EESTI
STATISTIKA

ANDMEHALDUSE JUHISED

Andmekvaliteedi juhis

Mai 2023

Versioon 1.3

Kommentaarid juhise kohta on oodatud:
Statistikaamet (andmehaldus@stat.ee),
Majandus- ja Kommunikatsiooniministeerium (andmed@mkm.ee).

Dokumendi ajalugu

ver	muutuse sisu	autor	kuupäev
1.1	August 2020 tehtud juhise sisu ülekandmine	Veiko Berendsen	03.03.2022
1.2	Toimetamine	Veiko Berendsen	nov-dets 2022
1.3	Osade lisamine, struktuurimuutused	Veiko Berendsen	märts 23
1.8	Töörühmale saatmine tagasisideks	Veiko Berendsen	mai 2023
1.9			
2.0	*uus versioon*		

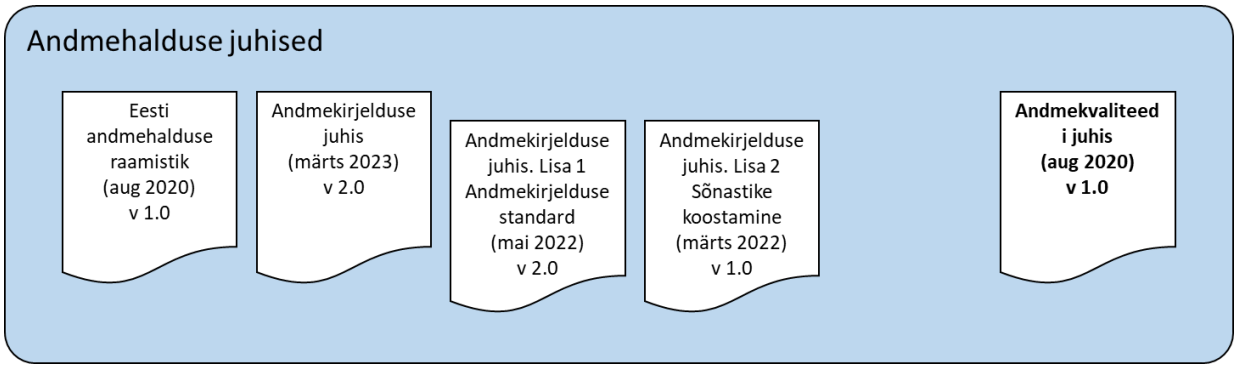
Sisukord

1 Sissejuhatus

1.1 Üldist

Statistikaamet koostööpartneritega koostab ja annab andmehalduse alaseid soovituslike juhiseid. Juhised moodustavad metoodilise ja praktilise raamistiku, kuidas asutuses ja üleriigiliselt andmehaldust korraldada. Käesolev andmekvaliteedi juhis on selle andmehalduse raamistiku ehk andmehaldust korraldava süsteemi osa.

Andmehalduse juhised (Joonis 1) on mõeldud asutustele andmehalduse sisse viimiseks ja korraldamiseks, samuti abistamaks andmehalduse sisulisi tegevusi.



Joonis 1: Andmehalduse juhised

Andmehaldus on organisatsiooni või asutuse tegevusvaldkond, mis võimaldab tal hallata oma andmeid varana. Tõhusast andmehaldusest organisatsioonile tulenev kasu seisneb andmetest täiendava väärtuse loomises ja paremate juhtimisotsuste langetamises. Nimetatud kasu on aga võimalik saavutada vaid siis, kui organisatsiooni andmed on kvaliteetsed. Seejuures on andmekvaliteedi haldamise eelduseks andmekirjelduse olemasolu. Andmekirjelduste koostamist on täpsemalt käsitletud andmekirjelduse juhisetes.

Juhis on koostatud selleks, et asutustel ja organisatsioonidel oleks andmekvaliteedi haldamiseks, reeglite seadmiseks, mõõtmiseks ja kvaliteedi parendamiseks olemas praktilised juhtnöörid.

Andmekvaliteedi rakendamise nõuded asutustele, kes on andmekogu vastutavad töötlejad, tulenevad avaliku teabe seadusest (AvTS) § 43^4 ja Vabariigi Valituse määrusest teenuste korraldamise ja teabehalduse alused (TKTA), mille § 16^1 sätestab andmehalduse täpsemad nõuded ja tingimused. Õigusaktidest tulenevaid nõudeid on täpsemalt kirjeldatud dokumendis: „Andmekvaliteedi haldus: Asutuse ülesanded andmekvaliteedi tagamisel“ (Mai 2023, Versioon 0.3).

1.2 Juhise koostamine

Juhis on koostatud kahes järgus ning juhis arvesta varasemate töödega.

Varasem töö:

- 2016. aastal valmis juhend pealkirjaga „Andmekvaliteedi tagamise juhend andmekogu omanikele“. Nimetatud juhend on suunatud eelkõige andmeomanikele ning pakub neile

head materjali andmekvaliteediga tegelemiseks. Käesolev juhis ei asenda varasemas juhendis toodut, vaid täiendab seda. Erinevalt varasemast juhendist on antud juhis suunatud eelkõige andmehalduritele (andmestikud), kuid pakub samas praktilisi oskusi ja taustinformatsiooni nii andeomanikule kui ka teistele andmekvaliteedi tagamisel ülesandeid omavatele rollidele.

- Varasem juhend pakub välja andmekvaliteedi juhtimise raamistiku andmekvaliteedi pidevaks parendamiseks ning seeläbi küpsustaseme tõstmiseks. Välja pakutud raamistik põhineb pideva parendamise tsüklil ehk Demingi rattal ning sisaldab järgmisi tegevusi: HINDAMINE, PLANEERIMINE, TEOSTAMINE, KONTROLLIMINE ja KORRIGEERIMINE. Kasutatud juhtimismeetod pakub üldisel tasemel andmekvaliteedi juhtimistegevuste kirjeldust. Täpsemat ülevaadet andmehalduse protsessidest, sh ka andmekvaliteedi protsessidest, pakub Eesti andmehalduse raamistik. Lisaks andmekvaliteedi juhtimise raamistikule on varasemas juhendis toodud **ülevaade andmekvaliteedi küpsusmudelist** ning juhised selle rakendamiseks. Neid käesolev juhis ei käsitle.
- Viimaks annab varasem juhend **ülevaate andmekvaliteedi dimensioonidest** (juhises nimetatakse neid tunnusteks). Varasemas juhises on käsitletavaid dimensioone üheksa: ÕIGSUS, TÄIELIKKUS, KOOSKÕLA, USALDUSVÄÄRSUS, AJAKOHASUS, REEGLIPÄRASUS, KONFIDENTSIAALSUS, ÜHEKORDSUS ja MITTELIIASUS. Iga käsitletud dimensiooni kohta on esitatud määratlus, dimensiooniga seotud kvaliteediprobleemide näited, ülevaade dimensiooni kvaliteedinõuete kohta käivatest õigusaktidest, lühikene kirjeldus meetoditest, mis aitavad tagada andmekvaliteedi vastavust nõuetele antud dimensiooni osas ning viimaks kontrollküsimused, mille eesmärgiks oli võimaldada andmeomanikul veenduda, et ta oleks arvestanud erinevate võimalike nõuete allikatega ja võimalike dimensiooni nõuetele vastavuse tagamise meetmetega.
- Käesolev juhis vaatleb andmekvaliteeti Eesti andmehalduse raamistiku kontekstis ning kirjeldab lahti andmehalduse raamistikus toodud andmekvaliteediga seotud peamised tegevused: KVALITEEDIREEGLITE HALDAMINE, ANDMEKVALITEEDI HALDAMINE, ANDMEKVALITEEDI MÕÕTMINE, ANDMEKVALITEEDI PROBLEEMIDE LAHENDAMINE. Kui varasem juhend käsitles üheksat dimensiooni, siis käesolevas juhises kasutusele võetud mudelis on dimensioone viis: õigsus, täielikkus, ajakohasus, reeglipärasus ja ühekordsus. Erinevalt varasemast juhisest on käesolevas juhises toodud terviklikud juhtnöörid andmekvaliteedi dimensioonidest rakendamise võtmes olulisemate mõõtmiseks.

