

**EUROPEAN COMMISSION  
DG REFORM**

**Eesti andmehalduse  
metoodikaprojekt**

Andmekvaliteedi juhis

Mai 2020

## Sisukord

|                                                           |    |
|-----------------------------------------------------------|----|
| Sisukord .....                                            | 1  |
| 1. Sissejuhatus .....                                     | 2  |
| 1.1 Lühendid ja mõisted.....                              | 2  |
| 1.2 Andmekvaliteedi juhise eesmärk .....                  | 3  |
| 1.3 Sihtrühmad.....                                       | 3  |
| 1.4 Juhise ulatus.....                                    | 3  |
| 2. Andmekvaliteedi juhtimine .....                        | 4  |
| 2.1 Andmekvaliteedi eesmärgistamine.....                  | 4  |
| 2.2 Andmekvaliteedi juhtimise protsess .....              | 5  |
| 3. Andmekvaliteedi juhtimise tegevused .....              | 8  |
| 3.1 Andmete profileerimine.....                           | 8  |
| 3.2 Andmekvaliteedi reeglite kirjeldamine .....           | 8  |
| 3.3 Andmekvaliteedi hindamine .....                       | 9  |
| 3.3.1 Andmekvaliteedi hindamise metamudel.....            | 11 |
| 3.3.2 Andmekvaliteedi probleemid .....                    | 13 |
| 3.3.3 Kvaliteediprobleemide tuvastamine .....             | 17 |
| 3.3.4 Dimensioonide ja indikaatorite mõõtmine.....        | 22 |
| 3.3.5 Andmekvaliteedi reeglile mõõdiku seadmine .....     | 26 |
| 3.3.6 Andmekvaliteedi aruandepõhjade väljatöötamine ..... | 29 |
| 3.3.7 Kvaliteediprobleemide prioritseerimine .....        | 29 |
| 3.4 Andmekvaliteedi juurpõhjuste analüüs .....            | 29 |
| 3.5 Kvaliteediprobleemide mõju hindamine.....             | 31 |
| 3.6 Andmekvaliteedi reeglite haldamine .....              | 36 |
| 3.6.1 Reeglite muutmine ja versioneerimine .....          | 36 |
| 3.6.2 Andmeelementide kaetus reeglitega.....              | 37 |
| 4. Rakendamise näited .....                               | 38 |
| 4.1 Andmekvaliteedi mõõtmise algoritmide näited .....     | 38 |
| 4.2 Andmekvaliteedi juhtimislaua mall.....                | 44 |
| 4.3 Soovituslikud töövahendid.....                        | 44 |

# 1. Sissejuhatus

Käesolev juhised on üks osa Andmehalduse raamistikust, mille eesmärk on aidata organisatsioonil kehtestada ühtseid põhimõtteid, protseduure ning paika seada vajalikke rolle ja mõõdikuid andmehalduse korraldamisel. Andmehaldus on organisatsiooni võimekus hallata andmeid varana, sh tagada andmete kvaliteet nõutaval tasemel.

## 1.1 Lühendid ja mõisted

Kõigil kolmel projekti tulemil saab lõpuks olema ühine mõistete sõnastik.

| Termin         | Määratlus ja selgitus                                                                                                                                                                                                                                                                                                                              |
|----------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Andmed         | Informatsiooni taastõlgendatav esitus formaliseeritud kujul, mis sobib edastuseks, tõlgenduseks või töötluks [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:<br>- andmete sisu, konteksti ja metaandmete eest vastutaja;<br>- kohustused sõltuvad kontekstist ja võivad osaliselt kattuda andmekäitleja omadega.<br>[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 <sup>1</sup> ] |
| AvTS           | Avaliku teabe seadus <a href="https://www.riigiteataja.ee/akt/115032019011">https://www.riigiteataja.ee/akt/115032019011</a>                                                                                                                                                                                                                       |
| 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<br><a href="https://akit.cyber.ee/term/5074-andmeolem-1-andmeuksus">https://akit.cyber.ee/term/5074-andmeolem-1-andmeuksus</a>                                                                         |

|                            |                                                                                                                                                                                           |
|----------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 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 <sup>6</sup> ] |
| SA                         | Statistikaamet <a href="http://www.stat.ee">http://www.stat.ee</a>                                                                                                                        |
| TKTA                       | VV määrus „Teenuste korraldamise ja teabehalduse alused“                                                                                                                                  |
| 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]                                                                                                            |

## 1.2 Andmekvaliteedi juhise eesmärk

Juhendi põhieesmärgiks on pakkuda praktilisi juhtnööre andmeomanikele, andmehalduritele ja teistele andmetega kokku puutuvatele asutuste töötajatele andmekvaliteedi juhtimiseks, mõõtmiseks ja kvaliteedi parendamise praktiliseks rakendamiseks.

Juhis pakub täiendust 2016. aastal loodud juhisele „Andmekvaliteedi tagamise juhend andmekogu omanikele“.

Andmekvaliteedi juhise koostamisel on lähtutud DAMA-DMBOK2 andmehalduse mudelist.

## 1.3 Sihtrühmad

Andmekvaliteedi juhise sihtühmad on andmeomanikud, äriprotsesside omanikud, andmehaldurid, andmeanalüütikud, andmeteadurid ja andmeinsenerid, ehk kõik osapooled, kes puutuvad kokku andmekvaliteedi parendamisega

## 1.4 Juhise ulatus

Andmekvaliteedi juhise annab esmalt lühiülevaate andmekvaliteedi eesmärgistamisest ja andmekvaliteedi juhtimisest. Seejärel on kirjeldatud peamised tegevused andmekvaliteediga seotud andmehalduse protsessides:

- Andmete profileerimine;
- Andmekvaliteedi reeglite kirjeldamine;
- Andmekvaliteedi hindamine;
- Andmekvaliteedi juurpõhjuste analüüs;
- Kvaliteediprobleemide ärilise mõju hindamine ja teavitamine;
- Andmekvaliteedi reeglite muutmine.

Viimasena esitatakse näited kirjeldatud tegevuste praktiliseks rakendamiseks.

## 2. Andmekvaliteedi juhtimine

### 2.1 Andmekvaliteedi eesmärgistamine

Usaldusväärsed ja kõrge kvaliteediga andmed klientide, toodete, teenuste ja äriprotsesside kohta 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. Peamine andmehalduse ning sealhulgas ka andmekvaliteedi tagamise eesmärk on võimaldada organisatsioonidel oma kogutud andmetest maksimaalset väärtust saada. Samamoodi nagu loob väärtust muude ressursside korrektne haldamine.

Andmekvaliteedi programmiga alustamise eesmärgid on järgmised:

- Organisatsiooni andmete väärtuse tõstmine;
- Uute võimaluste loomine andmete kasutamiseks;
- Halva andmekvaliteediga kaasnevate riskide ja kulude 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 erinevad riskid. Näiteks võib halb andmekvaliteet kahjustada organisatsiooni reputatsiooni, tuua kaasa rahalisi kaotusi ja negatiivseid meediakajastusi. Andmekvaliteediga on seotud ka mitmed otsesed kulud, näiteks:

- Suutmatus esitada korrektseid arveid;
- Suurenenud kliendiprobleemide arv ning vähenenud suutlikkus nende lahendamiseks;
- Suurem keerukus asutuste tegevuse ümberkorraldamisel;
- Väiksem võimekus pettuste tuvastamisel;
- Halva kvaliteediga andmete põhjal tehtud väärad 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 ressursi 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 protsess peaks keskenduma andmevigade ennetamisele ning andmete kasutatavust pärssivate tegurite vähendamisele. Kindlasti ei tohiks keskenduda vaid andmevigade parandamisele.

- **Probleemide juurpõhjuste lahendamine** – Andmekvaliteedi parandamine tähendab enamasti kui andmevigade parandamist. Andmekvaliteedi probleemide lahendamiseks tuleb tuvastada kvaliteediprobleemi algallikad, mitte keskenduda vaid nende tagajärgede likvideerimisele. Seejuures hõlmab andmekvaliteedi parandamine tihti protsesside ja süsteemide täiustamist, mis on levinud kvaliteediprobleemide algallikad.
- **Andmehaldus** – Andmehalduse tegevused peavad toetama kõrge kvaliteediga andmete teket ning andmekvaliteedi protsessi tegevused peavad toetama ja säilitama hallatavat andmekeskonda.
- **Sihttasemetest lähtuv** - kõik andmete elutsükli osapooled omavad tähelepanu vajavaid andmekvaliteedi reegleid. 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 protsessides järgitaks andmekvaliteedi standardeid.
- **Süsteemidesse juurutamine** - Süsteemide omanikud, ehk süsteemide eest vastutajad peavad tagama, et süsteemides järgitaks andmekvaliteedi standardeid.
- **Teenustasemega ühendamine** - Teenustaseme lepingud (*Service Level Agreements*) peaksid sisaldama andmekvaliteedist raporteerimist ja probleemide haldamist puudutavaid punkte.

## 2.2 Andmekvaliteedi juhtimise protsess

Andmekvaliteedi juhtimise protsessi võib vaadelda tuntud Demingi ringina ehk „kavanda-teosta-kontrolli-parenda” (ingl *PDCA – Plan-Do-Check-Act*) pideva parendamise tsükliks. Protsessi eesmärgiks on anda ülevaade andmekvaliteedi parandamise elutsüklist. Alljärgnevalt on antud lühiülevaade kõigist tsükli etappidest.



Joonis 1. Andmekvaliteedi parendamise elutsükkel

**Kavandamise** etapis hinnatakse andmekvaliteedi probleemide ulatust, mõju, määratakse prioriteedid ning töötatakse välja meetmed kvaliteediprobleemide lahendamiseks. Kvaliteediprobleemidele lahenduste pakkumise eeldusena peavad olema välja selgitatud probleemide juurpõhjused. Juurpõhjuste tuvastamise ning mõjuanalüüsi teostamise järel on võimalik hinnata kasu ja kulu, määrata täpsed prioriteedid ning töötada välja esialgne plaan muudatuste elluviimiseks.

