

EUROPEAN COMMISSION DG REFORM

Eesti andmehalduse metoodikaprojekt

Andmekvaliteedi juhised

August 2020



Funded by the European Commission's Structural Reform Support Programme and implemented by Ernst & Young Baltic AS in cooperation with the European Commission's Directorate-General for Structural Reform Support

This document has been produced under a contract with the Union and the opinions expressed are those of Ernst & Young Baltic AS and do not represent the official position of the European Commission

Sisukord

Sisukord	1
1. Sissejuhatus	3
1.1 Andmekvaliteedi juhise eesmärk	3
1.2 Sihtrühmad.....	3
1.2.1 Andmekvaliteediga seotud rollid andmehalduse raamistikust.....	3
1.2.2 Juhise sihtrühmad.....	4
1.3 Juhise ulatus.....	5
1.4 Andmekvaliteedi mudeli valik	6
1.5 Andmekvaliteedi varasemad käsitlused.....	6
1.6 Lühendid ja mõisted.....	7
2. Andmekvaliteedi juhtimise tegevused	9
2.1 Andmekvaliteedi eesmärkide määramine	9
2.2 Andmekvaliteedi reeglite kirjeldamine	10
2.3 Andmekvaliteedi mõõtmine	12
2.3.1 Andmekvaliteedi mõõtmise metamudel	13
2.3.2 Andmekvaliteedi probleemid	15
2.3.3 Kvaliteediprobleemide tuvastamine	21
2.3.4 Dimensioonide ja indikaatorite mõõtmine.....	26
2.3.5 Andmekvaliteedi reeglile mõõdiku seadmine	31
2.3.6 Andmekvaliteedi aruandepõhjade väljatöötamine	34
2.3.7 Kvaliteediprobleemide prioriseerimine	34
2.4 Andmekvaliteedi juurpõhjuste analüüs	34
2.5 Kvaliteediprobleemide mõju hindamine.....	37
2.6 Andmekvaliteedi reeglite haldamine	41
2.6.1 Reeglite muutmine ja versioneerimine	41
2.6.2 Andmeelementide kaetus reeglitega.....	42
3. Rakendamise näited	43
3.1 Pakettide paigaldamine ja näidisandmed	43
3.2 Andmete profileerimise rakendamise näide.....	46
3.3 Andmekvaliteedi mõõtmise näidisjuhtumid	47
3.3.1 Näide N1 (Täielikkus, puuduv väärtus)	47

3.3.2	Näide N2 (Ühekordsus, unikaalse väärtuse rikkumine).....	48
3.3.3	Näide N3 (Õigsus, väärtusvahemiku rikkumine).....	48
3.3.4	Näide N4 (Õigsus, väärtusvahemiku rikkumine).....	50
3.3.5	Näide N5 (Õigsus, süntaksi rikkumine)	51
3.3.6	Näide N6 (Õigsus, süntaksi rikkumine)	52
3.3.7	Näide N7 (Reeglipärasus, valdkonnakitsenduse rikkumine).....	53
3.3.8	Näide N8 (Õigsus, valdkonnakitsenduse rikkumine)	54
3.3.9	Näide N9 (Reeglipärasus, funktsionaalse sõltuvuse rikkumine).....	54
3.3.10	Näide N10 (Ajakohasus, aegunud väärtus).....	56
3.3.11	Näide N11 (Õigsus, süntaksi rikkumine)	57
3.4	Andmekvaliteedi juhtimislaua loomine	58
3.4.1	Andmeobjektide vastavus nõuetele	58
3.4.2	Parandamist vajavad andmeobjektid	60
3.4.3	Millised reeglid on kaetud	61
3.4.4	Andmeelementide kaetus reeglitega.....	61
3.4.5	Reeglite eksport ja import	63
3.5	Soovituslikud töövahendid.....	63

1. Sissejuhatus

Käesolev juhise on üks osa andmehalduse raamistikust¹. Andmehaldus on organisatsiooni või asutuse tegevusvaldkond, mis võimaldab tal hallata oma andmeid varana. Tõhusast andmehaldusest organisatsioonile tulenev kasu seisneb andmetest täiendava väärtuse loomises ja paremate juhtimisotsuste langetamises. Nimetatud kasu on aga võimalik saavutada vaid siis, kui organisatsiooni andmed on kvaliteetsed. Seejuures on andmekvaliteedi haldamise eelduseks andmekirjelduse olemasolu, mille loomiseks leiab juhised samuti andmehalduse raamistiku osana valminud andmekirjelduse juhise².

Andmekvaliteet näitab, mil määral andmekarakteristikud rahuldavad teadaolevaid või eeldatavaid vajadusi kasutamisel ettemääratud tingimustes. Andmekvaliteeti aitavad tagada andmehalduse raamistikus kirjeldatud andmekvaliteedi haldamise protsessid, mis katavad andmekvaliteedi reeglite haldamiseks, andmekvaliteedi mõõtmiseks ja seeläbi andmekvaliteedi raportite loomiseks ning andmete parandamiseks (andmehalduri poolt käsitsi või IKT osakonna poolt) tehtavad tegevused. Loetletud tegevuste täitmisel osaleb mitmeid erinevate rollide täitjaid, kellel tuleb oma vastutuste täitmiseks teha mitmesuguseid praktilisi ülesandeid. Käesolev juhise pakub praktilisi juhtnööre nende ülesannete täitmiseks.

1.1 Andmekvaliteedi juhise eesmärk

Juhise on loodud selleks, et asutustel ja organisatsioonidel oleks andmekvaliteedi reeglite haldamiseks, mõõtmiseks ja kvaliteedi parendamise rakendamiseks olemas praktilised juhtnöörid. Juhise järgimine aitab kaasa andmehalduse raamistiku terviklikule rakendamisele organisatsioonis.

Juhises käsitletav andmekvaliteedi mõõtmine on asutuse keskne ning riiklikul tasemel andmekogude võrdlust ei rakendata. Asutused on autonoomsed dimensioonide mõõtmiseks kasutatavate indikaatorite ja mõõdikute valimisel. Mõõtmisega alustamiseks pakub käesolev juhise nii teoreetilist tausta kui ka praktilisi näiteid.

1.2 Sihtrühmad

Andmaks ülevaadet juhise sihtrühmadest on esmalt toodud taustinformatsioon andmekvaliteedi raamistikus määratletud rollide, nende **andmekvaliteediga seotud** vastutuste ning tegevuste kohta milles nad osalevad. (Rollide kirjeldused koos täieliku vastutuste loendiga on toodud andmehalduse raamistikus.) Seejärel on andmehalduse raamistikus defineeritud rollidest ja nende vastutustes lähtuvalt kirjeldatud käesoleva juhise sihtrühmad.

1.2.1 Andmekvaliteediga seotud rollid andmehalduse raamistikust

Andmehalduse sponsor on asutuse juhtkonna liige, kes vastutab asutustes andmekvaliteedi parendamise propageerimise eest. Lisaks osaleb ta andmekvaliteedi projektide järelevalve teostamises.

Andmehalduse juht on andmehaldusorganisatsiooni juht ja asutuse või valdkonnaüleste tegevuste koordinaator. Tema vastutuseks on andmekvaliteedi aruannete ja mõõdikute väljatöötamine ja

¹ vt. EY, Andmehalduse raamistik (2020)

² Vt. EY, Andmekirjelduse juhise (2020)

jälgimine, andmekvaliteedi parendamise tasuvusanalüüside koostamine ning andmekvaliteedi projektide portfelli haldamine ning projektide tellimine ja järelevalve.

Andmeomanik on osakonna / valdkonna / teenuste juht või peakasutaja, kes on protsesside omanik, kindla huvigrupi esindaja andmekvaliteedi nõuete esitamisel ehk andmete tegelik omanik. Tema vastutuseks on andmekvaliteedi reeglite seadmine vastavalt reeglistikule, huvigruppide ja kasutajate andmekvaliteedi probleemide ja nõuete registreerimine, ootuste juhtimine ning andmekvaliteedi parendamise protsesside ja projektide algatamine ning järjestamine. Lisaks osaleb ta andmekvaliteedi reeglitele mõõdikute määramisel.

Andmehaldur (andmestikud) on andmete ekspert, kes omab parimat teadmist valdkonna andmestikest ja lähtesüsteemide andmetest, ühiskasutatavatest põhiandmetest ning teenuste ja mõõdikutega seotud andmetest. Tema vastutuseks on andmekvaliteedi reeglitele mõõdikute määramine, mõõtetulemuste kogumine ja raporteerimine ning andmekvaliteedi probleemide põhjuste väljaselgitamine ja sellest tegevusele tekkiva mõju hindamine. Lisaks osaleb ta andmekvaliteedi reeglite kirjeldamisel toetades selles tegevuses andmeomanikke.

Metaandmete analüütik on IT- ja metaandmete süsteemide tundja, kes omab ülevaadet kindla süsteemiga seotud andmetest ja kvaliteedi mõõtmise vahenditest. Tema vastutuseks on nõuete ja andmekvaliteedi reeglite IT süsteemidesse juurutamise nõustamine ning regulaarsete ja ühekordsete andmepäringute teostamine andmekvaliteedi probleemide põhjuste välja selgitamiseks. Lisaks osaleb ta andmekvaliteedi mõõdikute väljatöötamisel nõustajana.

Andmehaldur (andmed) tegeleb andmete füüsilise sisestamise ja korrigeerimisega (eelduseks toodangusüsteemide kõrgema taseme kasutajaõigused). Tema vastutuseks on andmekvaliteedi mõõdikute jälgimine. Lisaks osaleb ta andmekvaliteedi mõõdikute väljatöötamisel nõustajana.

1.2.2 Juhise sihtrühmad

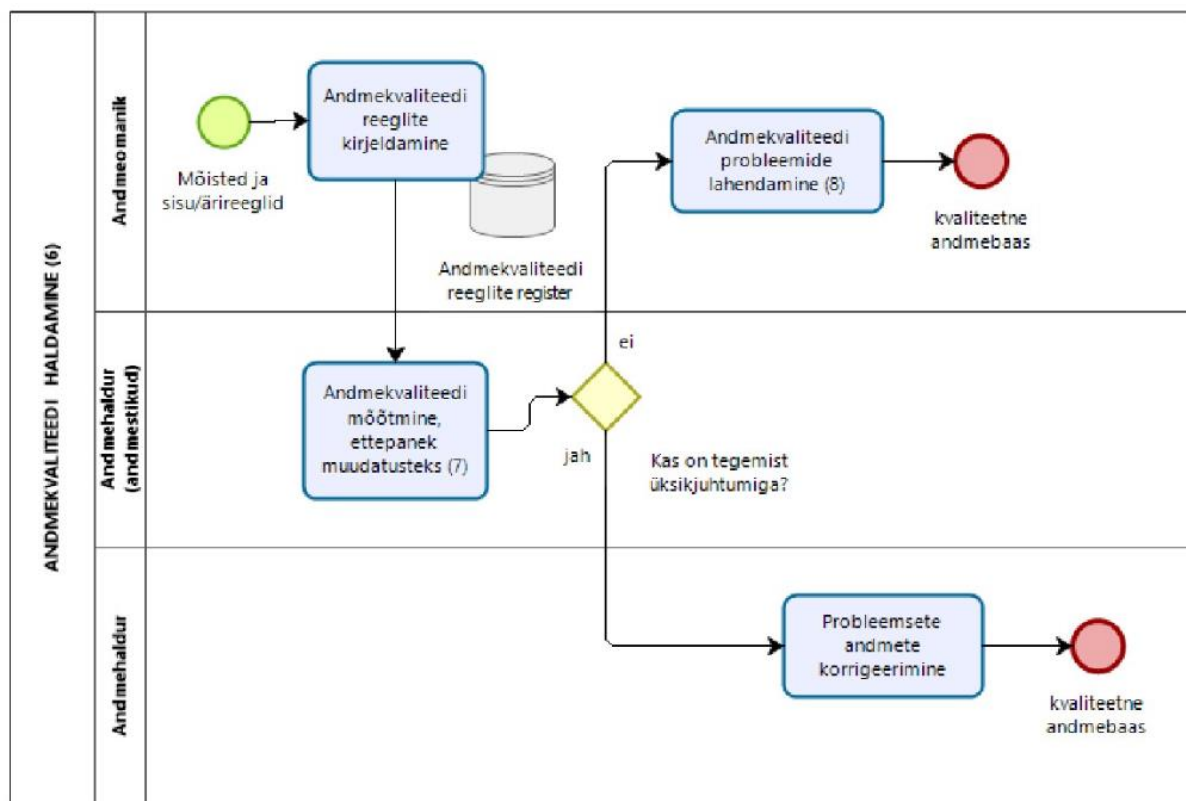
Juhis on eelkõige suunatud **andmehalduri (andmestikud)** rollis olevatele asutuste ja organisatsioonide töötajatele, kellele käesolev juhise pakub juhtnööre kõigi temaga seotud andmekvaliteeti puudutavate ülesannete täitmiseks: andmeomanike toetamiseks andmekvaliteedi reeglite kirjeldamisel rakendades profileerimist, andmekvaliteedi mõõtmiseks, andmekvaliteedi reeglitele mõõdikute määramiseks, mõõtetulemuste kogumiseks ja raporteerimiseks, andmekvaliteedi probleemide põhjuste väljaselgitamiseks ja kvaliteediprobleemide mõju hindamiseks.

Andmehalduse sponsor saab antud juhise vajalikku taustinformatsiooni andmekvaliteedi parendamise propageerimiseks ning andmekvaliteedi projektide järelevalves osalemiseks.

Andmehalduse juhile pakub käesolev juhise juhiseid andmekvaliteedi aruannete ja mõõdikute väljatöötamiseks ning andmekvaliteedi parendamise tasuvusanalüüside koostamiseks. **Andmeomanik** saab juhendist tuge andmekvaliteedi reeglite seadmiseks, andmekvaliteedi reeglitele mõõdikutele seadmisel osalemiseks ning konteksti ja üldist taustinformatsiooni oma ülejäänud andmekvaliteeti puudutava vastutuste täitmiseks. **Metaandmete analüütikule** pakub juhise teadmisi kvaliteedi mõõtmise vahenditest ning aitab tal täita oma andmekvaliteedi reeglite ja andmekvaliteedi probleemide põhjuste analüüsiga seotud vastutusi. Lisaks pakub juhise talle andmekvaliteedi mõõdikute väljatöötamisel nõustajana osalemiseks vajalikku taustinformatsiooni. **Andmehaldurile (andmed)** pakub juhise mõõdikute jälgimiseks ja andmekvaliteedi mõõdikute väljatöötamisel nõustajana osalemiseks vajalikku taustinformatsiooni.

1.3 Juhise ulatus

Andmekvaliteedi juhise annab esmalt lühiülevaate andmekvaliteedi eesmärgistamisest. Seejärel on toodud juhised peamiste andmehalduse raamistikus toodud ning praktilisi oskusi nõudvate andmekvaliteediga seotud tegevuste teostamiseks. Alljärgnevalt (Joonis 1) on esitatud andmehalduse raamistikus toodud andmekvaliteedi haldamise üldine protsess.



Joonis 1. Andmekvaliteedi haldamine (vt ka andmehalduse raamistik, seksioon 4.1.6)

Juhis kirjeldab lahti **andmekvaliteedi reeglite kirjeldamiseks** tehtavad tegevused ning pakub detailseid juhiseid **andmekvaliteedi mõõtmise** teostamiseks. Arvestades andmekvaliteedi mõõtmise keerukust ja alamtegevuste rohkust on oluline rõhk pandud just selle tegevuse juhendamisele. Andmekvaliteedi haldamise protsessi osaks olevast **andmekvaliteedi probleemide lahendamise** alamprotsessist on kirjeldatud kaks **andmehalduri (andmestikud)** peamist eriteadmisi nõudvat tegevust: andmekvaliteedi probleemide põhjuste analüüs ja kvaliteediprobleemide ärilise mõju hindamine. Eelnimetatud tegevuste käigus tekib hulk andmekvaliteedi reegleid, mis on pidevas muutumises. Seega võib andmekvaliteedi reeglite haldamine muutuda kiiresti keerukaks ning tekitada segadust. Probleemi lahendamiseks on juhendis esitatud juhised andmekvaliteedi reeglite haldamiseks.

Viimasena esitatakse praktilised näited eelnevalt lahti seletatud tegevuste praktiliseks rakendamiseks kasutades konkreetseid tööriistu. Juhendis on toodud näited profileerimise, andmekvaliteedi mõõtmise, andmekvaliteedi juhtimislaua koostamise ja andmekvaliteedi reeglite haldamise praktiliseks rakendamiseks.

Andmekvaliteedi juhise koostamisel on lähtutud nii DAMA-DMBOK2 andmehalduse mudelist³, andmekvaliteedi protsessi⁴ ning mõõtmist⁵ käsitlevast kirjandusest kui ka andmekvaliteedi probleemide taksonoomiast⁶.

1.4 Andmekvaliteedi mudeli valik

Andmete kõrge kvaliteedi tähtsustamine teoreetikute ja praktikute poolt ning kasu, mida sellest on saadud, on aidanud kaasa andmekvaliteedi raamistike paljususe tekkele. Lähtuvalt valdkondlikest iseärasustest, infosüsteemide eripäradest ja andmete kontekstist on loodud kümneid erinevaid andmekvaliteedi raamistikke. Eesti andmekvaliteedi juhises on võetud kasutusele mudel, milles on viis dimensiooni: täielikkus, ajakohasus, õigsus, reeglipärasus ja ühekordsus. Nimetatutest kolm esimest on eri raamistikes enimlevinud dimensioonid⁷. Reeglipärasus võeti mudelisse, sest see võimaldab jälgida klassifikaatorite ja infosüsteemide sisemiste loendite kasutamist ning põhiaandmete kasutamist. Et asutuse sees ei oleks dublitseerivaid andmeid, siis on raamistikus eraldi dimensioonina välja toodud ühekordsus.

Andmekvaliteedi mudelis on neid dimensioone kasutatud andmekvaliteedi indikaatorite, probleemide ja reeglite grupeerimiseks. See tagab andmekvaliteedi süsteemse käsitlemise ja lihtsustab andmekvaliteedi haldamist, sh andmekvaliteedi reeglite väljatöötamist. Andmekvaliteedi reeglite grupeerimiseks kasutame täiendavalt Oliveira jt⁸ poolt välja töötatud andmekvaliteedi probleemide taksonoomiat. Erinevalt teistest andmekvaliteedi probleemide raamistikest põhineb selles raamistikus toodud andmekvaliteedi probleemide taksonoomia laiapõhjalisel juhtumiuuringul, on formaliseeritud ning toetub samade autorite varasemale tööle⁹, mis pakub lisaks konkreetseid algoritme andmekvaliteedi probleemide tuvastamiseks ja klassifitseerimiseks. Andmekvaliteedi reeglite kirjeldamisel lähtume me tuvastatud andmekvaliteedi probleemidest ja juhtimisel andmekvaliteedi dimensioonidega seotud indikaatoritest. Seosed dimensioonide ja reeglite vahel tekivad läbi reeglite grupeerimise andmekvaliteedi probleemide alusel. Selline probleemipõhine liigitusskeem lihtsustab andmekvaliteedi reeglite jaotamist dimensioonidesse.

1.5 Andmekvaliteedi varasemad käsitlused

2016. aastal valmis varasem juhend pealkirjaga „Andmekvaliteedi tagamise juhend andmekogu omanikele“. Nimetatud juhend on suunatud eelkõige andmeomanikele ning pakub neile head materjali andmekvaliteediga tegelemiseks. Käesolev juhise ei asenda varasemas juhendis toodut, vaid täiendab seda. Erinevalt varasemast juhendist on antud juhise suunatud eelkõige andmehalduritele

³ <https://dama.org/sites/default/files/download/DAMA-DMBOK2-Framework-V2-20140317-FINAL.pdf>

⁴ McGilvray, Danette. Executing Data Quality Projects: Ten Steps to Quality Data and Trusted Information (Tm). 2008.

⁵ Batini, Carlo, and Monica Scannapieco. Data and Information Quality. 2016.

⁶ Oliveira, Paulo, Fátima Rodrigues, and Pedro Rangel Henriques. "A formal definition of data quality problems." ICIQ. 2005. <http://mitiq.mit.edu/ICIQ/Documents/IQ%20Conference%202005/Papers/AFormalDefinitionofDQProblems.pdf>

⁷ C. Cichy and S. Rass, "An Overview of Data Quality Frameworks," in *IEEE Access*, vol. 7, pp. 24634-24648, 2019, doi: 10.1109/ACCESS.2019.2899751.

⁸ Oliveira, Paulo, Fátima Rodrigues, and Pedro Rangel Henriques. "A formal definition of data quality problems." ICIQ. 2005. <http://mitiq.mit.edu/ICIQ/Documents/IQ%20Conference%202005/Papers/AFormalDefinitionofDQProblems.pdf>

⁹ Oliveira Paulo, Fátima Rodrigues, Pedro Henriques, and Helena Galhardas. "A taxonomy of data quality problems." In 2nd Int. Workshop on Data and Information Quality, pp. 219-233. 2005.

(andmestikud), kuid pakub samas praktilisi oskusi ja taustinformatsiooni nii andeomanikule kui ka teistele andmekvaliteedi tagamisel ülesandeid omavatele rollidele.

Varasem juhend pakub välja andmekvaliteedi juhtimise raamistiku andmekvaliteedi pidevaks parendamiseks ning seeläbi küpsustaseme tõstmiseks. Välja pakutud raamistik põhineb pideva parendamise tsüklil ehk Demingi rattal ning sisaldab järgmisi tegevusi: hindamine, planeerimine, teostamine, kontrollimine ja korrigeerimine. Kasutatud juhtimismeetod pakub üldisel tasemel andmekvaliteedi juhtimistegevuste kirjeldust. Täpsemat ülevaadet andmehalduse protsessidest, sh ka andmekvaliteedi protsessidest, pakub Eesti andmehalduse raamistik. Lisaks andmekvaliteedi juhtimise raamistikule on varasemas juhendis toodud ülevaade andmekvaliteedi küpsusmudelitest ning juhised selle rakendamiseks. Neid käesolev juhend ei käsitle.

Viimaks annab varasem juhend ülevaate üheksast andmekvaliteedi tunnusest (ehk dimensioonist). Varasemas juhises on käsitletavaid dimensioone üheksa: õigsus, täielikkus, kooskõla, usaldusväärsus, ajakohasus, reeglipärasus, konfidentsiaalsus, ühekordsus ja mitteliiasus. Iga käsitletud dimensiooni kohta on esitatud määratlus, dimensiooniga seotud kvaliteediprobleemide näited, ülevaade dimensiooni kvaliteedinõuete kohta käivatest õigusaktidest, lühikene kirjeldus meetoditest, mis aitavad tagada andmekvaliteedi vastavust nõuetele antud dimensiooni osas ning viimaks kontrollküsimused, mille eesmärgiks oli võimaldada andmeomanikul veenduda, et ta oleks arvestanud erinevate võimalike nõuete allikatega ja võimalike dimensiooni nõuetele vastavuse tagamise meetmetega.

Käesolev juhend vaatleb andmekvaliteeti Eesti andmehalduse raamistiku kontekstis ning kirjeldab lahti andmehalduse raamistikus toodud andmekvaliteediga seotud peamised tegevused: 4.1.5 Kvaliteedireeglite haldamine (5), 4.1.6 Andmekvaliteedi haldamine (6); 4.1.7 Andmekvaliteedi mõõtmine (7) ja 4.1.8 Andmekvaliteedi probleemide lahendamine (8). Kui varasem juhend käsitles üheksat dimensiooni, siis käesolevas juhises kasutusele võetud mudelis on dimensioone viis: õigsus, täielikkus, ajakohasus, reeglipärasus ja ühekordsus. Erinevalt varasemast juhiseist on käesolevas juhises toodud terviklikud juhtnöörid andmekvaliteedi dimensioonidest rakendamise võtmes olulisemate mõõtmiseks.

1.6 Lühendid ja mõisted

Termin	Määratlus ja selgitus
Andmed	Informatsiooni taastõlgendatav esitus formaliseeritud kujul, mis sobib edastuseks, tõlgenduseks või töötluseks [ISO/IEC 2382:2015]
Andmeelement	Elementaarüksusena käsitletav nimega seos käsitusvalla objektide ja neid esitavate sõnade vahel [ISO/IEC 2382-17:1999]
Andmehaldur	Äriprotsesse esindav roll andmehalduse alal: - andmete sisu, konteksti ja metaandmete eest vastutaja; - kohustused sõltuvad kontekstist ja võivad osaliselt kattuda andmekäitleja omadega. [https://akit.cyber.ee/term/2172-andmehaldur]
Andmehaldus	Juhtimis- ja kontrollitegevuste (planeerimine, seire ja kehtestamine) rakendamine andmevaradega seotud tegevuste üle (DAMA DMBOK2)
Andmekirjeldus	Andmeelemendi ning kõigi ta nime ja ta sõnu sisaldavate andmestruktuuride formaliseeritud kirjeldus [ISO/IEC 2382-17:1999]

Andmekogu	Andmekogu on riigi, kohaliku omavalitsuse või muu avalik-õigusliku isiku või avalikke ülesandeid täitva eraõigusliku isiku infosüsteemis töödeldavate korrastatud andmete kogum, mis asutatakse ja mida kasutatakse seaduses, selle alusel antud õigusaktis või rahvusvahelises lepingus sätestatud ülesannete täitmiseks [AvTS §41 ¹]
Andmekvaliteedi reegel	Reegel, millele andmed peavad vastama
Andmekvaliteedi indikaator	Numbriliselt mõõdetav andmekvaliteedi ühik
Vastasseis?	(i.k. confrontation) - andmete andmekvaliteedi reeglite või indikaatoritega vastandamise tulemused
Põhiandmed	Põhiandmed on riigi infosüsteemi kuuluvasse andmekogusse kogutavad andmekogu unikaalsed andmed, mis tekivad andmekogu haldaja avalike ülesannete täitmise käigus [AvTS §43 ⁶]
SA	Statistikaamet http://www.stat.ee
Andmestik	Andmete hulk, mis on avaldatud ja mida hallatakse kindla isiku poolt ning millele saab anda juurdepääsu või seda alla laadida ühes või enamas vormingus [DCAT]
Avaandmed	Avalik teave, mille üldist kasutamist ei ole seaduse alusel piiratud [AvTS §3]
Andmete elutsükel	Andmete elutsükel põhineb toote elutsüklil ning koosneb järgmistest tegevustest: andmete planeerimine, disain ja võimaluste loomine, andmete loomine ja hankimine, andmete talletamine ja säilitamine, andmete kasutamine, andmete täiustamine ja andmete kustutamine või arhiveerimine (DAMA DMBOK2)
AvTS	Avaliku teabe seadus https://www.riigiteataja.ee/akt/115032019011
Andmekvaliteet	Näitab, mil määral andmekarakteristikud rahuldavad teadaolevaid või eeldatavaid vajadusi kasutamisel ettemääratud tingimustes [ISO/IEC 25012]
Andmeobjekt	Andmeelement või määratletud andmeelemendikogum, mis on seotud üheainsa tähendust ja kompositsiooni määrava sildiga [ISO/IEC 18013-2] vt ka https://akit.cyber.ee/term/5074-andmeolem-1-andmeuksus

2. Andmekvaliteedi juhtimise tegevused

2.1 Andmekvaliteedi eesmärkide määramine

Usaldusväärsed ja kõrge kvaliteediga andmekogudes hoitavad andmed (näiteks rahvastiku ja avalike teenuste kohta hoitavad andmed) võimaldavad asutustel ja riigil teha paremaid otsuseid. Et paremate otsuste langetamine oleks võimalik peavad andmed olema usaldusväärsed. Halvasti hallatud andmete tõttu tekkivad probleemid sarnanevad halvasti hallatud finantside puhul tekkivatele probleemidele. Riigi ja asutuste peamine andmehalduse ning sealhulgas ka andmekvaliteedi tagamise eesmärk on maksimeerida andmetest saadavat väärtust. Samamoodi nagu loob väärtust muude ressursside korrektne haldamine. Andmete väärtuse all tuleb silmas pidada korrektsetest andmetest ja nende kasutamisest tõusvat kasu. Kvaliteetsetest andmetest tulenev kasu seisneb riigi võimes pakkuda kuluefektiivselt kvaliteetseid avalikke teenuseid ning langetada korrektselt riigi ja kodaniku jaoks olulisi otsuseid. Näiteks loovad õiged andmed võimekuse täpselt planeerida tulevast lasteaiakohtade arvu, omada täpset ülevaadet riigi reservväelaste hulgast ning tagavad ka korrektse sotsiaaltoetuste määramise.

