvalitātes

speciālistiem. Savukārt autores pētījums liecina par to, ka 96.4% ne-IT speciālistu nav dzirdējuši "datu kvalitātes dimensijas" jēdzienu. Kopumā vairākums eksistējošo risinājumu paredz datu kvalitātes un IT speciālistu iesaisti visos datu kvalitātes analīzes posmos, savukārt datu kvalitātes analīze atkarībā no konkrētā lietošanas piemēra, kā arī atvērto datu izmantošanas iespēja ar jebkuru ieinteresēto personu, paredz galalietotāja datu kvalitātes analīzes procesa kontroli, samazinot IT-speciālistu iesaisti.

Atbilstoši ar datu kvalitātes jautājumiem saistīto esošo pētījumu tēmu izpētei, neskatoties uz atvērto datu popularitāti, to kvalitātei tiek veltīts pārāk maz uzmanības. Lai atrisinātu literatūras izpētes un eksistējošo datu kvalitātes risinājumu analīzes rezultātā konstatētas problēmas, tika piedāvāta kardināli jauna datu objekta virzīta pieeja datu kvalitātes analīzei, kas būtiski atšķiras no eksistējošām, taču ievēro vispārpieņemtus ar datu kvalitātes jēdzienu saistītus pamatprincipus. Tā tika izstrādāta, ņemot vērā eksistējošo pieeju trūkumus. To var raksturot sekojoši:

1) ir piedāvāts kvalitātes modelis, kas sastāv no trīs pamatkomponentiem: (1) datu objekts, kura datu kvalitātes tiek vērtēta; (2) datu kvalitātes prasības, kas ir atkarīgas no konkrēta lietošanas piemēra jeb datu lietojuma; (3) datu kvalitātes pārbaudes process. Tie tiek definēti, izmantojot grafiskās diagrammas. Piedāvātā pieeja nesaista datu kvalitāti ar "datu kvalitātes dimensiju" jēdzienu, aizstājot to ar universālāku jēdzienu "datu kvalitātes prasība", kura ir uz datu kvalitāti attiecināmu datu kvalitātes dimensiju jēdzienu virskopa;
2) ir nodrošināta datu kvalitātes prasību specifikācijas (datu kvalitātes modeļa), kas ir formulēts *DSL* jēdzienos, izpildāmība. Tās izpildes rezultātā tiek iegūts datu kvalitātes pārbaužu izpildes protokols, kurā tiek reģistrēti visi konkrētām datu objektam raksturīgi datu defekti, kas turpmāk var tikt izmantots datu kvalitātes uzlabošanai;
3) piedāvātā pieeja ir lietotājorientētā pieeja, kur katru piedāvātās datu kvalitātes modeļa komponentu definē galalietotājs, pārbaudot konkrētas datu kopas datu kvalitāti saviem nolūkiem, fokusējoties tikai uz tiem datu objektu raksturojošiem parametriem, kas viņam konkrētas analīzes ietvaros ir svarīgi. Tādejādi datu kvalitātes analīzes rezultāti atbilst sākotnējām galalietotāja iecerēm, t.i. datu kvalitātes jēdziena izpratnei;
4) datu kvalitātes modelis var tikt formulēts un izmantots divos veidos: (a) neformāli (līdzīgi *PIM*), kur nepieciešamas pārbaužu darbības tiek aprakstītas dabiskā valodā – diagrammu simboli satur darbību tekstuālus aprakstus un (b)

formāli jeb izpildāmā veidā (līdzīgi *PSM*), kas var tikt panākts, pārveidojot neformālo modeli, aizstājot neformālus tekstus ar izpildāmiem tekstiem, piemēram, *C#* programmkodu, *SQL* vaicājumiem vai cita veida izpildāmiem objektiem. Datu kvalitātes modeļa formulēšana divos veidos un citi piedāvātās pieejas aspekti un tās "filozofija" ļauj salīdzināt un apskatīt piedāvāto pieeju no *MDA* skatupunkta.

Datu objekta un prasību specifikācijas definēšana neformālos jēdzienos ļauj iesaistīties datu kvalitātes analīzē lietotājiem bez padziļinātām zināšanām IT jomā, savukārt blokshēmām līdzīgo diagrammu izmantošana katra datu kvalitātes modeļa definēšanai veicina vairāku lietotāju savstarpēju mijiedarbību;

5) katrs piedāvātās pieejas komponents ir definējams, izmantojot relatīvi vienkāršo *DSL* valodu. Piedāvātie modeļi ir ātri un vienkārši veidojami, rediģējami un atkalizmantojami. Piedāvātā *DSL* platforma nodrošina plašu kvalitātes prasību valodu definēšanas spektru, ļaujot definēt kvalitātes prasības modulāri un pārbaudīt prasību izpildi dažādos datu apstrādes posmos, necenšoties iekļaut datu kvalitātes prasības vienā visaptverošā prasību specifikācijā, kur datu kvalitāte tiek pārbaudīta tikai datu uzkrāšanas gala posmā;
6) vairākums datu kvalitātes analīzes soļu neprasa no lietotājiem iepriekšējas padziļinātas zināšanas IT vai datu kvalitātes jomā. Datu kvalitātes analīzes process kļūst intuitīvs, kas ļauj veikt pieņēmumu, ka pieeja ir paredzēta plašam lietotāju lokam. IT specialistu iesaiste var kļūst nepieciešama tikai beidzamajā posmā, neformālas prasības pārveidojot par izpildāmām, tādejādi IT-specialisti veic atbalsta funkciju, neietekmējot datu kvalitātes analīzes pamatkomponenšu, t.i. datu objekta un uz to attiecināmu datu kvalitātes prasību definēšanu;
7) izstrādātā pieeja ir ārējs risinājums, kas ļauj analizēt "trešo pušu" datu kopu kvalitāti neatkarīgi no sistēmas, kurā dati tika uzkrāti, neprasot zināšanas par to uzkrāšanas un apstrādes mehānismiem. Piedāvātā pieeja ir piemērota gan strukturētu, gan daļēji strukturētu datu analīzei, to darbību demonstrējot, pielietojot to dažāda veida atvērtajiem datiem;
8) piedāvātā pieeja paredz iespēju veikt datu kvalitātes analīzi gan datu lietotājiem, t.i. lietotājiem, kas izmanto "trešo pušu" datus, gan datu turētājiem, t.i. tiem, kas uzkrāj un apstrādā datus savās IS;