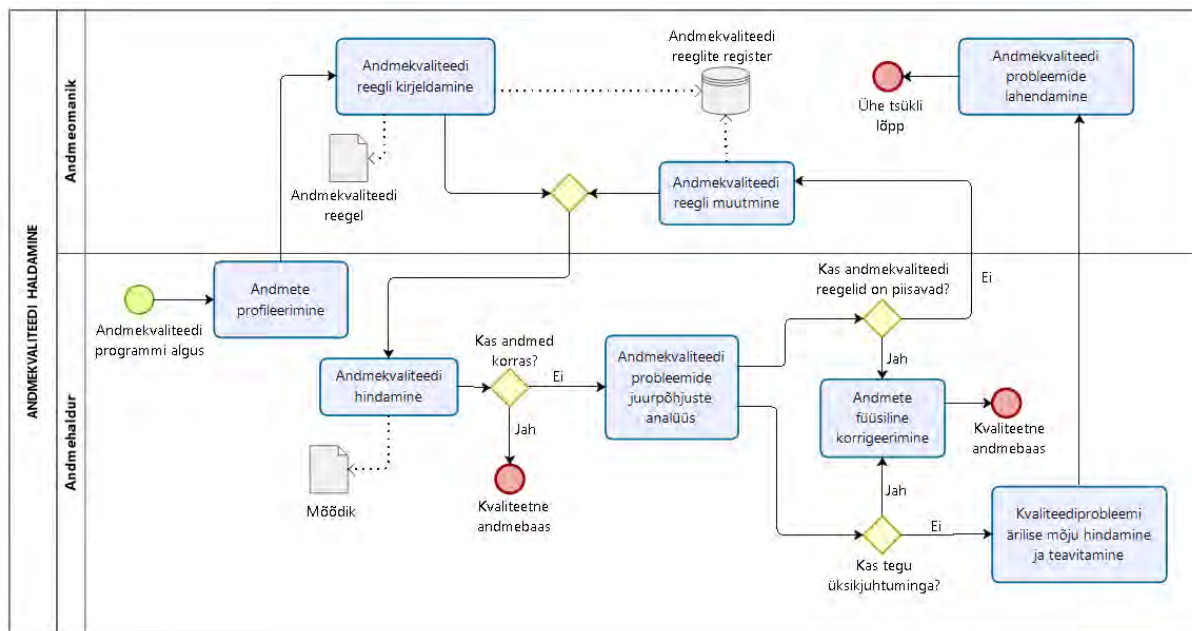
**Teostamise** etapis viiakse ellu muudatused probleemide juurpõhjuste lahendamiseks ning planeeritakse andmete edaspidist jälgimist. Kui probleemi algallikana on tuvastatud probleemne protsess tuleb muudatused rakendada koostöös konkreetse protsessi eest vastutajaga. Kui aga probleemi algallika lahendamine nõuab tehnilisi muudatusi tuleks seda teha koostöös tehnilise tiimiga. Seejuures tuleks tagada, et muudatused saaksid rakendatud korrektselt ning tehnilised muudatused ei tooks kaasa uusi vigu.

**Kontrollimise** etapis toimub andmete pidev jälgimine, mille käigus mõõdetakse andmete vastavust nõuetel. Täiendavad tegevused pole vajalikud kui andmed vastavad kehtestatud sihttasemele. Sellisel juhul on tegu kontrollitud protsessiga, mis vastab andmekvaliteedi reeglitele. Kui andmete jälgimine tuvastab, et andmed ei vasta kehtestatud sihttasemele tuleb alatatada edasised tegevused andmete kvaliteedi tõstmiseks vastuvõetavale tasemele.

**Parendamise** etapis tegeletakse esile kerkinud kvaliteediprobleemide lahendamisega. Tsükkel algab uuesti probleemide põhjuste hindamise ning lahenduste pakkumisega. Pidev parendamine saavutatakse uue tsükli alustamisega. Uus tsükkel algab kui:

- Jälgimise ja mõõtmise tulemuste põhjal ei vasta andmed enam kehtestatud kvaliteedikriteeriumitele.
- Esile kerkivad uued andmed.
- Tekivad uued nõuded olemasolevate andmete kvaliteedile.
- Ärireeglid, standardid või ootused muutuvad.

Alljärgnevalt (Joonis 2. Andmekvaliteedi haldamise protsess) on esitatud vaade juhendi kontekstis käsitletavale Eesti andmehalduse raamistikus toodud andmekvaliteedi haldamise protsessile. RR: Siin võiks see seos DGFIGa märksa selgem saada. Võiks olla viide konkreetsetele protsessidele ja alapeatükkidele; selgitada, et rollid on võetud DGFIGist ja seal täpsemalt lahti mõtestatud; võiks aru saada, et kas selle siinjoonistatud tegevusega võib tegeleda iseseisvalt või peab ootama mingit trigger-event'i DGFi juurutamise raames; kas siit läheb mingil kindlal kujul väljund tagasi DGFIGi?



Joonis 2. Andmekvaliteedi haldamise protsess



### 3. Andmekvaliteedi juhtimise tegevused

#### 3.1 Andmete profileerimine

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

Andmete profileerimine on protsess, mille eesmärgiks on uurida olemasolevaid andmeid (andmebaasist või konkreetsest failist) ning koguda statistikat ja informatiivseid kokkuvõtteid andmete koosseisu kohta. Näiteks tuvastatakse profileerimise käigus väärtuste sagedus, formaat, mustrid ja mitmed muud karakteristikuid. 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. Valik profileerimise teostamiseks sobivatest tööriistadest on kirjeldatud käesoleva juhendi soovituslike töövahendite sektsioonis. Kuigi profileerimistööriistad esitavad andmete kohta mitmesugust statistikat ja mõõdikuid pole profileerimise puhul tegu andmekvaliteedi hindamisega. Andmekvaliteedi hindamine 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 kontrole (näiteks summa leidmine, miinimumväärtuste leidmine ja maksimumväärtuste leidmine). Struktuuri põhine 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 telefoninumbreid.
- **Seoste põhise** profileerimise käigus tuvastatakse andmete omavahelised seosed, näiteks andmetabelite vahelised seosed või arvutustabelis (näiteks MS Exceli failis) hoitavate tabelite või väljade seosed.

#### 3.2 Andmekvaliteedi reeglite kirjeldamine

Olles profileerimise käigus saanud esmase ülevaate andmetest ning nende kvaliteedist tuleb jätkata andmekvaliteedi reeglite kirjeldamisega. 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 eesmärgiks on kirjeldada reeglid andmetele tagamaks andmete kasulikkus ja kasutatavus 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, programmikoodi jm kättesaadavat informatsiooni. Seejuures on andmekvaliteedi reeglite kirjeldamisel abiks eelnevalt profileerimise käigus andmete kohta kogutud informatsioon. Kirjeldamist aitab teostada ka käesoleva juhendi seksioonis 3.3.2 toodud andmekvaliteedi probleemide raamistik, mis esitab 21 tüüpilist andmekvaliteedi probleemi ning nende esinemist illustreerivad näited.

Juhendi 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 (data set). 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 hindamine, mille käigus teostatavad mõõtmised näitavad andmete vastavust andmekvaliteedi reeglile, näiteks „3% juhtudest pole väli väärtustatud, seega on andmete täielikkus 97%.“

### 3.3 Andmekvaliteedi hindamine

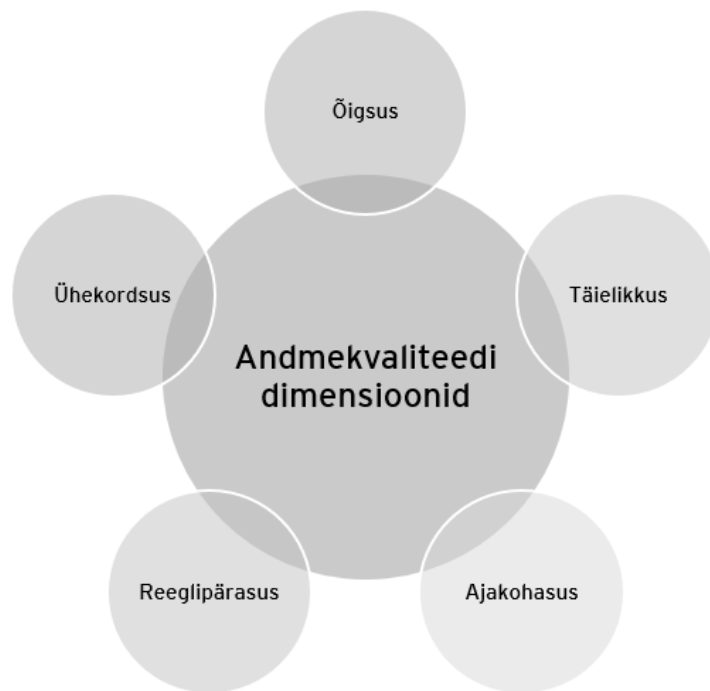
Andmekvaliteedi dimensioonid (Joonis 3) on mõõdetavad andmete omadused, mis võimaldavad kontrollida andmete kvaliteeti erinevatest aspektidest lähtuvalt. Eksisteerib palju erinevaid andmekvaliteedi dimensioonide käsitlusi, kuid antud juhendis keskendutakse 5-le TKTA<sup>1</sup> määruses määratletud kvaliteedidimensioonile ning nende hindamist toetavate indikaatorite kirjeldamisele. Kirjelduste koostamisel on lähtutud DAMA-DMBOK2 andmehalduse mudelist<sup>2</sup> ning andmekvaliteedi protsessi<sup>3</sup> ning mõõtmist<sup>4</sup> käsitlevast kirjandusest.

<sup>1</sup> VV määrus „Teenuste korraldamise ja teabehalduse alused“

<sup>2</sup> <https://dama.org/sites/default/files/download/DAMA-DMBOK2-Framework-V2-20140317-FINAL.pdf>

<sup>3</sup> McGilvray, Danette. Executing Data Quality Projects: Ten Steps to Quality Data and Trusted Information (Tm). 2008.

<sup>4</sup> Batini, Carlo, and Monica Scannapieco. Data and Information Quality. 2016.



Joonis 3. 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 „T6nu“ 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).