Andmekvaliteedi tõstmine ja hoidmine on järjepidev protsess, mitte ühekordne projekt. Andmekvaliteedi programmiga alustamise eesmärgid on järgmised:

- ▶ Organisatsiooni andmete väärtuse tõstmine;
- ▶ Uute võimaluste loomine andmete kasutamiseks (ehk andmete väärimine);
- ▶ Madalast andmekvaliteedist tulenevate riskide vähendamine;
- ▶ Organisatsiooni efektiivsuse ja produktiivsuse tõstmine;
- ▶ Organisatsiooni reputatsiooni kaitsmine ja tugevdamine.

Organisatsioonide jaoks, mis püüavad andmete abil väärtust luua, on kõrge kvaliteediga andmed oluliselt väärtuslikumad kui madala kvaliteediga andmed. Madala kvaliteediga andmetega kaasnevad lisaks ka kõrgemad riskid. Näiteks võib halb andmekvaliteet kahjustada organisatsiooni reputatsiooni, tuua kaasa rahalisi kaotusi ja negatiivseid meediakajastusi. Madala andmekvaliteediga on seotud ka mitmed otsesed kulud, näiteks:

- ▶ Suutmatus esitada korrektseid arveid;
- ▶ Suurenenud kliendiprobleemide arv ning vähenenud suutlikkus nende lahendamiseks;
- ▶ Suurem keerukus asutuste tegevuse (ümber)korraldamisel;
- ▶ Väiksem võimekus pettuste tuvastamisel;
- ▶ Madala või negatiivse mõjuga otsused.

Usaldusväärsed andmed mitte ainult ei maanda riske ja vähenda kulutusi, vaid toetavad ka efektiivsuse tõusu ning on üheks vahendiks organisatsiooni edu saavutamisel. Kvaliteetsed andmed aitavad töötajatel küsimustele kiiremini vastata, sest vähem aega kulub andmete õigsuse kontrollimisele. Seega jagub ajalist ressurssi rohkem andmete sisuliseks analüüsiks ja õigete otsuste langetamiseks.

Asutuse andmekvaliteedi protsess peaks juhinduma järgmistest põhimõtetest:

- ▶ **Kriitilisus** – andmekvaliteedi protsess peaks keskenduma kõige kriitilisematele andmetele. Muudatuste prioriseerimine peaks põhinema andmete kriitilisusel ning võtma arvesse võimalikke ebakorreksete andmetega kaasnevaid riske.

- ▶ **Elutsükli juhtimine** – andmekvaliteeti tuleb juhtida kogu andmete elutsükli jooksul. See hõlmab andmete liikumise haldamist nii protsessides, süsteemides kui ka eri süsteemide vahel. Näiteks peab iga andmeahela lüli (nii protsess kui süsteem) tagama andmete kõrge kvaliteedi.
- ▶ **Ennetamine** – andmekvaliteedi haldamise protsess peaks keskenduma andmevigade ennetamisele ning andmete kasutatavust pärssivate tegurite vähendamisele. Kindlasti ei tohiks keskenduda vaid andmevigade parandamisele.
- ▶ **Probleemide juurpõhjuste lahendamine** – andmekvaliteedi tõstmine tähendab enamasti kui andmevigade parandamist. Andmekvaliteedi probleemide lahendamiseks tuleb tuvastada probleemide algallikad, mitte keskenduda vaid tagajärgede likvideerimisele. Seejuures hõlmab andmekvaliteedi tõstmine tihti protsesside ja süsteemide täiustamist, mis on levinud kvaliteediprobleemide algallikad.
- ▶ **Andmehaldus** – andmehalduse tegevused peavad toetama kõrgekvaliteediliste andmete teket ning andmekvaliteedi protsessi tegevused peavad toetama ja säilitama hallatavat andmekeskonda.
- ▶ **Sihttasemetest lähtuv** - andmekvaliteedi reeglid kuuluvad kõikidele andmete elutsükli osapooltele. Nendele andmekvaliteedi reeglitele peaksid olema määratletud sihttasemed.
- ▶ **Objektiivne mõõtmine ja läbipaistvus** - andmekvaliteedi taset tuleb mõõta objektiivselt ja järjepidevalt ning mõõtetulemusi ja meetodeid tuleks jagada kõigi osapooltega.
- ▶ **Äriprotsessidesse juurutamine** - äriprotsesside omanikud ehk protsesside eest vastutajad peavad tagama, et ärireeglites sisalduvad andmekvaliteeti puudutavad reeglid.
- ▶ **Süsteemidesse juurutamine** - süsteemide omanikud, ehk süsteemide eest vastutajad peavad tagama, et süsteemides rakendatakse andmekvaliteedi mõõtmiseks andmekvaliteedi reegleid.
- ▶ **Teenustasemega ühendamine** - teenustaseme lepingud (*Service Level Agreements*) peaksid sisaldama andmekvaliteedist raporteerimist ja probleemide haldamist puudutavaid punkte.

2.2 Andmekvaliteedi reeglite kirjeldamine

Andmekvaliteedi programmiga alustades tuleb esmalt saada ülevaade olemasolevatest andmetest ja andmekvaliteedi hetkeseisust. Üks viis andmetest esmase ülevaate saamiseks on viia läbi andmete profileerimine.

Andmete profileerimine on protsess, mille eesmärgiks on uurida olemasolevaid andmeid (andmebaasist, konkreetsest failist jm) ning koguda statistikat ja informatiivseid kokkuvõtteid andmete koosseisu kohta. Näiteks tuvastatakse profileerimise käigus arväärtuste esinemissagedus, formaat, mustrid ja muud andmeid iseloomustavad omadused. Saadud info põhjal on võimalik küsida täiendavaid küsimusi, mis omakorda aitavad tuvastada andmekvaliteedi reegleid. Andmekvaliteedi reeglid on sisendiks andmekvaliteedi programmi hilisemas faasis toimuvale andmekvaliteedi hindamisele.

Andmete profileerimise teostamiseks kasutatakse üldjuhul profileerimistööriistu, mis annavad hea esmase ülevaate andmetest ja andmete kvaliteedist. Juhendi rakenduslikus osas on toodud juhised profileerimise teostamiseks Ehitisregistri näitel. Valik profileerimise teostamiseks sobivatest tööriistadest on leitav käesoleva juhise soovituslike töövahendite sektsioonis. Kuigi profileerimistööriistad esitavad andmete kohta mitmesugust statistikat ja mõõdikuid, pole profileerimise puhul tegu andmekvaliteedi hindamisega. Andmekvaliteedi mõõtmine on põhjalikum tegevus, mille käigus hinnatakse andmete vastavust andmekvaliteedi reeglitele.

Võimalik on eristada nelja tüüpi andmete profileerimist:

- ▶ **Struktuuripõhise** profileerimise käigus analüüsitakse andmete järjepidevust ja formaadilist korrektsust. Lisaks teostatakse matemaatilisi kontrolle (näiteks summa leidmine, miinimumväärtuste leidmine ja maksimumväärtuste leidmine). Struktuuripõhine profileerimine aitab tuvastada, kui hästi on andmed struktureeritud. Näiteks kui palju on vale pikkusega telefoninumbreid.
- ▶ **Sisupõhise** profileerimise käigus analüüsitakse konkreetseid andmekirjeid, mille tulemusena on võimalik tuvastada andmekirjetes esinevaid süstemaatilisi probleeme. Näiteks ilma suunakoodita telefoninumbrite esinemist.
- ▶ **Seostepõhise** profileerimise käigus tuvastatakse andmete omavahelised seosed, näiteks andmetabelite vahelised seosed või arvutustabelis (näiteks MS Exceli failis) hoitavate tabelite või väljade seosed.

Olles profileerimise käigus saanud esmase ülevaate andmetest ning nende kvaliteedist, on võimalik teostada andmekvaliteedi reeglite kirjeldamine. Andmekvaliteedi reeglid tuleks esmalt kirjeldada kõige olulisematele andmetele ehk alustada tuleks andmetest, mis loovad asutusele ning selle klientidele enim väärtust. Seetõttu on andmekvaliteedi reegleid tihti mõistlik kirjeldada esmalt põhiandmetele, mille Avaliku teabe seadus defineerib järgmiselt: "Põhiandmed on riigi infosüsteemi kuuluvasse andmekogusse kogutavad andmekogu unikaalsed andmed, mis tekivad andmekogu haldaja avalike ülesannete täitmise käigus."

Andmekvaliteedi reeglite kirjeldamise eesmärgiks on ilmutada nõuded, millele vastavus tagab andmete kasulikkuse ja kasutatavuse organisatsioonis. Osad andmekvaliteedi reeglid tulenevad ärireeglitest. Ärireeglid kirjeldavad protsesside sisemist toimimist eesmärgiga tagada äriine edu ja sobivus ärikeskkonnaga. Seejuures ei kajastu kõik andmekvaliteedi reeglid ärireeglites ning vastupidi, osad ärireeglid ei kajastu andmekvaliteedi reeglites.

Tihti puudub äri- ja andmekvaliteedi reeglite kohta selge dokumentatsioon, sellisel juhul on neid võimalik tuvastada analüüsides olemasolevaid äriprotsesse, töövooge, regulatsioone, eeskirju, standardeid, programmide lähtekoodi jms kättesaadavat informatsiooni. Seejuures on andmekvaliteedi reeglite kirjeldamisel abiks eelnevalt profileerimise käigus andmete kohta kogutud informatsioon. Kirjeldamist aitab teostada ka käesoleva juhise seksioonis 2.3.2 toodud andmekvaliteedi probleemide raamistik, mis esitab 21 tüüpilist andmekvaliteedi probleemi ning nende esinemist illustreerivad näited.

Juhise järgnevates seksioonides kirjeldatud andmekvaliteedi dimensioonid koos täpsustavate indikaatoritega toetavad samuti andmekvaliteedi reeglite kirjeldamist. Andmekvaliteedi probleemide ja dimensioonide seos on illustreeritud Tabel 1, kus on kirjeldatud ka millist andmestiku osa konkreetne probleem puudutab (näiteks atribuuti, veergu, kirjet või andmete vahelisi seoseid). Atribuudi ja veeru tasemel andmekvaliteedi reeglite kirjeldamine on pigem madala keerukusega. Näiteks täielikkuse dimensiooni kuuluvad andmekvaliteedi reeglid kirjeldavad, kas tegu on kohustusliku või valikulise veeruga. Valikulise veeru puhul peavad olema täpsustatud ka tingimused, millal antud veergu täita tuleb. Lisaks peaksid reeglid olema defineeritud andmestiku tasemel. Näiteks „Kõigis andmestikes peab soo tähis „M“ tähistama meessugu.“

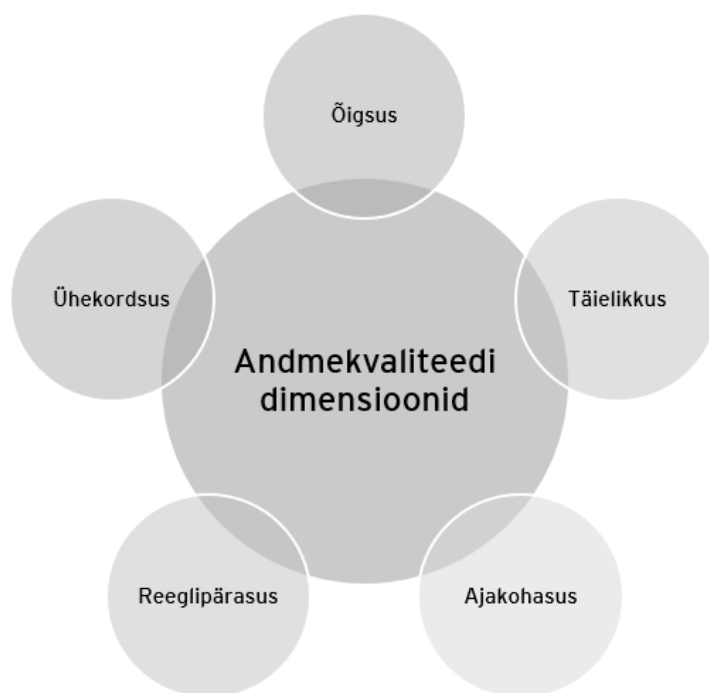
Vajadusel on reeglite kirjeldamisel abiks äriprotsesside sisendite ja väljundite täpsustamine äriprotsessi eri osapooltega. Samuti on kasulik uurida osapoolte probleeme. Näiteks täpsustada mis juhtub, kui andmed on valed või puuduvad ja kuidas tuvastatakse probleeme. Seejuures on kasulik meeles pidada,

et andmekvaliteedi hindamiseks pole vaja teada kõiki andmekvaliteedi reegleid. Reeglite tuvastamine ja täpsustamine on pidev protsess. Üks parimaid viise andmekvaliteedi reeglite kogumiseks on andmekvaliteedi hindamise tulemuste eri osapooltega jagamine. Tihti aitab tulemuste jagamine osapooli uute vaatenurkade leidmisel ning seeläbi uute reeglite sõnastamisel.

Eelpool kirjeldatud tegevuste tulemuseks on selgelt sõnastatud andmekvaliteedi reeglid, näiteks „Väli „SYNNIKUUPAEV“ on kohustuslik ning peab olema väärtustatud.“ Reeglite kirjeldamisele järgneb andmekvaliteedi mõõtmine, mille käigus teostatavad mõõtmised näitavad andmete vastavust andmekvaliteedi reeglile, näiteks „3% juhtudest pole väli väärtustatud, seega on andmete täielikkus 97%.“

2.3 Andmekvaliteedi mõõtmine

Andmekvaliteedi dimensioonid (Joonis 2) on mõõdetavad andmete omadused, mis väljendavad andmete kvaliteeti erinevatest aspektidest lähtuvalt. Eksisteerib palju erinevaid andmekvaliteedi dimensioonide käsitlusi, kuid antud juhises keskendutakse 5-le kvaliteedidimensioonile ning nende hindamist toetavate indikaatorite kirjeldamisele. Võimalik on kasutada ka teistsuguseid dimensioonide liigitusi, kuid konkreetsed andmekvaliteedi probleemid seejuures ei muutu. Teistsugust dimensioonide liigitust kasutades on vaja määrata andmekvaliteedi probleemide seosed dimensioonidega. Antud dimensioonide puhul on nimetatud seosed kirjeldatud käesoleva juhise seksioonis 2.3.2 toodud tabelis (Tabel 1).



Joonis 2. Andmekvaliteedi dimensioonid

Õigsus (Accuracy) näitab, mil määral vastavad andmed tegelikkusele. Andmete õigsus jaguneb süntaktiliseks ja semantiliseks õigsuseks. Süntaktiline õigsus kontrollib andmete vormilist korrektsust. Näiteks kui nimi „Tõnu“ on andmetes talletatud kui „Tõnu“ pole andmed süntaktiliselt õiged. Semantiline õigsus kontrollib andmete sisulist korrektsust ehk autentsust. Näiteks kui inimese nimi on „Tõnu“ aga tema sooks on märgitud „N“ (Naine).

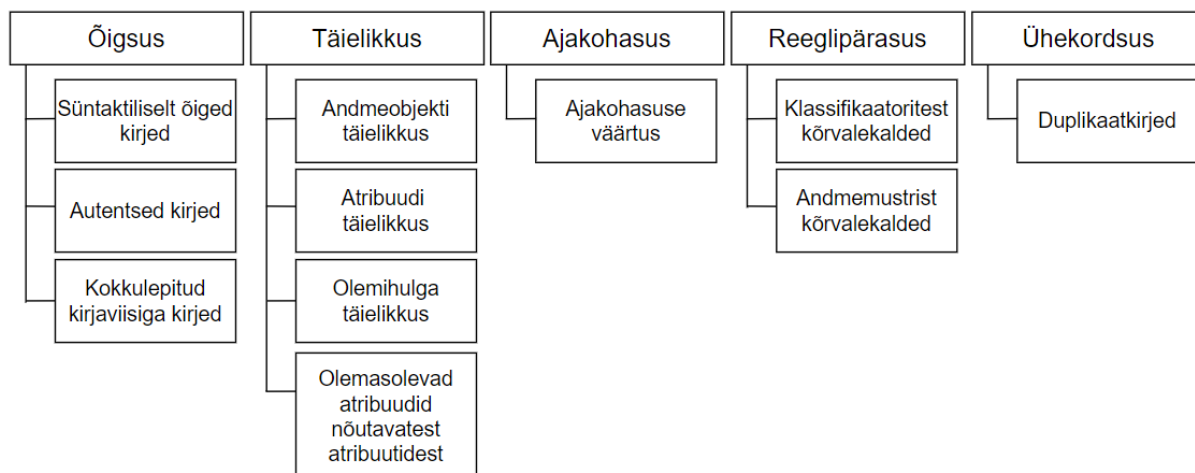
Täielikkus (Completeness) näitab, mil määral on olemas kõik nõutud andmed. Täielikkus on vaadeldav kahes osas: kirjade täielikkus ja kogumi ehk populatsiooni täielikkus. Kirjade täielikkus näitab, mil määral on andmekirje kõigil tunnustel olemas väärtused. Näiteks mil määral on andmetabeli veeru read (ehk atribuudid) täidetud. Populatsiooni täielikkus näitab kas kõik nõutavad kirjed on olemas. Näiteks andmetabeli puhul kõigi nõutud veergude olemasolu.

Ajakohasus (Timeliness) näitab, mil määral andmete värskus ja kättesaadavus vastab vajadustele ja nõuetele. Aja jooksul andmed muutuvad ning viide reaalsest sündmuste ning nende andmetes fikseerimise või andmete värskendamise vahel on vältimatu. Seetõttu on võimalik olukord, kus andmed on küll uuendatud, kuid nende tekkeks või värskendamiseks kuluv aeg muudab andmete kasutamise mõne konkreetse ülesande jaoks võimatuks. Näiteks võib ülikooli tunniplaan olla küll värske, kuid see pole ajakohane kui see jõuab tudengiteni alles pärast loengute algust.

Reeglipärasus (Orderliness) näitab, mil määral andmete formaat ja struktuur vastab nõuetele. Esiteks tähendab reeglipärasus kokkulepitud klassifikaatorite kasutamist (näiteks EMTAK-i kasutamist majandusliku tegevusala talletamiseks). Teiseks tähendab reeglipärasus kokkulepitud andmemustrite järgimist. Näiteks on kokkulepitud andmemuster (süntaks) kuupäeval ja isikukoodil ning need on seotud andmetüüpidega kuupäev (*date*) ja arv (*integer*). Reeglipärasuse alla kuulub ka andmete küsimine kokkulepitud põhiandmete allikast. Põhiandmete allikas võib olla nii asutuse sees (*master data*), kui ka üleriigiline. Üleriigilise põhiandmete allika puhul on tavaliselt tegu kokkulepitud klassifikaatoriga nagu aadressiandmed, katastritunnus või äriregistri kood.

Ühekordsus (Uniqueness) näitab, mil määral esineb andmetes duplikaatkirjeid. Ühekordsuse probleem tekib juhul, kui ühe reaalsest elust pärineva objekti kohta on andmetes talletatud kaks või enam kirjet. Näiteks kui ühe isiku kohta on andmetes talletatud mitu kirjet.

Dimensioonid on konkreetsemaks hindamiseks jagatud mõõdetavateks indikaatoriteks (Joonis 3).



Joonis 3. Dimensioonide jaotus indikaatoriteks

2.3.1 Andmekvaliteedi mõõtmise metamudel

Andmekvaliteeti kasutatakse eri tasemetel otsustamiseks nii andmehalduse enda kui muude asutuse või riigi tegevuste korraldamisel ja täideviimisel. Ühelt poolt kombineeritakse andmekvaliteedi reeglitele vastavused/mittevastavused indikaatorite väärtusteks ja need omakorda dimensioonide väärtusteks ja osad neist võivad saada võtmemõõdikuteks üksikutes asutustes või üle riigi. Teiselt poolt

agregeeritakse nimetatud tunnuseid andmeelementidelt andmestiku andmeobjekti liikide, andmestike ja asutuse või riigi tasemel otsuste tegemiseks.

Andmete kvaliteedi kasutamist iseloomustavad järgnevad stsenaariumid:

- SC-DQ-1: Selleks, et tagada ülevaade hoonete energiatõhususest tõstetakse Ehitisregistri energiamärgist puudutavate andmete süntaktiliselt õigete kirjade määr 95%-ni.
- SC-DQ-2: Selleks, et tagada riigi kodanike operatiivne teavitamine läbi digikanalite tõstetakse Eesti kodanike kontakttelefoninumbrite, e-postiaadresside ja kontaktaadresside täielikkus riigiüleselt 99%-ni.
- SC-DQ-3: Selleks, et tagada kodanike kirjade 100% täielikkus andmestikus X, on vaja parandada eesnimi või perekonnanimi 250 kirjes.
- SC-DQ-4: Selleks, et tagada piirkondlikele otsustele ühtlane kvaliteet, on vaja tõsta aadressi kirjade õigsus asutuste X, Y ja Z andmestikes 95%-ni.
- SC-DQ-5: Selleks, et võimaldada täielikult registripõhist rahvaloendust, on vaja tõsta rahvastikuregistri elukohaandmete täielikkus 100%-ni ja õigsus 90%-ni.

Taoliste stsenaariumite täideviimist toetab andmekvaliteedi mõõtmise metamudel (Joonis 4).

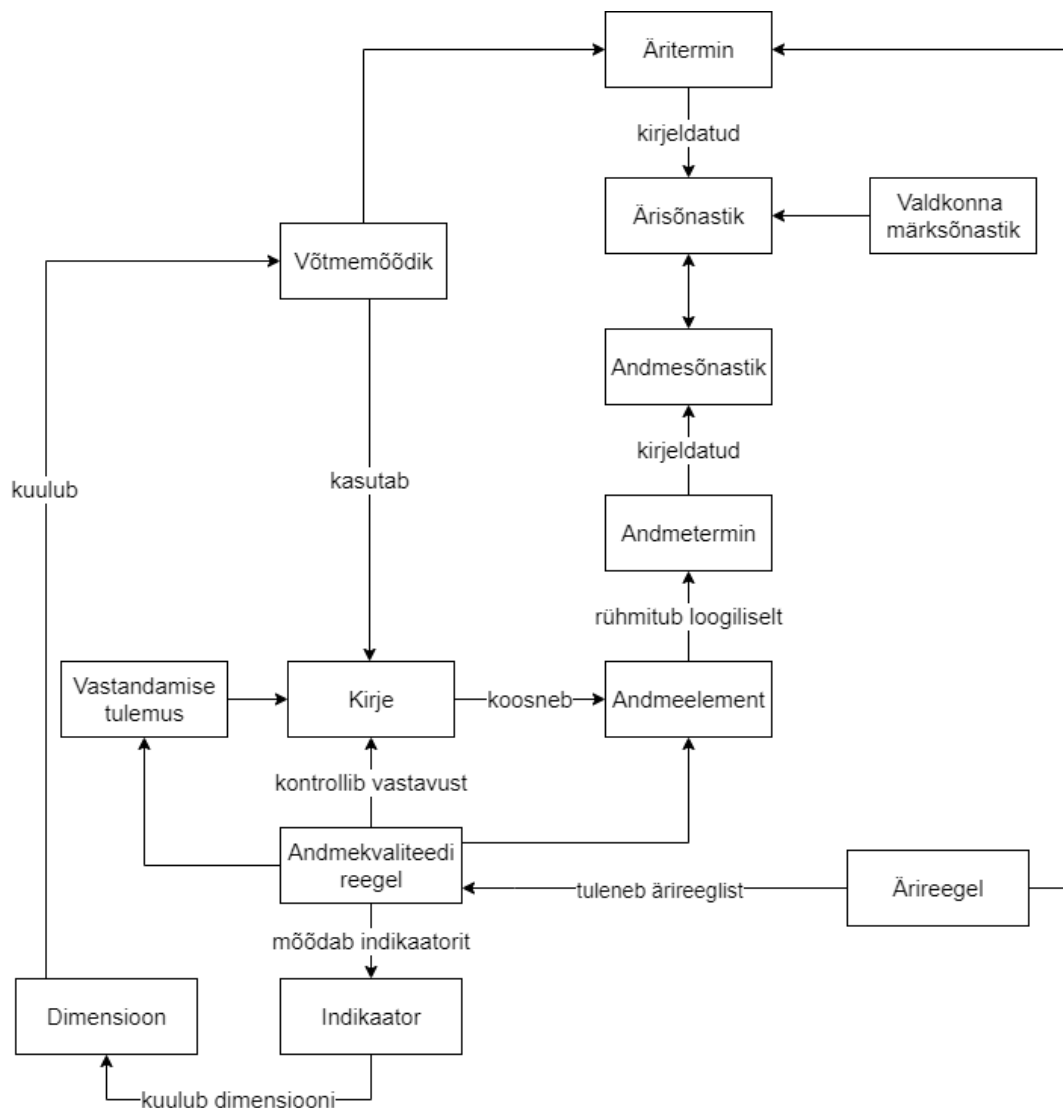
Andmekirjelduse olemasolu on andmekvaliteedi mõõtmise eelduseks. Seosed andmekirjelduse ja andmekvaliteedi olemite vahel on esitatud andmekvaliteedi mõõtmise metamudelis (Joonis 4). Täpsemad juhised andmekirjelduse loomiseks on esitatud andmekirjelduse juhises, mis valmis paralleelselt käesoleva juhise ja on samuti osa andmehalduse raamistikust.