9) datu objekta virzītās pieejas pielietošanas "trešo pušu" datiem, precīzāk atvērtajiem datiem, rezultātā autore (a) nodemonstrēja piedāvātās pieejas piemērotību "trešo pušu" datu kvalitātes analīzei, (b) konstatēja vairākas dažāda veida datu kvalitātes problēmas, kas ir raksturīgas gan Latvijas, gan ārzemju atvērtajiem datiem. Veicot datu kvalitātes analīzi datu kopām vienas datu kopas ietvaros, datu kvalitātes problēmas autore identificēja 83.3% datu kopās. Īpašu uzmanību prasa Latvijas atvērtie medicīnas dati, kuros ir novērojamas vairākas datu kvalitātes problēmas. Apkopojot tās un klasificējot pēc to rakstura, var secināt, ka daži datu kvalitātes problēmas veidi var tikt uzskatīti par plaši izplatītu tendenci;
10) pieejas paplašināšana ar kontekstuālo datu kvalitātes pārbaudes iespēju, nodrošinot iespēju veikt datu objekta kvalitātes analīzi vairāku datu objektu kontekstā, kas var tikt iegūti no dažādiem datu sniedzējiem, tādejādi veicot padziļināto kvalitātes analīzi, ļauj būtiski uzlabot datu kvalitātes analīzes rezultātus. Par to liecina tās pielietošanas rezultāti 18 datu kopām, datu kvalitātes problēmas konstatējot 17 no tām (94.4%). Parametru skaita ziņā, veicot datu kvalitātes analīzi 61 parametram, 83.6% parametros autore konstatēja vismaz dažas datu kvalitātes problēmas. Tas nodrošina arī datu objektu atkalizmantošanu;
11) neskatoties uz datu kvalitātes esamību 83.3% analizētājās datu kopās, vairākums datu kvalitātes problēmu varētu atrisināt, veicot nelielus labojumus, kas neprasa daudz resursu.

    Visplašāk izplatītas datu kvalitātes problēmas ir: (a) datu nepilnīgums pat primārajos datos (ir raksturīgs 77% analizētājām datu kopām); (b) kontekstuālās datu kvalitātes problēmas; (c) nederīgie datumi; (d) dažāda viena objekta notācija viena datu objekta un pat viena parametra ietvaros jeb vērtību pretrunīgums un dažādas vērtības viena reālā datu objekta apzīmēšanai; (e) datu kvalitātes problēmas savstarpēji saistīto parametru gadījumā.

Piedāvātā risinājuma specifika, t.sk. tā piemērotība "trešo pušu" datu kvalitātes analīzei ar galalietotājiem viņu nolūkiem, ļauj secināt, ka dotais risinājums ir viens no retajiem risinājumiem, kas pārklāj arī atvērto datu kvalitātes apgabalu.

Piedāvātās pieejas plašas iespējas un relatīva vienkāršība ļauj veikt pieņēmumu, ka piedāvātā pieeja tiks izmantota ne tikai ar galalietotājiem savām vajadzībām, bet arī ar atvērto datu kvalitātes entuziastiem, kas kļūst populāri visā pasaulē, tajā skaitā arī Latvijā. Vismaz Latvijas entuziastu iesaiste atvērto datu kvalitātes analīzē un atgriezeniskās saites izmantošana,

kas ir nodrošināta arī Latvijas Atvērto datu portālā, veicinās datu kvalitātes kopējo uzlabošanu nacionālajā līmenī, tuvojoties "absolūtai" datu kvalitātei – datu kvalitātei, kas apmierina visus iespējamus lietošanas piemērus.

Izstrādātās pieejas potenciāls tika nodemonstrēts, uz piedāvāto datu kvalitātes modeli balstot izpildlaika datu kvalitātes verifikācijas risinājumu, kas nodrošina datu kvalitātes pārbaudi gandrīz reālā laikā, veicot datu kvalitātes analīzi darījumprocesa izpildes laikā.

Datu objekta virzītās pieejas formalizācija ļāva piedāvāt daļēji formālu datu kvalitātes teoriju, kura, neskatoties uz vairākkārtējiem datu pētnieku mēģinājumiem pēdējo dekāžu laikā, tā arī netika piedāvāta.

Pētījuma rezultātā autore izstrādāja Specseminārus "Datu kvalitāte" (2019./ 2020. akad. gada rudens semestris) un "Atvērtie dati un datu kvalitāte" (2019./ 2020. akad. gada pavasara semestris), kurus vada bakalaura studiju "Datorzinātnes" programmas studentiem, savukārt šobrīd darba rezultāti tiek izmantoti 2019. gadā uzsāktā IT kompetences centra pētījumā Nr. 1.7. "Biznesa procesu modeļu lietojums pilnai informācijas sistēmas funkcionalitātes testēšanai".

Darba rezultāti ir publicēti 10 zinātniskajos rakstos, prezentēti piecās starptautiskajās konferencēs, kā arī ar Ekonomikas ministriju rīkotājā *"Digitālizācijas un Inovāciju foruma DIG-IN" POP^{up} Demo* centrā, savukārt pētījumu rezultātus, kas attiecas uz atvērto datu kvalitāti, 2020. gadā autore prezentēja Latvijas Atvērto Tehnoloģiju Asociācijas konferencē "Datu virzītā nācija" kā uzaicinātais lektors.

Darba autores pētījumu rezultāti tika augsti novērtēti ar ārzemju datu zinātniekiem, un par saviem datu kvalitātes pētījumu rezultātiem 2019. gadā darba autori nominēja uz vienu no pasaules prestižākajiem apbalvojumiem "*WDS Data Stewardship Award*", kas ikgadēji tiek piešķirts perspektīvākajiem datu zinātniekiem.

Visi darba sākumā izvirzītie uzdevumi tika izpildīti un mērķi tika sasniegti.

PATEICĪBAS

Paldies darba vadītājiem asoc. prof. Zanei Bičevskai un prof. Jānim Bičevskim par atbalstu darba tapšanas procesā, par vērtīgiem komentāriem un darba posmu novērtējumu. Esam daudz un cītīgi līdzdarbojušies trijos Eiropas Savienības struktūrfondu līdzfinansētos pētniecības projektos, sagatavojot vairākas kopīgās zinātniskās publikācijas.

Paldies prof. Marlonam Dumas (*Marlon Dumas*), prof. Markam Nisenam (*Marc Nyssen*), prof. Andrim Ambainim par vērtīgām piezīmēm par izstrādāto pieeju, pateicoties kurām ir izdevies gan precizēt dažas detaļas, gan paskatīties uz izstrādāto risinājumu no cita "leņķa", rezultātā saprotot kā tas varētu tikt pielāgots arī citiem nolūkiem, risinot problēmas, kuras līdz šīm netika aplūkotas.

Paldies arī maniem studentiem par brīvprātīgu un godprātīgu dalību rīkotājās aptaujās un aktīvu dalību izstrādātājā Specseminārā "Datu kvalitāte". Paldies Latvijas Universitātes Datorikas Fakultātes mācībspēkiem, kas ar savu attieksmi ir ļāvuši noticēt saviem spēkiem un iesaistīties gan pētnieciskajā, gan akadēmiskajā vidē.