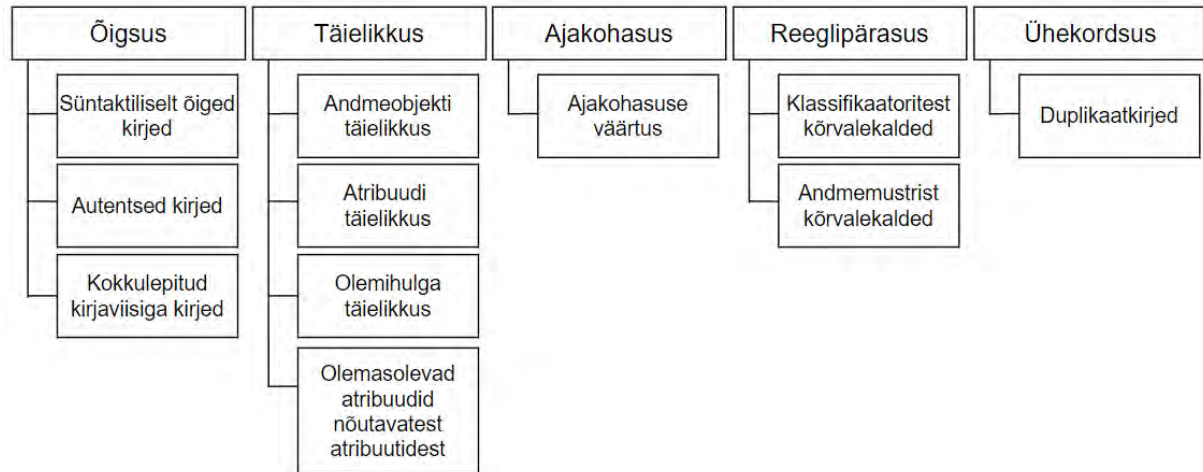
**Täielikkus (Completeness)** näitab, mil määral on olemas kõik nõutud andmed. Täielikkus on vaadeldav kahes osas: kirjade täielikkus ja kogumi ehk populatsiooni täielikkus. Kirjade täielikkus näitab, mil määral on andmekirje kõigil tunnustel olemas väärtused. Näiteks mil määral on andmetabeli veeru 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 reaalse sündmuse 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 4).



Joonis 4. Dimensioonide jaotus indikaatoriteks

### 3.3.1 Andmekvaliteedi hindamise metamudel

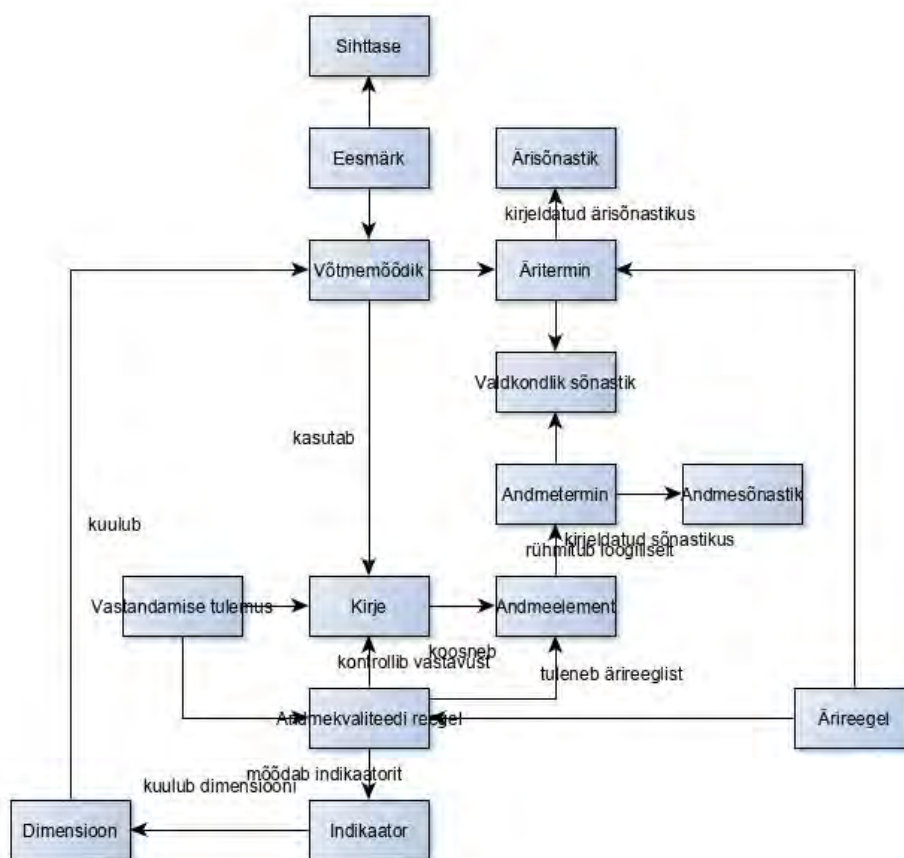
Andmekvaliteeti kasutatakse eri tasemetel otsustamiseks nii andmehalduse enda kui muude asutuse või riigi tegevuste korraldamisel ja täideviimisel. Ühelt poolt kombineeritakse andmekvaliteedi reeglitele vastavused/mittevastavused indikaatorite väärtusteks ja need omakorda dimensioonide väärtusteks ja osad neist saavad 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 vähendada liiklussurmade hulka 50 inimese võrra aastas, tõstetakse liiklusõnnetustega seotud rajatiste ja õnnetuste asukohaga seotud teiste keskkonda iseloomustavate andmete kvaliteet sellisele tasemele, et saab võimalikuks masinõppe rakendamine riskantsete asukohtade tuvastamise ja rajatiste ohutuse mudelite loomisel.
- SC-DQ-2: Selleks, et tõsta maapiirkondade populatsiooni läbi kõrge elukvaliteediga miljööväärtuslike alade info kättesaadavaks tegemise rakenduste kasutamise, selleks tõstetakse Ehitisregistri ehitiste ja nende tehniliste parameetrite andmete semantiliselt õigete kirjete määr 95%-ni.
- SC-DQ-3: Selleks, et tagada riigi kodanike operatiivne teavitamine läbi digikanalite, selleks tõstetakse Eesti kodanike kontakttelefoninumbrite, e-postiaadresside ja kontaktaadresside täielikkus riigiüleselt 99%-ni.
- SC-DQ-4: Selleks, et tagada kodanike kirjete 100% täielikkus andmestikus X, on vaja parandada eesnimi või perekonnanimi 250 kirjes.
- SC-DQ-5: Selleks, et tagada piirkondlikele otsustele ühtlane kvaliteet, on vaja tõsta aadressi kirjete õigsust asutuste X, Y ja Z andmestikes.

- SC-DQ-6: Selleks, et võimaldada elektroonilist rahvaloendust, on vaja ...

Taoliste stsenaariumite täideviimist toetab järgnev andmekvaliteedi mõõtmise metamudel (vt Joonis 5).



Joonis 5. Andmekvaliteedi mõõtmise metamudel

Andmekvaliteedi dimensioonid on osa organisatsiooni võtmemõõdikutest, mis on omakorda kirjeldatud ärisõnastikus. Võtmemõõdikute sihttasemed on määratud organisatsiooni eesmärkides ning võtmemõõdikute arvutamiseks kasutatakse organisatsiooni andmeid.

Andmekvaliteedi reeglid tulenevad ärireeglistest, mis on kirja pandud ärisõnastiku äritermineid kasutades (vt. ka Andmehalduse raamistik ptk. X.X). Samas ei kajastu kõik ärireeglid andmekvaliteedi reeglites ning vastupidi, osad andmekvaliteedi reeglid ei kajastu ärireeglistes.

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 esitatud andmete andmeelementidena (nt *ehr\_kood* on ehitise kirje andmeelement). Vastandamise tulemused kombineeritakse indikaatoriteks ja need omakorda dimensioonideks. Vaikimisi kasutatakse kombineerimise funktsioonina aritmeetilist keskmist, kuid vajadusel võib kasutada ka muid funktsioone kui nendes on üleriiklikult kokku lepitud. Üleriiklikult on vaja kokku leppida neis mõõdikutes, mis liidetakse kokku üleriiklikust andmehalduse olukorrast ülevaate saamiseks ja mida võrreldakse eri organisatsioonide, andmeterminite ja andmekomplektide lõikes.

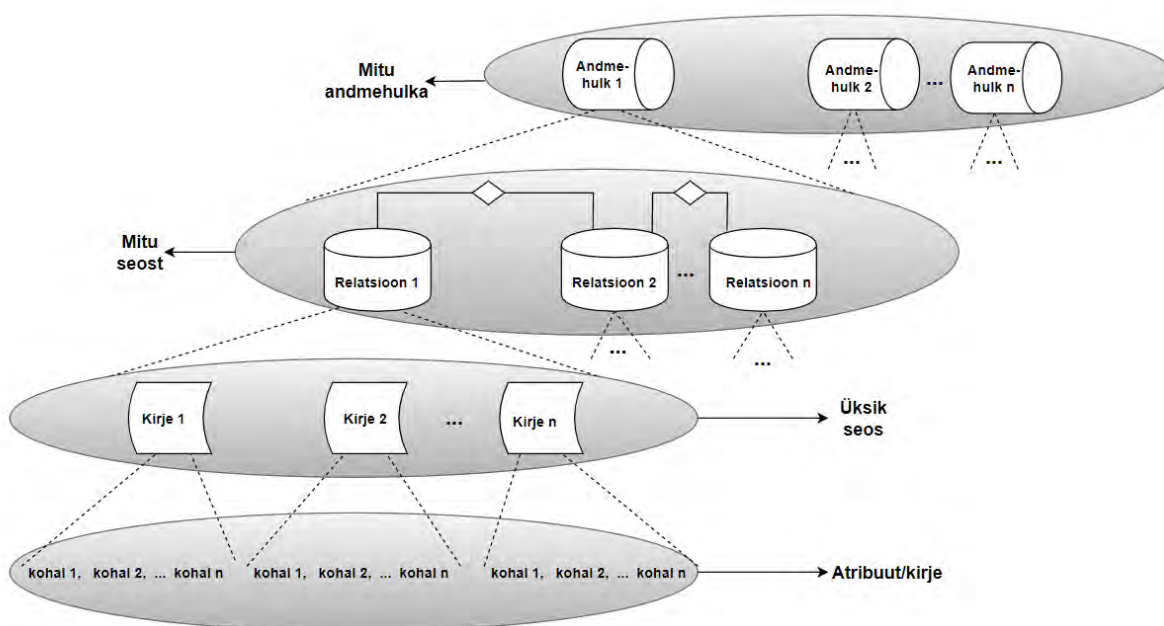
Andmeelemendid on grupeeritud andmeterminiteks ja kirjeldatud andmesõnastikus. Seosed äriterminite ja andmeterminite vahel tekivad läbi valdkondlike ontoloogiatega, kus on kirjeldatud seosed andmeterminite ja äriterminite vahel ning muud detailid, mis võimaldavad seostada äri ja IT.