Juhise koostamise versioonid on:

1. Statistikaameti tellimusel ja Euroopa Komisjoni Struktuursete Reformide Toetusteenistuse (SRSS) rahastatud projekti *„Support for the establishment of data governance services“* (Toetus andmehaldusele. Eesti andmehalduse metoodikaprojekt) raames 2019. a sügisest 2020. aasta suveni ettevõtte Ernst & Young eksperdid (projektijuht oli Siim Aben, eksperdid olid Kuldar Aas ja Raivo Ruusalepp).
2. Statistikaameti ja Majandus- ja Kommunikatsiooniministeeriumi koostöös, arvestades asutustelt saadud tagasisidet ja andmehalduse arenguid. Sealhulgas nende infosüsteemide arendusi, mis üleriigiliselt või asutustes on andmehaldust toetavatena kasutusel või arendamisel. Käesoleva versiooni koostasid Statistikaametist Veiko Berendsen ja Annika Uibopuu ning Majandus- ja Kommunikatsiooniministeeriumist Kuldar Aas.

Andmekvaliteedi juhise koostamisel on lähtutud nii DAMA-DMBOK2 andmehalduse mudelist, andmekvaliteedi protsessi ning mõõtmist käsitlevast kirjandusest kui ka andmekvaliteedi probleemide taksonoomiast.

1.3 Andmekvaliteedi juhise eesmärgid

Andmekvaliteedi juhise eesmärgiks on selgitada, mida andmekvaliteet tähendab ja hõlmab, kuidas andmekvaliteeti töökorralduslikult juurutada ehk millised on andmekvaliteedi sihtrühmad ja rollid, mis on andmekvaliteedi reeglid ning milline on andmekvaliteedi mudel.

Juhis seletab lahti andmekvaliteedi reeglite kirjeldamiseks tehtavad tegevused ning annab detailseid juhiseid andmekvaliteedi mõõtmise teostamiseks. Arvestades andmekvaliteedi mõõtmise keerukust ja alamtegevuste rohkust on oluline rõhk pandud just selle tegevuse lahti seletamisele. Samas ei ole juhise selles versioonis rakendamise tehnilisi näiteid. Nende jaoks koostatakse eraldi lisa.

Üheks juhise eesmärgiks on esitada rakendamiseks sobiv andmekvaliteedi mudel. Mudel on tehniliseks rakendamiseks.

Andmekvaliteedi käsitlus juhises on asutuse keskne ning riiklikul tasemel andmekogude võrdlust ei rakendata. Seega ei ole puudutatud andmete ühekordse küsimise teemat ega põhiandmete problemaatikat.

Juhise järgimine aitab kaasa andmehalduse raamistiku terviklikule rakendamisele organisatsioonis.

1.4 Sihtrühmad

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

Juhise käesolevas versioonis on rollid jäetud 2020. aasta andmehalduse raamistikuga samaks. Raamistiku rollide muutmisel muudetakse need ka selles juhises. Praktiliselt võivad ametinimetused ja rollid allpool kirjeldatust erineda.

1.4.1 Andmekvaliteediga seotud rollid andmehalduse raamistikust

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

Andmehalduse juht on andmehaldusorganisatsiooni juht ja asutuse või valdkonnaüleste tegevuste koordinaator. Tema vastutuseks on andmekvaliteedi aruannete ja mõõdikute väljatöötamine ja jälgimine, andmekvaliteedi parendamise tasuvusanalüüside koostamine ning andmekvaliteedi projektide portfelli haldamine ning projektide tellimine ja järelevalve.

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

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

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

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

1.4.2 Juhise sihtrühmad

Juhis on eelkõige suunatud andmehalduri (andmestikud) rollis olevatele asutuste ja organisatsioonide töötajatele, kellele juhis pakub juhtnööre kõigi temaga seotud andmekvaliteeti puudutavate ülesannete täitmiseks:

- andmeomanike toetamiseks andmekvaliteedi reeglite kirjeldamisel rakendades profileerimist;
- andmekvaliteedi mõõtmiseks;
- andmekvaliteedi reeglitele mõõdikute määramiseks, mõõtetulemuste kogumiseks ja raporteerimiseks;
- andmekvaliteedi probleemide põhjuste väljaselgitamiseks;
- kvaliteediprobleemide mõju hindamiseks.

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

1.5 Juhise ülesehitus

Andmekvaliteedi haldamise protsessi osaks olevast andmekvaliteedi probleemide lahendamise alamprotsessist on kirjeldatud kaks peamist eriteadmisi nõudvat planeerivat tegevust:

- andmekvaliteedi probleemide põhjuste analüüs;
- kvaliteediprobleemide mõju hindamine tegevusele.

Eelnimetatud planeerivate tegevuste käigus tekib hulk andmekvaliteedi reegleid, mis on pidevas muutumises. Seega võib andmekvaliteedi reeglite haldamine muutuda kiiresti keerukaks ning tekitada segadust. Selle vältimiseks on juhises esitatud andmekvaliteedi reeglite haldamiseks osa.

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

1.6 Andmekvaliteedi haldamise õiguslik raamistik

Andmekvaliteedi tagamise kohustus on asutustele pandud erinevate õigusaktidega.

Seaduse tasandil on õiguslik alus sätestatud avaliku teabe seaduses (AvTS). AvTS-i § 43^4 andmekogu vastutav ja volitatud töötleja, on lõikes 1^2 sätestatud, et andmehalduse täpsemad nõuded ja tingimused sätestab Vabariigi Valitsus või tema volitatud minister määrusega. [RT I, 15.03.2019, 2 - jõust. 01.04.2019] Selleks määruseks on teenuste korraldamise ja teabehalduse alused (TKTA), mille § 16^1 sätestab andmehalduse täpsemad nõuded ja tingimused.

Tulenevalt õigusaktide loogikast – AvTS-i andmekogude peatükk ja TKTA teabehalduse peatükk – on õiguslikult reguleeritud ainult üks osa andmetest, konkreetselt andmekogud. See aga ei tähenda, et nende digitaalselt hallatavate andmete, isegi digidokumentide, samuti andmestike ja sellistes infosüsteemides hallatavate andmete osas, mis ei ole andmekogud, ei peaks või ei saaks andmekvaliteedi reegleid määratleda, kehtestada, hallata ning andmekvaliteeti mõõta ja parandada.

Andmekogude osas on AvTS-is need osaks riigi infosüsteemist (§ 43^2) ning peavad olema riigi infosüsteemi haldussüsteemis (RIHA) registreeritud (§ 43^7). RIHA määrus [RT I 2008, 12, 84 - jõust. 08.03.2008] kohustab andmekogu vastutavat töötlejat koostama ja kooskõlastama RIHAs andmekogu dokumentatsiooni (ptk 2, §§ 6-7). RIHA §10 lg 3 sätestab, et andmekogu registreerimisel ja andmekogus kogutavate andmete koosseisu muutmise registreerimisel tuleb andmekogu vastutaval või volitatud töötlejal andmete koosseisule vastavad andmed aktualiseerida. See nõue sisaldab nii andmete kirjeldamist kui ka andmekvaliteeti.

2 Andmekvaliteet

2.1 Mis on andmekvaliteet?

Andmekvaliteedi mõiste lahti seletamiseks on vaja seletada mõlemat termini osa eraldi ja siis neid ka koos.