Paldies arī Rīgas Tehniskās Universitātes Attīstības Fondam, kas studiju gaitā vairākkārt ir atbalstījis darba autori, augsti vērtējot gan viņas akadēmiskus, gan pētnieciskus sasniegumus.

Paldies konferences *"Social Networks Analysis, Management and Security"(SNAMS)* organizatoriem, it īpaši Stefordšīras Universitātes (Staffordshire University) prof. *Elhadj Benkhelifa* un Jordānijas Zinātnes un Tehnoloģiju Universitātes (*Jordan University of Science and Technology*) asoc. prof. *Mohammad AL-Smadi*, par viņu draudzīgu atbalstu, kā arī aicinājumu darbseminära *Data Science Engineering and its Application (DSEA)* tehniskās programmas komitejā. Savas jomas darbu lasīšana un recenzēšana ir vērtīga pieredze, kas noteikti palīdz ne tikai iedziļināties citu autoru darbos, bet arī atskatīties uz saviem darbiem, rezultātā nosakot iespējas savu darbu uzlabošanai.

Atsevišķa pateicība ir izsakāma prof. Markam Nisenam (*Marc Nyssen*), kas ir augsti novērtējis darba autores sasniegumus izvēlētājā jomā, izceļot uz citu datu pētnieku fona un nominējot viņu uz vienu no pasaules prestižākajiem apbalvojumiem "*WDS Data Stewardship Award 2019*". Šīs novērtējums ir ļoti vērtīgs un nozīmīgs, kas ļauj noticēt ne tikai sava darba nozīmīgumā, bet arī savos spēkos, veicot pētniecisko darbu.

Īpašs paldies arī vecākiem par sniegtu morālo atbalstu! Paldies visiem, kas tiešajā vai netiešajā veidā ir atbalstījis darba autori viņas izvēlētājā ceļā!

IZMANTOTĀ LITERATŪRA UN AVOTI

(Aarshi et al., 2018) Aarshi, S., Malik, B. H., Habib, F., Ashfaq, K., Saleem, I., & Tariq, U. (2018). Dimensions of open government data web portals: A case of Asian countries. *International Journal of Advanced Computer Science and Applications*, *9*(6), 459-469.

(Acosta et al., 2013) Acosta, M., Zaveri, A., Simperl, E., Kontokostas, D., Auer, S., & Lehmann, J. (2013, October). Crowdsourcing linked data quality assessment. In *International semantic web conference* (pp. 260-276). Springer, Berlin, Heidelberg.

(Akehurst et al., 2002) Akehurst, D. H., Bordbar, B., Rodgers, P., & Dalgliesh, N. T. G. (2002). Automatic normalisation via metamodelling. In *ASE 2002 Workshop on Declarative Meta Programming to Support Software Development*.

(Andreassen et al., 2007) Andreassen, H. K., Bujnowska-Fedak, M. M., Chronaki, C. E., Dumitru, R. C., Pudule, I., Santana, S., ... & Wynn, R. (2007). European citizens' use of E-health services: a study of seven countries. *BMC public health*, *7*(1), 53.

(Arnold, 1992) Arnold, S. E. (1992). Information manufacturing: the road to database quality. Database, 15(5), 32-39.

(Ashkham et al., 2013) Askham, N., Cook, D., Doyle, M., Fereday, H., Gibson, M., Landbeck, U., Lee, R., Maynard, C., Palmer, G., Schwarzenbach, J. (2013). The six primary dimensions for data quality assessment. *DAMA UK Working Group*, 432-435.

(Attard et al., 2015) Attard, J., Orlandi, F., Scerri, S., & Auer, S. (2015). A systematic review of open government data initiatives. *Government Information Quarterly*, *32*(4), 399-418.

(Auguston et al., 2015) Auguston, M., Giammarco, K., Baldwin, W. C., & Farah-Stapleton, M. (2015). Modeling and verifying business processes with Monterey Phoenix. *Procedia Computer Science*, *44*, 345-353.

(Bārzdiņš et al., 2007) Bārzdiņš, J., Zariņš, A., Čerāns, K., Kalniņš, A., Rencis, E., Lāce, L., Liepiņš, R., Sproģis, A. (2007). GrTP: transformation based graphical tool building platform. In *The 10th International Conference on Model-Driven Engineering Languages and Systems, Models*.

(Basciani et al., 2016) Basciani, F., Di Rocco, J., Di Ruscio, D., Iovino, L., & Pierantonio, A. (2016, September). A customizable approach for the automated quality assessment of modelling artifacts. In *2016 10th International Conference on the Quality of Information and Communications Technology (QUATIC)* (pp. 88-93). IEEE.

(Batini et al., 2016) Batini, C., & Scannapieco, M. (2016). Data and information quality. *Cham, Switzerland: Springer International Publishing. Google Scholar*.

(Batini et al., 2009) Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM computing surveys (CSUR)*, *41*(3), 16.

(Batini et al., 2006) Batini, C., & Pernici, B. (2006, August). Data Quality Management and Evolution of Information Systems. In *IFIP World Computer Congress, TC 8* (pp. 51-62). Springer, Boston, MA.

(Bauer et al., 2011) Bauer, F., & Kaltenböck, M. (2011). Linked Open Data: The Essentials, edition mono/monochrom. *Vienna, Austria*.

(Beno et al., 2017) Beno, M., Figl, K., Umbrich, J., & Polleres, A. (2017, May). Open data hopes and fears: determining the barriers of open data. In *2017 Conference for E-Democracy and Open Government (CeDEM)* (pp. 69-81). IEEE.

(Bertossi et al., 2016) Bertossi, L., & Rizzolo, F. (2016). Contexts and data quality assessment.

(Bevan et al., 2012) Bevan, C., & Strother, D. (2012). Best practices for evaluating method validity, data quality and study reliability of toxicity studies for chemical hazard risk assessments. *Washington (DC): American Chemical Council, Centre for Advancing Risk Assessment Science and Policy*.

(Bicevskis et al., 2019a) Bicevskis, J., Nikiforova, A., Bicevska, Z., Oditis, I., Karnitis, G. (2019). A step towards a data quality theory. In *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*. IEEE (ieguldījums: 50%);

(Bicevskis et al., 2019b) Bicevskis, J., Bicevska, Z., Nikiforova, A., Oditis, I. (2019). Towards Data Quality Runtime Verification. In *2019 Federated Conference on Computer Science and Information Systems (FedCSIS)*. IEEE (ieguldījums: 40%)

(Bicevskis et al., 2018a) Bicevskis, J., Bicevska, Z., Nikiforova, A., Oditis, I. (2018). An Approach to Data Quality Evaluation. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)* (pp. 196-201). IEEE (ieguldījums: 50%)

(Bicevskis et al., 2018b) Bicevskis, J., Bicevska, Z., Nikiforova, A., Oditis, I. (2018). Data quality evaluation: a comparative analysis of company registers' open data in four European countries. In *FedCSIS Communication Papers* (pp. 197-204) (ieguldījums: 35%)

(Bizer, 2007) Bizer, C. (2007). *Quality-driven information filtering in the context of web-based information systems* (Doctoral dissertation, Wirtschaftswissenschaft, Freie Universität Berlin, 2007.