### 3.3.2 Andmekvaliteedi probleemid

Selleks, et andmekvaliteedi probleeme süsteemselt käsitleda ja katta andmekvaliteedi dimensioonid ühtlaselt andmekvaliteedi reeglitega, pakume välja Paolo jt [ORH05] poolt välja töötatud andmekvaliteedi probleemide raamistiku. Lisaks toome ilmutatud kujul välja samade autorite poolt välja pakutud [PRHG05] andmete ajakohasust puudutavad probleemid. Võrreldes teiste raamistikega [KCH+03, LPK11, LSB15] on Paolo jt lähenemise eelised järgmised:

1. tuvastatud probleemid tulenevad juhtumiuuringust;
2. tegu on formaalse raamistikuga, kus kirjeldatakse algoritmid [PRHG05] andmekvaliteedi probleemide tuvastamiseks ja klassifitseerimiseks.

Raamistik eristab probleeme andmehulkade arvu alusel (üks või rohkem), andmeobjektide vaheliste relatsioonide arvu (üks või rohkem) ning atribuutide/kirjete tasemel (vaata Joonis 6).



Joonis 6. Andmete korraldamise tüüpiline mudel

Andmekvaliteedi probleemid ja nende seosed andmekvaliteedi dimensioonidega on toodud alljärgnevalt (Tabel 1).

Tabel 1. Andmekvaliteedi probleemid ja nende seosed andmekvaliteedi dimensioonidega.

| Andmekvaliteedi probleem           | Atribuut/kirje |       |       | Seosed     |            |                 | Dimensioonid |            |            |               |            |
|------------------------------------|----------------|-------|-------|------------|------------|-----------------|--------------|------------|------------|---------------|------------|
|                                    | Atribuut       | Veerg | Kirje | Üksik seos | Mitu seost | Mitu andmehulka | Õigsus       | Täielikkus | Ajakohasus | Reeglipärasus | Ühekordsus |
| Puuduv väärtus                     | ✓              |       |       |            |            |                 |              | ✓          |            |               |            |
| Süntaksirike                       | ✓              |       |       |            |            |                 | ✓            |            |            |               |            |
| 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                |                |       |       |            |            | ✓               |              |            |            |               | ✓          |

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

#### 3.3.2.1 *Puuduv väärtus*

**Probleemi kirjeldus:** Puudub kohustuslikuks määratud atribuut. Valikuliste väärtuste puudumine ei ole andmekvaliteedi probleem.

**Näide:** Puudub kohustuslik atribuut "ISIKUKOOD."

#### 3.3.2.2 *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.

#### 3.3.2.3 *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.

#### 3.3.2.4 *Väärtusvahemiku rikkumine*

**Probleemi kirjeldus:** Väärtusvahemiku rikkumised avalduvad 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.

#### 3.3.2.5 *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").

#### 3.3.2.6 *Õigekirjaviga*

**Probleemi kirjeldus:** Tekstilise atribuudi väärtus sisaldab õigekirjavigasid.

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

#### 3.3.2.7 *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.

#### 3.3.2.8 *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.

### 3.3.2.9 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.

### 3.3.2.10 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 sama ametit.

### 3.3.2.11 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.

### 3.3.2.12 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 võib olla seotud ühe linnaga. Näiteks rikub funktsionaalset sõltuvust järgnev olukord: (POSTIINDEKS=10118; LINN="TALLINN") ja (POSTIINDEKS=10118 ja LINN="TARTU").

### 3.3.2.13 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.

### 3.3.2.14 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.

### 3.3.2.15 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.

3.3.2.16 *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".

3.3.2.17 *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.

3.3.2.18 *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.

3.3.2.19 *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.

3.3.2.20 *Esitluse mitmekesisus*

**Probleemi kirjeldus:** Erinevad andmestikud kasutavad sama infot talletavate atribuutide jaoks erinevaid väärtuste kogumeid sama informatsiooni esitamiseks.

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

3.3.2.21 *Homonüümide kasutus*

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

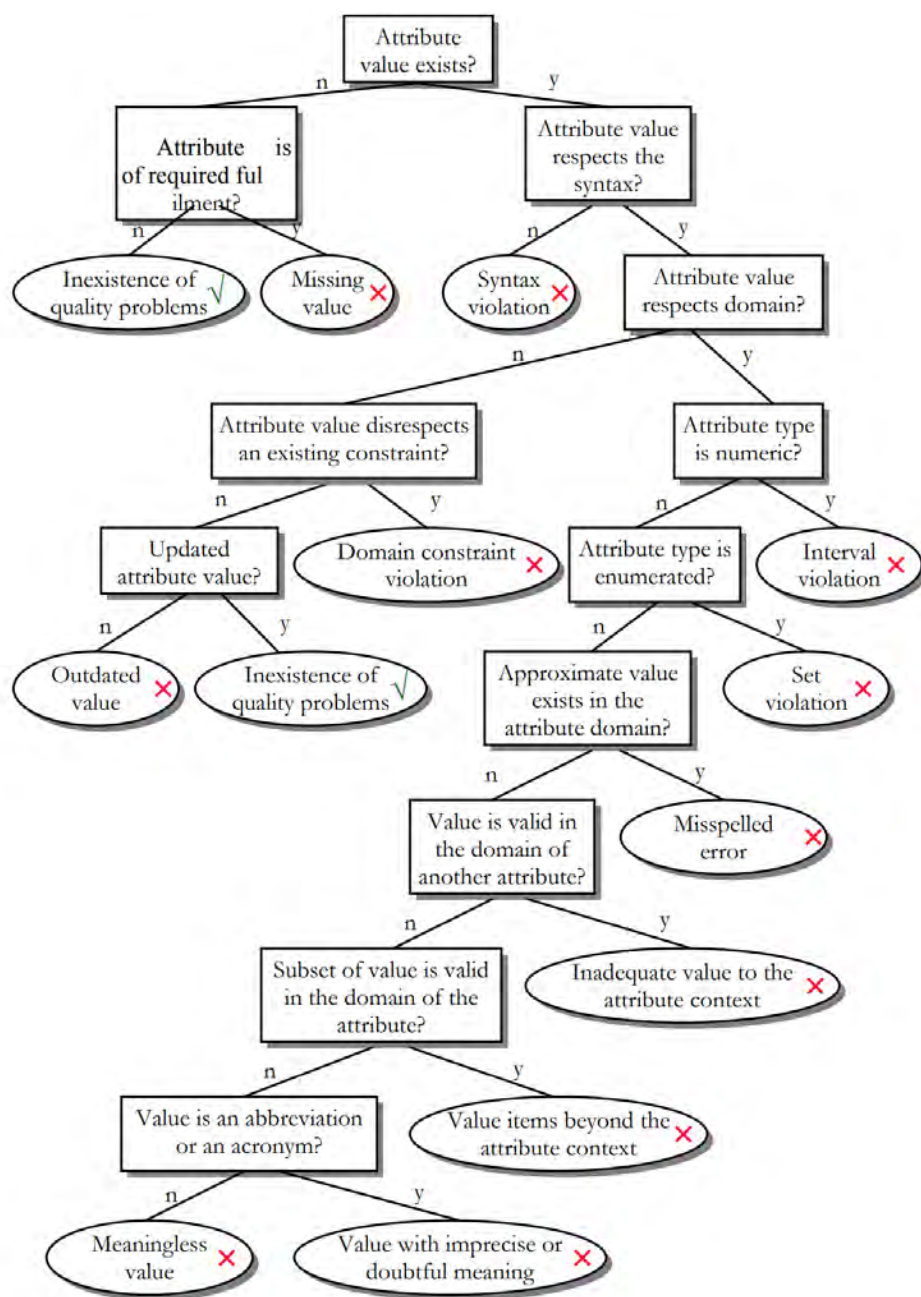
**Näide:** Andmestikus X kasutatakse "TOODE" tabelis terminit "Hiir" tähistamaks arvuti riistvaralist osutusseadet. Andmestikus Y kasutatakse "TOODE" tabelis terminit "Hiir" tähistamaks kodulooma.