Käesolevas juhises ei ole seletatud lahti, mis on andmed või metandmed, sest seda on tehtud andmekirjelduse juhises. Küll tuleb aga peatuda sellel mis on kvaliteet ja seda eelkõige andmekvaliteedi käsitluse võtmes. Selleks oleme lähtunud rahvusvahelise standardi ISO 8000 standardite perekonna definitsioonidest. ISO 8000 on andmekvaliteedi standardite perekond, mis koosneb mitmest osast, mille osa 2 on sõnastik (ISO 8000-2:2017, Data quality — Part 2: Vocabulary).

Üldisesse kvaliteedi käsitlusse [Joseph M. Juran, Quality Control Handbook, 1951] tuli andmekvaliteet 1970. aastatel (Jurani töö kolmas väljaanne 1974). Selle kohaselt võib andmeid pidada kõrgekvaliteetsekse, kui need on sobilikud eesmärgiks seautud kasutamiseks, mis võib olla kasutamine konkreetses töös, kasutamine otsustamiseks või kasutamine planeerimiseks (*Data can be considered of high quality when it is fit for its intended use in operations, decision – making and planning*.)

Viimastel aastakümnetel on tooni andmed kaks andmekvaliteedi koolkonda . Alates 1992. aastat **totaalse andmekvaliteedi halduse koolkond** (*Total Data Quality Management – TDQM*) ja viimasel kahekümnel aastal **täieliku andmekvaliteedi halduse koolkond** (*Complete Data Quality Management – CDQM*). Neid mõlemat iseloomustavad väga suured andmekvaliteedi mudelid ja rakendusraamistikud. CDQM koolkonna autorid Batini ja Scannapieco eristavad teoreetilise (research) ja rakendusliku (application domains) mudel osa. Selles on mudel ise teoreetiline ning rakenduslikkus on valdkondlik, näiteks tervishoid või statistika.

ISO 8000-2 esitab kvaliteedi üldise määratluse:

MÕISTE

Kvaliteet on määr, mille ulatuses objekti olemasolevad omadused vastavad nõudmistele.
Degree to which a set of inherent characteristics of an object fulfils requirements (ISO 8000-2:2017, 3.3.1)

ISO 8000-2 esitab ka andmekvaliteedi määratluse, mis on kvaliteedi määratlusele väga lähedane.

MÕISTE

Andmekvaliteet on määr, mille ulatuses andmete olemasolevad omadused vastavad nõudmistele.
Degree to which a set of inherent characteristics of data fulfils requirements (ISO 8000-2:2017, 3.3.8)

Nagu nendest määratlustest on näha, taandub kõik vastavatele ehk kehtestatud nõudmistele, mis osutab nii andmekvaliteedi haldusprotsessile kui ka andmete omadustele.

2.2 Mis on andmekvaliteedi haldamine?

Andmekvaliteedi haldamine on osa ehk üks tegevus andmehaldusest nagu seda on kujutatud DAMA-DMBOK2 andmehalduse käsiraamatus. ISO 8000-2 käsitluses on see osa juhtimisest ja kontrollist.

MÕISTE

Andmekvaliteedi haldus on organisatsiooni koordineeritud tegevus andmekvaliteedi juhtimiseks ja kontrollimiseks.
data quality management – Coordinated activities to direct and control an organization with regard to data quality events (ISO 8000-2:2017, 3.3.9)

Organisatsiooni koordineeritud tegevus on seega nii andmekvaliteedi juhtimise tegevused kui ka andmekvaliteedi mudeli ja selle osade rakendamine, aga mõistagi ka andmekvaliteedi mõõtmine, saadud tulemuste analüüs ja kvaliteedi parendamine.

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

Andmekvaliteedi mõõtmine ja haldamine võib toimuda mingil andmete kasutuse hetkel, aga väga sagedane on, et seda on vaja mitmel andmetöötluse etapil. Eri etappides võib olla vaja mõõta eri indikaatoreid. Samuti on protsessis võimalik, et andmed on kogumis hinnatavad kasutatavaks, näiteks on mõned vead suures hulgas andmetes, mis ei takista neid statistikas kasutada, kuid mõnel juhul on väiksemgi viga lubamatu, näiteks toiminguks mis nõuab ühemõtteliselt õigeid andmeid.

2.3 Andmekogu ja andmestik andmekvaliteedi haldamise objektina

Andmekvaliteedi mudelit saab rakendada väga laiale ja väga erinevatele andmete valdkondadele. Eesti avalikus sektoris on andmekvaliteedi haldust rakendatud eelkõige andmekogudele ja seda on pidanud tegema andmekogu omanikud (st vastutavad või volitatud töötlejad). Samuti on andmekvaliteedi teema oluline mitmetes valdkondades, Nii on näiteks Euroopa statistikasüsteemi (ESS) kvaliteedikäsitluses (ESS handbook for quality reports, 2014) kasutusel mudel, millel on kaks suuremat dimensiooni: väljundi/toote kvaliteedikriteeriumid ning andmetöötlusprotsessi kvaliteedikriteeriumid, millel on kokku üheksa mõõdetavat või hinnatavat indikaatorit (Asjakohasus, Täpsus ja usaldusväärsus, Ajakohasus ja õigeaegsus, Seostatavus ja võrreldvus, Kättesaadavus ja selgus, Levitamisvorming). Osa neist on ka kasutatud andmekogude andmekvaliteedi hindamiseks nende kasutamisel näiteks rahvaloenduseks. Samas ei ole see andmeanalüüsile ja andmete avaldamisele orienteeritud mudel andmekogude andmekvaliteedi haldamiseks neile enestele kõige sobivam.

Õigusaktides on sätestatud andmekvaliteedi halduse rakendamine andmekogudele. TKTA sätestab: riigi infosüsteemi kuuluva andmekogu vastutav töötleja dokumenteerib ja rakendab andmekvaliteedi seire ja haldamise protsessi, millega tagatakse riigi infosüsteemi kuuluvate andmekogude andmete kvaliteet vastavalt õigusaktidele.

See osa selgitab täpsemalt, kuidas asutus võiks endal andmehalduse juurutada juhul, kui ta ei lähtu ainult andmekogudest.

Andmekogul on avaliku teabe seaduses legaaldefinitsioon.

MÕISTE

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 § 43[1] lg 1)

Asutustel on lisaks andmekogudele või sageli andmekogude kõrval või asemel infosüsteemid, rakendused ning hulk olulisi faile analüüsiks, statistikaks jms. Ka andmebaasid on failid. Seega on kõrvuti andmekogudega mitmesuguseid andmestikke.

Andmestiku legaaldefinitsioon on riikliku statistika seaduses, kuid andmestikuna on nimetatud ka osa andmekogust (TIS § 21[1]).

MÕISTE

(1) andmestik on 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)

(2) andmestik on identifitseeritav ja hallatav andmete kogum (riikliku statistika seadus)

Andmestikke on eri tüüpi, millest selle dokumendi kontekstis on olulised järgmised:

- relatsioonilised andmebaasid
- arvutustabelid ja analüütilised andmestikud
- struktureeritud andmed ja andmestruktuurid nagu XML, JSON
- tekstidokumendid, tekstifailid, tekstikorpused ehk laiemalt mittestruktureeritud andmed
- veebilehed, wiki-d jms
- graafilise sisuga dokumendid nagu esitlused
- põhiliselt meediasisuga failid: pidi-, heli-, videofailid

Asutusele sisaldavad kõik ülal loetletud andmetike tüübid teavet. Küsimuse praktiliseks lahenduseks tuleb andmekvaliteedi haldus piiritleda vähemalt järgmiste andmestike tüüpidega ning need halduse alla võtta:

1. andmekogud – need on õiguslikult kõige enam reguleeritud osa andmetest, olenemata sellest, kas ollakse andmekogu vastutav või volitatud töötleja (eraldiga, kui volitatud töötleja on ainult andmdekogu tehniline haldaja);
2. asutuse infosüsteemid, millel on eraldi andmebaasid ning mis on seotud asutuse põhi- või tugiprotsesside täitmisega (ehk teenustega);
3. dokumendid ja failid mis on koostatud või saadud algandmete teises töötlemise tulemusel analüüsi käigus ning mis raportid, aruanded jms dokumendid, mille sisu on analüütiline üldistus või statistika;
4. andmestikud, mis on tehtud taaskasutatavaks avaandmetena;
5. peetavad klassifikaatorid ja (koodi)loendid, mis on kasutusel identifikaatorite (viidete, tähiste) ja/või nende väärtuste kasutamisel teistes andmestikes.