(Bojārs et al., 2014) Bojars, U., & Liepins, R. (2014). The State of Open Data in Latvia: 2014. *Baltic Journal of Modern Computing*, *2*(3), 160.

(Bovee et al., 2001) Bovee, M., Srivastava, R. P., & Mak, B. (2003). A conceptual framework and belief-function approach to assessing overall information quality. *International journal of intelligent systems*, *18*(1), 51-74.

(Bray et al., 2009) Bray, F., & Parkin, D. M. (2009). Evaluation of data quality in the cancer registry: principles and methods. Part I: comparability, validity and timeliness. *European journal of cancer*, *45*(5), 747-755.

(Brønnøysundregistrene, 2018) Brønnøysundregistrene. "Enhetsregisteret - Åpne Data" [Norvēģijas Uzņēmumu reģistrs] (norvēģu val.). [tiešsaiste]. - [atsauce 10.10.2019.]. Pieejams: http://data.brreg.no/oppslag/enhetsregisteret/enheter.xhtml

(Bula, 2019) Bula, I., Nacionālā Enciklopēdija, Matemātika, 2019, https://enciklopedija.lv/skirklis/1133

(Bullinger et al., 2012) Bullinger, A. C., Rass, M., Adamczyk, S., Moeslein, K. M., & Sohn, S. (2012). Open innovation in health care: Analysis of an open health platform. *Health policy*, *105*(2-3), 165-175.

(Cabitza et al., 2016) Cabitza, F., Batini, C. (2016). Information quality in healthcare. In Data and Information Quality (pp. 403-419). Springer, Cham.

(Cai et al., 2015) Cai, L., & Zhu, Y. (2015). The challenges of data quality and data quality assessment in the big data era. *Data science journal*, *14*.

(Cannon, 2016) Cannon, K., "Trillium Software Data Quality for Dynamics Trillium Software Trillium is a global provider and innovator of data quality solutions Part of.", [tiešsaiste].- [atsauce 10.10.2019.].Pieejams: http://slideplayer.com/slide/10388066/

(Caro et al., 2007) Caro, A., Calero, C., & Piattini, M. (2007, November). A Portal Data Quality Model For Users And Developers. In *ICIQ* (pp. 462-476).

(Carroll et al., 2006) Carroll, R., Fahy, C., Lehtihet, E., van der Meer, S., Georgalas, N., & Cleary, D. (2006, April). Applying the P2P paradigm to management of large-scale distributed networks using a Model Driven Approach. In *2006 IEEE/IFIP Network Operations and Management Symposium NOMS 2006* (pp. 1-14). IEEE.

(Castro et al., 2015) Castro, D., & Korte, T. (2015). Open Data in the G8: A Review of Progress on the Open Data Charter.

(Chen et al., 2016) Chen, D., Asaolu, B., & Qin, C. (2016). Big Data Analytics In The Public Sector: a Case Study of Neet Analysis For The London Boroughs. *IADIS International Journal on Computer Science & Information Systems*, *11*(2).

(Chen et al., 2014) Chen, H., Hailey, D., Wang, N., & Yu, P. (2014). A review of data quality assessment methods for public health information systems. *International journal of environmental research and public health*, *11*(5), 5170-5207.

(Chungoora et al., 2013) Chungoora, N., Young, R. I., Gunendran, G., Palmer, C., Usman, Z., Anjum, N. A., Cutting-Decelle, A., Harding, J., Case, K. (2013). A model-driven ontology approach for manufacturing system interoperability and knowledge sharing. *Computers in Industry*, *64*(4), 392-401.

(Colpaert et al., 2013) Colpaert, P., Joye, S., Mechant, P., Mannens, E., & Van de Walle, R. (2013). The 5 stars of open data portals. In *Proceedings of the 7th International Conference on Methodologies, Technologies and Tools Enabling E-Government (MeTTeG13), University of Vigo, Spain* (pp. 61-67).

(Coutinho et al., 2012) Coutinho, C., Cretan, A., & Jardim-Goncalves, R. (2012, September). Negotiations framework for monitoring the sustainability of interoperability solutions. In *International IFIP Working Conference on Enterprise Interoperability* (pp. 172-184). Springer, Berlin, Heidelberg.

(Czechowski et al., 2015) Czechowski, P., Badyda, A., & Majewski, G. (2013). Data mining system for air quality monitoring networks. *Archives of Environmental Protection*, *39*(4), 123-147.

(Dahbi et al., 2018) Dahbi, K. Y., Lamharhar, H., Chiadmi, D. (2018). Toward an Evaluation Model for Open Government Data Portals. In International Conference Europe Middle East & North Africa Information Systems and Technologies to Support Learning (pp. 502-511). Springer, Cham.

(English, 2009) English, L. P. (2009). *Information quality applied: Best practices for improving business information, processes and systems*. Wiley Publishing.

(Eppler, 2006) Eppler, M. J. (2006). *Managing information quality: Increasing the value of information in knowledge-intensive products and processes*. Springer Science & Business Media.

(European Data Portal, 2018) European Data Portal. LATVIA. State-of-Play on Open Data - 2018. [tiešsaiste]. - [atsauce 10.10.2019.]. Pieejams: https://www.europeandataportal.eu/sites/default/files/country-factsheet_latvia_2018.pdf

(FAO, 2019) FAO, Food and Agriculture Organization of the united Nations. Country codes/ names. [tiešsaiste]. - [atsauce 10.10.2019.]. Pieejams: http://www.fao.org/countryprofiles/iso3list/en/

(Färber et al., 2018) Färber, M., Bartscherer, F., Menne, C., & Rettinger, A. (2018). Linked data quality of dbpedia, freebase, opencyc, wikidata, and yago. *Semantic Web*, *9*(1), 77-129.

(Ferney et al., 2017) Ferney, M. M. J., Estefan, L. B. N., & Alexander, V. V. J. (2017). Assessing data quality in open data: A case study. In *2017 Congreso Internacional de Innovacion y Tendencias en Ingenieria (CONIITI)* (pp. 1-5). IEEE.

(Fisher et al., 2001) Fisher, C. W., & Kingma, B. R. (2001). Criticality of data quality as exemplified in two disasters. *Information & Management*, *39*(2), 109-116.

(Friedman et al., 2013) Friedman, T., & Judah, S. (2013). The state of data quality: Current practices and evolving trends. *Stamford: Gartner*.