Andmekirjelduse koostamisel ja haldamisel kasutatakse kolme tüüpi sõnastikke: **Valdkonna märksõnastik**, **Ärisõnastik** ja **Andmesõnastik**. **Andmeelemendid** rühmituvad **Andmeterminiteks**, mis omakorda on kirjeldatud **Andmesõnastikus**. **Äriterminid** on kirjeldatud **Ärisõnastikus**, lisaks kasutatakse **Ärisõnastike** loomisel ühe võimaliku terminite allikana **Valdkondlikke märksõnastikke**. **Andmeelemendid** rühmituvad loogiliselt (kontseptualiseerimine) **Andmeterminiteks** (näiteks **Andmetermin** „Isik“ koosneb vähemalt **Andmeelementidest** „eesnimi“, „perekonnanimi“ ja „isikukood“). **Andmeelemendid** on **Andmesõnastiku** kaudu seotud organisatsiooni tegevust kirjeldavate mõistetega ehk **Ärisõnastikuga**.

Käesolev andmekvaliteedi juhise kasutab järgmiseid andmekvaliteedi mõõtmist võimaldavaid olemeid: **Andmekvaliteedi reegel**, **Dimensioon** ja **Indikaator**. **Andmekvaliteedi reeglid** tulenevad **Ärireeglitest**, mis on kirja pandud **Ärisõnastikus** kirjeldatud **Äritermineid** kasutades (vt ka andmehalduse raamistik ptk. 4.1.5). Samas ei ole kõik **Ärireeglid** kajastatud **Andmekvaliteedi reeglites** ning vastupidi, osad **Andmekvaliteedi reeglid** ei kajastu **Ärireeglitest**. Andmekvaliteedi mõõtmisel rakendatakse **Andmekvaliteedi reegleid** kas üksikutele andmete **Kirjetele** või nende kogumitele ning tulemuseks saadakse **Andmekvaliteedi reeglitele** vastandamise tulemus (i.k. *confrontation*). **Andmekvaliteedi reeglites** kasutatakse muutujaid, mis on andmetes esitatud **Andmeelementidena** (nt „ehr_kood“ on kirje „Ehitis“ **Andmeelement**). Vastandamise tulemused kombineeritakse **Indikaatoriteks** ja need omakorda **Dimensioonideks**. (Iga **Indikaator** kuulub **Dimensiooni**, näiteks **Indikaator** „Klassifikaatoritest kõrvalekalde“ kuulub **Dimensiooni** „Reeglipärasus“.) Vaikimisi kasutatakse kombineerimise funktsioonina aritmeetilist keskmist, kuid vajadusel võib kasutada ka muid funktsioone, kui nendes on eelnevalt kokku lepitud. Juhul kui kokkulepe eksisteerib ning mõõtmine on

ühetaoline, on võimalik saada ülevaade andmehalduse olukorrast võrreldes andmekvaliteedi mõõtmise tulemusi eri andmekogude ja organisatsioonide lõikes.

Mõned andmekvaliteedi **Dimensioonid** võivad olla ka osa organisatsiooni võtmemõõdikutest, mis omakorda on **Äritermineid** kasutades kirjeldatud **Ärisõnastikus**. **Võtmemõõdikuteks** olevate **Dimensioonide** arvutamisel kasutatakse organisatsiooni andmeid.



Joonis 4. Andmekvaliteedi mõõtmise metamudel

2.3.2 Andmekvaliteedi probleemid

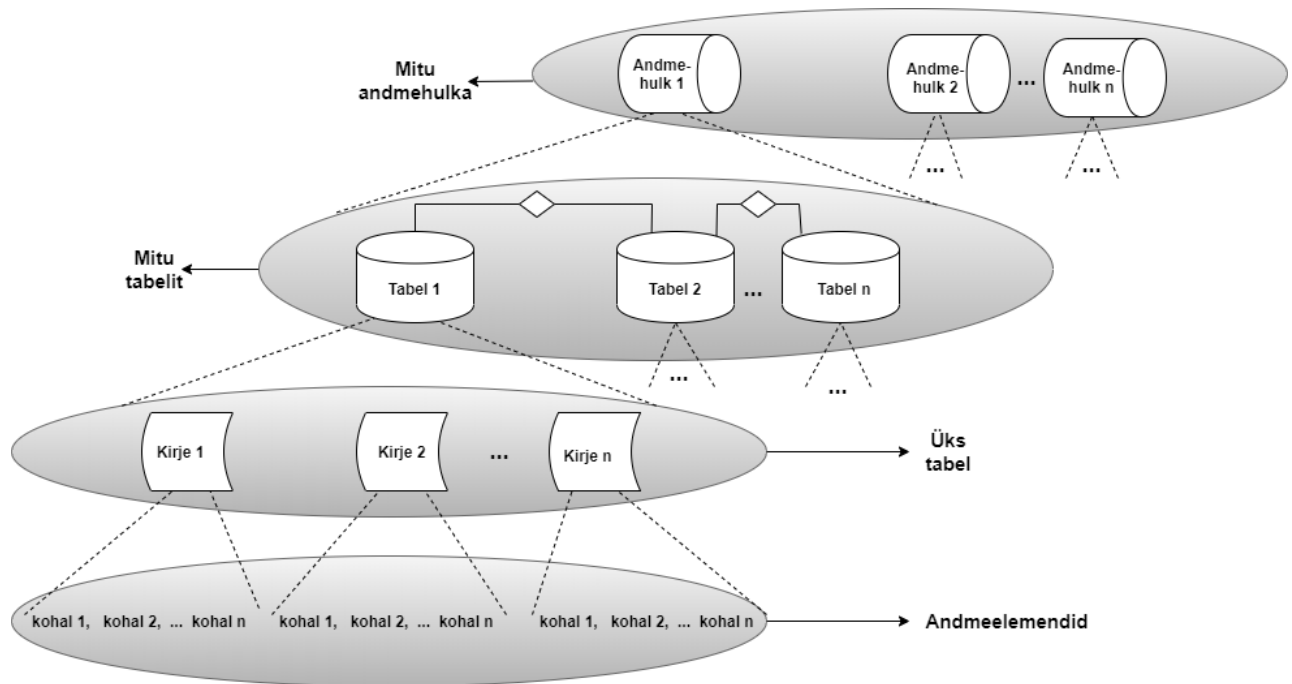
Selleks, et andmekvaliteedi probleeme süsteemselt käsitleda, lihtsustada andmekvaliteedi reeglite koostamist ja tagada andmekvaliteedi dimensioonide ühtlane kaetus andmekvaliteedi reeglitega oleme kasutanud Oliveira jt poolt välja töötatud andmekvaliteedi probleemide raamistikku¹⁰. Lisaks

¹⁰ Oliveira, Paulo, Fátima Rodrigues, and Pedro Rangel Henriques. "A formal definition of data quality problems." ICIQ. 2005. <http://mitiq.mit.edu/ICIQ/Documents/IQ%20Conference%202005/Papers/AFormalDefinitionofDQProblems.pdf>

toome ilmutatud kujul välja valitud raamistiku loojate poolt välja pakutud¹¹ andmete ajakohasust puudutavad probleemid. Võrreldes teiste sarnaste raamistikega^{12,13,14} on valitud lähenemise eelised järgmised:

1. tuvastatud probleemid tulenevad laiapõhjalisest juhtumiuuringust;
2. tegu on formaalse raamistikuga, kus on kirjeldatud konkreetsed algoritmid andmekvaliteedi probleemide tuvastamiseks ja klassifitseerimiseks.

Raamistik eristab andmekvaliteedi probleeme vastavalt sellele kas need ilmnevad eri andmehulkade, andmeobjektide vaheliste seoste või andmeelementide/kirjete tasemel (vaata Joonis 5).



Joonis 5. Andmete korraldamise tüüpiline mudel

Alljärgnevalt (Tabel 1) on konkreetsed andmekvaliteedi probleemid esitatud detailsuse taseme järgi vastavalt eelpool toodud andmete korraldamise tüüpilisele mudelile (Joonis 5). Lisaks on toodud andmekvaliteedi probleemide ja andmekvaliteedi dimensioonide vahelised seosed.

¹¹ Oliveira Paulo, Fátima Rodrigues, Pedro Henriques, and Helena Galhardas. "A taxonomy of data quality problems." In 2nd Int. Workshop on Data and Information Quality, pp. 219-233. 2005.
https://s3.amazonaws.com/academia.edu.documents/43292389/A_Taxonomy_of_Data_Quality_Problems20160302-6272-51xfc8.pdf

¹² Kim, W., Choi, B.J., Hong, E.K., Kim, S.K. and Lee, D., 2003. A taxonomy of dirty data. Data mining and knowledge discovery, 7(1), pp.81-99.
http://biosoft.kaist.ac.kr/bisl_2018/paperPDF/2003TaxonomyOfDirty.pdf

¹³ Li, L., Peng, T. and Kennedy, J., 2011. A rule based taxonomy of dirty data. GSTF Journal on Computing, 1(2).
<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.928.5766&rep=rep1&type=pdf>

¹⁴ N. Laranjeiro, S. N. Soydemir and J. Bernardino, "A Survey on Data Quality: Classifying Poor Data," 2015 IEEE 21st Pacific Rim International Symposium on Dependable Computing (PRDC), Zhangjiajie, 2015, pp. 179-188, doi: 10.1109/PRDC.2015.41.
https://www.researchgate.net/profile/Nuno_Laranjeiro/publication/281556889_A_Survey_on_Data_Quality_Classifying_Poor_Data/links/563b828208aec6f17dd4ddde.pdf

Tabel 1. Andmekvaliteedi probleemid korraldatud detailsuse taseme järgi ning seosed andmekvaliteedi dimensioonidega.

Andmekvaliteedi probleem	Detailsuse tase						Dimensioonid				
	Andme- element	Veerg	Kirje	Üks tabel	Mitu tabelit	Mitu andmehulka	Õigsus	Täielikkus	Ajakohasus	Reeglipärasus	Ühekordsus
Puuduv väärtus	✓							✓			
Süntaksi rikkumine	✓						✓				
Vale väärtus	✓						✓		✓		
Väärtusvahemiku rikkumine	✓						✓				
Sobimatu alamstring	✓						✓				
Õigekirjaviga	✓						✓				
Ebatäpne väärtus	✓						✓				
Valdkonnakitsenduse rikkumine	✓	✓	✓	✓	✓	✓	✓			✓	
Unikaalse väärtuse rikkumine		✓									✓
Sünonüümide kasutus		✓				✓				✓	
Pooltühi kirje			✓					✓			
Funktsionaalse sõltuvuse rikkumine			✓				✓			✓	
Ligikaudselt dubleerivad kirjed				✓		✓				✓	✓
Vastuoluliselt dubleerivad kirjed				✓		✓	✓				✓
Viiteterviklus					✓			✓			
Vale viide					✓		✓		✓		
Süntaksite mitmekesisus					✓	✓				✓	
Ringlus kirjete seas					✓			✓			
Mõõtmisühikute mitmekesisus						✓				✓	
Esitluse mitmekesisus						✓				✓	
Homonüümide kasutus						✓					✓

Et lihtsustada eeltoodud tabeli (Tabel 1) mõistmist on järgnevalt toodud tabeli veergude selgitused:

- ▶ **Andmeelement:** Andmekvaliteedi probleemid, mis ilmnevad konkreetse andmeelemendi tasemel ehk kui vaadeldakse korraga konkreetset väärtust.
- ▶ **Veerg:** Andmekvaliteedi probleemid, mis ilmnevad veeru tasemel ehk kui vaadeldakse veeru väärtusi üle kõikide kirjete.
- ▶ **Kirje:** Andmekvaliteedi probleemid, mis ilmnevad kirje tasemel ehk kui vaadeldakse korraga kirje ehk ühe tabeli rea kõiki väärtusi.
- ▶ **Üks tabel:** Andmekvaliteedi probleemid, mis ilmnevad ühe tabeli tasemel ehk kui vaadeldakse korraga ainult ühes tabelis olevaid andmeid.
- ▶ **Mitu tabelit:** Andmekvaliteedi probleemid, mis ilmnevad siis, kui omavahel on seotud mitu tabelit ehk kui vaadeldakse korraga mitmes tabelis olevaid andmeid.
- ▶ **Mitu andmehulka:** Andmekvaliteedi probleemid, mis ilmnevad siis, kui omavahel on seotud mitu andmehulka ehk kui vaadeldakse korraga mitmes andmehulgas olevaid andmeid. Näiteks Rahvastikuregistris ja Ehitusregistris hoitavad isikuandmed.
- ▶ **Dimensioonid (Õigsus, Täielikkus, Ajakohasus, Reeglipärasus, Ühekordsus):** Näitab millise dimensiooni alla konkreetne andmekvaliteedi probleem liigitub.

Järgnevalt on esitatud kõigi eelnevalt toodud andmekvaliteedi probleemide (Tabel 1) täpsemad kirjeldused ning konkreetse probleemi esinemist illustreerivad näited.

2.3.2.1 Puuduv väärtus

Probleemi kirjeldus: Puudub kohustuslikuks määratud atribuudi väärtus. Valikuliste väärtuste puudumine ei ole andmekvaliteedi probleem.

Näide: Kohustuslik atribuut "ISIKUKOOD" pole väärtustatud.

2.3.2.2 Süntaksi rikkumine

Probleemi kirjeldus: Atribuudi väärtuste mustrid erinevad kokkulepitud mustrist/süntaksist.

Näide: "TELLIMUSE_KUUPAEV" on esitatud kujul 2020/03/05. Tegelikult peaks see olema esitatud kokkulepitud kujul 03/05/2020.

2.3.2.3 Vale väärtus

Probleemi kirjeldus: Valede väärtuste all peame silmas väärtuseid, mis on aegunud ja ei kajasta enam tegelikkust või mille alamosal puudub tähendus kirje vaadeldava või mõne teise atribuudi väärtusena.

Näide: Atribuudi "TELLIMUSE_KUUPAEV" väärtus on 03/05/2020, kuid see peaks tegelikult olema 08/12/2020.

2.3.2.4 Väärtusvahemiku rikkumine

Probleemi kirjeldus: Väärtusvahemiku rikkumised avalduvad kahel viisil: kui atribuudi väärtus on väljaspool ettenähtud numbrilist vahemikku või kui see erineb ettenähtud loendi/klassifikaatori väärtustest.

Näide: Konkreetse isiku puhul on atribuudi "VANUS" väärtus negatiivne.

2.3.2.5 *Sobimatu alamstring*

Probleemi kirjeldus: Mõnikord kombineeritakse samasse tekstilisse atribuuti kokku erinevad andmeelemendid. Kui atribuudi väärtus sisaldab atribuudi skoobist väljaulatavaid elemente, siis on tegemist sobimatu alamstringiga.

Näide: Atribuut "NIMI" sisaldab ka akadeemilist tiitlit (näiteks "Urve Miller, PhD").

2.3.2.6 *Õigekirjaviga*

Probleemi kirjeldus: Tekstilise atribuudi väärtus sisaldab õigekirjavigasid.

Näide: Atribuut "LINN" väärtus on "Tallin," kuid see peaks tegelikult olema "Tallinn".

2.3.2.7 *Ebatäpne väärtus*

Probleemi kirjeldus: Ebatäpse väärtuse probleemid kerkivad esile siis, kui tekstilise väärtuse kodeerimisel kasutatakse lühendeid. Kui lühendite tähendus ja interpretatsioon aja jooksul või rakenduste lõikes muutub, siis muutub atribuudi väärtuse tähendus ebatäpseks.

Näide: Atribuut "NIMI" on väärtustatud kui „Ant,“ mis võib tähistada nimesid Anton ja Antonina.

2.3.2.8 *Valdkonnakitsenduse rikkumine*

Probleemi kirjeldus: Osade atribuutide tähendus ja kuju tuleneb valdkonnas kehtivatest kitsendustest või headest tavadest ning neist mitte kinnipidamisel väheneb andmete valdkondlik kasutatavus.

Näide: Atribuut "NIMI" peab sisaldama vähemalt kahte sõna.

2.3.2.9 *Unikaalse väärtuse rikkumine*

Probleemi kirjeldus: Kaks või enam kirjet, mis esindavad erinevaid olemeid, jagavad sama atribuudi väärtust, mis oleks pidanud olema atribuudi piires unikaalne.

Näide: Kahel isikul on atribuudi „ISIKUKOOD“ väärtus sama.

2.3.2.10 *Sünonüümide kasutus*

Probleemi kirjeldus: Sama tähendusega, kuid eri esitusega väärtuste suvaline kasutus sama tähenduse edastamiseks kas ühe või rohkema andmeallika piires.

Näide: Atribuut "AMET" sisaldab väärtusi "tarkvaraarendaja" ja "programmeerija," mis tähistavad antud kontekstis sama ametit.

2.3.2.11 *Pooltühi kirje*

Probleemi kirjeldus: Antud probleemi puhul on mitmed kirje atribuudid väärtustamata. Kui ületatakse (andmeomaniku poolt) määratletud lävi on tegu pooltühja kirjega.

Näide: Pooltühja kirjega on tegu kui väärtustamata on 60% kirje väärtustest. Näiteks kirje KLIENT(NIMI="Toomas Kask", AMET="", EMAIL="") puhul on atribuudid „AMET“ ja „EMAIL“ väärtustamata ehk 66% kirje väljadest on tühjad.

2.3.2.12 *Funktsionaalse sõltuvuse rikkumine*

Probleemi kirjeldus: Üks atribuut sõltub funktsionaalselt teise atribuudi väärtusest, kuid eksisteeriv seos ei vasta tegelikkusele.

Näide: Veergude "POSTIINDEKS" ja "LINN" vahel on funktsionaalne sõltuvus, sest iga postiindeks on seotud vaid ühe linnaga ja postiindeksi väärtusest tuleneb linna väärtus. Näiteks rikub funktsionaalset sõltuvust järgnev olukord: (POSTIINDEKS=10118; LINN="TALLINN") ja (POSTIINDEKS=10118 ja LINN="TARTU").

2.3.2.13 *Ligikaudselt dubleerivad kirjed*

Probleemi kirjeldus: Sama olem on esindatud võrdselt või samaväärselt kahes või enamas eri andmekogudest pärinevates kirjetes.

Näide: Kirje KLIENT(10, "Toomas Kask", "Tartu maantee", 123, 502899106) ning kirje KLIENT(72, "T. Kask", "Tartu mnt", 123, 502899106) on ligikaudselt dubleerivad kirjed.

2.3.2.14 *Vastuoluliselt dubleerivad kirjed*

Probleemi kirjeldus: Sama olemit esindavate kirjete ühe või enama atribuudi väärtustes on ebakõlad ja vastuolud.

Näide: Kirje KLIENT(10, "Toomas Kask", "Tartu maantee", 123, 502899106) ning kirje KLIENT(72, "Toomas Kask", "Liivalaia", 12, 502899106) on vastuoluliselt dubleerivad kirjed.

2.3.2.15 *Viiteterviklus*

Probleemi kirjeldus: Kirje välisvõtit hoiustava atribuudi väärtusele puudub vaste seotud relatsiooni primaarvõtmete hulgas.

Näide: Tabelis "KLIENT" oleva atribuudi "KLIENDI_POSTIINDEKS" väärtust on 5100, kuid nimetatud väärtust tabelis "POSTIINDEKS" ei eksisteeri.

2.3.2.16 *Vale viide*

Probleemi kirjeldus: Kirje välisvõtit hoiustava atribuudi väärtusele leidub vaste seotud relatsiooni primaarvõtmete hulgas, kuid väärtus ei vasta tegelikule olukorrale. Viide on kas vale või aegunud.

Näide: Tabelis "KLIENT" oleva atribuudi "KLIENDI_POSTIINDEKS" väärtust on 4415, kuid õige väärtus oleks 4445. Mõlemad postiindeksid eksisteerivad tabelis "POSTIINDEKS".

2.3.2.17 *Süntaksite mitmekesisus*

Probleemi kirjeldus: Eri relatsioonides on sama tüüpi atribuutide väärtuste esitamiseks kasutatud erinevaid süntakseid.

Näide: Tabeli "TELLIMUS" atribuudi "TELLIMUSE_KUUPAEV" puhul kasutatakse kuupäeva talletamiseks süntaksit dd/mm/yyyy. Tabelis "ARVE" atribuudi "ARVE_KUUPAEV" puhul kasutatakse kuupäeva talletamiseks süntaksit yyyy/mm/dd.

2.3.2.18 *Ringlus kirjete seas*

Probleemi kirjeldus: Tekkinud on tsüklilised seosed kahe (otsene ringlus) või enama (kaudne ringlus) kirje vahel.

Näide: Isik X on isaks isikule Y ja isik Y on samal ajal isaks isikule X.

2.3.2.19 *Mõõtmisühikute mitmekesisus*

Probleemi kirjeldus: Eri andmehulkades kasutatakse sama infot talletavate atribuutide puhul erinevaid mõõtühikuid.

Näide: Andmestikus X on atribuut "MAKSUMUS" dollarites ning andmestikus Y eurodes.

2.3.2.20 *Esitluse mitmekesisus*

Probleemi kirjeldus: Erinevad andmestikud kasutavad sama infot talletavate atribuutide jaoks erinevaid väärtuseid sama informatsiooni esitamiseks.

Näide: Andmestikus X kasutatakse soo tähisena väärtuseid "M" ja "N". Andmestikus Y kasutatakse soo tähisena väärtuseid 1 ja 0.

2.3.2.21 *Homonüümide kasutus*

Probleemi kirjeldus: Eri andmestike sama infot talletavates atribuutides on kasutatud süntaktiliselt võrdseid, kui erineva (semantilise) tähendusega väärtuseid.

Näide: Andmestikus X kasutatakse "TOODE" tabelis terminit "Hiir" tähistamaks arvuti riistvaralist osutusseadet. Andmestikus Y kasutatakse "TOODE" tabelis terminit "Hiir" tähistamaks kodulooma.

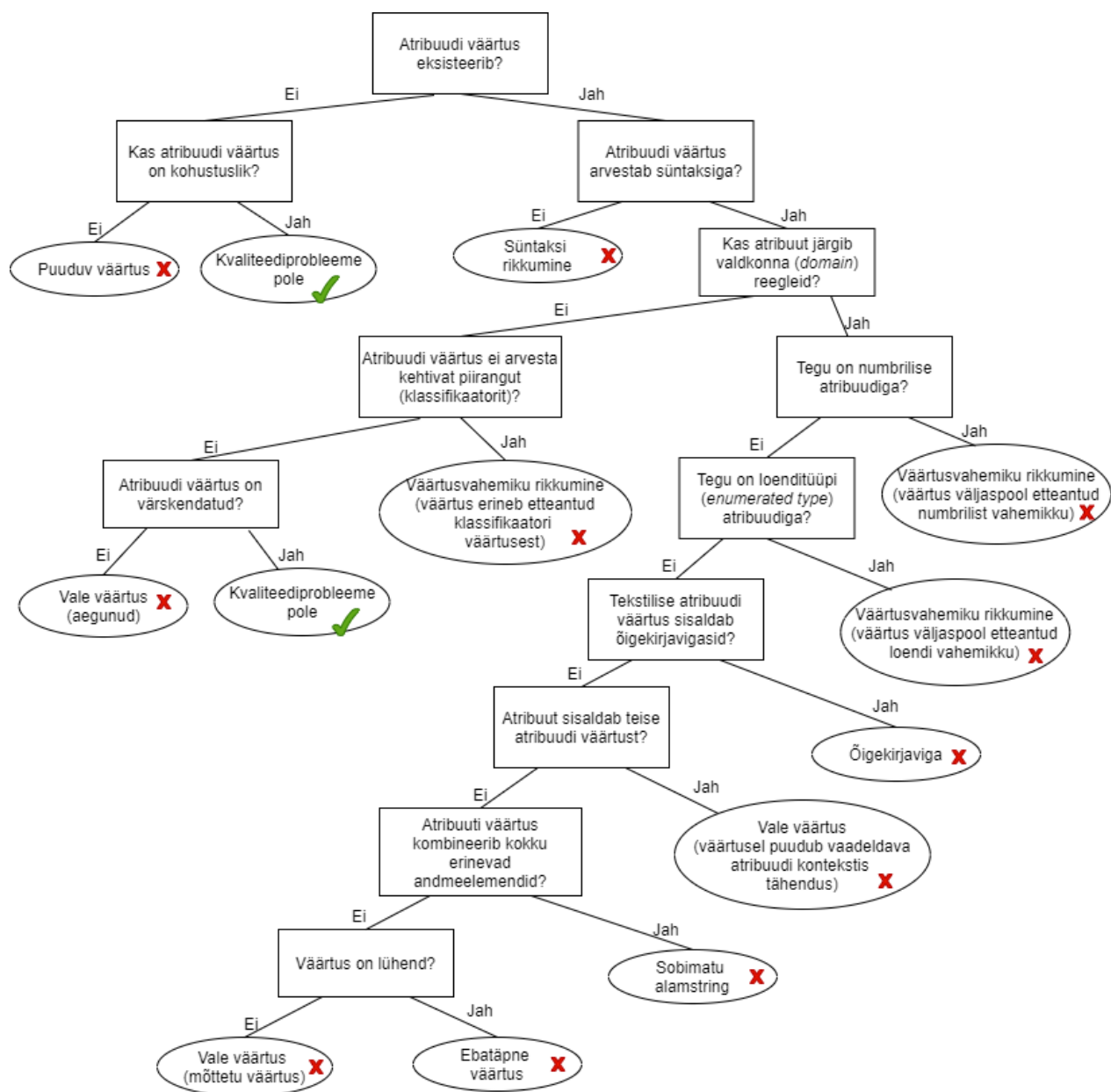
2.3.3 Kvaliteediprobleemide tuvastamine

Siin seksioonis esitame Oliveira jt poolt väljatöötatud meetodid¹⁵ andmekvaliteedi probleemide tuvastamiseks (identifitseerimine ja klassifitseerimine). Iga meetod on esitatud binaarse otsustuspuuna, mis võimaldab identifitseerida ja klassifitseerida probleeme andmeallikate hulga, andmeobjektide vaheliste relatsioonide hulga ja atribuutide/kirjete suhtes. Kvaliteediprobleemid on märgitud otsustuspuude lõppsõlmedes (tähiseks punane X). Lõppsõlme väärtus „Kvaliteediprobleeme pole“ tähistab olukorda, kus kvaliteediprobleeme ei tuvastatud.