## 3.3.3 Kvaliteediprobleemide tuvastamine

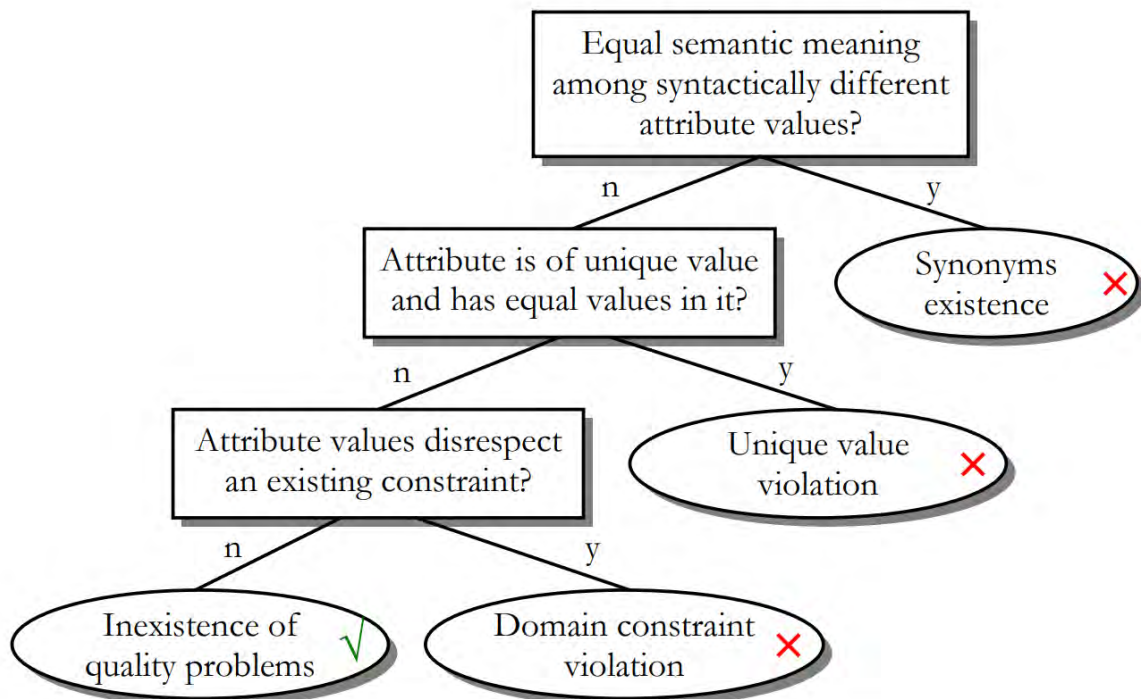
Siin sektsioonis esitame Paulo jt [PRHG05] poolt väljatöötatud meetodid andmekvaliteedi probleemide tuvastamiseks (identifitseerimine ja klassifitseerimine). Iga meetod on esitatud binaarse otsustuspuuna, mis võimaldab identifitseerida ja klassifitseerida probleeme andmeallikate hulga, andmeobjektide vaheliste relatsioonide hulga ja atribuutide/kirjete suhtes. Kvaliteediprobleemid on märgitud otsustuspuude lõppsõlmedes.

<alljärgnevad joonised on vaja tõlkida ja lisada täiendavad selgitused>

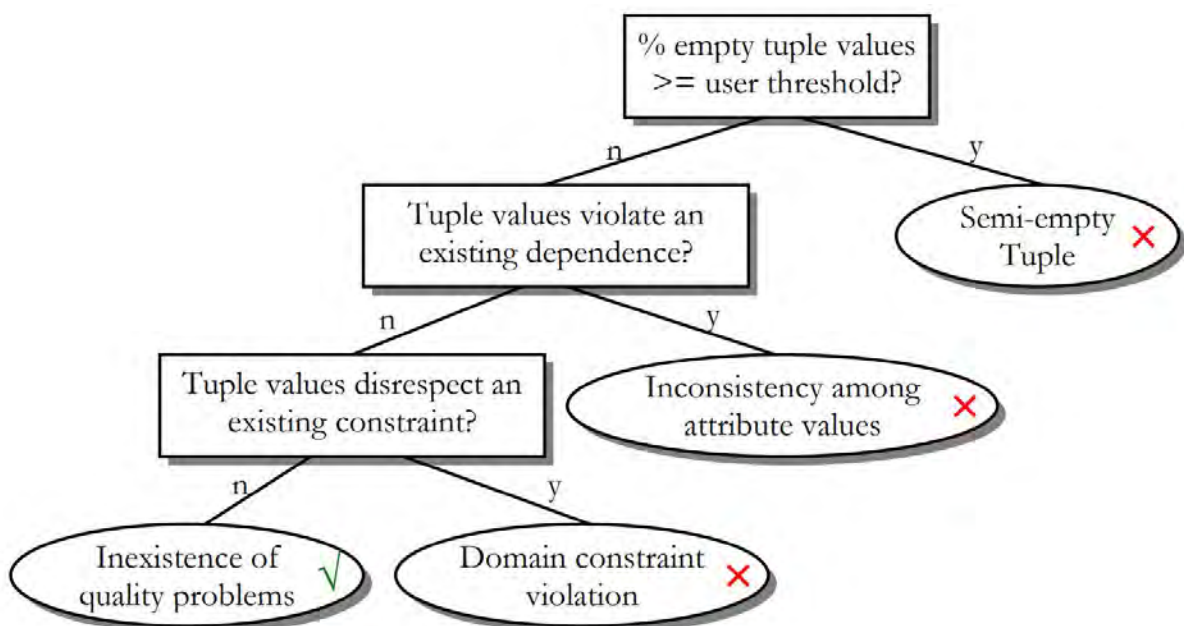
## Üksiku andmehulga üksiku kirje atribuudi väärtus



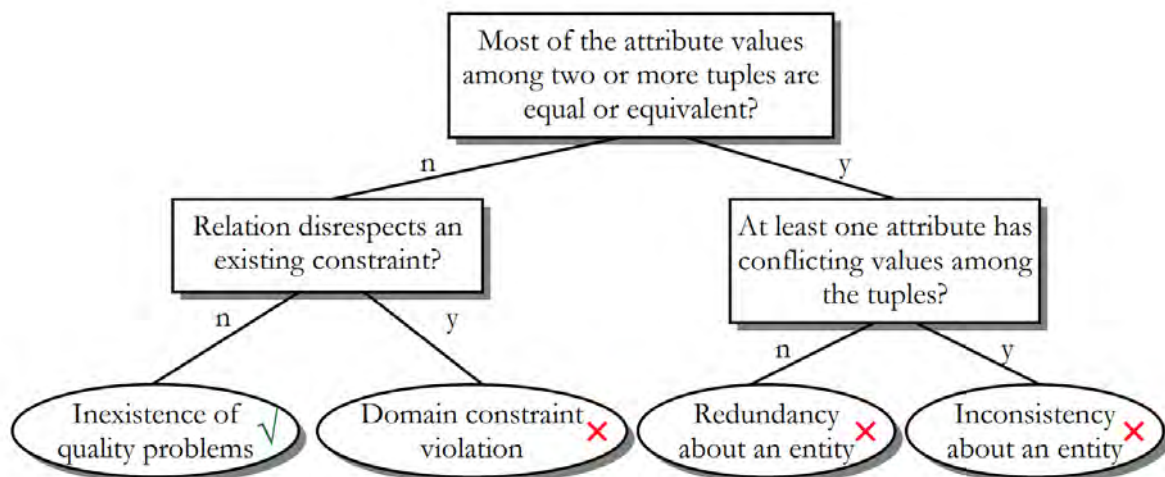
Üksiku andmehulga atribuudi väärtus



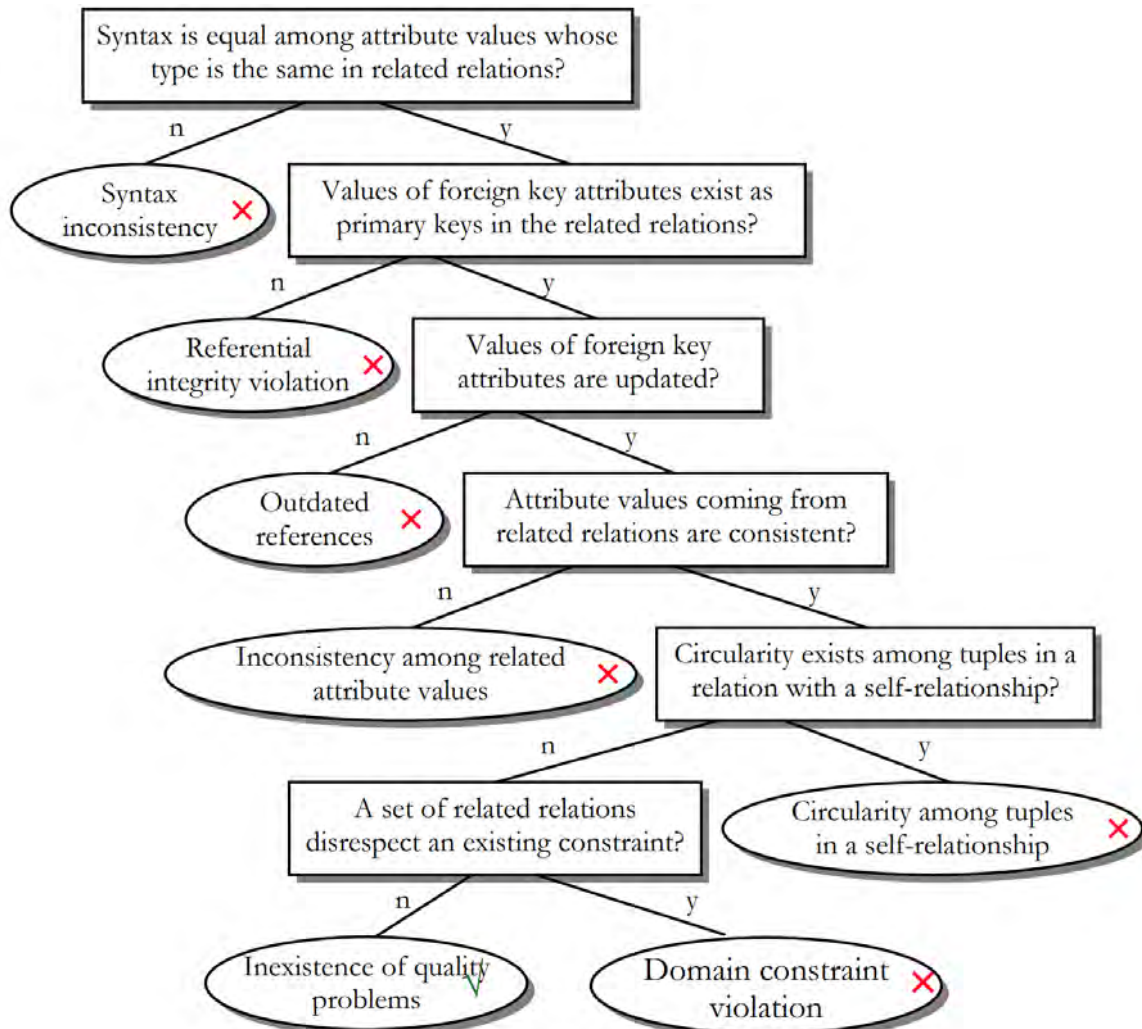
Üksiku andmehulga üksiku kirje atribuutide väärtused