3 Andmekvaliteedi juhtimise tegevused

3.1 Andmekvaliteedi eesmärkide määramine

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

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

- organisatsiooni andmete väärtuse tõstmine;
- uute võimaluste loomine andmete kasutamiseks (ehk andmete väärindamine);
- madalast andmekvaliteedist tulenevate riskide vähendamine;
- organisatsiooni efektiivsuse ja produktiivsuse tõstmine;
- organisatsiooni reputatsiooni kaitsmine ja tugevdamine.

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

- suutmatus esitada korrektseid arveid;
- suurenenud kliendiprobleemide arv ning vähenenud suutlikus nende lahendamiseks;
- suurem keerukus asutuste tegevuse (ümber)korraldamisel;
- väiksem võimekus pettuste tuvastamisel;
- madala või negatiivse mõjuga otsused.

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

Asutuse andmekvaliteedi protsess peaks juhinduma järgmistest põhimõtetest:

- **Kriitilisus** – andmekvaliteedi protsess peaks keskenduma kõige kriitilisematele andmetele. Muudatuste prioriseerimine peaks põhinema andmete kriitilisusel ning võtma arvesse võimalikke ebakorrektsete andmetega kaasnevaid riske.
- **Elutsükli juhtimine** – andmekvaliteeti tuleb juhtida kogu andmete elutsükli jooksul. See hõlmab andmete liikumise haldamist nii protsessides, süsteemides kui ka eri

süsteemide vahel. Näiteks peab iga andmeahela lüli (nii protsess kui süsteem) tagama andmete kõrge kvaliteedi.

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

3.2 Andmekvaliteedi reeglite kirjeldamine

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

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

Andmete profileerimise teostamiseks kasutatakse üldjuhul profileerimistööriistu, mis annavad hea esmase ülevaate andmetest ja andmete kvaliteedist. Juhendi rakenduslikus osas on toodud juhised profileerimise teostamiseks Ehitisregistri näitel. Valik profileerimise teostamiseks sobivatest tööriistadest on leitav käesoleva juhise soovituslike töövahendite sektsioonis. Kuigi profileerimistööriistad esitavad andmete kohta mitmesugust statistikat ja

mõõdikuid, pole profileerimise puhul tegu andmekvaliteedi hindamisega. Andmekvaliteedi mõõtmine on põhjalikum tegevus, mille käigus hinnatakse andmete vastavust andmekvaliteedi reeglitele.

Võimalik on eristada kolme tüüpi andmete profileerimist:

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

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

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

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

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

Vajadusel on reeglite kirjeldamisel abiks äriprotsesside sisendite ja väljundite täpsustamine äriprotsessi eri osapooltega. Samuti on kasulik uurida osapoolte probleeme. Näiteks täpsustada mis juhtub, kui andmed on valed või puuduvad ja kuidas tuvastatakse probleeme. Seejuures on kasulik meeles pidada, et andmekvaliteedi hindamiseks pole vaja teada kõiki andmekvaliteedi reegleid. Reeglite tuvastamine ja täpsustamine on pidev protsess. Üks parimaid viise andmekvaliteedi reeglite kogumiseks on andmekvaliteedi hindamise tulemuste eri osapooltega jagamine. Tihti aitab tulemuste jagamine osapooli uute vaatenurkade leidmisel ning seeläbi uute reeglite sõnastamisel.

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

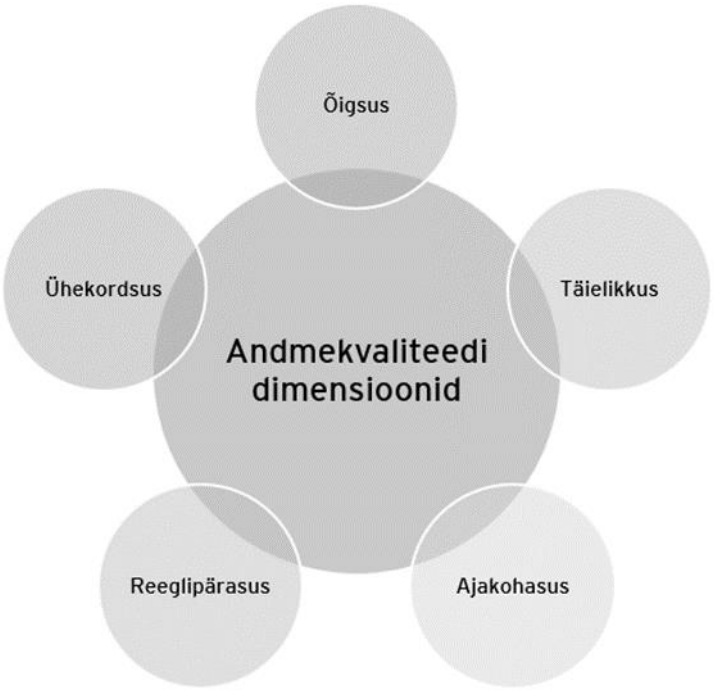
4 Andmekvaliteedi mudel

4.1 Andmekvaliteedi mudeli valik

Andmete kõrge kvaliteedi tähtsustamine teoreetikute ja praktikute poolt ning kasu, mida sellest on saadud, on aidanud kaasa andmekvaliteedi raamistike paljususe tekkele. Lähtuvalt valdkondlikest iseärasustest, infosüsteemide eripäradest ja andmete kontekstist on loodud kümneid erinevaid andmekvaliteedi raamistikke. Eesti andmekvaliteedi juhises on võetud kasutusele mudel, milles on viis dimensiooni:

- täielikkus,
- ajakohasus,
- õigsus,
- reeglipärasus ja
- ühekordsus.

Kolm esimest on eri raamistikes enimlevinud dimensioonid. Reeglipärasus võeti mudelisse, sest see võimaldab jälgida klassifikaatorite ja infosüsteemide sisemiste loendite kasutamist ning põhiandmete kasutamist. Et asutuse sees ei oleks dublitseerivaid andmeid, siis on raamistikus eraldi dimensioonina välja toodud ühekordsus.



Joonis 2: Andmekvaliteedi mudel

Andmekvaliteedi mudelis on neid dimensioone kasutatud andmekvaliteedi indikaatorite, probleemide ja reeglite grupeerimiseks. See tagab andmekvaliteedi süsteemse käsitlemise ja lihtsustab andmekvaliteedi haldamist, sh andmekvaliteedi reeglite väljatöötamist.

Andmekvaliteedi reeglite grupeerimiseks kasutame täiendavalt Oliveira jt poolt välja töötatud andmekvaliteedi probleemide taksonoomiat. (Oliveira, Paulo, Fátima Rodrigues, and Pedro Rangel Henriques. "A formal definition of data quality problems." ICIQ. 2005) Erinevalt teistest andmekvaliteedi probleemide raamistikest põhineb selles raamistikus toodud andmekvaliteedi probleemide taksonoomia laiapõhjalisel juhtumiuuringul, on formaliseeritud ning toetub samade autorite varasemale tööle , mis pakub lisaks konkreetseid algoritme andmekvaliteedi probleemide tuvastamiseks ja klassifitseerimiseks. Andmekvaliteedi reeglite kirjeldamisel lähtume me tuvastatud andmekvaliteedi probleemidest ja juhtimisel andmekvaliteedi dimensioonidega seotud indikaatoritest. Seosed dimensioonide ja reeglite vahel tekivad läbi reeglite grupeerimise andmekvaliteedi probleemide alusel. Selline probleemipõhine liigitusskeem lihtsustab andmekvaliteedi reeglite jaotamist dimensioonidesse.