Meetod andmekvaliteedi probleemide tuvastamiseks üksiku kirje ühe atribuudi väärtuses on esitatud alljärgnevalt (

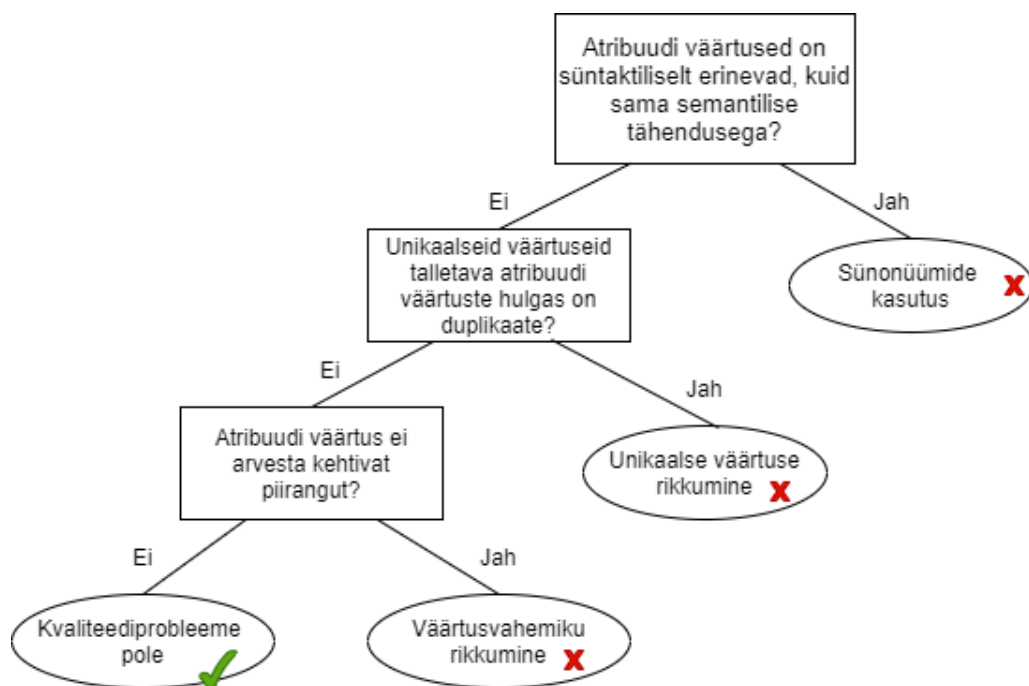
Joonis 6). Antud meetod tuvastab kvaliteediprobleeme ühes andmehulgas. Näiteks andmetabeli puhul saab antud meetodiga tuvastada ühe tabeli rea (üksiku kirje) ühe välja (ühe atribuudi väärtuse) kvaliteediprobleemid.

¹⁵ Oliveira Paulo, Fátima Rodrigues, Pedro Henriques, and Helena Galhardas. "A taxonomy of data quality problems." In 2nd Int. Workshop on Data and Information Quality, pp. 219-233. 2005.
https://s3.amazonaws.com/academia.edu.documents/43292389/A_Taxonomy_of_Data_Quality_Problems20160302-6272-51xfc8.pdf



Joonis 6. Üksiku andmehulga üksiku kirje atribuudi väärtus.

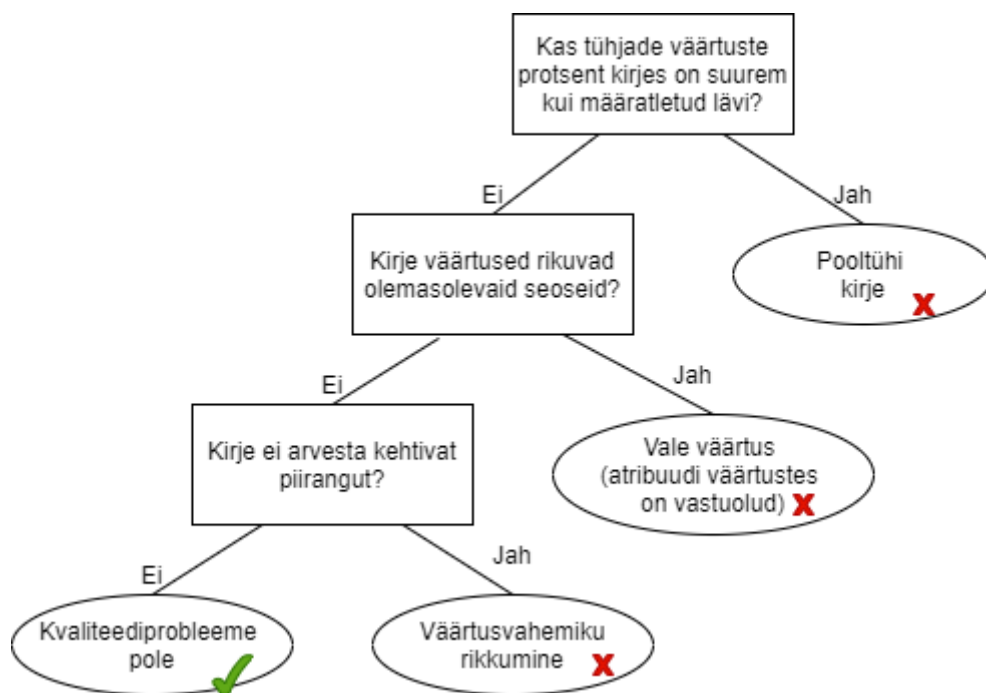
Järgnevalt on esitatud meetod atribuudi väärtustes esinevate kvaliteediprobleemide tuvastamiseks (Joonis 7). Antud meetod tuvastab probleeme ühes andmehulgas. Näiteks andmetabeli puhul saab antud meetodiga tuvastada ühe veeru (atribuudi) väärtustes esinevad andmekvaliteedi probleemid.



Joonis 7. Üksiku andmehulga atribuudi väärtus.

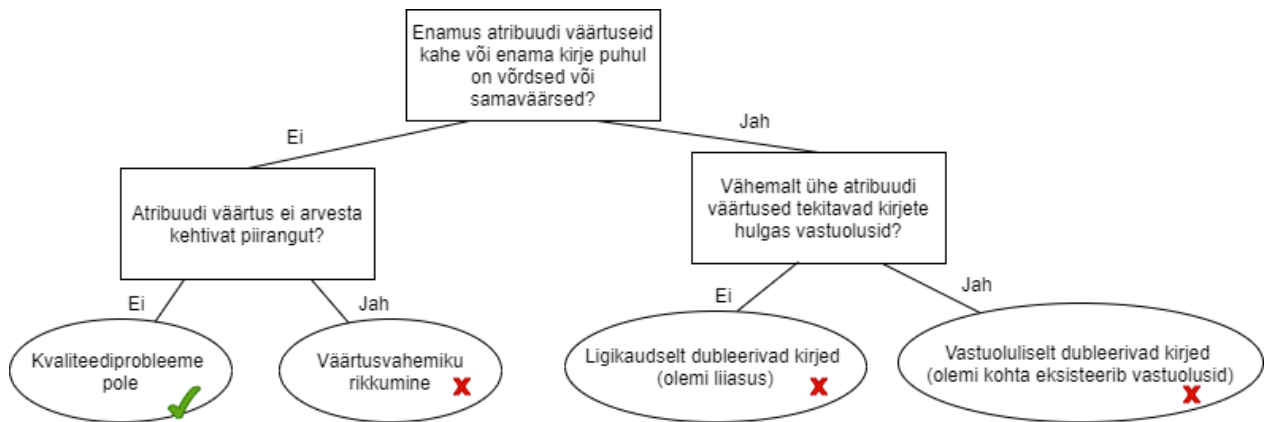
Järgnev meetod aitab tuvastada üksiku kirje atribuutide väärtuste kogumis esinevad andmekvaliteedi probleemid (

Joonis 8). Antud meetod tuvastab probleeme ühes andmehulgas. Näiteks andmetabeli puhul saab antud meetodiga tuvastada ühe tabeli rea (kirje) kõigi väljade väärtuseid (atribuudi väärtuste kogumit) puudutavad kvaliteediprobleemid.



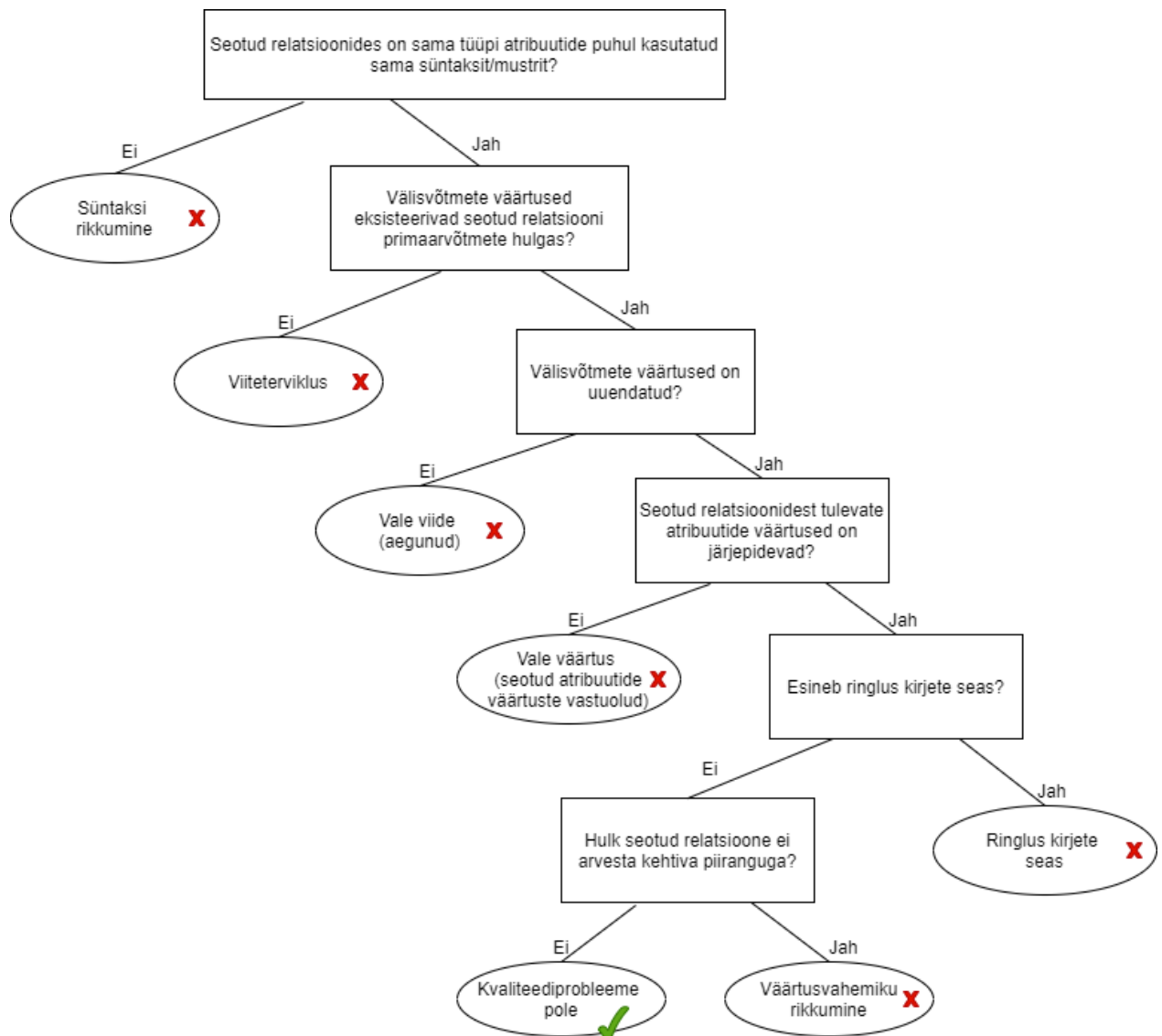
Joonis 8. Üksiku andmehulga üksiku kirje atribuutide väärtused.

Meetod andmekvaliteedi probleemide tuvastamiseks mitme kirje atribuudi väärtustes on toodud alljärgnevalt (Joonis 9). Antud meetod tuvastab kvaliteediprobleeme ühes andmehulgas. Näiteks andmetabeli puhul saab antud meetodiga tuvastada mitmete ridade (kirjete) väljade väärtustes esinevad andmekvaliteedi probleemid.



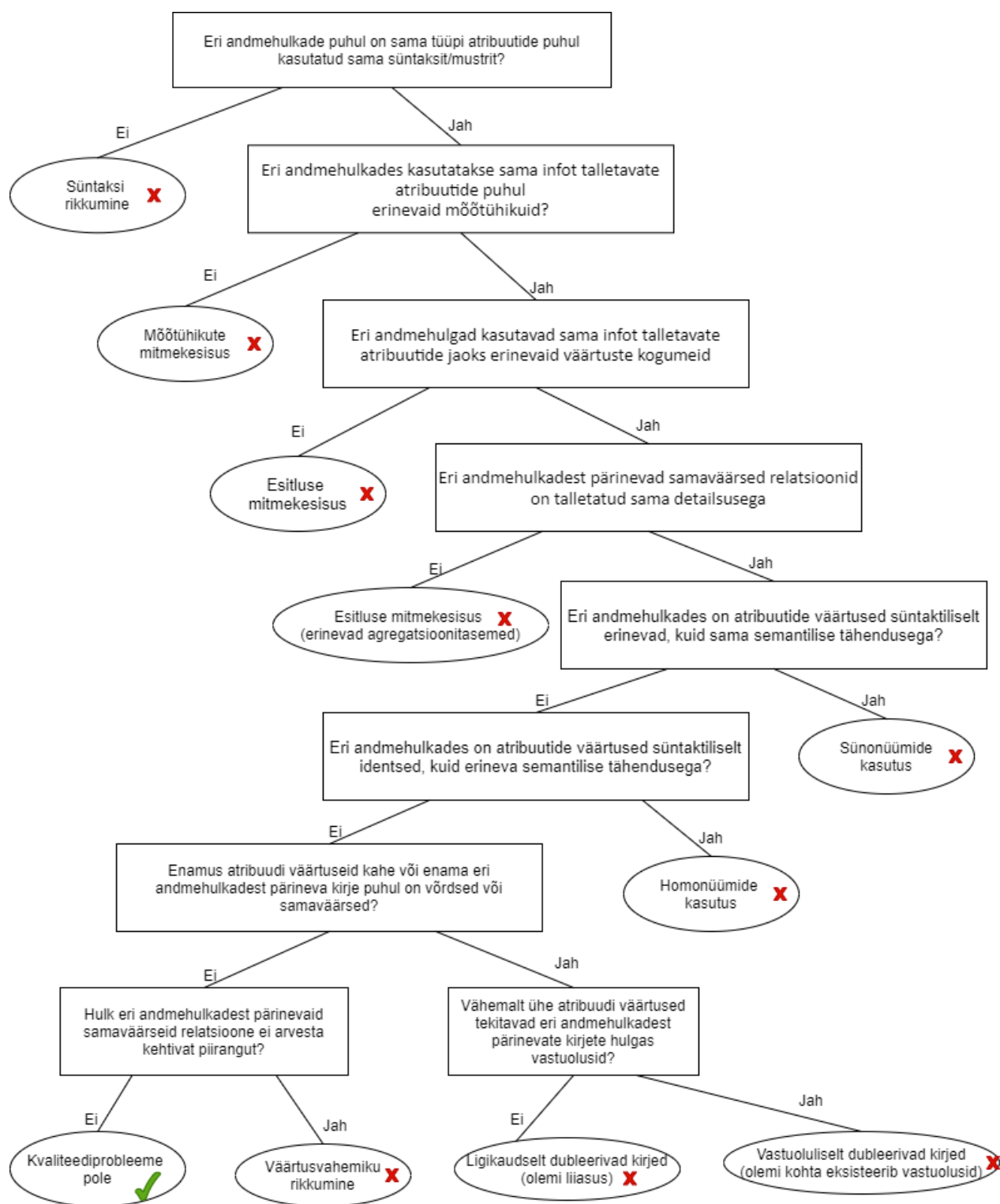
Joonis 9. Üksiku andmehulga mitme kirje atribuudi väärtused.

Järgnev meetod aitab tuvastada tabelite vahelisi seoseid puudutavad andmekvaliteedi probleemid (Joonis 10). Näiteks tabeli X ja tabeli Y vahelistes seostes eksisteerivad probleemid.



Joonis 10. Relatsioonide vahelised seosed.

Järgnevalt esitatud meetod aitab tuvastada mitut andmehulka puudutavad andmekvaliteedi probleemid, kui mitut andmehulka vaadeldakse kui ühte tervikut (Joonis 11). Näiteks kui organisatsioon talletab andmeid andmehulkades X, Y ja Z aitab antud meetod tuvastada andmehulki puudutavad andmekvaliteedi probleemid.



Joonis 11. Mitu andmehulka.

2.3.4 Dimensioonide ja indikaatorite mõõtmine

Andmekvaliteedi mõõtmisega alustamisel tuleb valida milliseid dimensioone esmajärjekorras hinnata. Valiku tegemisel on kasulik lähtuda allpool toodud kahest põhimõttest.

- ▶ Dimensiooni tuleks hinnata vaid juhul, kui selle hindamine loob kvaliteedi parandamiseks kasutatavat informatsiooni ning see on seotud äriliste vajadustega.
- ▶ Dimensiooni tuleks hinnata vaid juhul, kui selle hindamine on võimalik ja praktiline. Mõnikord pole võimalik andmeid konkreetse dimensiooni kontekstis hinnata või on hindamiseks tehtavad kulutused ebamõistlikult suured.

Selgitamaks kvaliteedidimensioonide mõõtmist ning nende jaotust mõõdetavateks indikaatoriteks on alljärgnevalt esitatud dimensioonide mõõtmise täpsemad kirjeldused ning indikaatorite arvutamist illustreerivad näited.

2.3.4.1 Õigsus (Accuracy)

Õigsus näitab mil määral vastavad andmed tegelikkusele ning see jagatakse süntaktiliseks ja semantiliseks õigsuseks.

Süntaktilist õigsust saab defineerida kui kaugust väärtuse v ja domeeni väärtusi sisaldava hulga D elemendi vahel. Näiteks kui tegelikkuses on inimese nimi $v' = \text{Tõnu}$ ning andmeväärtus $v = \text{Toomas}$ siis on tegu süntaktiliselt korrektsete andmetega, sest nimi Toomas kuulub korrektsete nimede hulka ehk hulka D . Süntaktilist õigsust mõõdetakse spetsiaalsete funktsioonide abil, mida nimetatakse võrdlusfunktsioonideks (*comparison functions*). Võrdlusfunktsioonid hindavad kaugust hinnatava väärtuse v ning domeeni väärtusi sisaldava hulga D väärtuste vahel. Lihtne näide võrdlusfunktsiooni illustreerimiseks on väärtuste kauguste hindamise funktsioon (*edit distance function*), mis hindab minimaalset märkide sisestamiste, kustutamiste ja asendamiste arvu, et teisendada väärtus s väärtuseks s' . Leidub ka keerukamaid võrdlusfunktsioone, näiteks terminite sarnast kõla arvestavad võrdlusfunktsioonid.

Semantilist õigsust saab defineerida kui kaugust andmeväärtuse v ja tegelikkust tähistava väärtuse v' vahel. Näiteks kui tegelikkuses on inimese nimi $v' = \text{Tõnu}$ siis sellisel juhul on andmeväärtus $v = \text{Toomas}$ vigane. Semantilist õigsust saab kontrollida võrreldes eri allikatest (näiteks eri andmebaasidest) pärinevaid samu andmeid ning seeläbi tuvastada õiged väärtused. Õigsuse kontrollimiseks peab teada olema õige väärtus v või peab olema võimalik muu teadaoleva informatsiooni põhjal tuletada, kas väärtus v on õige või mitte. Seega on semantilist õigsust keerulisem mõõta, kui süntaktilist õigsust. Kui on teada, et vigade hulk on madal ning vigade põhjuseks on üldiselt kirjavead, võib süntaktiline õigsus kokku langeda semantilisega. Sellisel juhul on semantilist õigsust võimalik saavutada asendades süntaktiliselt väärad väärtused lähima väärtusega domeenist D .

Eelnevast tulenevalt oleme õigsuse dimensioon jaganud mõõdetavateks indikaatoriteks järgmiselt:

- Süntaktiliselt õigete kirjade määr;
- Autentsete kirjade määr;
- Kokkulepitud kirjaviisi järgivate kirjade määr.

Süntaktiliselt õigete kirjade määr näitab tunnuste ehk muutujate väärtuste tehnilist korrektsust. Näiteks kirjavigade puudumist, ühtset suure ja väikese algustähe kasutust, ühesugust asutuste nimede kasutust jne.

Autentsete kirjade määr näitab kas andmed ehk tunnused ehk muutujad on nimetatud nende kokkulepitud või tegelike nimedega. Näiteks eesnime, perekonnaseisu, puude või vanuse õigsus (näiteks vanus väärtus 220 pole realistlik inimeste korral). Tegu on osaga semantilisest õigsusest.

Kokkulepitud kirjaviisi järgivate kirjade määr näitab kuivõrd on kirjetes kasutatud ühtset ja korrektset kirjaviisi. Näiteks võib selle indikaatori hindamisel kontrollida kuidas on transliteeritud võõrkeelsed nimed ja nimetused (näiteks “Ельцин” transliteeritud kuju on “El’cin”) või kas lühendeid on kasutatud läbivalt kõigis kirjetes. Tegu on osaga semantilise õigsusest.

2.3.4.2 Täielikkus (Completeness)

Täielikkus näitab, mil määral on olemas kõik nõutud andmed. Täielikkust on võimalik täpsemalt kirjeldada toetudes ühele konkreetsele andmemudelile, mistõttu on järgnev ülevaade täielikkusest toodud relatsioonilise andmemudeli kontekstis.

Relatsioonilises andmemudelil saab täielikkust mõõta kahel viisil: tühje väärtusi (*NULL values*) arvestades või mitte arvestades. Tühje väärtusi mitte arvestades on täielikkust võimalik arvutada leides olemasolevate andmete osakaal kõigist nõutud andmetest. Oletame näiteks, et Eestis on 1 300 000 elanikku. Kui meie andmetabelis on andmed 1 000 000 elaniku kohta siis on meie andmete täielikkus 0,8 ($1\,000\,000/1\,300\,000=0,8$).

Võttes arvesse tühje väärtusi saab täielikkuse jagada kolmeks osaks:

- Andmeobjekti täielikkus näitab tühjade väärtuste esinemist võttes arvesse kõiki andmeobjekti väljasid (rea veerge). Andmeobjekti täielikkuse arvutamiseks jagatakse väärtustatud väljade arv väljade koguarvuga. Näiteks (Tabel 2) töötajal id-ga 2 on andmeobjekti täielikkus 1 (4/4), sest kõik andmeobjekti (rea) väljad on väärtustatud. Töötajal id-ga 1 on andmeobjekti täielikkus 0,75 (3/4) ning töötajal id-ga 3 on andmeobjekti täielikkus 0.5 (2/4).
- Atribuudi täielikkus näitab tühjade väärtuste esinemist relatsiooni ühe atribuudi (tabeli veeru) ulatuses. Atribuudi täielikkuse arvutamiseks tuleb väärtustatud atribuutide arv jagada atribuutide koguarvuga. Näiteks (Tabel 2) atribuudi „Vanus“ täielikkus on 0.7 (2/3).
- Relatsiooni ehk olemihulga täielikkus näitab tühjade väärtuste esinemist kogu relatsiooni (tabeli) ulatuses. Relatsiooni täielikkuse arvutamiseks tuleb väärtustatud relatsiooni väärtuste arv jagada relatsiooni väärtuste koguarvuga. Näiteks (Tabel 2) näidisrelatsiooni täielikkus on 9/12.

Tabel 2. Täielikkuse näidistabel

Töötaja_ID	Eesnimi	Perekonnanimi	Vanus
1	Tõnu	NULL	54
2	Toomas	Mets	28
3	John	NULL	NULL

Eelnevast tulenevalt oleme Täielikkuse dimensiooni jaganud mõõdetavateks indikaatoriteks järgmiselt:

- Andmeobjekti täielikkuse määr;
- Atribuudi täielikkuse määr;
- Olemihulga täielikkuse määr;
- Olemasolevate atribuutide määr nõutavatest atribuutidest.

Andmeobjekti täielikkuse määr näitab ühe andmeobjekti (näiteks tabeli rea) täielikkust ehk tühjade väärtuste esinemist võttes arvesse kõiki andmeobjekti atribuute. Näiteks kui andmetabelis on rea (andmeobjekti) kaks veergu (atribuuti) kolmest väärtustatud on andmeobjekti täielikkus 2/3 ehk 66%.

Atribuudi täielikkuse määr näitab atribuudi (näiteks tabeli ühe veeru) täielikkust ehk tühjade väärtuste esinemist ühe atribuudi ulatuses. Näiteks kui tabelis on neli rida (andmeobjekti) ning veerg (atribuut) X on väärtustatud kahel real neljast on atribuudi täielikkus $2/4$ ehk 50%.

Olemihulga täielikkuse määr näitab kogu olemihulga (näiteks tabeli) täielikkust ehk tühjade väärtuste esinemist kogu olemihulga (tabeli) ulatuses. Näiteks oletame, et tabelis on 3 veergu ja 4 rida ehk kokku 12 väärtust. Neist 12 väärtusest 9 on väärtustatud. Sellisel juhul on tabeli täielikkus $9/12$ ehk 75%.

Olemasolevate atribuutide määr nõutavatest atribuutidest näitab, kas kõigil kirjetel/andmeobjektidel eksisteerib konkreetne tunnus või see puudub. Näiteks kui andmetabeli puhul on olemas 3 nõutud veerust (atribuudist) 2 siis on antud indikaatori väärtuseks $2/3$ ehk 66%.