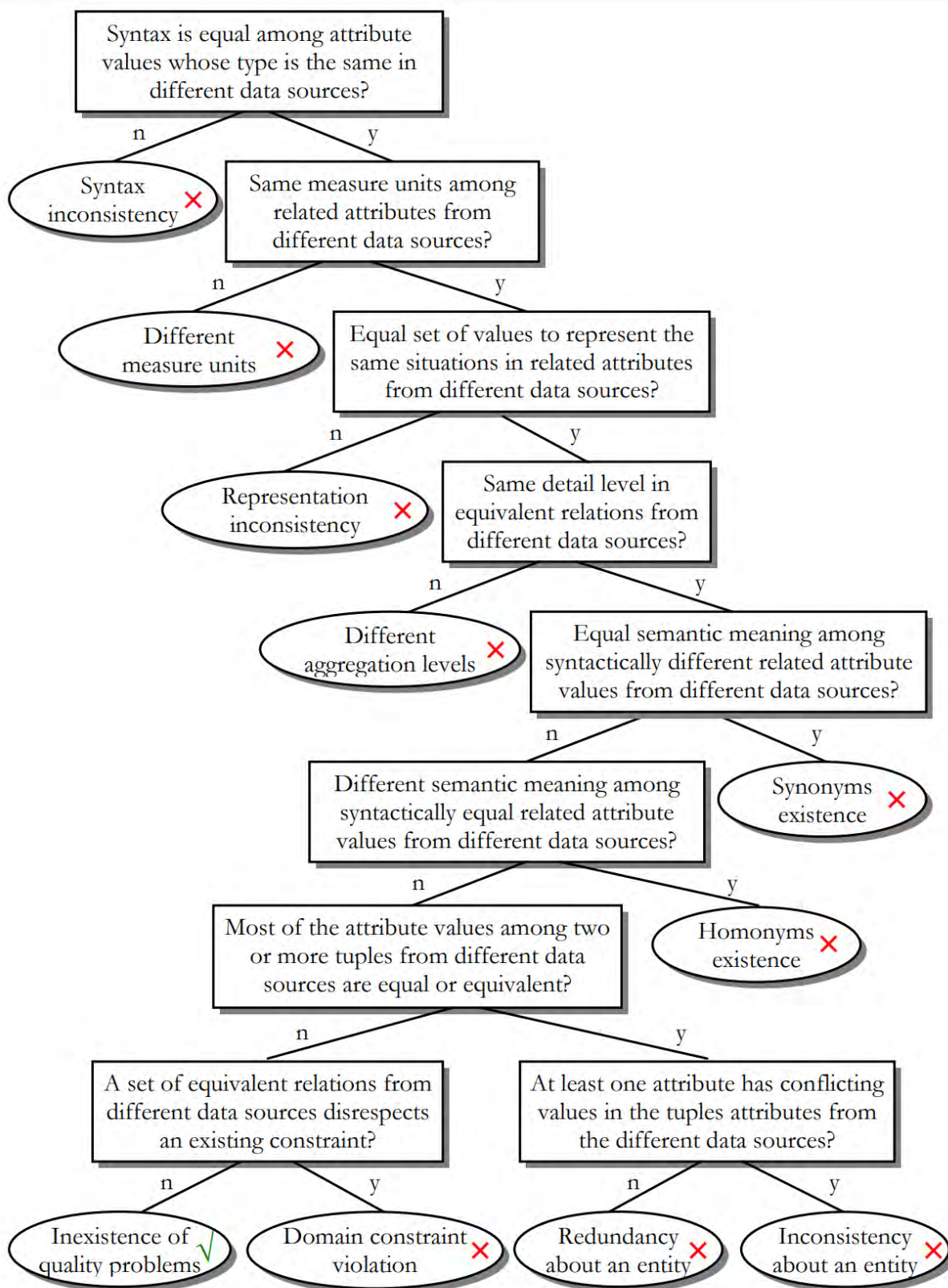
Üksiku andmehulga mitme kirje atribuutide väärtused



Mitu seost





3.3.3.1 *Mitu andmehulka*

### 3.3.4 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.

#### 3.3.4.1 Õigsus (Accuracy)

Nagu eelpool mainitud näitas õigsus andmete vastavust tegelikkusele ning seda sai jagada 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 võimalik). 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 vöörkeelsed nimed ja nimetused (näiteks “Ельцин” transliteeritud kuju on “El’cin”) või kas lühendeid on kasutatud läbivalt kõigis kirjetes. Tegu on osaga semantilisest õigsusest.

### 3.3.4.2 Täielikkus (Completeness)

Täielikkus näitab, mil määral on olemas kõik nõutud andmed. Täielikkust on võimalik täpsemalt kirjeldada toetudes ühele konkreetsele andmemudelile, mistötu 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 ( $1\,000\,000/1\,300\,000=0,8$ ).

Vöttes 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 (**Error! Reference source not found.**) 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 (**Error! Reference source not found.**) 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 (**Error! Reference source not found.**) 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%.

### 3.3.4.3 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 (**Error! Reference source not found.**).

Võrrand 1. Värskus (*Currency*)

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

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

Volatiilsus on defineeritud kui andmete kehtivuse periood. Sellest tulenevalt saame ajakohasuse defineerida järgmiselt (**Error! Reference source not found.**).

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 (**Error! Reference source not found.**).

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 (**Error! Reference source not found.**) 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 (**Error! Reference source not found.**)  $\max(0, 1 - (5/1440)) = 1 - 0.003 = 0.997$ . 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.

#### 3.3.4.4 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.

#### 3.3.4.5 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 on 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 emaille 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.

### 3.3.5 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.

| 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.<br><br>Numbreid sisaldavate nimede arv: 100<br>Isikute koguarv: 5 000                                              | $(5000-100)/5000$<br>$*100 = 98\%$<br><br>98% isikute nimedest ei sisaldab numbreid.                                | 100%     |
|            | Autentsed kirjed                 | Isiku elukoht peab vastama rahvastikuregistri andmetele.                                           | Teostada isikute ja elukohtade väljavõte rahvastikuregistrist ning võrrelda seda hinnatavate andmetega.<br><br>Rahvastikuregistriga mitte kattuvate elukohaandmete arv: 10<br>Kirjete koguarv: 1000 | $(1000-10)/1000$<br>$*100 = 99\%$<br><br>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).<br><br>Ebakorrektsete transkriptsioonide arv: 450<br>Venekeelsete nimede koguarv: 500         | $(500-450)/500*100 = 10\%$<br><br>10% teostatud transkriptsioonidest järgib korrektset vene-eesti transkriptsiooni. | 100%     |

| 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.<br><br>Väärtustatud väljade arv: 4<br>Andmeid talletavate väljade arv: 5 | $4/5 * 100 = 80\%$<br><br>80% konkreetse isiku kohta käivatest andmetest on väärtustatud. | 100%     |
|            | Atribuudi täielikkus                              | Kõigil isikutel peab olema telefoninumber                                                        | Loendada kokku isikute arv, kus tunnus „Telefoninumber“ on tühi ning isikute koguarv.<br><br>Telefoninumbrita isikute arv: 30<br>Isikute koguarv: 150                                                | $30/150 * 100 = 20\%$<br><br>80% isikutest on telefoninumber olemas.                      | 100%     |
|            | Olemihulga täielikkus                             | Kõik isiku kohta käivad andmed peavad olema väärtustatud.                                        | Loendada kokku isiku tabeli väärtustatud väljade arv ning väljade koguarv.<br><br>Isiku tabeli väärtustatud väljade arv: 600<br>Isiku tabeli väljade koguarv: 1000                                   | $600/1000 * 100 = 60\%$<br><br>60% isiku kohta käivatest andmeväljadest on väärtustatud.  | >80%     |
|            | Olemasolevad atribuudid nõutavatest atribuutidest | Isiku kohta peab olema võimalik talletada eesnime, perekonnanime, isikukoodi ja rahvuse andmeid. | Leida olemasolevate atribuutide arv ning võrrelda seda nõutavate atribuutide arvuga.<br><br>Olemasolevate atribuutide arv (eesnimi, perekonnanimi, isikukood): 3<br>Nõutavate atribuutide arv: 4     | $3/4 * 100 = 75\%$<br><br>75% nõutud atribuutidest on olemas.                             | 100%     |

| Dimensioon | Indikaator          | Reegel                                                                           | Meede                                                                                                             | Hetketase                | Sihttase |
|------------|---------------------|----------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------|--------------------------|----------|
| Ajakohasus | Ajakohasuse väärtus | Riigieksamite tulemused peavad olema kasutatavad enne sisseastumisperioodi lõppu | Arvutada ajakohasuse väärtus vastavalt ajakohasuse definitsioonile ( <b>Error! Reference source not found.</b> ). | Ajakohasuse väärtus: 0.9 | >0.8     |

| 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.<br><br>EMTAK-ist kõrvalekallete arv: 200<br>EMTAK koodide arv: 1771                                                             | 200/1771 * 100 = 11%<br><br>11% 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 isikukoodi sisaldavate väljade koguarv.<br><br>Andmemustrile mittevastavate isikukoodide arv: 400<br>Isikukoodide koguarv: 700 000 | 400/700 000 * 100 = 0.05%<br><br>0.05% isikukoodidest ei vasta nõuetele           | 0%       |

| Dimensioon | Indikaator      | Reegel                               | Meede                                                                                                                         | Hetketase                                                                        | Sihttase |
|------------|-----------------|--------------------------------------|-------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------|----------|
| Ühekordsus | Duplikaatkirjed | Iga postiindeks peab olema unikaalne | Loendada kokku duplikaatide arv ning postiindeksite koguarv.<br><br>Duplikaatide arv: 10 000<br>Postiindeksite arv: 1 000 000 | 10 000/<br>1 000 000*100 = 1.0%<br><br>1% postiindeksi kirjetest on duplikaadid. | 0%       |

### 3.3.6 Andmekvaliteedi aruandepõhjade väljatöötamine

<Lisandub järgmises versioonis>

### 3.3.7 Kvaliteediprobleemide prioritseerimine

On tõenäoline, et eelnevalt kirjeldatud kvaliteedidimensioonide hindamise käigus tuvastati mitmeid erinevaid kvaliteediprobleeme. Esmalt tuleks põhjalikumalt tegeleda vaid kõige olulisemate kvaliteediprobleemide edasise analüüsi ja probleemide lahendamiseks. Kvaliteediprobleemi olulisus sõltub äri vajadustest ehk esmalt tuleks lahendada äriliselt kõige olulisemad probleemid. Seetõttu tuleks kvaliteediprobleemide esmasel prioriseerimisel kasutada mõnd ärilise mõju hindamise tehnikat. 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 konkreetset andmed. Kasutada võib ka viis korda „Miks?“ küsimise tehnikat eesmärgiga jõuda kvaliteediprobleemi tegeliku põhjuni. Põhjalikum mõju hindamine toimub protsessi hilisemas faasis pärast andmekvaliteedi juurpõhjuste analüüsi. Siis toimub ka KPI-de ja andmekvaliteedi sidumine.