4.2 Andmekvaliteedi mudeli dimensioonid

Andmekvaliteedi dimensioonid (Joonis 2) on mõõdetavad andmete omadused, mis väljendavad andmete kvaliteeti erinevatest aspektidest lähtuvalt. Eksisteerib palju erinevaid andmekvaliteedi

dimensioonide käsitlusi, kuid antud juhises keskendutakse viiele kvaliteedidimensioonile ning nende hindamist toetavate indikaatorite kirjeldamisele. Võimalik on kasutada ka teistsuguseid dimensioonide liigitusi, kuid konkreetsed andmekvaliteedi probleemid seejuures ei muutu. Teistsugust dimensioonide liigitust kasutades on vaja määrata andmekvaliteedi probleemide seosed dimensioonidega. Antud dimensioonide puhul on nimetatud seosed kirjeldatud käesoleva juhise tabelis 1.

Õigsus (Accuracy) näitab, mil määral vastavad andmed tegelikkusele. Andmete õigsus jaguneb süntaktiliseks ja semantiliseks õigsuseks. Sünktaktiline 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).

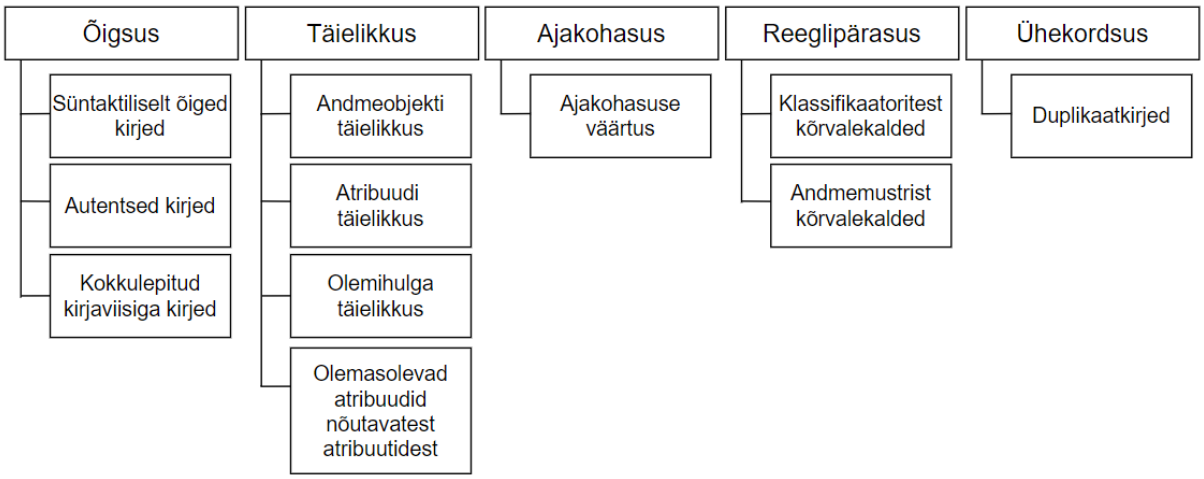
Täielikkus (Completeness) näitab, mil määral on olemas kõik nõutud andmed. Täielikkus on vaadeldav kahes osas: kirjete täielikkus ja kogumi ehk populatsiooni täielikkus. Kirjete täielikkus näitab, mil määral on andmekirje kõigil tunnustel olemas väärtused. Näiteks mil määral on andmetabeli veeru read (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 reaalsete sündmuste ning nende andmetes fikseerimise või andmete värskendamise vahel on vältimatu. Seetõttu on võimalik olukord, kus andmed on küll uuendatud, kuid nende tekkeks või värskendamiseks kuluv aeg muudab andmete kasutamise mõne konkreetse ülesande jaoks võimatuks. Näiteks võib ülikooli tunniplaan olla küll värske, kuid see pole ajakohane kui see jõuab tudengiteni alles pärast loengute algust.

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

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

Dimensioonid on konkreetsemaks hindamiseks jagatud mõõdetavateks indikaatoriteks (Joonis 3). Edasipidi on indikaatorid esitatud konkreetsete andmekvaliteedi probleemidena. See võimaldab probleeme tuvastada, seda nende kindlakstegemiseks ja andmekvaliteedi tuvastamiseks ja kontrolliks reegleid. Probleemide tuvastamise, reeglite seadmise ja mõõtmise süsteemi on nimetatud andmekvaliteedi mõõtmise metamudeliks ja seda on käsitletakse järgmises peatükis.



Joonis 3: Dimensioonide jaotus indikaatoriteks

4.3 Andmekvaliteedi mõõtmise metamudel

Andmekvaliteeti kasutatakse eri tasemetel otsustamiseks nii andmehalduse enda kui muude asutuse või riigi tegevuste korraldamisel ja täideviimisel. Ühelt poolt kombineeritakse andmekvaliteedi reeglitele vastavused/mittevastavused indikaatorite väärtusteks ja need omakorda dimensioonide väärtusteks ja osad neist võivad saada võtmemõõdikuteks üksikutes asutustes või üle riigi. Teiselt poolt agregeeritakse nimetatud tunnuseid andmeelementidelt andmestiku andmeobjekti liikide, andmestike ja asutuse või riigi tasemel otsuste tegemiseks.

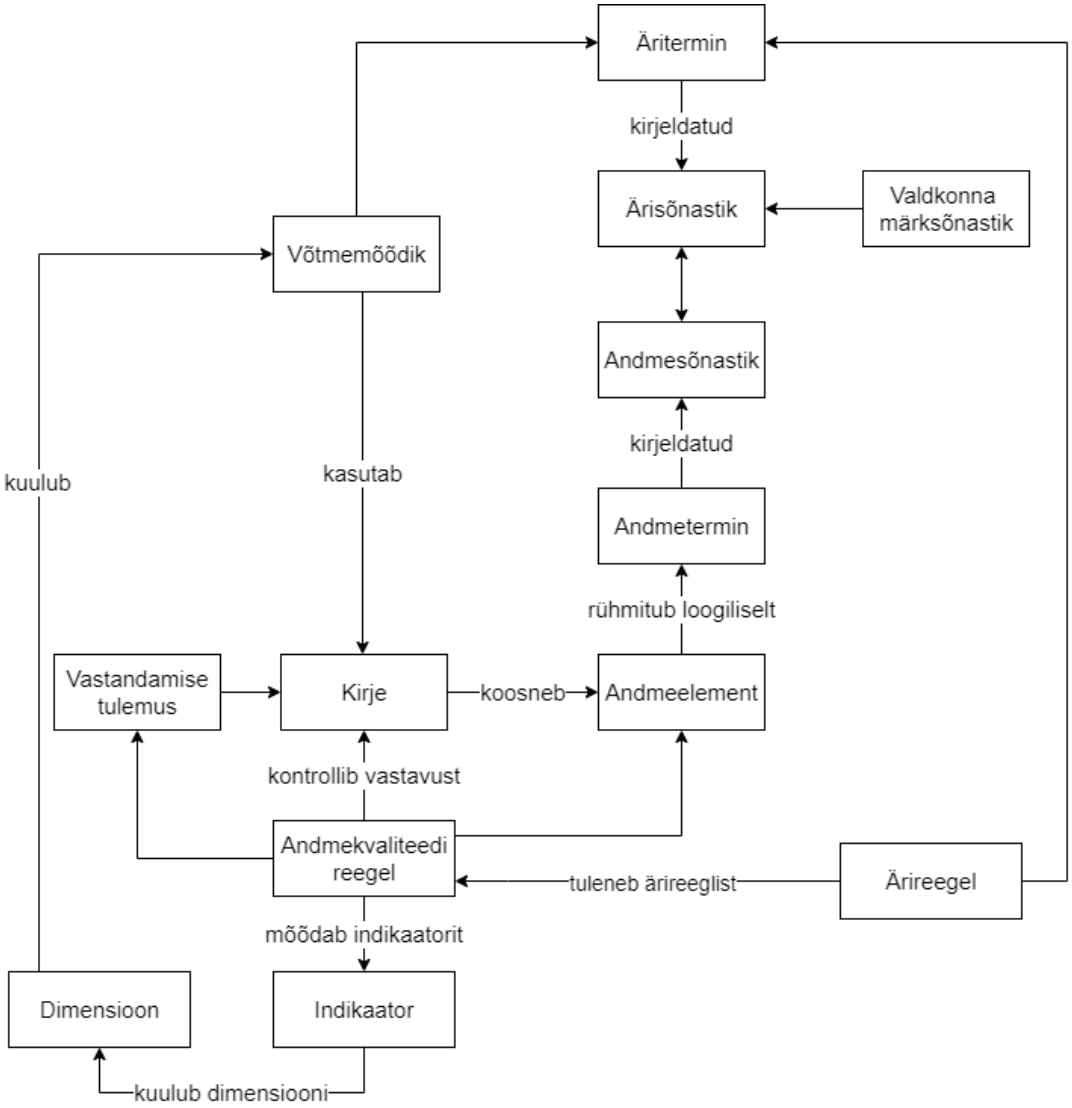
Andmete kvaliteedi kasutamist iseloomustavad järgnevad stsenaariumid:

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

Taoliste stsenaariumite täideviimist toetab andmekvaliteedi mõõtmise metamudel (Joonis 4). Andmekirjelduse olemasolu on andmekvaliteedi mõõtmise eelduseks. Seosed andmekirjelduse ja andmekvaliteedi olemite vahel on esitatud andmekvaliteedi mõõtmise metamudelis. Täpsemad juhised andmekirjelduse loomiseks on esitatud andmekirjelduse juhises, mis valmis paralleelselt käesoleva juhisega ning on samuti osa andmehalduse raamistikust.

Andmekirjelduse koostamisel ja haldamisel kasutatakse kolme tüüpi sõnastikke: VALDKONNA MÄRKSÕNASTIK, ÄRISÕNASTIK ja ANDMESÕNASTIK. ANDMEELEMENDID rühmituvad ANDMETERMINITEKS, mis omakorda on kirjeldatud ANDMESÕNASTIKUS. ÄRITERMINID on kirjeldatud ÄRISÕNASTIKUS, lisaks kasutatakse ÄRISÕNASTIKE loomisel ühe võimaliku terminite allikana VALDKONNA SÕNASTIKKE. ANDMEELEMENDID rühmituvad loogiliselt

(kontseptualiseerimine) ANDMETERMINITEKS (näiteks Andmetermin „Isik“ koosneb vähemalt Andmeelementidest „eesnimi“, „perekonnanimi“ ja „isikukood“). ANDMEELEMENDID on ANDMESÕNASTIKU kaudu seotud organisatsiooni tegevust kirjeldavate mõistetega ehk ÄRISÕNASTIKUGA.



Joonis 4: Andmekvaliteedi mõõtmise metamudel

Käesolev andmekvaliteedi juhis kasutab järgmiseid andmekvaliteedi mõõtmist võimaldavaid olemeid: **Andmekvaliteedi reegel**, **Dimensioon** ja **Indikaator**. Andmekvaliteedi reeglid tulenevad ÄRIREEGLITEST, mis on kirja pandud ÄRISÕNASTIKUS kirjeldatud ÄRITERMINEID kasutades. Samas ei ole kõik ÄRIREEGLID kajastatud ANDMEKVALITEEDI REEGLITES ning vastupidi, osad ANDMEKVALITEEDI REEGLID ei kajastu ÄRIREEGLITES. Andmekvaliteedi mõõtmisel rakendatakse ANDMEKVALITEEDI REEGLEID kas üksikutele andmete KIRJETELE või nende kogumitele ning tulemuseks saadakse ANDMEKVALITEEDI REEGLITELE vastandamise tulemus (confrontation). ANDMEKVALITEEDI REEGLITES kasutatakse muutujaid, mis on andmetes esitatud ANDMEELEMENTIDENA (nt „ehr_kood“ on kirje „Ehitis“ ANDMEELEMENT). Vastandamise tulemused kombineeritakse **Indikaatoriteks** ja need omakorda **Dimensioonideks**. (Iga Indikaator kuulub Dimensiooni, näiteks Indikaator „Klassifikaatoritest kõrvalekalded“ kuulub Dimensiooni „Reeglipärasus“.) Vaikimisi kasutatakse kombineerimise funktsioonina aritmeetilist keskmist, kuid vajadusel võib kasutada ka muid funktsioone, kui nendes on eelnevalt kokku lepitud. Juhul kui kokkulepe eksisteerib ning mõõtmine on ühetaoline,

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

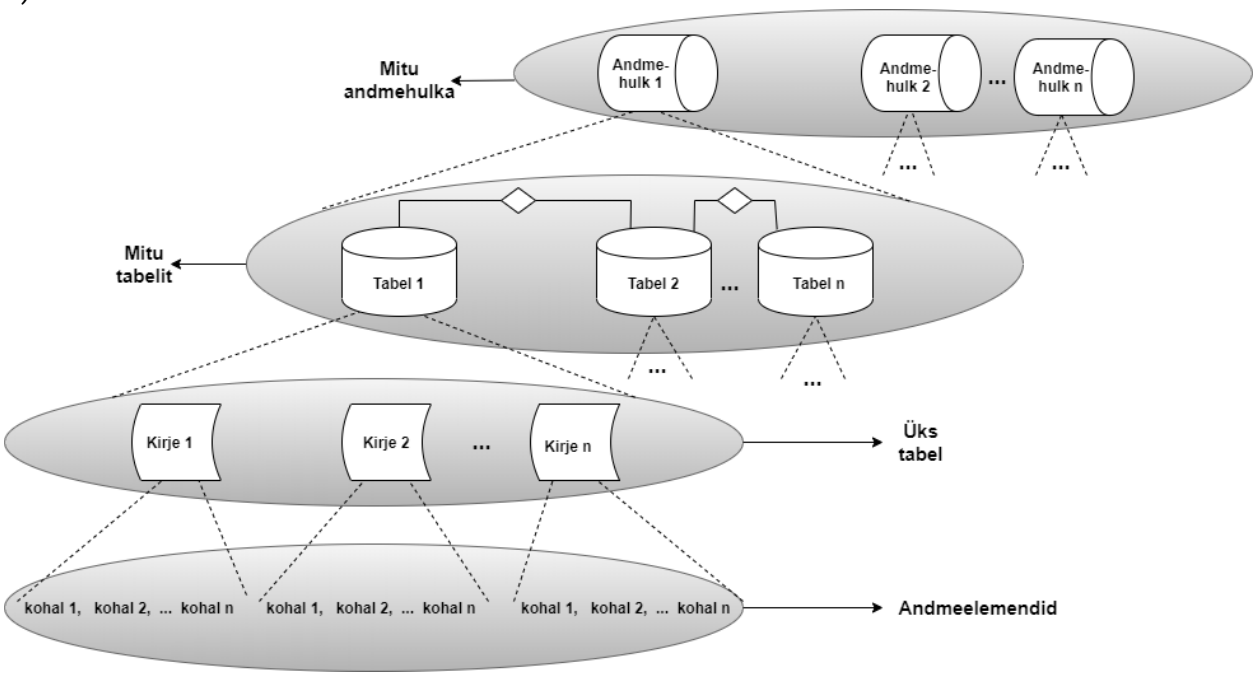
Mõned andmekvaliteedi **Dimensioonid** võivad olla ka osa organisatsiooni võtmemõõdikutest, mis omakorda on ÄRITERMINEID kasutades kirjeldatud ÄRISÕNASTIKUS. Võtmemõõdikuteks olevate **Dimensioonide** arvutamisel kasutatakse organisatsiooni andmeid.