(Friedman et al., 2011) Friedman, T., & Smith, M. (2011). Measuring the business value of data quality. *Gartner, Stamford, 464*.

(Gabernet et al., 2017) Gabernet, A., Limburn, J. "Breaking the 80/20 rule: How data catalogs transform data scientists' productivity". IBM. 2017. [tiešsaiste]. - [atsauce 10.10.2019.]. Pieejams: https://www.ibm.com/blogs/bluemix/2017/08/ibm-data-catalog-data-scientists-productivity/

(Gandhi, 2016) Gandhi, G. "What is SSIS? Its advantages and disadvantages". 2016. [tiešsaiste]. - [atsauce 10.10.2019.]. Pieejams: https://www.sarjen.com/ssis-advantages-disadvantages/

(Global Open Data Index, 2018) Global Open Data Index. [tiešsaiste]. - [atsauce 10.10.2019.]. Pieejams: https://index.okfn.org/

(Guha-Sapir et al., 2002) Guha-Sapir, D., & Below, R. (2002). *The quality and accuracy of disaster data: a comparative analyses of three global data sets*. World Bank, Disaster Management Facility, ProVention Consortium.

(Hashem et al., 2015) Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Khan, S. U. (2015). The rise of "Big data" on cloud computing: Review and open research issues. *Information systems*, *47*, 98-115.

(Haubold et al., 2010) Haubold, T., Beier, G., Golubski, W., & Herbig, N. (2010). The GeneSEZ approach to model-driven software development. In *Advanced Techniques in Computing Sciences and Software Engineering* (pp. 395-400). Springer, Dordrecht.

(ISO, 2008) ISO/IEC 25012:2008: Software engineering - Software product Quality Requirements and Evaluation (SQuaRE) - Data quality model, International Organization for Standartization. [tiešsaiste]. - [atsauce 10.11.2019.]. Pieejams: https://www.iso.org/standard/35736.html

(ISO/IEC 25024, 2015) ISO/IEC 25024:2015: Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Measurement of data quality. [tiešsaiste]. - [atsauce 4.12.2019.]. Pieejams: https://www.iso.org/obp/ui/#iso:std:iso-iec:25024:ed-1:v1:en

(Janssen et al., 2012) Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, adoption barriers and myths of open data and open government. *Information systems management*, *29*(4), 258-268.

(Jarke, 1999) Jarke, M., Jeusfeld, M. A., Quix, C., & Vassiliadis, P. (1999). Architecture and quality in data warehouses: An extended repository approach. *Information Systems*, *24*(3), 229-253.

(Jayawardene et al., 2015) Jayawardene, V., Sadiq, S., & Indulska, M. (2015). An analysis of data quality dimensions.

(Jayawardene et al., 2013) Jayawardene, V., Sadiq, S., & Indulska, M. (2013). The curse of dimensionality in data quality. In *24th Australasian Conference on Information Systems (ACIS)* (pp. 1-12). RMIT University.

(Jetzek, 2017) Jetzek, T. (2017). Innovation in the open data ecosystem: Exploring the role of real options thinking and multi-sided platforms for sustainable value generation through open data. In *Analytics, Innovation, and Excellence-Driven Enterprise Sustainability* (pp. 137-168). Palgrave Macmillan, New York.

(Juran, 1995) Juran, J. M. (1995). *Managerial breakthrough: The classic book on improving management performance*. McGraw-Hill.

(Karel, 2015) Karel R., "The "All In" Costs of Poor Data Quality. It goes beyond dollars and cents", ComputerWorld, 2015. [tiešsaiste]. - [atsauce 10.10.2019.]. Pieejams: https://www.computerworld.com/article/2949323/the-all-in-costs-of-poor-data-quality.html

(Kelly, 2009) Kelly, J. (2009). Poor data quality costing companies millions of dollars annually. *SearchDataManagement. com. Aug*. [tiešsaiste]. - [atsauce 10.05.2019.]. Pieejams: https://searchdatamanagement.techtarget.com/news/1365965/Poor-data-quality-costing-companies-millions-of-dollars-annually

(Kerr et al., 2007a) Kerr, K., & Norris, T. (2007). The development of a health data quality programme. In *Information quality management: Theory and applications* (pp. 94-118). IGI Global.

(Kerr et al., 2007b) Kerr, K., Norris, T., & Stockdale, R. (2007). Data quality information and decision making: a healthcare case study. *ACIS 2007 Proceedings*, 98.

(Kessler et al., 2010) Kessler, C. W., Schamai, W., & Fritzson, P. (2010, February). Platform-independent modeling of explicitly parallel programs. In *23th International Conference on Architecture of Computing Systems 2010* (pp. 1-11). VDE.

(Khurshid et al., 2003) Khurshid, S., Păsăreanu, C. S., & Visser, W. (2003). Generalized symbolic execution for model checking and testing. In International Conference on Tools and

Algorithms for the Construction and Analysis of Systems (pp. 553-568). Springer, Berlin, Heidelberg.

(Kitchin, 2014) Kitchin, R. (2014). *The data revolution: Big data, open data, data infrastructures and their consequences*. Sage.

(Klein et al., 2018) Klein, R. H., Klein, D. B., & Luciano, E. M. (2018). Open Government Data: Concepts, Approaches and Dimensions over Time. *Revista Economia & Gestão*, *18*(49), 4-24.

(Kleppe, 2008) Kleppe, A. (2008). *Software language engineering: creating domain-specific languages using metamodels*. Pearson Education.

(Kleppe et al., 2003) Kleppe, A. G., Warmer, J., Warmer, J. B., & Bast, W. (2003). MDA explained: the model driven architecture: practice and promise. Addison-Wesley Professional.

(Konstante, 2016) Konstante, R. (2016). Sekundārās veselības aprūpes infrastruktūras plānošana Latvijā, promocijas darbs, Latvijas Universitāte, Latvija.

(Kontokostas et al., 2014) Kontokostas, D., Westphal, P., Auer, S., Hellmann, S., Lehmann, J., Cornelissen, R., & Zaveri, A. (2014, April). Test-driven evaluation of linked data quality. In *Proceedings of the 23rd international conference on World Wide Web* (pp. 747-758). ACM.

(Kučera et al., 2013) Kučera, J., Chlapek, D., & Nečaský, M. (2013, August). Open government data catalogs: Current approaches and quality perspective. In *International conference on electronic government and the information systems perspective* (pp. 152-166). Springer, Berlin, Heidelberg.

(Kuk et al., 2011) Kuk, G., & Davies, T. (2011). The roles of agency and artifacts in assembling open data complementarities.

(Lano, 2005) Lano, K. (2005). *Advanced systems design with Java, UML and MDA*. Elsevier.