## 3.4 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.

Juurpõhjuste analüüs võib vajalikuks osutuda pärast andmekvaliteedi mõõtmist, et selgitada välja tuvastatud kvaliteediprobleemi algallikas. Sellisel juhul peaks enne juurpõhjuste analüüsi olemasolevat 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 raames, 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. Juurpõhjuse tuvastamiseks võib olla vajalik elutsükli detailne uurimine. Alljärgnevalt on kirjeldatud kolme 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.

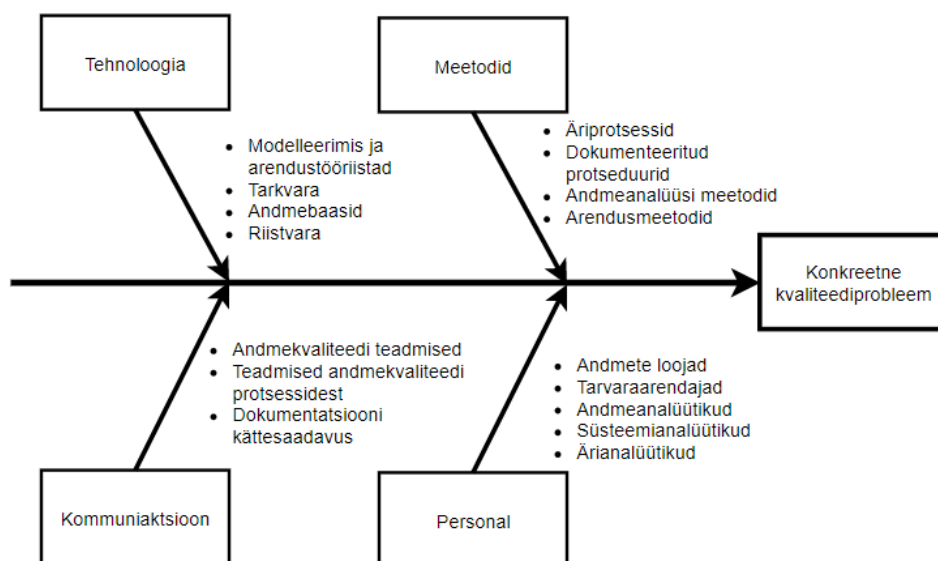
Nimetatud kolm meetodit andmekvaliteedi juurpõhjuste analüüsimiseks on järgmised:

- 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 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 juurpõhjust, soovitusi nende lahendamiseks, meetodi käigus läbitud järeldusteni jõudmise protsessi ning tuvastatud otseseid ja kaudseid ärilisi mõjusid.

Jälitamise (*Track and Trace*) meetodit saab kasutada tuvastamiseks probleemi esmakordset ilmnemist informatsiooni elutsükli. Nagu ka eelnevalt kirjeldatud meetodi puhul on esimeseks sammuks tuvastatud kvaliteediprobleemi selge sõnastamine. Järgmiseks tuleb kindlaks määrata informatsiooni elutsükkel 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 juurpõhjust, soovitusi nende lahendamiseks, meetodi käigus läbitud järeldusteni 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 liikmetega ajurünnak või kasutada kombinatsiooni mitmest nimetatud viisist. Koosoleku liikmete küsitlemist tuleb jätkata kuni on jõutud probleemi juurpõhjuseni. Koosolekul leitud juurpõhjuste tuleb dokumenteerida nende edasiseks kasutamiseks. Tuvastatud juurpõhjuste põhjal tuleb seejärel koostada soovitused kvaliteediprobleemi juurpõhjuste lahendamiseks. Viimaks tuleb taas kõik tulemused dokumenteerida. Dokumentatsioon peaks kindlasti kirjeldama juurpõhjuste, soovitusi nende lahendamiseks, meetodi käigus läbitud järeldusteni jõudmise protsessi ning tuvastatud otseseid ja kaudseid ärilisi mõjusid.



Joonis 7. Põhjus ja tagajärg diagrammi näidis.

### 3.5 Kvaliteediprobleemide mõju hindamine

Peale andmekvaliteedi probleemide tuvastamist ning probleemide juurpõhjuste analüüsi tuleb identifitseerida ja prioriseerida probleemide lahendamiseks tehtavad muudatused. Selleks tuleb esmalt teostada kvaliteediprobleemide ärilise mõju hindamine. Teostatud analüüsi tulemused tuleb edastada andmeomanikule andmekvaliteedi parendamise 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 äriliselt olulised. Keskenduda tuleks vaid äriliselt 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 tuleks 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 parendamiseks, ülevaadet kvaliteediprobleemide juurpõhjustest ning muid analüüsi käigus ilmnenud olulisi tulemusi. Koostatud ülevaade tuleb edastada andmeomanikule andmekvaliteedi parendamise 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 äriks, 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 gruppis 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 asetus 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 8).

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

Joonis 8. 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. Jälgimise 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 arutellu 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 3)

Tabel 3. Hindamise ja prioriseerimise näidistabel.

| Andmekvaliteedi mõju hindamine 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 kaasna 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 määratleda 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. Investeeringu tasuvuse (ROI) valem.Võrrand 3). Investeeringust tulenev tulu ja alginvesteering tuleb määratleda 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.

Andmekvaliteedi mõju on võimalik vaadelda ka seoses asutuse võtmeindikaatoritega (*KPI ehk Key Performance Indicator*). Andmekvaliteedi dimensioone ja võtmeindikaatorite seoseid tihti koos ei vaadata. Samas on üheks peamiseks võtmeindikaatorite juurutamise läbikukkumise põhjuseks puudulik andmete kvaliteet. Lisaks on andmekvaliteedist kasu kui see võimaldab toetada organisatsiooni teisi tegevusi peale mõõtmise. Andmehaldusega seotud kulusid on lihtsam planeerida kui on selge seos andmekvaliteedi ja selle mõjule asutuses ja riigis.

## 3.6 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 mõistetavus. 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 tohi tuvastada arvamise teel. Reegleid tuleb testida reaalsel andmetel. Tihti toob selline testimine välja andmetes eksisteerivad probleemid ning aitab tuvastada ka andmekvaliteedi reeglites eksisteerivaid puuduseid.
- **Valdkonna eksperdi poolt heakskiidetud.** Andmekvaliteedi reeglite eesmärgiks on näidata millised andmed peavad olema. Tihti on reeglite õigeks kirjeldamiseks vajalikud teadmised konkreetse ärivaldkonna protsessidest. Neid teadmisi tuleks koguda konkreetse valdkonna eksperdilt, kelle ülesandeks oleks kinnitada kirjeldatud ärireeglid või selgitada andmeanalüüsi tulemusi.
- **Andmete kasutajatele kättesaadavad.** Kõigil andmete kasutajatel peaks olema juurdepääs dokumenteeritud andmekvaliteedi reeglitele. Juurdepääsu 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.

Nagu eelnevalt kirjeldatud aitas andmete profileerimine ja analüüs tuvastada andmekvaliteedi reeglid. 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.

### 3.6.1 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 millisest 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.

### 3.6.2 Andmeelementide kaetus reeglitega

Et saada ülevaadet andmeelementide kaetusest andmekvaliteedi reeglitega peab see olema dokumenteeritud. Dokumentatsioon võib olla esitatud näiteks tabelina (Tabel 4). Toodud tabelile toetudes on võimalik tuvastada andmeelemendid mille kvaliteeti ei mõõdeta, 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 juhendi rakenduslikus osas (näites N10).

Tabel 4. Andmeelementide reeglitega kaetuse näidistabel.

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



## 4. Rakendamise näited

### 4.1 Andmekvaliteedi mõõtmise algoritmide näited

Järgnevad andmekvaliteedi mõõtmise näited kasutavad *validate* [LJ19] paketti, mis on arendatud Mark van der Loo ja Edwin de Jonge poolt.

*validate* teek on realiseeritud R-i jaoks andmete kvaliteedi mõõtmise lihtsustamiseks samal ajal hoides seda hallatavana ja taasesitamist võimaldavana. Teek võimaldab:

- testida andmekomplekti vastu eeldefineeritud reeglite, 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.

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

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

#### 4.1.1 Näidisjuhtum (Riikliku ehitisregistri avaandmed)

Ehitisregistri avaandmed on pärit lehelt <https://opendata.riik.ee/andmehulgad/ehitisregister/>.

Andmekvaliteedi mõõtmiseks loeme esmalt sisse andmed

```
filename <- "ehitis_katasrtiysus_1994-01_2020-02-09.csv"
ehitis_katastriysus <- fread(filename, encoding="UTF-8")
```

<Peep: lisa ehitise andmemudel ka selgituseks>

#### Näide N0 (täielikkus, atribuudi täielikkus)

Dimensioon: täielikkus

Indikaator: atribuudi täielikkus

Reegli kirjeldus: Kõikidel tabeli kirjetel peab olema identifikaator

Selgitus: Ehitisregistri tabelis eh\_ehitised on unikaalseks identifikaatoriks veerg id. Täielikkuse testimiseks loome reeglite komplekti completeness\_rules, kuhu lisame reegli exists\_id (kontrollib kas andmeobjektidel on identifikaator):

```
completeness_rules <- validator(
```

```
exists_id = !any(is.na(id))
)
```

#### Näide N1 (ühekordsus, kirjete unikaalsus)

Dimensioon: ühekordsus

Indikaator: kirjete unikaalsus

Reegli kirjeldus: Kõikidel tabeli kirjetel peab olema identifikaator, mis on unikaalne

Selgitus: Ehitisregistri tabelis eh\_ehitised on unikaalseks identifikaatoriks veerg id. Unikaalsuse testimiseks loome reeglite komplekti uniqueness\_rules, kuhu lisame reegli unique\_id (kontrollib kas andmeobjekti identifikaator on unikaalne):

```
uniqueness_rules <- validator(
  unique_id = all(is_unique(id))
)
```

#### Näide N2 (reeglipärasus, kõrvalekalded klassifikaatorist)

Dimensioon: reeglipärasus

Indikaator: kõrvalekalded klassifikaatorist

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

Selgitus: Ehitisregistri tabelis eh\_ehitised võivad elemendi omandi liik väärtused olla järgnevad:

EHITIS\_OMANDI\_LIIK\_VALLAS, EHITIS\_OMANDI\_LIIK\_KINNIS, EHITIS\_OMANDI\_LIIK\_VALLASRO .