4.4 Andmekvaliteedi probleemid

Selleks, et andmekvaliteedi probleeme süsteemselt käsitleda, lihtsustada andmekvaliteedi reeglite koostamist ja tagada andmekvaliteedi dimensioonide ühtlane kaetus andmekvaliteedi reeglitega oleme kasutanud Oliveira jt poolt välja töötatud andmekvaliteedi probleemide raamistikku. Lisaks toome ilmutatud kujul välja valitud raamistiku loojate poolt välja pakutud andmete ajakohasust puudutavad probleemid. (Oliveira Paulo, Fátima Rodrigues, Pedro Henriques, and Helena Galhardas. "A taxonomy of data quality problems." In 2nd Int. Workshop on Data and Information Quality, pp. 219-233. 2005) Võrreldes teiste sarnaste raamistikega, on valitud lähenemise eelised järgmised:

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

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



Joonis 5: Andmete korraldamise tüüpiline mudel

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

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

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

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

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

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

Puuduv väärtus

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

Näide: Kohustuslik atribuut "ISIKUKOOD" pole väärtustatud.

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.

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.

Väärtusvahemiku rikkumine

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

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

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").

Õigekirjaviga

Probleemi kirjeldus: Tekstilise atribuudi väärtus sisaldab õigekirjavigasid.

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

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.

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.

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.

Sünonüümide kasutus

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

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

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.

Funktsionaalse sõltuvuse rikkumine

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

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

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.

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.

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.

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".

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.

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.

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.

Esitluse mitmekesisus

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

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

Homonüümide kasutus

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

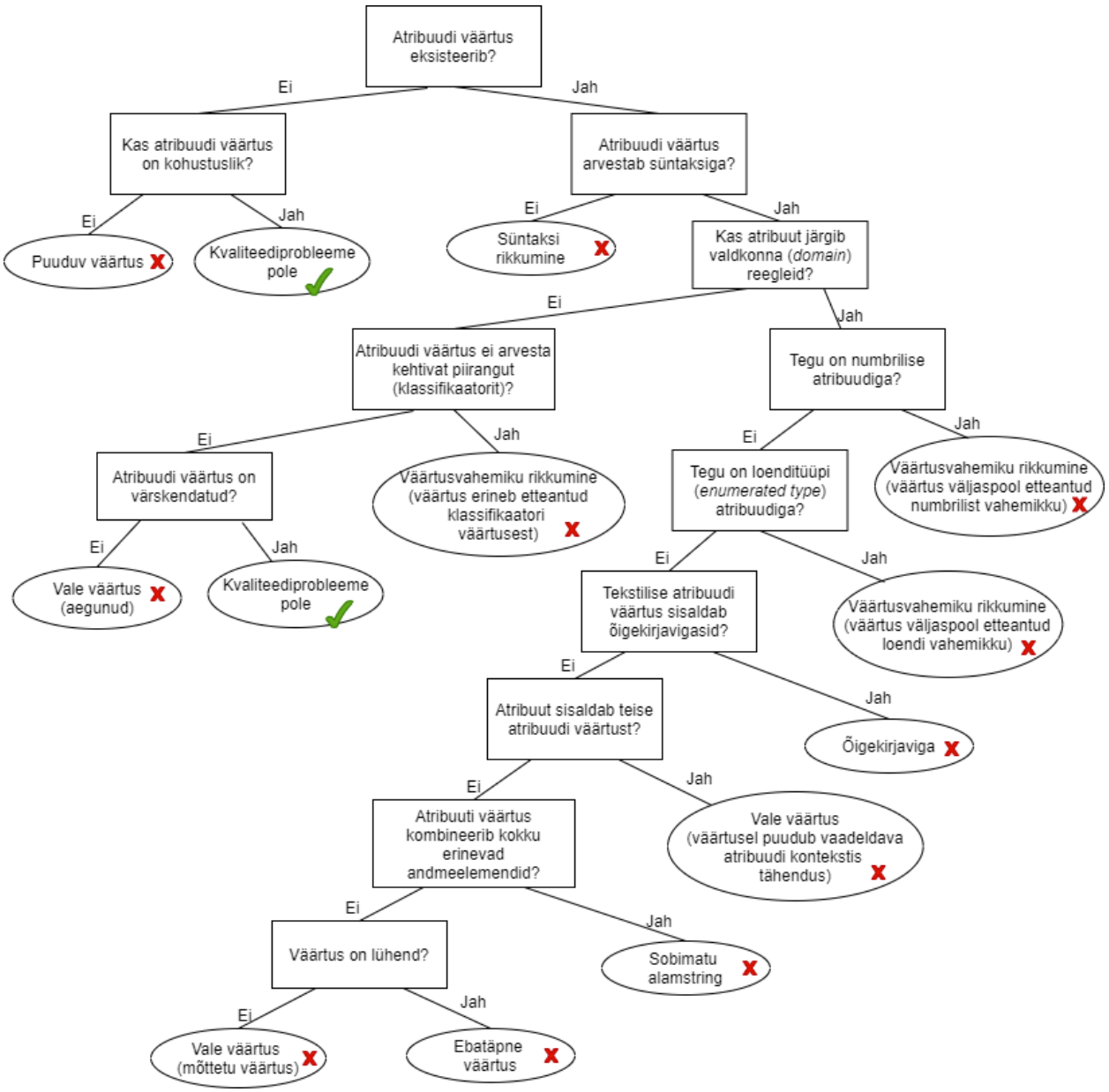
Näide: Andmestikus X kasutatakse "TOODE" tabelis terminit "Hiir" tähistamaks arvuti riistvaralist osutusseadet. Andmestikus Y kasutatakse "TOODE" tabelis terminit "Hiir" tähistamaks kodulooma.

4.5 Kvaliteediprobleemide tuvastamine

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

Meetod andmekvaliteedi probleemide tuvastamiseks üksiku kirje ühe atribuudi väärtuses on esitatud joonisel 6. Antud meetod tuvastab kvaliteediprobleeme ühes andmehulgas. Näiteks andmetabeli puhul saab antud meetodiga tuvastada ühe tabeli rea (üksiku kirje) ühe välja (ühe atribuudi väärtuse) kvaliteediprobleemid.

Sellise binaarse puu meetodiga on võimalik tuvastada andmekvaliteedi probleeme erinevate andmehulkade korral alates üksikust andmeelemendist kuni mitme andmehulgani, mida on vaja kas siduda või koos töödelda.



Joonis 6: Üksiku andmehulga üksiku kirje atribuudi väärtus

4.6 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.

Õigsus (Accuracy)

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

Süntaktilist õigsust saab defineerida kui kaugust väärtuse v ja domeeni väärtusi sisaldava hulga D elemendi vahel. Näiteks kui tegelikkuses on inimese nimi v' = Tõnu ning andmeväärtus v = 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' = Tõnu siis sellisel juhul on andmeväärtus v = 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 kirjete määr;
- autentsete kirjete määr;
- kokkulepitud kirjaviisi järgivate kirjete määr.

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

Autentsete kirjete 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 realistik inimeste korral). Tegu on osaga semantilisest õigsusest.