(Larsen et al., 2009) Larsen, I. K., Småstuen, M., Johannesen, T. B., Langmark, F., Parkin, D. M., Bray, F., & Møller, B. (2009). Data quality at the Cancer Registry of Norway: an overview of comparability, completeness, validity and timeliness. European journal of cancer, 45(7), 1218-1231.

(Latvijas Atvērto datu portāls, 2018a) Latvijas Atvērto datu portāls. "Definīcijas". [tiešsaiste]. - [atsauce 10.10.2019.]. Pieejams: https://data.gov.lv/lv/buj

(Latvijas Atvērto datu portāls, 2018b) Latvijas Atvērto datu portāls. "Statistika par saziņu ar Rīgas pašvaldību". [tiešsaiste]. - [atsauce 10.10.2019.]. Pieejams: https://data.gov.lv/dati/lv/dataset/statistika-par-sazinu-ar-pasvaldibu

(Latvijas Atvērto datu portāls, 2018c) Latvijas Atvērto datu portāls. "Valsts informācijas sistēmu reģistrs". [tiešsaiste]. - [atsauce 10.10.2019.]. Pieejams: https://data.gov.lv/dati/lv/dataset/visr

(Latvijas Republikas Uzņēmumu reģistrs, 2018) Latvijas Republikas Uzņēmumu reģistrs. "Index of /register". [tiešsaiste]. - [atsauce 10.10.2019.]. Pieejams: http://dati.ur.gov.lv/register

(Laudenschlager et al., 2017) Laudenschlager D., Milener G., Guyer C., Roth J. "Data Quality Services Features and Tasks", 2017. [tiešsaiste]. - [atsauce 10.10.2019.]. Pieejams: https://docs.microsoft.com/en-us/sql/data-quality-services/data-quality-services-features-and-tasks?view=sql-server-2017

(Laudenschlager et al., 2012a) Laudenschlager D., Milener G., Guyer C., Roth J. "DQS Knowledge Bases and Domains", Microsoft 2012, [tiešsaiste]. - [atsauce 10.10.2019.]. Pieejams: https://docs.microsoft.com/en-us/sql/data-quality-services/dqs-knowledge-bases-and-domains?view=sql-server-2017

(Laudenschlager et al., 2012b) Laudenschlager D., Milener G., Guyer C., Roth J. "DQS Security". Microsoft Docs, 2012, [tiešsaiste]. - [atsauce 10.10.2019.]. Pieejams: https://docs.microsoft.com/en-us/sql/data-quality-services/dqs-security

(Lebied, 2018) M. Lebied, "The Ultimate Guide to Modern Data Quality Management (DQM) For An Effective Data Quality Control Driven by The Right Metrics", The Data Pine Blog, 2018, [tiešsaiste]. - [atsauce 10.10.2019.]. Pieejams: https://www.datapine.com/blog/data-quality-management-and-metrics/

(Lee et al., 2009) Lee, Y. W., Pipino, L. L., Funk, J. D., & Wang, R. Y. (2009). *Journey to data quality*. The MIT Press.

(Lehmannm et al., 2016) Lehmann C., Roy K., Winter. B., "The State of Enterprise Data Quality: 2016 Perception, Reality and the Future of DQM", 2016, [tiešsaiste]. - [atsauce 10.10.2019.]. Pieejams: https://siliconangle.com/files/2016/01/Blazent_State_of_Data_Quality_Management_2016.pdf

(Leonard et al., 2014) Leonard, A., Mitchell, T., Masson, M., Moss, J., & Ufford, M. (2014). Data correction with data quality services. In *SQL Server Integration Services Design Patterns* (pp. 101-123). Apress, Berkeley, CA.

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

(Loshin, 2001) Loshin, D. (2001). *Enterprise knowledge management: The data quality approach*. Morgan Kaufmann.

(Martin, 2014) Martin, C. (2014). Barriers to the open government data agenda: Taking a multi-level perspective. *Policy & Internet*, *6*(3), 217-240.

(Martin et al., 2013) Martin, S., Foulonneau, M., Turki, S., & Ihadjadene, M. (2013). Risk analysis to overcome barriers to open data. *Electronic Journal of e-Government*, *11*(1), 348.

(Masson, 2011) Masson, A., "Overview of the DQS Cleansing Transform Microsoft, SSIS Team Blog, 2011, [tiešsaiste]. - [atsauce 10.10.2019.]. Pieejams: https://techcommunity.microsoft.com/t5/SQL-Server-Integration-Services/Overview-of-the-DQS-Cleansing-Transform/ba-p/387792

(McGilvray, 2013) McGilvray, D. (2013). Data Quality Projects and Programs. In *Handbook of Data Quality* (pp. 41-73). Springer, Berlin, Heidelberg.

(Mellor et al., 2004) Mellor, S. J., Scott, K., Uhl, A., & Weise, D. (2004). MDA distilled: principles of model-driven architecture. Addison-Wesley Professional.

(Miller et al., 2003) Miller, J., & Mukerji, J. (2003). MDA Guide Version 1.0. 1. *Object Management Group*, *234*, 51.

(Moore, 2017) Moore, S. (2017). How to Create a Business Case for Data Quality Improvement. *Retrieved January*, *27*, 2018.

(Moraga et al., 2009) Moraga, C., Moraga, M. Á., Calero, C., & Caro, A. (2009, August). SQuaRE-aligned data quality model for web portals. In *2009 Ninth International Conference on Quality Software* (pp. 117-122). IEEE.

(Naumann, 2003) Naumann, F. (2003). *Quality-driven query answering for integrated information systems* (Vol. 2261). Springer.

(Neumaier, 2015) Neumaier, S. (2015). Open Data Quality: Assessment and Evolution of (Meta-) Data Quality in the Open Data Landscape. *Technische Universität Wien*.

(Ngomo et al., 2014) Ngomo, A. C. N., Auer, S., Lehmann, J., & Zaveri, A. (2014, September). Introduction to linked data and its lifecycle on the web. In *Reasoning Web International Summer School* (pp. 1-99). Springer, Cham.

(Nikiforova, 2019a) Nikiforova, A. (2019). Analysis of open health data quality using data object-driven approach to data quality evaluation*:* insights from a Latvian context. In *IADIS International Conference e-Health 2019, Part of the IADIS Multi Conference on Computer Science and Information Systems, MCCSIS 2019, July 16 - 19, 2019* (pp. 119-126). IADIS.

(Nikiforova, 2019b) Ņikiforova, A. (2019). Izpildāmu modeļu lietojums datu kvalitātes novērtēšanai (maģistra darbs).

(Nikiforova, 2018a) Nikiforova, A. (2018). Open Data Quality Evaluation: A Comparative Analysis of Open Data in Latvia. *Baltic Journal of Modern Computing*, *6*(4), 363-386.