Elemendi 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 .

Selleks kirjeldame vastavalt reeglid unique\_type ja state\_set:

```
ehitised_rules <- validator(
  unique_type = omandi_liik %in% c('EHITIS_OMANDI_LIIK_VALLAS',
    'EHITIS_OMANDI_LIIK_KINNIS')
, state_set = 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')
)
```

Rakendame reeglit

```
check <- confront(eh_ehitised, ehitised_rules)
```

Vaatame tulemuste kokkuvõtet

```
> summary(check)
name      items  passes fails nNA error warning
1 unique_type 1058629 1056570 2059   0 FALSE  FALSE
2 state_set   1058629 1058628     1   0 FALSE  FALSE
```

Väljavõtte näitab, et 1058629-st kirjest ei ole vastavalt 2059-s ja 1-s kirjes omandi liik ja seisund õigesti kodeeritud.

Vaatame lähemalt milline kirje ei vasta reeglile

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

```
      id      name value
1: 1268480 unique_type FALSE
2: 1268482 unique_type FALSE
3: 1650270 unique_type FALSE
4: 5537838 unique_type FALSE
5: 5628359 unique_type FALSE
---
2056: 5626874 unique_type FALSE
2057: 5625965 unique_type FALSE
2058: 5626720 unique_type FALSE
2059: 5626503 unique_type FALSE
2060: 5622695   state_set FALSE
```

### Näide N3 (õigsus, süntaktiliselt õiged kirjed)

Dimensioon: õigsus

Indikaator: süntaktiliselt õigete kirjete protsent

Reegli kirjeldus:

Selgitus: Kontrolli kas element esmane\_kasutus on on neljakohaline aastanumber kasutades mustreid.

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

Rakendame reeglit

```
check <- confront(eh_ehitised, pattern_rules)
```

Vaatame tulemuste kokkuvõtet

```
> summary(check)
      name      items  passes  fails nNA error warning
1 pattern_year 1058629 435294 623335   0 FALSE  FALSE
...

```

### Näide N4 (õigsus, semantiliselt õiged kirjed)

Dimensioon: õigsus

Indikaator: semantiliselt õiged kirjed

Osad reeglid kirjeldavad kitsendusi erinevate andmeveergude vahel. Need kitsendused on kasulikud ärireeglite esitamisel.

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

Selgitus: Määruses "Ehitise tehniliste andmete loetelu ja pindade arvestamise alused"

(<https://www.riigiteataja.ee/akt/107102014003>) 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 mittekanvate tarindite alune pind;
- § 22(2) Hoone suletud netopind on kõigi korruste suletud netopindade summa.

Kontrollime, 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
)
```

Rakendame reeglit

```
check <- confront(eh_ehitised, conditional_rules)
```

Vaatame tulemuste kokkuvõtet

```
> summary(check)
      name  items passes fails  nNA error warning
1 condition1 1058629 993968 10145 54516 FALSE  FALSE
...

```

#### Näide N5 (reeglipärasus, kõrvalekalded klassifikaatorist)

Dimensioon: reeglipärasus

Indikaator: kõrvalekalded klassifikaatorist

Reegli kirjeldus: Omavalitsus on seotud vaid ühe maakonnaga.

Selgitus: Kasutame selle mõõtmiseks reeglit, mis kirjeldab funktsionaalset sõltuvust omavalitsuse ja maakonna vahel

```
dependency_rules <- validator(
  dependency1 = omavalitsus ~ maakond
)
```

Rakendame reeglit

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

Vaatame tulemuste kokkuvõtet

```
> summary(check)
      name  items passes fails nNA error warning
1 dependency1 1058629 1002331 56298  0 FALSE  FALSE

```

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

Vaatame lähemalt milline kirje ei vasta reeglile

```
> output <- as.data.frame(check)
> output[output$value == FALSE,]
      id      name value      expression
1: 2303468 dependency1 FALSE omavalitsus ~ maakond
2: 2302324 dependency1 FALSE omavalitsus ~ maakond
---
56297: 5625860 dependency1 FALSE omavalitsus ~ maakond
56298: 5626247 dependency1 FALSE omavalitsus ~ maakond
```

Vaatame viimast kirjet lähemalt:

```
> eh_ehitised[eh_ehitised$id == 5626247, c('id', 'nimetus', 'maakond',
'omavalitsus')]
0809      Tori vald      0068      Pärnu maakond
```

Vaatame millist maakonna koodi kasutatakse veel sama omavalitsuse kirjetes

```
> eh_ehitised[eh_ehitised$omavalitsus == 809, c('id', 'nimetus',
'maakond', 'omavalitsus')]
      id      nimetus maakond omavalitsus
1: 3162780      alajaam      67          809
2: 2501880      Elamu      68          809
---
15591: 5625477      Aiamaja      68          809
15592: 5626247 Üksikelamu      68          809
```

Omavalitsuse 809 (Tori vald) maakonnaks on andmetes 67 ja 68, peab olema 68 (Pärnu maakond).

...

#### Näide N6 (õigus, süntaktiliselt õigete kirjete protsent)

Dimensioon: õigus

Indikaator: süntaktiliselt õigete kirjete protsent

Reegli kirjeldus: testi kas kuupäevad (eh\_alust\_kp ja date\_created) on andmetes õigesti esitatud.

Selgitus: ...

```
date_rules <- validator(
  start_date = is.Date(eh_alust_kp),
  created_date = is.Date(date_created)
)
```

Rakendame reeglit

```
check <- confront(eh_ehitised, date_rules)
```

Vaatame tulemuste kokkuvõtet

```
summary(check)
```

...

**Näide N7 (õigsus, semantiliselt õiged kirjed)**

Dimensioon: õigsus

Indikaator: semantiliselt õiged kirjed

Reegli kirjeldus: Hoone suletud netopind on kõigi korruste suletud netopindade (kasuliku pinna) summa.

Selgitus: Määruses "Ehitise tehniliste andmete loetelu ja pindade arvestamise alused"

(<https://www.riigiteataja.ee/akt/107102014003>) on defineeritud hoone suletud netopind järgnevalt:

- § 22(1) Korruse suletud netopind ehk kasulik pind ...
- § 22(2) Hoone suletud netopind on kõigi korruste suletud netopindade summa.

Kontrollime kas hoone suletud netopind (kasulik\_pind) on õigesti arvutatud kui korruste suletud netopinnad on teada.

```
domain_rules <- validator(
  hoone_suletud_netopind := <valem siia>
  , hoone_suletud_netopind == kasulik_pind
)
```

Rakendame reeglit

```
check <- confront(eh_ehitised, x_rules)
```

Vaatame tulemuste kokkuvõtet

```
summary(check)
```

...

**Näide N8 (õigsus, semantiliselt õiged kirjed)**

Dimensioon: õigsus

Indikaator: semantiliselt õiged kirjed

Reegli kirjeldus:

Selgitus: Kasutame korrelatsiooni tabeli veergude vahel tuvastamiseks kas veergude vahel on soovitud seos. Ehitisregistri vaha otsida need näited

```
correlation_rules <- validator(
  correlation1 = cor(maakond, omavalitsus) > 0.9
)
```

Rakendame reeglit

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

Vaatame tulemuste kokkuvõtet

```
summary(check)
```

...

**Näide N10 (metataseme dimensioon x, metataseme indikaator x)**

Reegli kirjeldus: Reeglid peavad katma kõik mittetehnilised andmetabeli veerud.

Selgitus: Analüüsime mil määral olemasolevad kvaliteedikontrolli reeglid katavad tabeli veerud.

Esmalt kombineerime reeglid

```
rules <- dependency_rules + dependency_rules + correlation_rules +
conditional_rules + pattern_rules + uniqueness_rules +
ehitised_rules + date_rules
```

Esitame reeglite ja veergude (allpool on toodud välja 4 esimest) kaetuse maatriksina

```
> variables(rules, as='matrix')
rule          omavalitsus maakond koetav_pind kasulik_pind
dependency1    TRUE      TRUE      FALSE      FALSE
correlation1   TRUE      TRUE      FALSE      FALSE
condition1     FALSE     FALSE      TRUE      TRUE
pattern_year   FALSE     FALSE     FALSE     FALSE
```

Esita veerud, mille kohta pole reegleid

```
> names(eh_ehitised)[!names(eh_ehitised) %in% variables(rules)]
[1] "ehit_id"                "ehr_kood"
[3] "kaos_id_peamine"        "tahis"
[5] "nimetus"                "rajatis_hoone"
...
[49] "abiruumide_pind"        "toimiku_nr"
[51] "sygavus"
```

// kirjeldatks näidisjuhtumid ja konkreetseid SQL päringud või R protseduurid andmekvaliteedi mõõtmise algoritmide realiseerimiseks.

## 4.2 Andmekvaliteedi juhtimislaua mall

// luuakse Google Data Studio's avalik juhtimislaua, mis visualiseerib andmekvaliteedi indikaatoreid.

Vt <https://datastudio.google.com/u/0/reporting/78f6718d-f678-4a1f-bd77-5d3f260714c5/page/FK09>

## 4.3 Soovituslikud töövahendid

// kirjeldatakse valik töövahendeid erinevate andmekvaliteedi mõõtmise tööloikude realiseerimiseks.

Viited

[LJ19] <https://cran.r-project.org/web/packages/validate/>, ...