2.3.4.3 Ajakohasus (*Timeliness*)

Ajakohasus näitab, mil määral andmete värskus ja kättesaadavus vastab vajadustele ja nõuetele.

Kuna ajakohased andmed peavad olema nii värsked, kui ka neid kasutava sündmuse jaoks õigeaegselt kättesaadavad, koosneb ajakohasuse mõõtmine kahest osast. Esimeseks sammuks on andmete värskuse hindamine. Teise sammuna tuleb kontrollida andmete kättesaadavust ehk seda, kas andmed on kättesaadavad enne planeeritud kasutamise aega. Üks võimalus ajakohasuse täpsemaks mõõtmiseks on eristada andmete värskust (*currency*), volatiilsust (*volatility*) ning ajakohasust (*timeliness*). Andmete värskust on defineeritud järgmiselt (Võrrand 1).

Võrrand 1. Värskus (*Currency*)

$$\text{Värskus} = \text{Vanus} + (\text{EdastamiseAeg} - \text{SisestusAeg})$$

Siinkohal tähistab *Vanus* andmete vanust nende vastuvõtmise hetkel. *EdastamiseAeg* tähistab ajahetke, kui andmed jõuavad kasutajani ning *SisestusAeg* andmete sisestamise ajahetke. Ehk teisisõnu näitab muutujate *EdastamiseAeg* ja *SisestusAeg* vahe kui kaua on andmed olnud infosüsteemis.

Volatiilsus on defineeritud kui andmete kehtivuse periood. Sellest tulenevalt saame ajakohasuse defineerida järgmiselt (Võrrand 2).

Võrrand 2. Ajakohasus (*Timeliness*)

$$\max\left\{0, 1 - \frac{\text{Värskus}}{\text{Volatiilsus}}\right\}$$

Ajakohasuse väärtus on vahemikus nullist üheni, kus null näitab ajakohasuse puudumist ning üks ideaalset ajakohasust. Siinkohal on oluline märkida, et andmete värskuse olulisus sõltub volatiilsusest, sest väga volatiilsed andmed peavad olema värsked samas kui madala volatiilsusega andmete puhul pole värskus niivõrd oluline.

Tulenevalt eelnevast oleme ajakohasuse dimensiooni indikaatori sõnastanud järgmiselt:

- ▶ Ajakohasuse määr skaalal 0-1 vastavalt ajakohasuse definitsioonile (Võrrand 2).

Näitena vaatleme digiretsepti ajakohasust. Oletame, et arsti vastuvõtul vaadatakse üle 19.03.2020 kell 08:00 toimunud uuringu tulemused. Tulemuste põhjal määratakse patsiendile retseptiravim, mille määramine leiab aset 19.03.2020 kell 13:00. Seega on andmete vanus sel hetkel 5 tundi ehk *Vanus*=5. Arst märgib retsepti andmed süsteemi ning loob digiretsepti 19.03.2020 kell 14:00. Proviisiori poolt kasutatavasse süsteemi jõuavad digiretsepti andmed 19.03.2020 kell 15:00. Hetkeks kui andmed jõuavad proviisorini on need süsteemis olnud ühe tunni (*EdastamiseAeg* – *SisestusAeg* = 15:00 – 14:00

ehk 1 tund). Sellest tulenevalt saame värskuse definitsiooni põhjal (Võrrand 1) öelda, et andmete värskus on 6 tundi ($Värskus = 5 + 1 = 6$). Digiretsepti kehtivus on 60 päeva ehk 1440 tundi. Järelikult on andmete ajakohasus vastavalt ajakohasuse definitsioonile (Võrrand 2) $\max(0, 1 - (5/1440)) = 1 - 0.003 = 0.997$. Väärtus 0.997 näitab, et digiretsepti ajakohasus on peaaegu täiuslik. Negatiivse stsenaariumi puhul jõuavad andmed süsteemi vahetult enne või pärast kehtivuse lõppu ning sellisel juhul läheneks andmete ajakohasuse väärtus nullile.

2.3.4.4 Reeglipärasus (Orderliness)

Reeglipärasus näitab, mil määral andmete formaat ja struktuur on esitatud süsteemselt ja korrapäraselt. Andmete formaadi osas kontrollitakse esmalt kas andmete esitamisel kasutatakse sobivaid andmetüüpe, seejärel kas andmed on esitatud konkreetsele andmetüübile vastavas vormingus ning lõpuks kas andmed vastavad standardite ja rakenduste poolt seatud nõuetele. Üheks tihtiesinevaks nõudeks on joondus valitud referentssüsteemiga, olgu selleks siis loendid, klassifikaatorid või andmete väärtus mõnes teises andmekogumikus.

Reeglipärasuse dimensiooni oleme indikaatoriteks jaganud järgmiselt:

- ▶ Kokkulepitud klassifikaatorite kasutamisest kõrvalekallete määr;
- ▶ Kokkulepitud andmemustritest kõrvalekallete määr.

Kokkulepitud klassifikaatorite kasutamisest kõrvalekallete määra abil saab näiteks hinnata, kuivõrd aadressiandmed ja majanduslik tegevusala on talletatud vastavalt kokkulepitud klassifikaatorile.

Kokkulepitud andmemustritest kõrvalekallete määr on seotud andmemustritega nagu kuupäev (*date*), arv (*integer*) või loogiline jah/ei (*boolean*) väli. Näiteks saab kokkulepitud andmemustritest kõrvalekallet hinnata isikukoodi või kuupäeva puhul.

2.3.4.5 Ühekordsus (Uniqueness)

Ühekordsus näitab, mil määral esineb duplikaatkirjeid. Teisisõnu on andmed ühekordselt talletatud siis, kui igale unikaalsele kirjele vastab üks võtmeväärtus mistõttu tuleb ühekordsuse mõõtmisel keskenduda võtmeväärtustele ning nende seostele. Ühekordsuse probleemi ei tohiks esineda, kui kasutatakse relatsioonilist andmebaasi struktuuri ning määratud on korrektsed primaarvõtmed. Seejuures on aga oluline, et primaarvõtme määramise protseduur oleks usaldusväärne. Duplikaatkirjete probleem on suurem teiste andmestruktuuride puhul, kus unikaalsete võtmete määramine pole võimalik (näiteks MS Exceli tabelis). Näitlikustamaks ühekordsuse probleemi tingitud lisakulu võime vaadelda olukorda, kus ühe kliendi kohta on andmetes talletatud mitu kirjet. Saates nende andmete põhjal klientidele emaili saavad dubleeritud kirjetega kliendid mitu kirja, mis mõjutab nii saatmise kulu kui ka asutuse reputatsiooni kliendi silmis.

Eelnevast tulenevalt oleme ühekordsuse indikaatori määranud järgmiselt:

- ▶ Duplikaatkirjete määr.

Duplikaatkirjete määr näitab duplikaatkirjete osakaalu kirjete koguarvus ning on oluline näiteks aadresside, kontaktisikute ja klassifikaatorite talletamisel.

2.3.5 Andmekvaliteedi reeglile mõõdiku seadmine

Järgnevalt on esitatud kvaliteedidimensioonide ja indikaatorite kasutamist illustreerivad näidisjuhtumid. Iga indikaatori kohta on võimalik kirjeldada hulk andmekvaliteedi reegleid. Näidisjuhtumites on kirjeldatud iga indikaatori kohta üks näidisreegel, kirjeldatud reegli hindamiseks vajalikud meetmed, illustreeritud hetketaseme hindamist ning toodud näidis sihttasemest.

Tabel 3. Andmekvaliteedi reeglile mõõdiku seadmine – Õigsus.

Dimensioon	Indikaator	Reegel	Meede	Hetketase	Sihttase
Õigsus	Süntaktiliselt õiged kirjed	Isiku nimi ei tohi sisaldada numbreid.	Loendada kokku veergude arv, kus isiku nimedes on numbreid ning kirjete koguarv. Numbreid sisaldavate nimede arv: 100 Isikute koguarv: 5 000	$(5000-100)/5000 * 100 = 98\%$ 98% isikute nimedest ei sisalda numbreid.	100%
	Autentsed kirjed	Isiku elukoht peab vastama Rahvastikuregistri andmetele.	Teostada isikute väljavõte ning elukohtade väljavõte Rahvastikuregistrist ning võrrelda seda hinnatavate andmetega. Rahvastikuregistriga mitte kattuvate elukohaandmete arv: 10 Kirjete koguarv: 1000	$(1000-10)/1000 * 100 = 99\%$ 99% elukohaandmetest vastab Rahvastikuregistri andmetele.	100%
	Kokkulepitud kirjaviisiga kirjed	Venekeelsed isikute nimed peavad olema talletatud järgides korrektset vene-eesti transkriptsiooni.	Võrrelda vene keele tähestikus kirjutatud nimesid (Зернов) ning transkriptsioonide tulemusi (Zernov). Ebakorrektsete transkriptsioonide arv: 450 Venekeelsete nimede koguarv: 500	$(500-450)/500 * 100 = 10\%$ 10% teostatud transkriptsioonidest järgib korrektset vene-eesti transkriptsiooni.	100%

Tabel 4. Andmekvaliteedi reeglile mõõdiku seadmine - Täielikkus.

Dimensioon	Indikaator	Reegel	Meede	Hetketase	Sihttase
Täielikkus	Andmeobjekti täielikkus	Riigi peaprokuröri kohta peavad olema täidetud kõik isikuandmed.	Loendada kokku riigi peaprokuröri ametiga seotud isiku kohta andmeid talletavate väljade arv ning väärtustatud väljade arv. Väärtustatud väljade arv: 4 Andmeid talletavate väljade arv: 5	$4/5 * 100 = 80\%$ 80% konkreetse isiku kohta käivatest andmetest on väärtustatud.	100%
	Atribuudi täielikkus	Kõikide isikute kohta peab olema teada nende telefoninumber.	Loendada kokku isikute arv, kus tunnus „Telefoninumber“ on tühi ning isikute koguarv. Telefoninumbrita isikute arv: 30 Isikute koguarv: 150	$30/150 * 100 = 20\%$ 80% isikute kohta on teada telefoninumber.	100%
	Olemihulga täielikkus	Kõik isiku kohta käivad andmed peavad olema väärtustatud.	Loendada kokku isiku tabeli väärtustatud väljade arv ning väljade koguarv. Isiku tabeli väärtustatud väljade arv: 600 Isiku tabeli väljade koguarv: 1000	$600/1000 * 100 = 60\%$ 60% isiku kohta käivatest andmeväljadest on väärtustatud.	>80%
	Olemasolevad atribuudid nõutavatest atribuutidest	Isiku kohta peab olema võimalik talletada eesnimi, perekonnanime, isikukoodi ja rahvuse andmeid.	Leida olemasolevate atribuutide arv ning võrrelda seda nõutavate atribuutide arvuga. Olemasolevate atribuutide arv (eesnimi, perekonnanime, isikukood): 3 Nõutavate atribuutide arv: 4	$3/4 * 100 = 75\%$ 75% nõutud atribuutidest on olemas.	100%

Tabel 5. Andmekvaliteedi reeglile mõõdiku seadmine – Ajakohasus.

Dimensioon	Indikaator	Reegel	Meede	Hetketase	Sihttase
Ajakohasus	Ajakohasuse määr	Riigieksamite tulemused peavad olema kasutatavad enne sisseastumisperioodi lõppu.	Arvutada ajakohasuse väärtus vastavalt ajakohasuse definitsioonile (Võrrand 2).	Ajakohasuse väärtus: 0.9	>0.8

Tabel 6. Andmekvaliteedi reeglile mõõdiku seadmine - Reeglipärasus.

Dimensioon	Indikaator	Reegel	Meede	Hetketase	Sihttase
Reeglipärasus	Klassifikaatoritest kõrvalekalded	Majanduslik tegevusala peab olema talletatud vastavalt EMTAK-ile.	Loendada kokku juhtumite arv, kus tegevusala ei vasta EMTAK-i klassifikatsioonile. EMTAK-ist kõrvalekallete arv: 200 EMTAK koodidega kirjade arv: 1771	$200/1771 * 100 = 11\%$ 11% kirjade majanduslikest tegevusaladest ei vasta EMTAK-ile.	0%
	Andmemustritest kõrvalekalded	Isikukood peab olema 11 kohaline täisarv.	Loendada kokku väärtused, kus isikukood ei vasta kokkulepitud tingimustele ning isikukode sisaldavate väljade koguarv. Andmemustrile mittevastavate isikukoodide arv: 400 Isikukoodide koguarv: 700 000	$400/700\ 000 * 100 = 0.05\%$ 0.05% isikukoodidest ei vasta nõuetele	0%

Tabel 7. Andmekvaliteedi reeglile mõõdiku seadmine – Ühekordsus.

Dimensioon	Indikaator	Reegel	Meede	Hetketase	Sihttase
Ühekordsus	Duplikaatkirjed	Iga postiindeks peab olema unikaalne.	Loendada kokku duplikaatide arv ning postiindeksite koguarv. Duplikaatide arv: 10 000 Postiindeksite arv: 1 000 000	$10\ 000 / 1\ 000\ 000 * 100 = 1\%$ 1% postiindeksi kirjetest on duplikaadid.	0%

2.3.6 Andmekvaliteedi aruandepõhjade väljatöötamine

Andmekvaliteedi aruanded aitavad tuua välja andmete kvaliteediga seotud kitsaskohad ning kommunikeerida edasiminekuks, selleks, et tõenduspõhiselt juhtida andmekvaliteediga seotud andmehalduse protsesse.

Kuigi aruannete põhjad tulenevad paljuski sellest kuidas on asutuses andmehaldus korraldatud, on osad näitajad kasulikud laiemalt ja eri asutustes.

Näidisenä oleme esitanud need vaated andmete kvaliteedi seiramiseks ja andmekvaliteedi reeglitest ülevaate saamiseks, mis võiksid olla kasutuses enamikes asutustes. Näidis on juurdepääsetav siin: <https://datastudio.google.com/u/0/reporting/e34c9f7b-7c05-4870-a5b7-2cb1e6b6299c/page/FKO9>

Kindlasti ei pretendeeri vaadete komplekt täielikule andmekvaliteedi aruande vaadete komplektile. Aruanded annavad vastused järgmistele küsimustele:

- ▶ Kui suur osa hallatud andmeobjektidest vastab andmekvaliteedi nõuetele?
- ▶ Milline on konkreetse ärimõistega seotud andmete andmekvaliteedi hetkeseis dimensioonide lõikes?
- ▶ Kuidas on konkreetse ärimõistega seotud andmete andmekvaliteet muutunud ajas dimensioonide lõikes?
- ▶ Millised andmeelemendid ja mil määral on kaetud andmekvaliteedi reeglitega?
- ▶ Millised andmekvaliteedi reeglid on kirjeldatud?
- ▶ Milliste andmeobjektide kvaliteeti on vaja tõsta?

Loetletud küsimustele vastava aruande loomist toetavad juhised on toodud sektsioonis 4.4.

2.3.7 Kvaliteediprobleemide prioriseerimine

On tõenäoline, et eelnevalt kirjeldatud kvaliteedidimensioonide hindamise käigus tuvastati mitmeid erinevaid kvaliteediprobleeme. Esmalt on mõistlik põhjalikumalt tegeleda nende kvaliteediprobleemide analüüsimise ja lahendamiseks, mis on asutusele suurima mõjuga. Seetõttu tuleks kvaliteediprobleemide esmasel prioriseerimisel kasutada ärilise mõju hindamise tehnikaid. Esmaseks hindamiseks sobivad hästi lihtsamad tehnikad, näiteks võib koguda stsenaariumeid, mis kirjeldavad halva andmekvaliteedi mõju äriks. Teine lihtne tehnika ärilise mõju on hindamiseks on koostada loend konkreetseid andmeid kasutavatest asutustest ja protsessidest. Mida rohkem on andmeid kasutavaid asutusi ja protsesse seda olulisemad on konkreetsed andmed. Kasutada võib ka viis korda „Miks?“ küsimise tehnikat eesmärgiga jõuda kvaliteediprobleemi tegeliku mõjuni. Põhjalikum mõju hindamine toimub protsessi hilisemas faasis pärast andmekvaliteedi juurpõhjuste analüüsi.

2.4 Andmekvaliteedi juurpõhjuste analüüs

Juurpõhjuste analüüsi käigus uuritakse avastatud andmekvaliteedi probleeme eesmärgiga tuvastada nende algallikad. Tihti tegeletakse andmekvaliteedi probleemi avastamisel probleemi sümptomitega, kuid ei tuvastata ja lahendada probleemi juurpõhjust. Andmekvaliteedi juurpõhjuste analüüsi peamiseks eesmärgiks on tuvastada kvaliteediprobleemi tekkepõhjused ning määrata tegevused probleemi edaspidiseks ennetamiseks.

Juurpõhjuse analüüs võib vajalikuks osutuda pärast andmekvaliteedi mõõtmist, et selgitada välja tuvastatud kvaliteediprobleemi algallikas. Sellisel juhul peaks enne juurpõhjuste analüüsi olema lõpetatud vähemalt ühe kvaliteedidimensiooni mõõtmine. Juurpõhjuste analüüsi on mõistlik teostada vaid äriliselt olulistele andmekvaliteedi probleemidele.

Juurpõhjuste analüüs võib vajalikuks osutuda ka olukorras, kus konkreetne probleem mõjutab ootamatult asutuse toimimist. Tihti tekitavad kvaliteediprobleemid kriitilisi olukordi, mis vajavad kiiret lahendust. Näiteks võib kvaliteediprobleem põhjustada olukorra, kus pole võimalik osutada teenust, võtta vastu või edastada informatsiooni. Kui kiireloomuline probleem on lahendatud, tuleb tagada, et konkreetne probleem ei korduks ehk tuleb tuvastada probleemi juurpõhjus ning ka see lahendada. Ka võib juurpõhjuste tuvastamine vajalik olla olukorras, kus kvaliteediprobleem ei põhjustanud kriitilist olukorda ja kõik teavad probleemist ning suhtuvad selle lahendamiseks vajalikesse tegevustesse kui paratamatusesse. Ka selliste kvaliteediprobleemide puhul on juurpõhjuste analüüsist kasu, kuna see aitab probleemi algallika tuvastada, probleemi lahendada ning seeläbi vähendada jooksvatele andmeparandustele kuluva aja ja muude ressursside ebavajalikku kulutamist.

Juurpõhjuste tuvastamist on käsitletud informatsiooni elutsükli raames¹⁶, mis kajastab ka andmete elutsükli. Informatsiooni elutsükli mudelis on kõik ressursid, näiteks raha, inventar, inimressurss või informatsioon, hallatud kogu elutsükli vältel. Vaid hästi hallatud ressurssidest on võimalik saada maksimaalset kasu. Informatsiooni elutsükkel koosneb kuuest faasist:

- ▶ Planeerimine – andmeressursi vastuvõtmiseks valmistumine.
- ▶ Hankimine – andmeressursi omandamine.
- ▶ Talletamine ja jagamine – elektrooniliselt või muul viisil andmeressursi talletamine ja jagamine.
- ▶ Säilitamine – andmeressursi kasutatavuse tagamine.
- ▶ Rakendamine – andmeressursi kasutamine eesmärkide saavutamiseks.
- ▶ Kustutamine või arhiveerimine – andmeressursi kasutusest eemaldamine.

Tundes informatsiooni elutsükli on võimalik kindlaks teha, millises faasis viga tekib ning seeläbi ka kvaliteediprobleemi juurpõhjus tuvastada. Juurpõhjuste tuvastamiseks võib olla vajalik elutsükli detailne uurimine. Alljärgnevalt on kirjeldatud kolm täpsemat meetodit andmekvaliteedi juurpõhjuste analüüsimiseks. Sõltuvalt avastatud probleemide keerukusest ja kiireloomulisusest võib olla vajalik meetodite omavaheline kombineerimine või eelistatav hoopis ühe konkreetse meetodi valik. Näiteks võib kiireloomulise probleemi esmaseks analüüsiks sobida lihtne viis korda „Miks?“ meetod.

Kolm meetodit andmekvaliteedi juurpõhjuste analüüsimiseks on järgmised:

- ▶ Küsi 5 korda „Miks?“ (Five „Whys“)
- ▶ Jälita (Track and Trace)
- ▶ Põhjus ja tagajärg diagramm (Cause-and-Effect/ Fishbone Diagram)

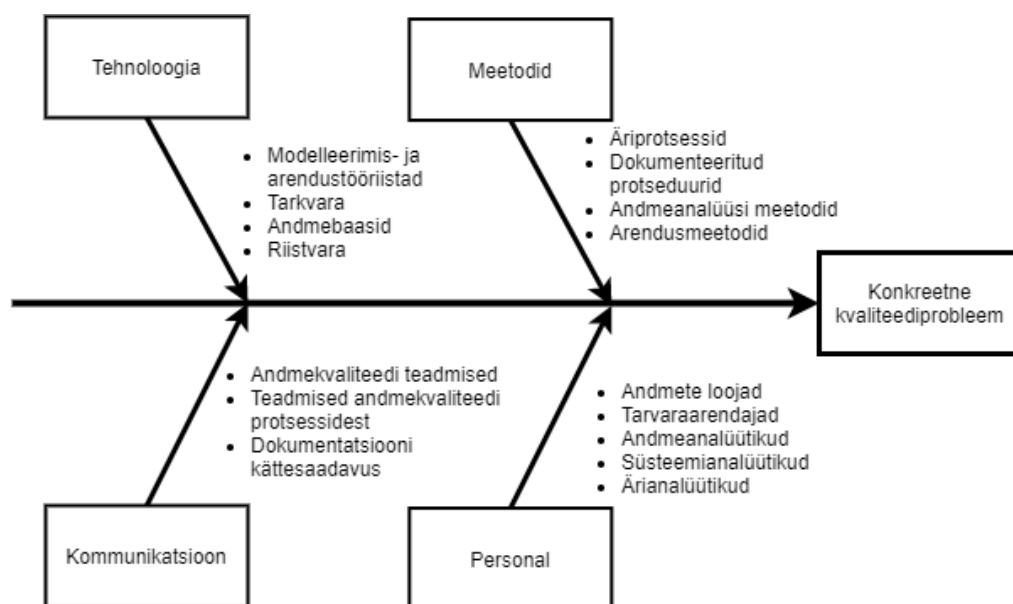
Esimene meetod on küsida 5 korda „Miks?“ (*Five „Whys“*). Meetodit kasutatakse laialdaselt tööstuses probleemide juurpõhjuste leidmiseks. Meetodit rakendades tuleb kõigepealt selgelt sõnastada tuvastatud andmekvaliteedi probleem. Mida selgemalt probleem on sõnastatud, seda lihtsam on leida tuvastada probleemi juurpõhjust. Järgmiseks tuleb viis korda küsida „Miks?“. Alustada tuleks sõnastatud kvaliteediprobleemist küsimusega „Miks selline tulemus saadi?“ või „Miks selline

¹⁶ McGilvray, Danette. Executing Data Quality Projects: Ten Steps to Quality Data and Trusted Information (Tm). 2008.

probleemne situatsiooni tekkis?“. Saadud vastuse peal tuleb küsimist korrata 5 korda. Edasi tuleb saadud tulemusi analüüsida. Kasulik on seejuures küsida: „Kas juurpõhjuseid on mitu?“ ja „Kas eri juurpõhjustel on ühiseid jooni?“. Järgmiseks tuleb anda konkreetseid soovitusi juurpõhjuse lahendamiseks. Vajadusel tuleks rakendada ka soovitude prioriseerimist. Viimaks tuleb meetodi rakendamise tulemid dokumenteerida. Dokumentatsioon peaks kindlasti kirjeldama juurpõhjused, soovitusi nende lahendamiseks, meetodi käigus läbitud järeldesteni jõudmise protsessi ning tuvastatud otseseid ja kaudseid ärilisi mõjusid.

Jälitamise (*Track and Trace*) meetodit saab kasutada tuvastamiseks probleemi esmakordset ilmnemist informatsiooni elutsüklis. Nagu ka eelnevalt kirjeldatud meetodi puhul on esimeseks sammuks tuvastatud kvaliteediprobleemi selge sõnastamine. Järgmiseks tuleb kindlaks määrata informatsiooni elutsükel ning informatsiooni jälitamise marsruut. Kui informatsiooni elutsükli on juba varasemalt kirjeldatud saab kasutada olemasolevat informatsiooni ning seda vastavalt vajadusele täiendada. Probleemi esmakordse esinemise tuvastamiseks tuleb järgnevalt võrrelda iga elutsükli sammu puhul sisendandmeid ning sammu teostamise järgset väljundit. Üks võimalus kirjeldatud võrdluse teostamiseks on sisend- ja väljundandmete profileerimine. Võrdluse tulemusena tuvastatakse samm protsessis, kus sisendandmed on õiged, kuid väljundandmed valed. Järgmiseks tuleb välja selgitada vajaminevad muudatused, et andmed oleksid õiged ka pärast probleemse sammu läbimist. Siinkohal võib täiendavalt rakendada teisi meetodeid (Küsi 5 korda „Miks?“ või Põhjus ja tagajärg diagramm), et tuvastada probleemses protsessis tehtavad andmekvaliteeti mõjutavad tegevused. Viimaks tuleb tulemused dokumenteerida. Dokumentatsioon peaks kindlasti kirjeldama juurpõhjused, soovitusi nende lahendamiseks, meetodi käigus läbitud järeldesteni jõudmise protsessi ning tuvastatud otseseid ja kaudseid ärilisi mõjusid.