(Nikiforova, 2018b) Nikiforova, A. (2018). Open Data Quality. In *Doctoral Consortium/Forum@ DB&IS* (pp. 151-160)

(Nikiforova et al., 2018) Nikiforova, A., & Bicevska, Z. (2018). Application of LEAN Principles to Improve Business Processes: a Case Study in Latvian IT Company. *Baltic Journal of Modern Computing*, *6*(3), 247-270.

(Nikiforova et al., 2020) Nikiforova, A., Bicevskis, J., Bicevska, Z., Oditis, I., (2020). User-Oriented Approach to Data Quality Evaluation. *Journal of Universal Computer Science*, *26*(1), 107-126.

(Nikiforova et al., 2019) Nikiforova, A., Bicevskis, J. (2019). An Extended Data Object-driven Approach to Data Quality Evaluation: Contextual Data Quality Analysis. *Proceedings of the 21st International Conference on Enterprise Information Systems (ICEIS 2019)*, 274-281. DOI: 10.5220/0007838602740281

(Oliveira et al., 2016) Oliveira, M. I. S., de Oliveira, H. R., Oliveira, L. A., & Lóscio, B. F. (2016, June). Open government data portals analysis: the Brazilian case. In *Proceedings of the 17th International Digital Government Research Conference on Digital Government Research* (pp. 415-424). ACM.

(Olson, 2003) Olson, J. E. (2003). *Data quality: the accuracy dimension*. Elsevier.

(Open Government Partnership, 2015) Open Government Partnership. Latvia National Action Plan 2015-2017. [tiešsaiste]. - [atsauce 10.10.2019.]. Pieejams: https://www.opengovpartnership.org/documents/latvia-national-action-plan-2015-2017-original/

(Ostadzadeh et al., 2008) Ostadzadeh, S. S., Aliee, F. S., & Ostadzadeh, S. A. (2008). An MDA-based generic framework to address various aspects of enterprise architecture. In *Advances in Computer and Information Sciences and Engineering* (pp. 455-460). Springer, Dordrecht.

(Parkin et al., 2009) Parkin, D. M., Bray, F. (2009). Evaluation of data quality in the cancer registry: principles and methods Part II. Completeness. European journal of cancer, 45(5), 756-764.

(Pauker et al., 2016) Pauker, F., Frühwirth, T., Kittl, B., & Kastner, W. (2016). A systematic approach to OPC UA information model design. *Procedia CIRP*, *57*, 321-326.

(Paulheim et al., 2014) Paulheim, H., Bizer, C. (2014). Improving the quality of linked data using statistical distributions. International Journal on Semantic Web and Information Systems (IJSWIS), 10(2), 63-86.

(Petychakis et al., 2014) Petychakis, M., Vasileiou, O., Georgis, C., Mouzakitis, S., & Psarras, J. (2014). A state-of-the-art analysis of the current public data landscape from a functional, semantic and technical perspective. *Journal of theoretical and applied electronic commerce research*, *9*(2), 34-47.

(Prakash, 2016) Prakash, A. "Introduction to Data Quality Services (DQS) of SQL Server" Microsoft, 2016, [tiešsaiste]. - [atsauce 10.10.2019.]. Pieejams: https://community.dynamics.com/crm/b/crmhuntsdynamicscrmblog/archive/2016/01/19/introduction-to-data-quality-services-dqs-of-sql-server

(Price et al., 2005) Price, R., & Shanks, G. (2016). A semiotic information quality framework: development and comparative analysis. In *Enacting Research Methods in Information Systems* (pp. 219-250). Palgrave Macmillan, Cham.

(Price et al., 2004) Price, R., & Shanks, G. (2004). A semiotic information quality framework. In *Proceedings of the International Conference on Decision Support Systems DSS04* (pp. 658-672).

(Prieto Rodrigues et al., 2018) Prieto Rodríguez, J. D., Suárez Hurtado, V. (2018). Medical Records Digital Perspective Colombian: Safety, Quality, and Management of Data. *Journal of Alternative Perspectives in the Social Sciences*, *9*(3).

(Pyle, 1999) Pyle, D. (1999). Data preparation for data mining. morgan kaufmann.

(Raghupathi et al., 2014) Raghupathi, W., &Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. *Health information science and systems*, *2*(1), 3.

(Redman, 2001) Redman, T. C. (2001). *Data quality: the field guide*. Digital press.

(RD IKSD, 2019) RD IKSD, Rīgas domes izglitības, kultūras un sporta departaments, Iestāžu katalogs. "Atvērtie dati", [tiešsaiste]. - [atsauce 10.10.2019.]. Pieejams: http://dev-dati.e-skola.lv/lv/open-data

(RIK, 2018) RIK. Centre of Registers and information Systems. "Open data: Commercial Register". [tiešsaiste]. - [atsauce 10.10.2019.]. Pieejams: https://www.rik.ee/en/open-data

(Roscoe, 1975) Roscoe, J. T. (1975). *Fundamental research statistics for the behavioral sciences [by] John T. Roscoe*.

(Ross, 2017) Ross, J. E. (2017). *Total quality management: Text, cases, and readings*. Routledge.

(Ruijer et al., 2019) Ruijer, E., Meijer, A. (2019). Open Government Data as an Innovation Process: Lessons from a Living Lab Experiment. *Public Performance & Management Review*, 1-23.

(Ruiz, 2018) Ruiz, M. (2018). *TraceME: A Traceability-Based Method for Conceptual Model Evolution*. Springer International Publishing.

(Sasse et al., 2017) Sasse, T., Smith, A., Broad, E., Kennison, J., Wells, P., Atz, U. (2017) "Recomendations for Open Data Portals: from Setup to sustainability", Disponível na WWW:
https://www.europeandataportal.eu/sites/default/files/edp_s3wp4_sustainability_recommendations.pdf.

(Sáez Martín et al., 2016) Sáez Martín, A., Rosario, A. H. D., & Pérez, M. D. C. C. (2016). An international analysis of the quality of open government data portals. *Social Science Computer Review*, *34*(3), 298-311.

(Scannapieco et al., 2005) Scannapieco, M., Missier, P., & Batini, C. (2005). Data quality at a glance. *Datenbank-Spektrum*, *14*(January), 6-14.

(Scannapieco et al., 2002a) Scannapieco, M., & Catarci, T. (2002). Data quality under a computer science perspective. *Archivi & Computer*, *2*, 1-15.

(Scannapieco et al., 2002b) Scannapieco, M., Pernici, B., & Pierce, E. M. (2002). IP-UML: Towards a Methodology for Quality Improvement Based on the IP-MAP Framework. In *ICIQ* (pp. 279-291).

(Schmidt et al., 2015) Schmidt, M., Schmidt, S. A. J., Sandegaard, J. L., Ehrenstein, V., Pedersen, L., & Sørensen, H. T. (2015). The Danish National Patient Registry: a review of content, data quality, and research potential. Clinical epidemiology, 7, 449.