Põhjus ja tagajärg diagrammi (*Cause-and-Effect/ Fishbone Diagram*) abil identifitseeritakse, uuritakse ja esitatakse graafiliselt kõik võimalikud kvaliteediprobleemi põhjused. Meetodit on laialdaselt kasutatud tootmises ning see sobib hästi ka andmekvaliteedi juurpõhjuste analüüsimiseks. Meetodit on kasulik rakendada pärast jälitamise meetodi abil probleemi asukoha tuvastamist. Põhjus ja tagajärg diagramm võtab arvesse enam kui vaid ilmseid probleeme ning toob suurimat kasu kui seda rakendada grupis. Seega on esimeseks sammuks tiimi komplekteerimine ning koosoleku kokkukutsumine. Enne koosolekut tuleks tiimi liikmetele kättesaadavaks teha probleemiga seotud teadaolev info (näiteks jälitamise meetodi abil tuvastatud informatsioon), et osalejad jõuaksid koosoleku jaoks valmistud. Koosolekul tuleb selgelt välja tuua andmekvaliteedi poolt põhjustatud probleem. Probleem on meetodis tähistatud kui „tagajärg“ ning see on diagrammi esimene komponent. Järgnevalt tuleb diagrammile lisada probleemi peamised põhjused ehk kategooriad. Selleks võib diagrammile lisada enimlevinud üldised põhjused, kasutada varasema analüüsi käigus tuvastatud põhjuseid (näiteks jälitamise meetodi rakendamisel kogutud info), korraldada koosoleku liimetega ajurünnak või kasutada kombinatsiooni mitmest nimetatud viisist. Koosoleku liikmete küsitlemist tuleb jätkata kuni on jõutud probleemi juurpõhjuseni. Koosolekul leitud juurpõhjused tuleb dokumenteerida nende edasiseks kasutamiseks. Tuvastatud juurpõhjuste põhjal tuleb seejärel koostada soovitusel kvaliteediprobleemi juurpõhjuse lahendamiseks. Viimaks tuleb taas kõik tulemused dokumenteerida. Dokumentatsioon peaks kindlasti kirjeldama juurpõhjused, soovitusi nende lahendamiseks, meetodi käigus läbitud järeldesteni jõudmise protsessi ning tuvastatud otseseid ja kaudseid ärilisi mõjusid.



Joonis 12. Põhjus ja tagajärg diagrammi näidis.

2.5 Kvaliteediprobleemide mõju hindamine

Peale andmekvaliteedi probleemide tuvastamist ning probleemide juurpõhjuste analüüsi tuleb identifitseerida ja prioriseerida probleemide lahendamiseks tehtavad muudatused. Selleks teostatakse esmalt kvaliteediprobleemide ärilise mõju hindamine. Teostatud analüüsi tulemused tuleb edastada andmeomanikule andmekvaliteedi parandamise protsessi või projekti algatamiseks.

Kvaliteediprobleemide ärilise mõju suuruse hindamine on oluline, sest see näitab andmekvaliteedi parandamisest asutusele ja riigile tõusvat otsest tulu. Enimlevinud andmekvaliteedi probleemid ei pruugi olla äriselt olulised. Keskenduda tuleks vaid äriselt olulistele andmekvaliteedi probleemidele. Andmekvaliteedi ärilise mõju hindamiseks leidub mitmeid erinevaid tehnikaid, mille rakendamisel tuleb arvestada eelnevates sammudes kogutud informatsiooniga. Näiteks tuleks arvesse võtta tuvastatud kvaliteediprobleemide ulatuslikkust ja juba varasemalt määratletud prioriteete. Alljärgnevalt on toodud 3 tehnikat koos juhistega nende rakendamiseks. Tehnikad on järjestatud lihtsamalt keerukamale vastavalt tehnika rakendamise hinnangulisele ajakulule. Tehnika valimisel tuleb esmalt nendega tutvuda ning valida konkreetse situatsiooni ja kvaliteediprobleemi jaoks sobivaim. Seejuures tuleks arvestada, et aeganõudvama tehnika kasutamine ei taga tingimata paremaid tulemusi ning häid tulemusi on võimalik saavutada ka vähem aeganõudvate tehnikate rakendamisel. Peale tehnika rakendamist ning ärilise mõju kindlaksmääramist tuleb analüüsida kõiki kogutud tulemusi. Mõju analüüsi tulemuste põhjal tuleb koostada soovitusel andmekvaliteedi parandamiseks. Viimaks tuleb kõigi eelnevalt läbitud sammude tulemused ühtlustada ja dokumenteerida. Dokumentatsioon peaks sisaldama andmekvaliteedi ärilisest mõjust tulenevaid soovitusi andmekvaliteedi parandamiseks, ülevaadet kvaliteediprobleemide juurpõhjustest ning muid analüüsi käigus ilmnenu olulisi tulemusi. Koostatud ülevaade tuleb edastada andmeomanikule andmekvaliteedi parandamise protsessi või projekti algatamiseks.

Järgnevalt on kirjeldatud tehnikad andmekvaliteedi probleemi mõju hindamiseks.

Kasu ja kulu maatriks illustreerib kasu ja kulu vahelisi seoseid. Tehnika rakendamisel tuleb esmalt määrata isikud, kes osalevad prioriteetide määramisel. Seejärel tuleb valida meetod prioriteetide arutamiseks ja talletamiseks. Sobivad kõik meetodid, mis võimaldavad prioriteetide kiiret muutmist

arutelu käigus. Näiteks võib kasutada märkmepabereid või tahvlit. Arutelu käigus prioriseeritavate kvaliteediprobleemide nimekiri peab olema selgelt välja toodud ja kättesaadav arutelus osalejatele. Lisaks tuleb selgelt defineerida mida iga telg kasu ja kulu maatriksil tähistab. Kasu võib tähendada positiivset mõju ärile, mis kaasneb konkreetse kvaliteediprobleemi lahendamisega. Kasu võib olla ka kvaliteediprobleemi lahendamisega kaasnev produktiivsuse tõus või mis tahes muu konkreetse asutuse jaoks oluline muutus. Kulu võib olla kvaliteediprobleemi lahendamise maksumus rahas, selleks tehtava töö maht (ajakulu) või muu konkreetse asutuse jaoks oluline kulutus. Arutelu alguses tuleks tuua mõni näide, et meetodi kasutamine oleks kõigile arusaadav. Järgmiseks tuleb iga prioriseeritava kvaliteediprobleemi jaoks määrata hindamiskriteerium. Hindamiskriteerium võib olla kvalitatiivne (näiteks kliendimugavus) või kvantitatiivne (näiteks ajakulu). Näiteks kui kliendimugavus on oluline kasu tuleks hindamisel küsida: „Mis on arutatavate kvaliteediprobleemide nimekirjast probleemi number 1 mõju kliendimugavusele (skaalal madal kuni kõrge)?“. Kui kvaliteediprobleemi lahendamiseks kuluv aeg on oluline kulu tuleks küsida: „Kui palju aega kulub arutatavate kvaliteediprobleemide nimekirjast probleemi number 1 lahendamiseks (skaalal madal kuni kõrge)?“. Hindamisel on võimalik arvestada ka mitut kriteeriumit, kuid valitud kriteeriumite arv peaks olema piisavalt väike, et hindamisprotsess oleks hallatav.

Järgmise sammuna tuleb eelnevalt seatud kriteeriumite põhjal määrata kvaliteediprobleemidele hinnangud. Üks võimalus selle teostamiseks on lasta kõigil arutelul osalejatel panna oma hinnang kirja ning seejärel neid grupis arutada. Teine võimalus on asetada eri võimalused otse maatriksile nii, et need oleksid kõigile nähtavad ning seeläbi jõuda arutelu käigus ühise kokkuleppeni kvaliteediprobleemi lõpliku asukoha osas maatriksil. Kui kõik arutatavate kvaliteediprobleemide nimekirjas olnud probleemid on hinnatud ning maatriksile kantud, tuleb antud hinnangud veelkord üle vaadata ning kontrollida, kas eri kvaliteediprobleemide asetused maatriksil on põhjendatud. Viimaks tuleb tulemused dokumenteerida. Dokumentatsioon peaks kindlasti sisaldama põhjendusi kuidas konkreetsete hinnanguteni jõuti. Järgnevalt on toodud kasu ja kulu näidismaatriks maatriksi kasutamise illustreerimiseks (Joonis 13).

Kõrge			
Kasu		1. Kõrge kasu/ madal kulu Kvaliteediprobleemid, mida on mõistlik lahendada esmajärjekorras. Toovad kiirest palju kasu.	2. Kõrge kasu/ kõrge kulu Olulised, kuid kallid kvaliteediparandused. Neid võib lahendada enne sektoris 1 toodud probleeme, kui saadav kasu kaalub üle tehtavad kulutused.
		3. Madal kasu/ madal kulu Siin sektoris olevad kvaliteediprobleemid peaks olema kolmas valik. Samas tuleks kontrollida, et sektoris toodud probleemid poleks seotud sektorites 1 ja 2 toodud probleemidega.	4. Madal kasu/ kõrge kulu Kõige madalama prioriteediga kvaliteediprobleemid. Nende parandamisest võib loobuda või parandada need järjekorras viimasena.
Madal		Kulu	
			Kõrge

Joonis 13. Kasu ja kulu näidismaatriks.

Hindamise ja prioriseerimise tehnika järjestab andmekvaliteedi probleemid vastavalt äriprotsessidele avaldatavale mõjule. Andmete olulisuse indikaatoriks on andmete kasutus ning sellega kaasnevad riskid ja võimalused. Andmete olulisus varieerub sõltuvalt konkreetsetest andmetest ning andmete kasutamise viisidest. Tehnika annab parimaid tulemusi, kui selle rakendamisse on kaasatud need kes andmeid kasutavad, või kes disainivad uusi andmete kasutamise äriprotsesse ja praktikaid.

Esimene samm tehnika rakendamisel on identifitseerida äriprotsessid ning prioriseeritavate andmete kasutusviisid. Keskenduda tuleb äriprotsessidele, mis kasutavad ja loovad andmeid. Nagu varasemalt kirjeldatud, koosnes andmete elutsüklil planeerimisest, hankimisest, talletamisest ja jagamisest, säilitamisest, rakendamisest ning kustutamisest või arhiveerimisest. Hindamise ja prioriseerimise tehnika keskendub andmete rakendamise faasile, kus toimub andmeressursside kasutamine eesmärkide saavutamiseks. Andmeid võib järjestada ja hinnata nii konkreetsete kirjete põhjal kui ka andmete gruppidega, mis sisaldavad mitmeid seotud elemente. Näiteks tarnijale tasumiseks peab olemas olema täielik ja õige arve info.

Järgmine samm on määratleda prioriseerimisel osalejad. Vastavalt eelnevalt tuvastatud äriprotsessidele ja andmete kasutusele tuleb otsustada, keda prioriteetide seadmisel kaasata. Eri valdkondade inimeste arutelu kaasamine toetab ühtset asutusesisest mõistmist andmete kasutusest, andmete olulisusest, andmete kvaliteedist ning toetab seeläbi üldiselt andmete kvaliteedi paranemist. Enne arutelu tuleb valida meetod järjestuse loomiseks ning selle kiireks muutmiseks arutelu käigus. Järgmiseks tuleb arutelu alguses kokku leppida protsessides ja andmetes, mida mõju alusel järjestama hakatakse. Osalistele tuleb enne alustamist selgitada hindamise protsessi ning tuua näiteid. Hindamisel kasutatav skaala on toodud alljärgnevalt:

- ▶ Hinne A - halvast andmekvaliteedist tulenev protsessi täielik läbikukkumine.
- ▶ Hinne B - protsessi toimimine on takistatud ning probleemiga kaasnevad tõsised majanduslikud tagajärjed.
- ▶ Hinne C - vähesed majanduslikud tagajärjed.
- ▶ Hinne D – minimaalsed majanduslikud tagajärjed.
- ▶ N/A – kvaliteediprobleem ei mõjuta protsessi.

Näiteks arve saatmisel ei takista õigekirjaviga inimese nimes kirja kohaletoimetamist. Seega võib selle kvaliteediprobleemi hindeks panna C või D. Kui viga on aga korteri numbris takistab see kirja kohaletoimetamist ning seetõttu on tegu olulise probleemiga, mille hindeks tuleks panna A.

Järgmiseks sammuks on sarnaselt näitega hinnata konkreetsete andmete mõju igale äriprotsessile. Iga protsessi jaoks tuleks küsida küsimust: „Kui antud andmed oleksid puudu või valed siis kuidas see mõjutaks äriprotsessi?“. Kuigi tegu on subjektiivsete hinnangutega, on antud meetod väga efektiivne ning aitab tuvastada äriliselt olulised andmed ning nende mõju. Viimaks tuleb hinnatavatele andmetele omistada lõplik üldine hinnang. Lõplik üldine hinnang on kõrgeim konkreetsetele andmetele antud hinnang. Näiteks kui protsessis X oli andmetele antud hindeks C, kuid protsessis Y oli andmetele antud hinnang A siis on andmete lõplik üldine hinnang A. Viimaks tuleb analüüsida kogu kogutud informatsiooni ning antud hinnangute järgi tuvastada äriliselt kõige olulisemad andmed. Lõpuks tuleb taas tulemused dokumenteerida. Alljärgnevalt on esitatud näidistabel illustreerimaks tehnika rakendamist (Tabel 8)

Tabel 8. Hindamise ja prioriseerimise näidistabel.

Andmekvaliteedi mõju äriprotsessidele			
	Klientide teavitamine	Raporteerimine	Lõplik hinne
Email	B	C	B
Kontaktisiku nimi	A	A	A
Aadress	A	C	A
Amet	D	A	A

Tasuvusanalüüs on standardne meetod, mida kasutatakse finantsotsuste tegemisel. Andmekvaliteedi parandamisega võivad kaasneda märkimisväärsed kulutused. Seetõttu võib vajalikuks osutuda tasuvusanalüüsi abil juhtidele sisendi pakkumine, et oleks võimalik otsustada konkreetse kulutuse mõistlikkuse üle. Enamikel juhtudel pole niivõrd põhjaliku meetodi rakendamine mõistlik. Seda tuleks teha vaid väga suurte investeeringute puhul.

Tasuvusanalüüs hindab kas andmekvaliteedi parandamiseks tehtavast investeeringust tõusev kasu kaalub üle selle teostamiseks tehtavad kulutused. Selle teadasaamiseks arvutatakse investeeringu tasuvuse (ROI) väärtus, mis näitab investeeringust tõusvat kasu protsendina.

Tasuvusanalüüsi teostamiseks tuleb esmalt identifitseerida kõik kvaliteediprobleemi parandamiseks tehtavad kulutused (ehk tuvastada alginvesteering). Näiteks kulutused tööjõule, väljaõppele, riistvarale ja tarkvarale. Järgmiseks tuleb identifitseerida kvaliteediprobleemi parandamisega kaasnev potentsiaalne rahaline kasu (ehk investeeringu tulu). Seejärel tuleb leida kvaliteedi parandamisest tulenev kokkuhoid, mille arvutamiseks tuleb leida kasu ja kulu vahe. Lisaks tuleb määratlada eeldatavate tulude ja kulude ajakava. Hinnata tuleks ka kasu ja kulu mida pole võimalik kvantifitseerida. Kuna neid pole võimalik arvutada tuleks need lisada kommentaaridena.

Järgmiseks tuleb arvutada investeeringu tasuvus ehk ROI (Võrrand 3). Investeeringust tulenev tulu ja alginvesteering tuleb määratlada toetudes eelnevalt kirjeldatud analüüsi tulemustele.

Võrrand 3. Investeeringu tasuvuse (ROI) valem.

$$ROI = \frac{(Investeeringu\ tulu - Alginvesteering)}{Alginvesteering}$$

ROI väärtus ehk investeeringu tasuvus peab olema positiivne. Lisaks tuleb ROI väärtust võrrelda teiste kvaliteediprobleemide ROI väärtustega. Peale tasuvusanalüüsi lõpetamist tuleb taas tulemused dokumenteerida.

2.6 Andmekvaliteedi reeglite haldamine

Andmekvaliteedi reeglid on oluline metaandmete vorm. Et andmekvaliteedi reeglid oleksid efektiivsed tuleks neid ka hallata kui metaandmeid. Andmekvaliteedi reeglid peaksid olema:

- ▶ **Järjepidevalt dokumenteeritud.** Andmekvaliteedi reeglite dokumenteerimiseks tuleb luua selge mall, et tagada reeglite ühtne formaat ning mõistetus. Dokumentatsioon peaks kindlasti sisaldama andmekvaliteedi reegli unikaalset identifikaatorit ning reegli versiooni numbrit.
- ▶ **Seostatud andmekvaliteedi dimensioonidega.** Andmekvaliteedi dimensioonid aitavad inimestel mõista mida mõõdetakse. Andmekvaliteedi dimensioonide järjepidev rakendamine toetab mõõtmise ning probleemide haldamise protsesse.
- ▶ **Seotud ärilise mõjuga.** Andmekvaliteedi dimensioonid aitavad mõista levinud andmekvaliteedi probleeme, kuid dimensioonide kasutamine ja mõõtmine pole eesmärk omaette. Andmekvaliteedi reeglid peavad omama otsest mõju organisatsiooni edule. Seega pole äriprotsessidega mitteseotud mõõtmised vajalikud.
- ▶ **Andmeanalüüsi poolt toetatud.** Andmekvaliteedi reegleid ei peaks kirjeldama subjektiivsete arvamuste alusel. Reegleid tuleb testida reaalsel andmetel. Tihti toob selline testimine välja andmetes eksisteerivad probleemid ning aitab objektiivselt tuvastada ka andmekvaliteedi reeglites eksisteerivaid puuduseid.
- ▶ **Valdkonna eksperdi poolt heakskiidetud.** Andmekvaliteedi reeglite eesmärgiks on ilmutada nõuded andmetele. Tihti on reeglite õigeks kirjeldamiseks vajalikud teadmised konkreetse ärivaldkonna protsessidest. Neid teadmisi tuleks koguda konkreetse valdkonna eksperdilt, kelle ülesandeks on kinnitada kirjeldatud ärireeglid või selgitada andmeanalüüsi tulemusi.
- ▶ **Andmete kasutajatele kättesaadavad.** Kõigil andmete kasutajatel peaks olema juurdepääs dokumenteeritud andmekvaliteedi reeglitele. Juurdepääs reeglitele aitab andmete kasutajatel andmeid paremini mõista ning aitab samas tagada, et reeglid on täielikud ja õiged. Lisaks peab olema võimalus küsida reeglite kohta küsimusi ning anda tagasisidet.

Andmekvaliteedi reeglite tuvastamist lihtsustab andmete profileerimine ja analüüs. Koos andmekvaliteedi praktika küpsuse tõusuga peaks selline reeglite kirjeldamine liikuma süsteemide arendamise ja parandamise protsessi, sest andmekvaliteedi reeglite varases faasis kirjeldamine loob:

- ▶ Selged ootused andmete kvaliteedinäitajatele.
- ▶ Nõuded tarkvarasüsteemidele, mille rakendamise tulemusena välditakse andmekvaliteedi probleemide teket.
- ▶ Andmekvaliteedi nõuded partnerorganisatsioonidele ja muudele välistele osapooltele.
- ▶ Aluse pidevaks andmete kvaliteedi mõõtmiseks ja aruandluseks.

2.6.1 Reeglite muutmine ja versioneerimine

Andmekvaliteedi reeglid võivad muutuda kui:

- ▶ Andmekvaliteedi mõõtmise ja juurpõhjuste analüüsi tulemusena tuvastatakse, et hetkel kehtivad reeglid pole piisavad.
- ▶ Ebakvaliteetsete andmete tõttu on tekkinud probleem.
- ▶ Toimuvad muudatused ärinõuetes, regulatsioonides või mujal.

Kui andmehaldur on tuvastanud vajaduse andmekvaliteedi reegli muutmiseks, tuleb tal esitada andmeomanikule muudatusettepanek. Muudatusettepanek peaks sisaldama vähemalt muudatuse kirjeldust, muudatuse vajajaid või muudatusest kasusaajaid (näiteks konkreetne osakond või klient) ning muudatusettepaneku esitaja poolt defineeritud muudatuse prioriteetsust.

Et jälgida andmekvaliteedi reeglite muutumist ajas peab olema võimalik tuvastada konkreetse reegli versioone nii minevikust kui ka tulevikus. Teisisõnu tähendab see, et peab olema võimalik tuvastada reegli algset versiooni, kõiki arenduse käigus kasutatud versioone ning hetkel kasutusel olevat versiooni. Korrektse versioneerimise tulemusena on võimalik saada ülevaade, kas konkreetne reegel juba eksisteerib ja millises andmekvaliteedi reeglist on varasemalt lähtutud. Lisaks on versioneerimine oluline muudatuste haldamiseks. Näiteks on seeläbi võimalik paremini ennustada reegli muutmiseks vajamineva töö mahtu ning lihtsustada reeglite hilisemat taaskasutust, näiteks uuele süsteemile üleminekul. Probleemide tekkimisel toetavad korrektselt versioneeritud reeglid juurpõhjuste analüüsi läbiviimist aidates tuvastada probleemi algallikat. Versioneerimise toetamiseks peab igal andmekvaliteedi reeglil olema unikaalne identifikaator, talletatud konkreetse reegli kehtivust tähistavad kuupäevad ning versiooni number.

2.6.2 Andmeelementide kaetus reeglitega

Et saada ülevaadet andmeelementide kaetusest andmekvaliteedi reeglitega peab see olema dokumenteeritud. Dokumentatsioon võib olla esitatud näiteks tabelina (Tabel 9). Toodud tabelile toetudes on võimalik tuvastada andmeelemendid mille kvaliteeti ei mõõdata, kui ka üksteist dubleerivad andmekvaliteedi reeglid. Reeglitega kaetuse tabelit on võimalik genereerida automaatselt peale andmekvaliteedi mõõtmist. Üks võimalik viis nimetatud tabeli genereerimisest on esitatud antud juhise rakenduslikus osas.

Tabel 9. Andmeelementide reeglitega kaetuse näidistabel.

Andmeelement	Andmekvaliteedi reeglid
Omavalitsus	DQR1, DQR5
Maakond	
Köetav pind	DQR14, DQR2, DQR6
Kasulik pind	DQR14

3. Rakendamise näited

Antud sektsioonis on kirjeldatud profileerimise, andmekvaliteedi mõõtmise, andmekvaliteedi juhtimislaua loomise ja reeglite haldamise rakenduslikud näited. Näited on esitatud programmeerimiskeeles R¹⁷, mida kasutatakse laialdaselt andmete analüüsiks. Näidisalgoritmide käivitamise eelduseks on, et arvutisse on paigaldatud programmeerimiskeele R interpretaator¹⁸ ning arendustööriist RStudio¹⁹.

Lisaks kasutavad järgnevalt esitatud andmekvaliteedi mõõtmise näited Mark van der Loo ja Edwin de Jonge poolt arendatud *validate*²⁰ paketti. *Validate* pakett lihtsustab R keeles andmete kvaliteedi mõõtmist ning andmekvaliteedi reeglite haldamist.

Pakett võimaldab:

- testida andmekomplekti vastu eeldefineeritud reegleid, kas siis andmekomplekti sees või nende üleselt;
- importida ja eksportida defineeritud andmekvaliteedi reegleid;
- uurida ja visualiseerida andmekvaliteedi tulemusi;
- teostada lihtsat reeglite haldust;
- defineerida ja hallata andmete kvaliteedi indikaatoreid andmetest eraldiseisvalt.

Paketi peamised objektid:

- *validator* - objekt, mis esitab hulka reegleid, millele andmed peavad vastama
- *indicator* - objekt, mis esitab hulka numbrilisi kvaliteedi indikaatoreid
- *confrontation* - objekt, mis esitab andmete andmekvaliteedi reeglite või indikaatoritega vastandamise tulemusi

Paketis on ka meetod *confront*, mis rakendab andmetele andmekvaliteedi reegleid või andmekvaliteedi indikaatoreid.

Lisaks *validate* pakatile kasutatakse rakenduslikes näidetes paketti *dlookr*²¹ andmete profileerimiseks ning pakette *data.table*²², *stringr*²³ ja *tidyr*²⁴ andmete töötluse lihtsustamiseks.

3.1 Pakettide paigaldamine ja näidisandmed

Esimese sammuna on vaja arvutisse paigaldada näidetes kasutatavad paketid. Selleks tuleb jooksutada alljärgnevaid käskke.

¹⁷ <https://www.r-project.org/>

¹⁸ <https://cran.r-project.org/bin/windows/base/>

¹⁹ <https://rstudio.com/products/rstudio/download/>

²⁰ <https://cran.r-project.org/web/packages/validate/vignettes/introduction.html>

²¹ <https://cran.r-project.org/web/packages/dlookr/>

²² <https://cran.r-project.org/web/packages/data.table/>

²³ <https://cran.r-project.org/web/packages/stringr/>

²⁴ <https://cran.r-project.org/web/packages/tidyr/>

```
install.packages("dlookr")
install.packages("validate")
install.packages("data.table")
install.packages("stringr")
install.packages("tidyr")
```

Juhul kui eelnevalt toodud käskude käivitamine ebaõnnestub võib põhjuseks olla Rtools40 tööriista puudumine. Sellisel juhul tuleb järgmise sammuna nimetatud tööriist alla laadida ning arvutisse paigaldada. Tööriista saab alla laadida siit: <https://cran.rstudio.com/bin/windows/Rtools/>. Peale tööriista paigaldamist tuleb uus tööriistakomplekt seadistada. Selleks tuleb RStudios käivitada alljärgnev käsk.

```
writeLines('PATH="{RTOOLS40_HOME}\\usr\\bin;{PATH}"', con =
"~/.Renviron")
```

Peale kirjeldatud tegevuste teostamist tuleb eelnevalt toodud pakettide installeerimise käsud uuesti käivitada (juhul kui need eelnevalt ebaõnnestusid).

Järgmiseks tuleb äsja paigaldatud paketid sisse lugeda.

```
library(dlookr)
library(validate)
library(data.table)
library(stringr)
library(tidyr)
```

Viimaseks ettevalmistavaks sammuks on näidisandmete sisse lugemine. Antud juhul on näidisandmetena kasutatud Eesti Avaandmete Portaalist avalikult kättesaadavaid ehitisregistri²⁵ andmeid ning samuti Eesti Avaandmete Portaalist kättesaadavaid aadressiandmeid²⁶. Näidisandmete komplekt koosneb kaheksast failist ning seda on võimalik tellida nimetatud portaalist või alla laadida siit: <https://www.dropbox.com/sh/bf886osqo7kdy7/AACgIn1TbYv7JBROVTV5TJyDa?dl=0>

Allalaetud fail tuleks arvutis salvestada sobivasse kausta. Näite puhul on failide asukohaks arvuti töölaual asuv kaust „Andmed“ ("C:\\Users\\Margus\\Desktop\\Andmed\\").

Järgmiseks tuleb RStudios määrata failide kataloog järgmise käsuga.

```
filedir <- "C:\\Users\\Margus\\Desktop\\Andmed\\"
```

Seejärel anname failidele nimed käivitades järgnevad käsud.

²⁵ <https://opendata.riik.ee/andmehulgad/ehitisregister/>

²⁶ <https://opendata.riik.ee/andmehulgad/aadressiandmed-aadressid-koos-komponentide-ja-ajalooga/>

```

eh_ehitised_file <- paste0(filedir, "eh_ehitised_1994-01_2020-02-09.csv")
eh_tehna_file <- paste0(filedir, "eh_tehna_1994-01_2020-02-09.csv")
kl_tehna_file <- paste0(filedir, "kl_tehna_1994-01_2020-02-09.csv")
aadressid_file <- paste0(filedir, "AADRESSID.csv")
eh_ehitis_osad_file <- paste0(filedir, "eh_ehitis_osad_1994-01_2020-02-09.csv")
ehitis_aadress_file <- paste0(filedir, "ehitis_aadress_1994-01_2020-02-09.csv")
hoone_energia_margised_file <- paste0(filedir, "hoone_energia_margised_1994-01_2020-02-09.csv")
ad_aadress_file <- paste0(filedir, "ad_aadress_1994-01_2020-02-09.csv")

```

Peale failidele nimede andmist loeme andmed sisse. Selleks tuleb käivitada järgmised käsud.

```

#ehitise andmed
h_ehitised <- fread(eh_ehitised_file, encoding="UTF-8")

#ehitise tehnilised andmed
eh_tehna <- fread(eh_tehna_file, encoding="UTF-8")

#tehniliste andmete klassifikaatorid
kl_tehna <- fread(kl_tehna_file, encoding="UTF-8")

#ehitise osad
eh_ehitis_osad <- fread(eh_ehitis_osad_file, encoding="UTF-8")

#energiamärgised
hoone_energia_margised <- fread(hoone_energia_margised_file, encoding="UTF-8")

#seosed ehitise ja aadressi vahel
ehitis_aadress <- fread(ehitis_aadress_file, encoding="UTF-8")

#kohalik koopia aadressandmete tabelist
ad_aadress <- fread(ad_aadress_file, encoding="UTF-8")

#Aadressid aadressiandmete süsteemist
aadressid <- fread(aadressid_file, encoding="UTF-8")

```

3.2 Andmete profileerimise rakendamise näide

Andmekvaliteedi programmiga alustades tuleb esmalt saada ülevaade olemasolevatest andmetest. Selleks on mõislik toetada andmete profileerimine. Profileerimist on võimalik teostada mitmesuguste tööriistadega, kuid antud näide illustreerib *dlookr* paketi kasutamist. Paketi paigaldamiseks tehtavad tegevused on kirjeldatud eelnevas seksioonis.

Ehitiste andmetabeli põhjal esmase andmekvaliteedi raporti (*Data Quality Diagnosis Report*) genereerimiseks tuleb käivitada järgnev käsk. Sarnaselt saab genereerida raporteid ka teiste andmetabelite põhjal.

```
eh_ehitised %>% diagnose_report(output_format = "html", output_file =  
"Diagn_eh_ehitised.html")
```

Kui arvutisse on eelnevalt paigaldatud TeX²⁷ saab raportit genereerida ka pdf formaadis. Selleks tuleb käivitada alljärgnev käsk.

```
diagnose_report(eh_ehitised)
```

Käsu käivitamisel genereeritakse fail nimega *Diagn_eh_ehitised.html* (või fail nimega *DataDiagnosis_Report.pdf* juhul kui käivitati pdf formaadis faili genereerimise käsk). Kirjeldatud tegevuse eesmärgiks on saada esmane ülevaade andmetest. Saadud tulemused on abiks ka andmekvaliteedi reeglite kirjeldamisel.

dlookr paketi üldise andmekvaliteedi raporti dokumentatsiooniga on võimalik tutvuda siin:

<https://cran.r-project.org/web/packages/dlookr/vignettes/diagnosis.html>

Genereeritud raport annab põhjaliku esmase vaate olemasolevatele andmetele. Näiteks võimaldab raporti osana genereeritud arvvaartuste diagnoosimise tabel (Joonis 14) tuvastada võimalikke arvvaartuste kvaliteediprobleeme ning sõnastada andmekvaliteedi reegleid. Alltoodud ehitise andmete põhjal genereeritud näitest näeme, et andmeelemendi „ehituslane_pind“ miinimumvaartus on negatiivne (-24). See võib viidata, et andmetes esineb kvaliteediprobleeme. Lisaks on tuvastatud probleemid toeks andmekvaliteedi reeglite sõnastamisel. Näiteks on kirjeldatud näite puhul võimalik sõnastada järgnev reegel: „Ehitise ehituslane pind peab olema positiivne.“

variables	min	median	max	minus	minus ratio(%)
korgus	-300.0	2.6	15,205.0	7,205	0.681
sygavus	-150.0	0.0	832.4	2,106	0.199
pikkus	-275.0	9.2	5,361,019.0	1,166	0.110
maaaluste_korruste_arv	-2.0	0.0	17.0	593	0.056
max_korruste_arv	-23.0	1.0	120.0	10	0.001
laius	-65.0	3.0	60,482.0	9	0.001
abs_0_korgus	-16.0	37.0	7,700.0	4	0.000
maht_bruto	-10.0	150.0	10,452,091.0	3	0.000
ehituslane_pind	-24.0	64.0	13,432,000.0	2	0.000
koetav_pind	-1.6	80.6	360,540.0	1	0.000

Joonis 14. Näidis *dlookr* üldisest andmekvaliteedi raportist

²⁷ <https://miktex.org/>

Lisaks eelpool toodud raportile võimaldab *dlookr* pakett genereerida ka andmeanalüüsi raportit (*Exploratory Data Analysis Report*). Nimetatud raporti ehitiste andmete põhjal genereerimiseks tuleb käivitada järgnev käsk.

```
eh_ehitised %>% eda_report(output_format = "html", output_file =  
"EDA_Report_eh_ehitised.html")
```

Kirjeldatud käsu tulemusena genereeritakse fail nimega EDA_Report_eh_ehitised.html. Sarnaselt eelpool kirjeldatuga saab ka antud raportit genereerida pdf formaadis, juhul kui arvutisse on eelnevalt installeeritud TeX²⁸. Selleks tuleb käivitada järgnev käsk.

```
eda_report(eh_ehitised)
```

Käsu käivitamise tulemusena genereeritakse fail nimega EDA_Report.pdf.

dlookr andmeanalüüsi raporti dokumentatsiooniga on võimalik tutvuda siin:

<https://cran.r-project.org/web/packages/dlookr/vignettes/EDA.html>

Kirjeldatud raportite genereerimine ja tulemustega tutvumine võimaldab saada hea esmase ülevaate andmetest. Seejuures on oluline meeles pidada, et profileerimine ei ole andmekvaliteedi mõõtmine ning siit tuleks edasi liikuda andmekvaliteedi süsteemse mõõtmise suunas.

3.3 Andmekvaliteedi mõõtmise näidisjuhtumid

3.3.1 Näide N1 (Täielikkus, puuduv väärtus)

Dimensioon: Täielikkus.

Probleem: Puuduv väärtus

Reegli kirjeldus: Igal ehitisel peab olema määratud tema tunnus (ehitisregistri kood).

Selgitus: Ehitisregistri tabelis *eh_ehitised* peab igal ehitisel olema määratud ehitisregistri kood *ehr_kood*. Täielikkuse testimiseks loome reeglite komplekti *existance_rules*, kuhu lisame reegli *exists_id*, mis kontrollib kas andmeobjektil on tunnus *ehr_kood* määratud.

```
existance_rules <- validator(  
  exists_id = !is.na(ehr_kood)  
)
```

Reeglile vastavuse kontrollimiseks ja tulemuste kuvamiseks tuleb käivitada järgmised kaks käsku.

```
check <- confront(eh_ehitised, existance_rules, key="ehr_kood")  
summary(check)
```

Käsu käivitamise tulemusena kuvatakse RStudio konsooli järgnev tulemuste kokkuvõte.

²⁸ <https://miktex.org/>

	name	items	passes	fails	nNA	error	warning	expression
1	exists_id	1058629	1058629	0	0	FALSE	FALSE	!is.na(ehr_kood)

Tulemustest on näha, et kõigil ehitistel on ehitisregistri kood määratud (*fails* ehk reeglile mittevastavate kirjete arv on 0).

3.3.2 Näide N2 (Ühekordsus, unikaalse väärtuse rikkumine)

Dimensioon: Ühekordsus.

Probleem: Unikaalse väärtuse rikkumine

Reegli kirjeldus: Iga ehitise tunnus (ehitisregistri kood) peab olema unikaalne

Selgitus: Ehitisregistri tabelis *eh_ehitised* peab iga ehitise ehitusregistri kood olema unikaalne. Unikaalsuse testimiseks loome reeglite komplekti *uniqueness_rules*, kuhu lisame reegli *unique_id*, mis kontrollib kas andmeobjektidel on tunnus *ehr_kood* unikaalne.

```
uniqueness_rules <- validator(
  unique_id = is_unique(ehr_kood)
)
```

Reeglile vastavuse kontrollimiseks ja tulemuste kuvamiseks tuleb käivitada järgmised käsud.

```
check <- confront(eh_ehitised, uniqueness_rules, key="ehr_kood")
summary(check)
```

Käsu käivitamise tulemusena kuvatakse RStudio konsooli järgnev tulemuste kokkuvõte.

	name	items	passes	fails	nNA	error	warning	expression
1	unique_id	1058629	1058629	0	0	FALSE	FALSE	is_unique(ehr_kood)

Tulemustest on näha, et kõigil ehitistel on unikaalne tunnus (ehitisregistri kood).

3.3.3 Näide N3 (Õigsus, väärtusvahemiku rikkumine)

Dimensioon: Õigsus.

Probleem: Väärtusvahemiku rikkumine

Reegli kirjeldus: Omandi liigi ja seisundi väärtused peavad vastama etteantud loendi väärtutele.

Selgitus: Ehitisregistri tabelis *eh_ehitised* võivad atribuudi *omandi_liik* väärtused olla järgnevad:

- EHITIS_OMANDI_LIIK_VALLAS;
- EHITIS_OMANDI_LIIK_KINNIS;
- EHITIS_OMANDI_LIIK_VALLASRO.
-

Atribuudi *seisund* väärtused võivad olla järgnevad:

- EHITIS_SEISUND_EHITAM_LUBA;
- EHITIS_SEISUND_EHITAM_LUBA_EIKEH;
- EHITIS_SEISUND_EHITAMISEL;
- EHITIS_SEISUND_KASUT_LUBA_EIKEH;
- EHITIS_SEISUND_KASUT_MAAS;
- EHITIS_SEISUND_KASUT_OSALINE;
- EHITIS_SEISUND_KASUTUSEL;
- EHITIS_SEISUND_LAMMUT_LUBA;

- EHITIS_SEISUND_LAMMUTAMISEL;
- EHITIS_SEISUND_LAMMUTATUD;
- EHITIS_SEISUND_MAARAMATA;
- EHITIS_SEISUND_MENETLUSES;
- EHITIS_SEISUND_REG_OBJ_LOPP.

Seega kirjeldame vastavalt eeltoodud selgitusele reeglite komplekti *ehitised_rules*, kuhu lisame reeglid *property_type* ja *state_type*.

```
ehitised_rules <- validator(
  property_type = omandi_liik %in% c('EHITIS_OMANDI_LIIK_VALLAS',
  'EHITIS_OMANDI_LIIK_KINNIS', 'EHITIS_OMANDI_LIIK_VALLASRO'),
  state_type = seisund %in% c('EHITIS_SEISUND_EHITAM_LUBA',
  'EHITIS_SEISUND_EHITAM_LUBA_EIKEH', 'EHITIS_SEISUND_EHITAMISEL',
  'EHITIS_SEISUND_KASUT_LUBA_EIKEH', 'EHITIS_SEISUND_KASUT_MAAS',
  'EHITIS_SEISUND_KASUT_OSALINE', 'EHITIS_SEISUND_KASUTUSEL',
  'EHITIS_SEISUND_LAMMUT_LUBA', 'EHITIS_SEISUND_LAMMUTAMISEL',
  'EHITIS_SEISUND_LAMMUTATUD', 'EHITIS_SEISUND_MAARAMATA',
  'EHITIS_SEISUND_MENETLUSES', 'EHITIS_SEISUND_REG_OBJ_LOPP')
)
```

Reeglile vastavuse kontrollimiseks ja tulemuste kuvamiseks tuleb käivitada järgmised käsud.

```
check <- confront(eh_ehitised, ehitised_rules, key="ehr_kood")
summary(check)
```

Käsu käivitamise tulemusena kuvatakse RStudio konsooli järgnev tulemuste kokkuvõte.

	name	items	passes	fails	nNA	error	warning
1	property_type	1058629	1057500	1129	0	FALSE	FALSE
2	state_type	1058629	1058628	1	0	FALSE	FALSE

Väljavõte näitab, et 1058629-st kirjest ei ole vastavalt 1129-s ja 1-s kirjes omandi liik ja seisund õigesti kodeeritud.

Järgnevalt vaatame lähemalt millised kirjed ei vasta reeglile. Selleks konverteerime esmalt tulemuse kokkuvõtte sobivale kujule ning filtreerime tulemused nii, et alles jäävad vaid reeglitele mittevastavad kirjed. Kirjeldatu teostamiseks tuleb käivitada järgnevad kaks käsku.

```
output <- as.data.frame(check)
output[output$value == FALSE, ]
```

Käskude käivitamise tulemusena kuvatakse RStudio konsoolis ülevaade probleemsetest kirjetest.

	ehr_kood	name	value
1:	120773369	property_type	FALSE
2:	120739741	property_type	FALSE
3:	120713061	property_type	FALSE
4:	120713124	property_type	FALSE
5:	120713112	property_type	FALSE

1126:	120834655	property_type	FALSE
1127:	120827890	property_type	FALSE
1128:	120713778	property_type	FALSE
1129:	120826592	property_type	FALSE
1130:	221294327	state_type	FALSE

Tuvastatud konkreetseid kirjeid saab ka täpsemalt uurida. Näiteks on alljärgnevalt toodud käsk kirje *ehr_kood*-iga 120773369 kuvamiseks.

```
eh_ehitised[eh_ehitised$ehr_kood == 120773369, c('id', 'nimetus',
'seisund', 'omandi_liik')]
```

Käsu käivitamise tulemusena kuvatakse RStudio konsooli probleemse kirje atribuutide *id*, *nimetus*, *seisund* ja *omandi_liik* väärtused.

	id	nimetus	seisund	omandi_liik
1:	5537838	Kelder	EHITIS_SEISUND_KASUTUSEL	

Tulemusest on näha, et antud kirje puhul on *omandi_liik* väärtustamata ehk antud kirje ei vasta eelnevalt defineeritud andmekvaliteedi reeglile.

3.3.4 Näide N4 (Õigsus, väärtusvahemiku rikkumine)

Dimensioon: Õigsus.

Probleem: Väärtusvahemiku rikkumine.

Reegli kirjeldus: Ehitise tehniliste andmete kodeerimisel tuleb kasutada tehniliste andmete klassifikaatorit.

Selgitus: Ehitise tehniliste andmete kodeerimisel kasutatakse tehniliste andmete klassifikaatorit. Antud näites testime muutujaid eraldi tabelis esitatud klassifikaatori vastu. Ehitise tehnilised andmed on tabelis *eh_tehna*. Ehitise tehniliste andmete klassifikaator on tabelis *kl_tehna*. *eh_tehna* tabeli atribuut *tena_id* esitab seost *kl_tehna* tabeli atribuudiga *id* (*eh_tehna.tena_id* <-> *kl_tehna.id*).

```
tech_classifier_rules <- validator(
  tech_data_conformance = tena_id %in% codelist
)
```

Reeglile vastavuse kontrollimiseks ja tulemuste kuvamiseks tuleb käivitada järgmised käsud.

```
check <- confront(eh_tehna, tech_classifier_rules, ref = list(codelist =
kl_tehna$id), key="tena_id")
summary(check)
```

Käsu käivitamise tulemusena kuvatakse RStudio konsooli järgnev tulemuste kokkuvõte.

	name	items	passes	fails	nNA	error	warning	expression
1	tech_data_conformance	19114165	19114165	0	0	FALSE	FALSE	tena_id ...

Tulemustest on näha, et ühtegi reegli rikkumist ei tuvastatud ning järelkult on tehniliste andmete kodeerimisel kasutatud korrektselt tehniliste andmete klassifikaatorit.

3.3.5 Näide N5 (Õigsus, süntaksi rikkumine)

Dimensioon: Õigsus.

Probleem: Süntaksi rikkumine

Reegli kirjeldus: Kuupäev peab vastama etteantud vormingule: Ehituse alustamise kuupäev (*eh_alust_kp*) peab olema vormingus yyyy-MM-dd ning kirje loomise kuupäev (*date_created*) vormingus yyyy-MM-dd'HH:mm:ss.SSSSSS.

Selgitus: Kuupäevad ehitiste tabelis *eh_ehitised* peavad vastama etteantud vormingule. Atribuudi *eh_alust_kp* kuupäeva õige vorming on yyyy-MM-dd (näiteks 2008-09-20) ja *date_created* õige vorming on yyyy-MM-dd'HH:mm:ss.SSSSSS (näiteks 2019-10-14 14:12:46.666714). Et kontrollida kuupäevade vormingule vastamist loome reeglite komplekti *date_rules*, kuhu lisame reegli *start_date_syntax* atribuudi *eh_alust_kp* väärtuste kontrollimiseks ning reegli *created_date_syntax* atribuudi *date_created* väärtuste kontrollimiseks.

```
date_rules <- validator(
  start_date_syntax = grepl("^([0-9]{4}-[0-9]{2}-[0-9]{2})$",
eh_alust_kp),
  created_date_syntax = grepl("^([0-9]{4}-[0-9]{2}-[0-9]{2})\\ ([0-9]{2}\\\\:([0-9]{2}\\\\:([0-9]{2}\\\\.([0-9]{6})$)", date_created)
)
```

Reeglile vastavuse kontrollimiseks ja tulemuste kuvamiseks tuleb käivitada järgmised käsud.

```
check <- confront(eh_ehitised, date_rules, key="ehr_kood")
summary(check)
```

Käsu käivitamise tulemusena kuvatakse RStudio konsooli järgnev tulemuste kokkuvõte.

	name	items	passes	fails	nNA	error	warning	expression
1	start_date_syntax	1058629	59942	998687	0	FALSE	FALSE	grepl("^([0-...
2	created_date_syntax	1058629	1055767	2862	0	FALSE	FALSE	grepl("^([0-...

Tulemustest on näha, et reeglile *start_date_syntax* ei vasta 998 687 kirjet ning reeglile *created_date_syntax* ei vasta 2862 kirjet. Järgnevalt vaatame taas lähemalt millised kirjed ei vasta reeglile. Selleks konverteerime esmalt tulemuse kokkuvõtte sobivale kujule ning filtreerime tulemused nii, et alles jäävad vaid reeglitele mittevastavad kirjed.

```
output <- as.data.frame(check)
output[output$value == FALSE, ]
```

Käskude käivitamise tulemusena kuvatakse RStudio konsoolis ülevaade probleemsetest kirjetest.

	ehr_kood	name	value	expression
1:	120231238	pattern_year	FALSE	grep("^[0-9]{4}\$", esmane_kasutus)
2:	115011733	pattern_year	FALSE	grep("^[0-9]{4}\$", esmane_kasutus)
3:	108034400	pattern_year	FALSE	grep("^[0-9]{4}\$", esmane_kasutus)
4:	108028019	pattern_year	FALSE	grep("^[0-9]{4}\$", esmane_kasutus)
5:	120295322	pattern_year	FALSE	grep("^[0-9]{4}\$", esmane_kasutus)

623331:	121293239	pattern_year	FALSE	grep("^[0-9]{4}\$", esmane_kasutus)
623332:	221296656	pattern_year	FALSE	grep("^[0-9]{4}\$", esmane_kasutus)
623333:	221303569	pattern_year	FALSE	grep("^[0-9]{4}\$", esmane_kasutus)
623334:	121303570	pattern_year	FALSE	grep("^[0-9]{4}\$", esmane_kasutus)
623335:	121303683	pattern_year	FALSE	grep("^[0-9]{4}\$", esmane_kasutus)

Tuvastatud konkreetseid kirjeid saab ka täpsemalt uurida. Näiteks on alljärgnevalt toodud käsk kirje *ehr_kood*-iga 121303683 kuvamiseks.

```
eh_ehitised[eh_ehitised$ehr_kood == 121303683, c('id', 'eh_alust_kp',
'date_created')]
```

Käsu käivitamise tulemusena kuvatakse RStudio konsooli probleemse kirje atribuutide *id*, *eh_alust_kp* ja *date_created* väärtused. Nagu näha on antud probleemse kirje puhul *eh_alust_kp* väärtustamata.

	id	eh_alust_kp	date_created
1:	5626327		2019-09-09 08:09:22.756168

3.3.6 Näide N6 (Õigsus, süntaksi rikkumine)

Dimensioon: Õigsus.

Probleem: Süntaksi rikkumine.

Reegli kirjeldus: Esmase kasutuse aasta talletamiseks peab olema kasutatud neljakohalist aastanumbrit (vastama etteantud vormingule: yyyy).

Selgitus: Tabelis *eh_ehitised* veerus *esmane_kasutus* on talletatud esmase kasutuse kuupäev ja kellaaeg. Kuupäeva süntaktilise õigsuse kontrollimiseks loome reeglite komplekti *pattern_rules*, kuhu lisame reegli *pattern_year*. Antud reegel kontrollib, et kasutatud oleks neljakohalist aastanumbrit.

```
pattern_rules <- validator(
  pattern_year = grep("^[0-9]{4}$", esmane_kasutus)
)
```

Reeglile vastavuse kontrollimiseks ja tulemuste kuvamiseks tuleb käivitada järgmised käsud.

```
check <- confront(eh_ehitised, pattern_rules, key="id")
summary(check)
```

Käsu käivitamise tulemusena kuvatakse RStudio konsooli järgnev tulemuste kokkuvõte.

	name	items	passes	fails	nNA	error	warning	expression
1	pattern_year	1058629	435294	623335	0	FALSE	FALSE	grep("^[0-9]...

Saadud tulemusest on näha, et 623 335 kirjet ei vasta kontrollitavale reeglile. Konkreetsete probleeme saab lähemalt uurida näidetes N2 ja N4 näidatud viisil.

3.3.7 Näide N7 (Reeglipärasus, valdkonnakitsenduse rikkumine)

Osad reeglid kirjeldavad kitsendusi erinevate andmeveergude vahel. Need kitsendused on kasulikud ärireeglite esitamisel. Antud näide illustreerib sellist juhtu.

Dimensioon: Reeglipärasus.

Probleem: valdkonnakitsenduse rikkumine.

Reegli selgitus: Alates 01.01.2020 on liginullenergia nõue kohustuslik (v.a. erandid) üle 220 m² väikeelamutele. Liginullenergiahoone nõuet ei kohaldata väikeelamule köetava pinnaga kuni 220 m² (kehtivad madalenergia ehk B klassi nõuded). Samuti ei rakendata liginullenergiahoonet (ka eramule suurema köetava pinnaga kui 220 m²) juhul, kui päikeseenergiasüsteemi paigaldamine ei ole majanduslikult põhjendatud või tehniliselt teostatav.

Selgitus: Hoonete energiatõhusust reguleerivad kolm määrust^{29,30,31}. Neil määrustel põhinev reegel esitab hulga tingimusi, millele *eh_ehitised* tabelis olevad kirjed peavad vastama. Et kontrollida kirjeldatud reeglile vastavust konverteerime atribuudi *esmane_kasutus* numbriliseks väärtuseks. Selle tegevuse eesmärgiks on muuta võimalikuks andmete võrdlemine kasutades validator objekti. Järgmiseks loome reeglite komplekti *energy_class_rules*. Sinna lisame reegli *mandatory_energy_class*, mis kontrollib kõiki reeglis toodud tingimusi: kui esmane kasutus on 2020 või hilisem, hoone tüüp on väikeelamu ning köetav pind on suurem kui 220 m² peab ehitise energiaklass olema A.

Esimese sammuna konverteerime atribuudi *esmane_kasutus* numbriliseks väärtuseks.

```
eh_ehitised$esmane_kasutus <- as.numeric(eh_ehitised$esmane_kasutus)
```

Järgmiseks loome eelnevalt kirjeldatud reeglite komplekti *energy_class_rules*.

```
energy_class_rules <- validator(  
  mandatory_energy_class = if (esmane_kasutus >= 2020 & hoone_tyypp ==  
    'VAIKEELAMUD' & koetav_pind > 220) energia_klass == 'ENERGIACL_A'  
)
```

Reeglile vastavuse kontrollimiseks ja tulemuste kuvamiseks tuleb käivitada järgmised kaks käsku.

```
check <- confront(hoone_energia_margised, energy_class_rules, key="id")  
summary(check)
```

Käsu käivitamise tulemusena kuvatakse RStudio konsooli järgnev tulemuste kokkuvõte.

	name	items	passes	fails	nNA	error	warning	expression
1	mandatory_energy_class	30649	30644	1	4	FALSE	FALSE	!(...

²⁹ <https://www.riigiteataja.ee/akt/122082019002>

³⁰ <https://www.riigiteataja.ee/akt/122082019005>

³¹ <https://www.riigiteataja.ee/akt/122082019004>

Tulemusest on näha, et üks kirje ei vasta reeglile ning neli kirjet on väärtustamata (nNA). Konkreetsete probleeme saab lähemalt uurida näidetes N2 ja N4 näidatud viisil.

3.3.8 Näide N8 (Õigsus, valdkonnakitsenduse rikkumine)

Dimensioon: Reeglipärasus

Probleem: Valdkonnakitsenduse rikkumine.

Reegli kirjeldus: Kui ehitise köetav pind on määratud, siis peab see olema väiksem või võrdne kasuliku pinnaga.

Selgitus: Määruses³² "Ehitise tehniliste andmete loetelu ja pindade arvestamise alused" on kirjeldatud hoone köetav pind järgnevalt:

- § 28(2) Hoone köetav pind on hoone kõigi sisekliima tagamisega ruumide suletud netopindade summa;
- § 22(1) Korruse suletud netopind ehk kasulik pind on korruse suletud brutopind, millest on maha arvatud korruse välistarindite alune pind, sisetarindite alune pind ja mittekandvate tarindite alune pind;
- § 22(2) Hoone suletud netopind on kõigi korruste suletud netopindade summa.

Sellest tulenevalt kontrollime järgnevalt, et kui köetav pind on olemas, siis see on väiksem või võrdne kasuliku pinnaga.

```
conditional_rules <- validator(  
  condition1 = if (!is.na(koetav_pind)) koetav_pind <= kasulik_pind  
)
```

Reeglile vastavuse kontrollimiseks ja tulemuste kuvamiseks tuleb käivitada järgmised käsud.

```
check <- confront(eh_ehitised, conditional_rules, key="id")  
summary(check)
```

Käsu käivitamise tulemusena kuvatakse RStudio konsooli järgnev tulemuste kokkuvõte.

	name	items	passes	fails	nNA	error	warning	expression
1	condition1	1058629	993968	10145	54516	FALSE	FALSE	!(is.na(...