(Selic, 2009) Selic, B. (2009, July). The theory and practice of modeling language design for model-based software engineering—a personal perspective. In *International Summer School on Generative and Transformational Techniques in Software Engineering* (pp. 290-321). Springer, Berlin, Heidelberg.

(Shi et al., 2005) Shi, X., Han, W., Huang, Y., & Li, Y. (2005, September). Service-oriented business solution development driven by process model. In *The Fifth International Conference on Computer and Information Technology (CIT'05)* (pp. 1086-1092). IEEE.

(Sigurdardottir et al., 2012) Sigurdardottir, L. G., Jonasson, J. G., Stefansdottir, S., Jonsdottir, A., Olafsdottir, G. H., Olafsdottir, E. J., & Tryggvadottir, L. (2012). Data quality at the Icelandic Cancer Registry: comparability, validity, timeliness and completeness. Acta oncologica, 51(7), 880-889.

(Singh, 2017) Singh, A. J., “Inside Big Data. The Hidden Costs of Bad Data, 2017”, [tiešsaiste]. - [atsauce 10.10.2019.]. Pieejams: https://insidebigdata.com/2017/05/05/hidden-costs-bad-data/

(Soley, 2000) Soley, R. (2000). Model driven architecture. OMG white paper, 308: p. 308, 5.

(Sprogis et al., 2013) Sprogis, A., Barzdins, J. (2013). Specification, Configuration and Implementation of DSL Tool. In Databases and Information Systems VII: Selected Papers from the Tenth International Baltic Conference, DB & IS 2012 (Vol. 249, p. 330). IOS Press.

(Sproģis, 2014) Sproģis, A. (2014). Domēnspecifisku rīku konfigurācijas valoda un tās realizācija, promocijas darbs, Latvijas Universitāte, Latvija.

(Sunlight Foundation, 2017) Sunlight Foundation. “Ten Principles For Opening Up Government Information”. [tiešsaiste]. - [atsauce 10.10.2019.]. Pieejams: https://sunlightfoundation.com/policy/documents/ten-open-data-principles/

(Tayi et al., 1998) Tayi, G. K., & Ballou, D. P. (1998). Examining data quality. *Communications of the ACM*, *41*(2), 54-57.

(Tinholt, 2013) Tinholt, D. (2013). The Open Data Economy: Unlocking Economic Value by Opening Government and Public Data. *Capgemini Consulting*.

(Tomic et al., 2015) Tomic, K., Sandin, F., Wigertz, A., Robinson, D., Lambe, M., & Stattin, P. (2015). Evaluation of data quality in the National Prostate Cancer Register of Sweden. European journal of cancer, 51(1), 101-111.

(Ubaldi, 2013) Ubaldi, B. (2013). Open government data.

(Umbrich et al., 2015) Umbrich, J., Neumaier, S., & Polleres, A. (2015, March). Towards assessing the quality evolution of open data portals. In *Proceedings of ODQ2015: Open Data Quality: from Theory to Practice Workshop, Munich, Germany*.

(Van den Berghe et al., 2017) Van den Berghe, S., & Van Gaeveren, K. (2017). Data quality assessment and improvement: a Vrije Universiteit Brussel case study. *Procedia Computer Science*, *106*, 32-38.

(VARAM, 2016) Vides aizsardzības un reģionālās attīstības ministrija. “Atvērtie dati”. [tiešsaiste]. - [atsauce 10.10.2019.]. Pieejams: http://www.varam.gov.lv/lat/darbibas_veidi/e_parv/atvertie_dati/?doc=20449

(Vetrò et al., 2016) Vetrò, A., Canova, L., Torchiano, M., Minotas, C. O., Iemma, R., & Morando, F. (2016). Open data quality measurement framework: Definition and application to Open Government Data. *Government Information Quarterly*, *33*(2), 325-337.

(VismaLatvia, 2015) VismaLatvia, “4 iemesli pievērsties uzņēmuma rīcībā esošo datu kvalitātes uzlabošanai”, 2015. [tiešsaiste]. - [atsauce 10.10.2019.]. Pieejams:

https://www.visma.lv/blogs/4-iemesli-pieversties-uznemuma-riciba-esoso-datu-kvalitates-uzlabosanai/

(Wand et al, 1996) Wand, Y., & Wang, R. Y. (1996). Anchoring data quality dimensions in ontological foundations. *Communications of the ACM*, *39*(11), 86-95.

(Wang et al., 1996) Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of management information systems*, *12*(4), 5-33.

(Wang et al., 1993a) Wang, R. Y., & Kon, H. B. (1993, January). Toward total data quality management (TDQM). In *Information technology in action* (pp. 179-197). Prentice-Hall, Inc..

(Wang et al., 1993b) Wang, Y. Y. R., Storey, V., & Firth, C. (1993). *Data quality research: a framework, survey, and analysis*. Total Data Quality Management Research Program, Sloan School of Management, Massachusetts Institute of Technology.

(Wanner et al., 2018) Wanner, M., Matthes, K. L., Korol, D., Dehler, S., & Rohrmann, S. (2018). Indicators of data quality at the cancer registry Zurich and Zug in Switzerland. *BioMed research international*, *2018*.

(Weiskopf et al., 2013) Weiskopf, N. G., & Weng, C. (2013). Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. Journal of the American Medical Informatics Association, 20(1), 144-151.

(Witchalls, 2014) Witchalls, C. (2014). Gut & gigabytes: Capitalising on the art & science in decision making: PwC. *Retrieved April*, *24*, 2017.

(Yi, 2019) Yi, M. (2019). Exploring the quality of government open data: Comparison study of the UK, the USA and Korea. *The Electronic Library*, *37*(1), 35-48.

(Zaveri et al., 2016) Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., & Auer, S. (2016). Quality assessment for linked data: A survey. *Semantic Web*, *7*(1), 63-93.

(Zhang et al., 2014) Zhang, R., Jayawardene, V., Indulska, M., Sadiq, S., & Zhou, X. (2014). A data driven approach for discovering data quality requirements.

(Zhao et al., 2003) Zhao, W., Bryant, B. R., Raje, R. R., Auguston, M., Gray, J. G., Burt, C. C., & Olson, A. M. (2003). *A generative and model driven framework for automated software product generation*. Alabama Univ in Birmingham Dept Of Computer And Information Sciences.

(Zuiderwijk et al., 2014) Zuiderwijk, A., & Janssen, M. (2014). Barriers and development directions for the publication and usage of open data: A socio-technical view. In *Open government* (pp. 115-135). Springer, New York, NY.

(Zuiderwijk et al., 2012) Zuiderwijk, A., Janssen, M., Choenni, S., Meijer, R., & Alibaks, R. S. (2012). Socio-technical Impediments of Open Data. *Electronic Journal of e-Government*, *10*(2).