Tulemusest on näha, et üks 10 145 kirjet ei vasta reeglile ning neli kirjet on väärtustamata (nNA). Konkreetsete probleeme saab lähemalt uurida näidetes N2 ja N4 näidatud viisil.

3.3.9 Näide N9 (Reeglipärasus, funktsionaalse sõltuvuse rikkumine)

Dimensioon: Reeglipärasus.

Probleem: funktsionaalse sõltuvuse rikkumine.

Reegli kirjeldus: Omavalitsus on seotud vaid ühe maakonnaga.

Selgitus: Kontrollimaks, et omavalitsus on seotud vaid ühe maakonnaga loome reeglite komplekti *dependency_rules*. Sinna lisame reegli *dependency_county*, mis kirjeldab funktsionaalset sõltuvust omavalitsuse ja maakonna vahel.

³² <https://www.riigiteataja.ee/akt/107102014003>

```
dependency_rules <- validator(
  dependency_county = omavalitsus ~ maakond
)
```

Reeglile vastavuse kontrollimiseks ja tulemuste kuvamiseks tuleb käivitada järgmised käsud.

```
check <- confront(eh_ehitised, dependency_rules, key="ehr_kood")
summary(check)
```

Käsu käivitamise tulemusena kuvatakse RStudio konsooli järgnev tulemuste kokkuvõte.

	name	items	passes	fails	nNA	error	warning	expression
1	dependency_county	1058629	1002331	56298	0	FALSE	FALSE	omavalitsus ...

Tulemused näitavad, et 1058629-st kirjest ei ole 56298-s kirjes omavalitsus üheselt maakonnaga seotud.

Järgmiseks vaatame lähemalt milline kirje ei vasta reeglile. Selleks tuleb käivitada järgnev käsk.

```
output <- as.data.frame(check)
output[output$value == FALSE,]
```

Käsu käivitamise tulemusena kuvatakse RStudio konsooli järgnev ülevaade kirjetest.

	ehr_kood	name	value	expression
1:	114011772	dependency_county	FALSE	omavalitsus ~ maakond
2:	114011505	dependency_county	FALSE	omavalitsus ~ maakond
3:	220838926	dependency_county	FALSE	omavalitsus ~ maakond
4:	121309384	dependency_county	FALSE	omavalitsus ~ maakond
5:	221309392	dependency_county	FALSE	omavalitsus ~ maakond

56295:	121302502	dependency_county	FALSE	omavalitsus ~ maakond
56296:	121300602	dependency_county	FALSE	omavalitsus ~ maakond
56297:	221300400	dependency_county	FALSE	omavalitsus ~ maakond

Vaatame viimast kirjet lähemalt. Selleks tuleb käivitada järgnev käsk.

```
eh_ehitised[eh_ehitised$ehr_kood == 121303567, c('id', 'nimetus',
  'maakond', 'omavalitsus')]
```

Käsu käivitamise tulemusena kuvatakse RStudio konsooli järgnev tulemuste kokkuvõte.

	id	nimetus	maakond	omavalitsus
1:	5626247	Üksikelamu	68	809

Järgmiseks vaatame millist maakonna koodi kasutatakse veel sama omavalitsuse kirjetes.

```
eh_ehitised[eh_ehitised$omavalitsus == 809, c('id', 'nimetus',
  'maakond', 'omavalitsus')]
```

Käsu käivitamise tulemusena kuvatakse RStudio konsooli järgnev tulemuste kokkuvõte.

	id	nimetus	maakond	omavalitsus
1:	3162780	alajaam	67	809
2:	2501880	Elamu	68	809
3:	2531178	Üksikelamu	68	809
4:	5548142	Üksikelamu	68	809
5:	5630137	Puurkaev	68	809

15588:	5624925	üksikelamu	68	809
15589:	5624926	Puurkaev	68	809
15590:	5629827	Kuur	68	809
15591:	5625477	Aiamaja	68	809
15592:	5626247	Üksikelamu	68	809

Tulemustest selgub, et omavalitsuse 809 (Tori vald) maakonnaks on andmetes 67 ja 68. Tegelikult peab seal olema vaid 68 (Pärnu maakond).

3.3.10 Näide N10 (Ajakohasus, aegunud väärtus)

Dimensioon: Ajakohasus.

Probleem: vale väärtus (aegunud).

Reegli kirjeldus: Ehtisregistrisse imporditud aadresside lühiaadress ja täisaadress peab vastama aadressiandmete süsteemis hallatavate aadresside samadele andmeelementidele.

Selgitus: Antud näites võrdleme sama andmehulga eri versioone tuvastamaks lokaalse koopia tabelis *ad_address* ajakohasust võrreldes tabeliga *aadressid*.

Kõigepealt mestime kahe andmehulga (*ad_address* ja *aadressid*) aadresside *adr_id* väärtuse alusel ja loome referentstabeli *ads_address*, mis sisaldab vaid lähiaadressi (*lahiaadress*) ja täisaadressi (*taisaadress*) kirjeid.

```
adsdiff = merge(x=ad_address, y=aadressid, by.x="adr_id",
by.y="ADR_ID", all.x=TRUE)
ads_address <- adsdiff[,c('LAHIAADDRESS', 'TAISAADDRESS')]
colnames(ads_address) <- c('lahiaadress', 'taisaadress')
ehr_address <- adsdiff[,c('lahiaadress', 'taisaadress')]
```

Seejärel loome reegli, mis kontrollivad vastavalt kas andmetabeli elemendid *lahiaadress* ja *taisaadress* vastavad referentstabeli (*ads_reference*) analoogsetele väärtustele.

```
timeliness_rules <- validator(
  timeliness_lahiaadress = lahiaadress == ads_reference$lahiaadress,
  timeliness_taisadress = taisaadress == ads_reference$taisaadress
)
```

Tabeli *ehr_address* reeglitele vastavuse kontrollimiseks ja tulemuste kuvamiseks tuleb käivitada järgmised käsud, millega ütleme, et *ehr_address* tabeli referentstabelina (*ads_reference*) kasutatakse *ads_address* andmeid.

```
check <- confront(ehr_aadress, timeliness_rules, ref =
list(ads_reference = ads_aadress))
summary(check)
```

Käivitatud käsu tulemusena kuvatakse RStudio konsooli järgnev tulemuste kokkuvõte.

	name	items	passes	fails	nNA	error	expression
1	timeliness_lahiaddress	2464347	2413787	77	50483	FALSE	lahi...
2	timeliness_taisaddress	2464347	2413785	79	50483	FALSE	tais...

3.3.11 Näide N11 (Õigsus, süntaksi rikkumine)

Antud näites kontrollime kuupäevade süntaksile vastavust mitme muutuja puhul korraga, laiendades näitest N5 tuttava reegli *start_date_syntax* kasutust.

Dimensioon: Õigsus.

Probleem: Süntaksi rikkumine

Reegli kirjeldus: Kuupäev peab vastama etteantud vormingule: Ehituse alustamise kuupäev (*eh_alust_kp*), kultuurimälestiseks muutmise aeg (*kultuurimalestis_kp*), kavandatav kasutusele võtmise aeg (*kav_kasutus_kp*), ajutise ehitise kasutuse lõpp (*ajeh_kasutlopp_kp*) ja ajutise ehitise kasutusele võtmise algus (*ajeh_kasutalg_kp*) peavad olema vormingus yyyy-MM-dd.

Selgitus: Kuupäevad ehitiste tabelis *eh_ehitised* peab vastama etteantud vormingule. Kuupäeva õige vorming on yyyy-MM-dd (näiteks 2008-09-20). Antud näiteks kontrollitakse atribuute *eh_alust_kp*, *kultuurimalestis_kp*, *kav_kasutus_kp*, *ajeh_kasutlopp_kp* ja *ajeh_kasutalg_kp*. Et kontrollida kuupäevade vormingule vastamist loome reeglite komplekti *date_rules*, kuhu lisame reegli *start_date_syntax*.

```
date_rules <- validator(
eh_alust_kp_syntax = grepl("[0-9]{4}-[0-9]{2}-[0-9]{2}$",
eh_alust_kp),
kultuurimalestis_kp_syntax = grepl("[0-9]{4}-[0-9]{2}-[0-9]{2}$",
kultuurimalestis_kp),
kav_kasutus_kp_syntax = grepl("[0-9]{4}-[0-9]{2}-[0-9]{2}$",
kav_kasutus_kp),
ajeh_kasutlopp_kp_syntax = grepl("[0-9]{4}-[0-9]{2}-[0-9]{2}$",
ajeh_kasutlopp_kp),
ajeh_kasutalg_kp_syntax = grepl("[0-9]{4}-[0-9]{2}-[0-9]{2}$",
ajeh_kasutalg_kp)
)
```

Nagu eeltoodud näitest näha kasvab atribuutide lisandumisel oluliselt vajaminevat koodiridade arv. Sellise olukorra vältimiseks saab alternatiivina kasutada lühemat ja loetavamalt, kuid samaväärset käsku.

```
date_rules <- validator(
  G := var_group(eh_alust_kp, kultuurimalestis_kp, kav_kasutus_kp,
    ajeh_kasutlopp_kp, ajeh_kasutalg_kp), grepl("^([0-9]{4}-[0-9]{2}-[0-9]{2})$", G)
)
```

Reeglile vastavuse kontrollimiseks ja tulemuste kuvamiseks tuleb endiselt käivitada järgmised käsud.

```
check <- confront(eh_ehitised, date_rules, key="ehr_kood")
summary(check)
```

Käsu käivitamise tulemusena kuvatakse RStudio konsooli järgnev tulemuste kokkuvõte.

	name	items	passes	fails	nNA	error	warning	expression
1	V2.1	1058629	59942	998687	0	FALSE	FALSE	grepl("^([0-9]{4}...
2	V2.2	1058629	4	1058625	0	FALSE	FALSE	grepl("^([0-9]{4}...
3	V2.3	1058629	118499	940130	0	FALSE	FALSE	grepl("^([0-9]{4}...
4	V2.4	1058629	6015	1052614	0	FALSE	FALSE	grepl("^([0-9]{4}...
5	V2.5	1058629	2	1058627	0	FALSE	FALSE	grepl("^([0-9]{4}...

3.4 Andmekvaliteedi juhtimislaua loomine

Eeltoodud reeglite näidatud viisil realiseerimise tulemusi on lihtne esitada andmekvaliteedi juhtimislaua. Juhtimislaua loomise eesmärgiks on tuua välja andmete kvaliteediga seotud kitsaskohad ning kommunikeerida edasiminekuid. Vaata ka juhtimislaua näidet aadressilt:

<https://datastudio.google.com/u/0/reporting/e34c9f7b-7c05-4870-a5b7-2cb1e6b6299c/page/FK09>.

Allpool kirjeldame sellise juhtimislaua loomist, mis aitab anda vastused järgmistele küsimustele:

- ▶ Kui suur osa hallatud andmeobjektidest vastab andmekvaliteedi nõuetele?
- ▶ Millised andmeelemendid ja mil määral on kaetud andmekvaliteedi reeglitega?
- ▶ Millised andmekvaliteedi reeglid on kirjeldatud?
- ▶ Milliste andmeobjektide kvaliteeti on vaja tõsta?

3.4.1 Andmeobjektide vastavus nõuetele

Selleks, et visualiseerida kõikide andmeobjektide nõuetele vastavust tuleb kokku koondada kõik andmekvaliteedi reeglid. Antud puhul koondame kokku ehitiste andmeobjektide kohta käivad reeglid (*eh_ehitised_rules*).

```
eh_ehitised_rules <- existence_rules + uniqueness_rules +
ehitised_rules + pattern_rules + conditional_rules + dependency_rules
```

Ehitise andmete vastavust eelnevalt kirjeldatud reeglitele on võimalik korraga kontrollida käivitades järgnevalt toodud käsud.

```
check <- confront(eh_ehitised, eh_ehitised_rules, key="ehr_kood")
summary(check)
```

Käsu käivitamise tulemusena kuvatakse RStudio konsooli järgnev tulemuste kokkuvõte.

	name	items	passes	fails	nNA	error	warning
1	exists_id	1058629	1058629	0	0	FALSE	FALSE
2	unique_id	1058629	1058629	0	0	FALSE	FALSE
3	property_type	1058629	1057500	1129	0	FALSE	FALSE
4	state_type	1058629	1058628	1	0	FALSE	FALSE
5	pattern_year	1058629	435294	623335	0	FALSE	FALSE
6	condition_koetav_pind	1058629	993968	10145	54516	FALSE	FALSE
7	dependency_county	1058629	1002331	56298	0	FALSE	FALSE

Järgmiseks käivitame käsud reeglite tabelikujule viimiseks.

```
output <- as.data.frame(check)
head(output)
```

Käsu käivitamise tulemusena kuvatakse RStudio konsoolis loodud tabelkuju kuut esimest kirjet.

	ehr_kood	name	value	expression
1:	220258718	exists_id	TRUE	!is.na(ehr_kood)
2:	220258678	exists_id	TRUE	!is.na(ehr_kood)
3:	221267603	exists_id	TRUE	!is.na(ehr_kood)
4:	120741879	exists_id	TRUE	!is.na(ehr_kood)
5:	120231238	exists_id	TRUE	!is.na(ehr_kood)
6:	120535518	exists_id	TRUE	!is.na(ehr_kood)
		...		

Andmeobjektide kõikidele nõuetele vastavuse visualiseerimiseks võtame näidisenä 100 000 kirjet (tuleneb Google Sheets, mida lihtsustamise huvides kasutame, kitsendustest) ning kirjutame need CSV faili nimega data-conformance-long.csv. Selleks tuleb käivitada alljärgnevad käsud.

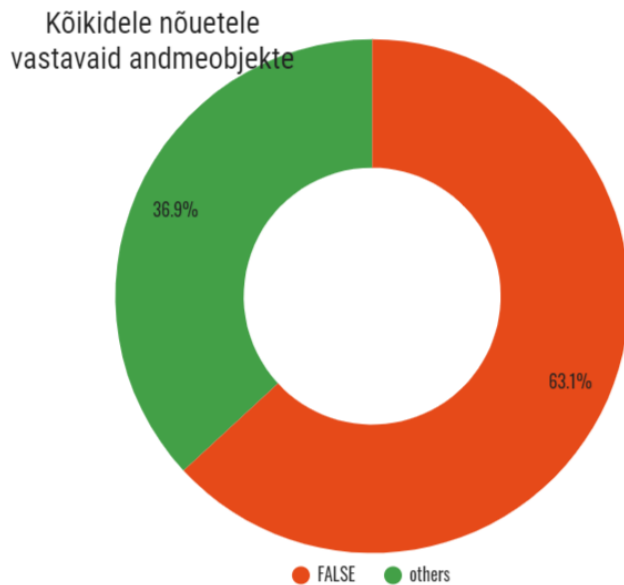
```
output <- output[sample(.N, 100000)]
write.csv(output[,c('ehr_kood', 'name', 'value')], paste0(filedir,
"data-conformance-long.csv"), row.names = FALSE, na="")
```

Käsu käivitamise tulemusena genereeritud CSV faili saab kasutada nõuetele vastavate reeglite visualiseerimiseks kasutades näiteks MS Excel-i, Google Sheets-i, Google Data Studio-t vmt tarkvara.

Näites ettevalmistatud andmete põhjal koostatud näidisraportiga on võimalik tutvuda siin:

<https://datastudio.google.com/u/0/reporting/e34c9f7b-7c05-4870-a5b7-2cb1e6b6299c/page/FK09>

Eeltoodud tegevuste tulemusena ettevalmistatud andmete põhjal on Google Data Studio abil koostatud graafik pealkirjaga „Kõikidele nõuetele vastavaid andmeobjekte“, mis on toodud ka alljärgneval joonisel (Joonis 15).



Joonis 15. Kõikidele nõuetele vastavaid andmeobjekte

3.4.2 Parandamist vajavad andmeobjektid

Selleks, et saaksime ülevaatlilikult tuua välja konkreetsed andmeobjektid, mille kvaliteet vajab tõstmist viime andmed nn laiale kujule, milleks tuleb käivitada järgmised käsud.

```
output <- as.data.frame(check)
data_wide <- spread(output[, c('ehr_kood', 'name', 'value')], name, value)
```

Seejärel võtame näitlikustamiseks 60000 suvalist kirjet ning kirjutame need CSV faili nimega data-conformance-wide.csv.

```
data_wide <- data_wide[sample(.N, 60000)]
write.csv(data_wide, paste0(filedir, "data-conformance-wide.csv"),
row.names = FALSE, na="")
```

Genereeritud CSV faili saab kasutada andmekvaliteedile vastavuse visualiseerimiseks kasutades näiteks MS Excel-i, Google Sheets-i, Google Data Studio-t vmt tarkvara.

Näidistulemusi on võimalik grupeeritult vaadata näidisjuhtimislaualt (vaata ka Joonis 16):

<https://datastudio.google.com/u/0/reporting/e34c9f7b-7c05-4870-a5b7-2cb1e6b6299c/page/H0bVB>

ehr_kood	exists_id	unique_id	conditio...	depende...	property_ty...	pattern_year	state_type
101000002	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE
101000021	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
101000051	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
101000072	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE
101000114	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE
101000134	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
101000151	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
101000193	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE
101000196	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
101000243	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE
101000278	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE
101000282	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE
101000285	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
101000318	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE

Joonis 16. Ehitise andmete vastavus andmekvaliteedi reeglitele

3.4.3 Millised reeglid on kaetud

Reeglitest ülevaate saamiseks saab neid esitada tabelkujul (Joonis 17), tuues välja milliste ärimõistete, andmetabelite ja andmekvaliteedi dimensioonidega on nad seotud. Vaata ka:

<https://datastudio.google.com/u/0/reporting/e34c9f7b-7c05-4870-a5b7-2cb1e6b6299c/page/TWQLB>

Id	Ärimõiste	Andmetabel	Dimensio...	Probleemi liik	Reegli kirjeldus	Reegel (R)
exists_id	Ehitis	ehr.eh_ehitised	Täielikkus	Puuduv väärtus	Igal ehitisel peab olema määratud tema tunnus (ehitis...	is_unique(ehr_kood)
unique_id	Ehitis	ehr.eh_ehitised	Ühekordsus	Unikaalse väärtuse rikkumine	Iga ehitise tunnus (ehitisregistri kood) peab olema uni...	lis.na(ehr_kood)
property_type	Ehitis	ehr.eh_ehitised	Õigsus	Väärtusvahemiku rikkumine	Ehitise omandi liik vastab etteantud loendile	omandi_liik %in% c('E...
state_type	Ehitis	ehr.eh_ehitised	Õigsus	Väärtusvahemiku rikkumine	Ehitise seisund vastab etteantud loendile	seisund %in% c('EHIT...
tech_data_conforman...	Ehitise tehnilise...	ehr.eh_tehna	Õigsus	Väärtusvahemiku rikkumine	Ehitise tehniliste andmete kodeerimisel kasutatakse t...	tena_id %in% codelist
start_date_syntax	Ehitis	ehr.eh_ehitised	Õigsus	Süntaksi rikkumine	Ehituse alustamise ajatempel peab olema esitatud va...	grepl("^[0-9]{4}-[0-9]{2}...
created_date_syntax	Ehitis	ehr.eh_ehitised	Õigsus	Süntaksi rikkumine	Kirje loomise kuupäev peab olema esitatud vastavalt ...	grepl("^[0-9]{4}-[0-9]{2}...
pattern_year	Ehitis	ehr.eh_ehitised	Õigsus	Süntaksi rikkumine	Esmase kasutamise aastaarv peab vastama ettenäht...	grepl("^[0-9]{4}\$", es...
mandatory_energy_cla...	Energiamärgis	ehr.hoone_energia_margis...	Reeglipäras...	Valdkonnakäsitluse rikkumine	Alates 01.01.2020 on liginullenergia nõue kohustuslik ...	if (esmane_kasutus =...
condition_koetav_pind	Ehitis	ehr.eh_ehitised	Reeglipäras...	Valdkonnakäsitluse rikkumine	Kui koetav pind on olemas, siis koetav_pind <= kasulik...	if (lis.na(koetav_pind)...
dependency_county	Ehitis	ehr.eh_ehitised	Reeglipäras...	Funktsionaalse sõltuvuse rikkum...	Omavalitsus saab olla vaid ühe maakonna koosseisus	omavalitsus ~ maak...
timeliness_lahiaaddress	Aadress	ehr.ad_aadress	Ajakohasus	Aegunud väärtus	EHRI lähiaadress erineb ADS lähiaadressist	lahiaadress == ads_r...
timeliness_taisaaddress	Aadress	ehr.ad_aadress	Ajakohasus	Aegunud väärtus	EHRI täisaadress erineb ADS täisaadressist	taisaaadress == ads_r...

Joonis 17. Kaetus andmekvaliteedi reeglitega

3.4.4 Andmelementide kaetus reeglitega

Et teha kindlaks kas kõik olulised andmeelemendid on kaetud andmekvaliteedi reeglitega, koondame esmalt kokku kõik eelnevalt loodud reeglid.

```
all_rules <- existence_rules + uniqueness_rules + ehited_rules +
tech_classifier_rules + date_rules + pattern_rules + energy_class_rules
+ conditional_rules + dependency_rules + timeliness_rules
```

Reeglite ja muutujate maatrikskujul vaatamiseks tuleb käivitada alljärgnev käsk.

```
variables(all_rules, as='matrix')
```

Käsu tulemusena kuvatakse atribuutide kaetus reeglitega maatriksina. Allpool on näidisenä toodud 4 esimest.

rule	ehr_kood	omandi_liik	seisund	tena_id	codelist
exists_id	TRUE	FALSE	FALSE	FALSE	FALSE
unique_id	TRUE	FALSE	FALSE	FALSE	FALSE
property_type	FALSE	TRUE	FALSE	FALSE	FALSE
state_type	FALSE	FALSE	TRUE	FALSE	FALSE
tech_data_conformance	FALSE	FALSE	FALSE	TRUE	TRUE
...					

Järgmiseks tuvastame need andmekvaliteedi veerud, mille koha andmekvaliteedi reeglid puuduvad. Selleks tuleb käivitada alljärgnev käsk.

```
names(eh_ehitised)[!names(eh_ehitised) %in%
variables(eh_ehitised_rules)]
```

Käsu käivitamise tulemusena kuvatakse atribuudid ehk veerud, mis pole veel reeglitega kaetud.

[1] "id"	"ehit_id"	...
[6] "rajatis_hoone"	"max_korruste_arv"	...
[11] "kav_kasutus_kp"	"ajeh_kasutalg_kp"	...
[16] "energia_valjast_kp"	"energia_margise_kp"	...
[21] "date_created"	"date_updated"	...
[26] "mitteeluruumide_arv"	"korterite_arv"	...
[31] "kultuurimalestis_viit"	"abs_0_korgus"	...
[36] "elamispind"	"eluruumi_pind"	...
[41] "laius"	"lift"	...
[46] "pind"	"suletud_netopind"	...
[51] "abiruumide_pind"	"toimiku_nr"	...

Eelnev võimaldab küll analüüsida andmeelementide kaetust reeglitega, kuid pole mugav viis ülevaate saamiseks. Parema ülevaate saamiseks võib samasid andmeid esitada andmekvaliteedi juhtimislaua. Selleks tuleb käivitada järgmised andmeid ettevalmistavad käsud.

```
df <- data.frame(variables(all_rules, as='matrix'))
df[names(eh_ehitised)[!names(eh_ehitised) %in%
variables(eh_ehitised_rules)]] = FALSE
t_df <- transpose(df)
colnames(t_df) <- rownames(df)
rownames(t_df) <- colnames(df)
write.csv(t_df, paste0(filedir, "variable-coverage-with-
rules.csv"), row.names = TRUE, na="")
```

Käskude tulemusena genereeritakse varem vaadeldud reeglitega kaetuse andmed maatrikskujul CSV faili nimega variable-coverage-with-rules.csv. Nimetatud failis olevaid andmeid saab kasutada juhtimislaua koostamisel (Joonis 18). Juhtimislaua näide on juurdepääsetav siit:

<https://datastudio.google.com/u/0/reporting/e34c9f7b-7c05-4870-a5b7-2cb1e6b6299c/page/yISVB>

field_name	exists...	unique...	dependenc...	property_t...	tech_data_confo...	timelines...	state_type	pattern_y...	mandatory...	condition_koetav_pind
ehr_kood	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
omandi_liik	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
seisund	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE
tena_id	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
codelist	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
eh_alust_kp	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
kultuurimalestis_kp	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
kav_kasutus_kp	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
ajeh_kasutlopp_kp	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
ajeh_kasutalg_kp	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
esmane_kasutus	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE
hoone_tyyp	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
koetav_pind	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE
energia_klass	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE

Joonis 18. Juhtimislaua koostamiseks kasutatavad andmed

3.4.5 Reeglite eksport ja import

Võib tekkida olukordi, kus kirja pandud reegleid on vaja jagada. Sellisel juhul on reeglid võimalik eksportida reeglite faili käivitades alljärgnevalt toodud käsud. Käskude käivitamisel genereeritakse yml formaadis fail nimega rules.yml.

```
rule_filename <- paste0(filedir, "rules.yml")
export_yaml(all_rules, file=rule_filename)
```

Võib tekkida ka olukord, kus reeglid on vaja failist sisse lugeda. Seda saab teha järgneva käsuga.

```
rules2 <- validator(.file=rule_filename)
```

3.5 Soovituslikud töövahendid

Järgnevalt on toodud nimekiri tööriistades, mis on abiks andmekvaliteediga seotud eri tööloikude realiseerimisel.

Profileerimine:

- ▶ RStudio (<https://rstudio.com/products/rstudio/>)
- ▶ R paketid profileerimise toetamiseks:
 - ▶ DataExplorer R pakett (<https://cran.r-project.org/web/packages/DataExplorer/>)
 - ▶ dlookr R pakett (<https://cran.r-project.org/web/packages/dlookr/>)

Kvaliteedi mõõtmine:

- ▶ Validate R pakett (<https://cran.r-project.org/web/packages/validate/>)

Andmekvaliteedi reeglite haldamine:

- ▶ Confluence (<https://www.atlassian.com/software/confluence>)
- ▶ Jira (<https://www.atlassian.com/software/jira>)