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

Täielikkus (Completeness)

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

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

Võttes arvesse tühje väärtusi saab täielikkuse jagada kolmeks osaks:

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

Tabel 2. Täielikkuse näidistabel

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

Eelnevast tulenevalt oleme Täielikkuse dimensiooni jaganud mõõdetavateks indikaatoriteks järgmiselt:

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

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

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

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

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

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 värkust on defineeritud järgmiselt (Võrrand 1).

Võrrand 1. Värskus (*Currency*)

$$Värskus = Vanus + (EdastamiseAeg - SisestusAeg)$$

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

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

Võrrand 2. Ajakohasus (*Timeliness*)

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

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

Tulenevalt eelnevast oleme ajakohasuse dimensiooni indikaatori sõnastanud järgmiselt:

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

Näitena vaatleme digiretsepti ajakohasust. Oletame, et arsti vastuvõtul vaadatakse üle 19.03.2020 kell 08:00 toimunud uuringu tulemused. Tulemuste põhjal määratakse patsiendile retseptiravim, mille määramine leiab aset 19.03.2020 kell 13:00. Seega on andmete vanus sel hetkel 5 tundi ehk Vanus=5. Arst märgib retsepti andmed süsteemi ning loob digiretsepti 19.03.2020 kell 14:00. Proviisiori poolt kasutatavasse süsteemi jõuavad digiretsepti andmed 19.03.2020 kell 15:00. Hetkeks kui andmed jõuavad proviisorini on need süsteemis olnud ühe tunni ($\text{EdastamiseAeg} - \text{SisestusAeg} = 15{:}00 - 14{:}00 \text{ ehk } 1 \text{ tund}$). Sellest tulenevalt saame värskuse definitsiooni põhjal (Võrrand 1) öelda, et andmete värskus on 6 tundi ($\text{Värskus} = 5 + 1 = 6$). Digiretsepti kehtivus on 60 päeva ehk 1440 tundi. Järelikult on andmete ajakohasus vastavalt ajakohasuse definitsioonile (Võrrand 2) $\max(0{,}1 - (5/1440) = 1 - 0.003) = 0.997$. Väärtus 0.997 näitab, et digiretsepti ajakohasus on peaaegu täiuslik. Negatiivse stsenaariumi puhul jõuavad andmed süsteemi vahetult enne või pärast kehtivuse lõppu ning sellisel juhul läheneks andmete ajakohasuse väärtus nullile.

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.

Ühekordsus (Uniqueness)

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

4.7 Andmekvaliteedi reeglile mõõdiku seadmine

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

Tabel 3. Andmekvaliteedi reeglile mõõdiku seadmine – Õigsus.

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

Tabel 4. Andmekvaliteedi reeglile mõõdiku seadmine - Täielikkus.

Indikaator	Reegel	Meede	Hetketase	Sihttase
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.	4/5*100 = 80% 80% konkreetse isiku kohta käivatest andmetest on väärtustatud.	100%

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

Tabel 5. Andmekvaliteedi reeglile mõõdiku seadmine – Ajakohasus.

Indikaator	Reegel	Meede	Hetketase	Sihttase
Ajakohasuse määr	Riigieksamite tulemused peavad olema kasutatavad enne sisseastumis-perioodi lõppu.	Arvutada ajakohasuse väärtus vastavalt ajakohasuse definitsioonile (**Error! Reference source not found.**).	Ajakohasuse väärtus: 0.9	>0.8

Tabel 6. Andmekvaliteedi reeglile mõõdiku seadmine - Reeglipärasus.

Indikaator	Reegel	Meede	Hetketase	Sihttase
Klassifikaatoritest kõrvalekalded	Majanduslik tegevusala peab olema talletatud vastavalt EMTAK-ile.	Loendada kokku juhtumite arv, kus tegevusala ei vasta EMTAK-i klassifikatsioonile. EMTAK-ist kõrvalekallete arv: 200 EMTAK koodidega kirjete arv: 1771	200/1771 * 100 = 11% 11% kirjete 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 isikukoode sisaldavate väljade koguarv. Andmemustrile mittevastavate isikukoodide arv: 400 Isikukoodide koguarv: 700 000	400/700 000 * 100 = 0.05% 0.05% isikukoodidest ei vasta nõuetele	0%

Tabel 7. Andmekvaliteedi reeglile mõõdiku seadmine – Ühekordsus.

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

4.8 Kvaliteediprobleemide prioriseerimine

On tõenäoline, et eelnevalt kirjeldatud kvaliteedidimensioonide hindamise käigus tuvastati mitmeid erinevaid kvaliteediprobleeme. Esmalt on mõistlik põhjalikumalt tegeleda nende kvaliteediprobleemide analüüsimise ja lahendamisega, mis on asutusele suurima mõjuga. Seetõttu tuleks kvaliteediprobleemide esmasel prioriseerimisel kasutada ärilise mõju hindamise tehnikaid. Esmaseks hindamiseks sobivad hästi lihtsamad tehnikad, näiteks võib koguda stsenaariumeid, mis kirjeldavad halva andmekvaliteedi mõju ärile. Teine lihtne tehnika ärilise mõju on hindamiseks on koostada loend konkreetseid andmeid kasutavatest asutustest ja protsessidest. Mida rohkem on andmeid kasutavaid asutusi ja protsesse seda olulisemad on

konkreetsed andmed. Kasutada võib ka viis korda „Miks?“ küsimise tehnikat eesmärgiga jõuda kvaliteediprobleemi tegeliku mõjuni. Põhjalikum mõju hindamine toimub protsessi hilisemas faasis pärast andmekvaliteedi juurpõhjuste analüüsi.

4.9 Andmekvaliteedi aruandepõhjade väljatöötamine

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

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

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

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

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

4.10 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 peaks kirjeldama subjektiivsete arvamuste alusel. Reegleid tuleb testida reaalsetel andmetel. Tihti toob selline testimine välja andmetes eksisteerivad probleemid ning aitab objektiivselt tuvastada ka andmekvaliteedi reeglites eksisteerivaid puuduseid.

- **Valdkonna eksperdi poolt heakskiidetud.** Andmekvaliteedi reeglite eesmärgiks on ilmutada nõuded andmetele. Tihti on reeglite õigeks kirjeldamiseks vajalikud teadmised konkreetse ärivaldkonna protsessidest. Neid teadmisi tuleks koguda konkreetse valdkonna eksperdilt, kelle ülesandeks on kinnitada kirjeldatud ärireeglid või selgitada andmeanalüüsi tulemusi.
- **Andmete kasutajatele kättesaadavad.** Kõigil andmete kasutajatel peaks olema juurdepääs dokumenteeritud andmekvaliteedi reeglitele. Juurdepääs reeglitele aitab andmete kasutajatel andmeid paremini mõista ning aitab samas tagada, et reeglid on täielikud ja õiged. Lisaks peab olema võimalus küsida reeglite kohta küsimusi ning anda tagasisidet.

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

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

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 kastutusel 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.

5 Andmekvaliteedi põhjuste analüüs ja mõjude hindamine

5.1 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 juurpõhjuse analüüsi peamiseks eesmärgiks on tuvastada kvaliteediprobleemi tekkepõhjused ning määrata tegevused probleemi edaspidiseks ennetamiseks.

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

Juurpõhuse 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 juurpõhjuse 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 juurpõhjuse analüüsist kasu, kuna see aitab probleemi algallika tuvastada, probleemi lahendada ning seeläbi vähendada jooksvatele andmeparandustele kuluva aja ja muude ressursside ebavajalikku kulutamist.

Juurpõhjuse tuvastamist on käsitletud informatsiooni elutsükli raames , mis kajastab ka andmete elutsüklit. 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 elutsüklit 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 kolm täpsemat meetodit andmekvaliteedi juurpõhjuste analüüsimiseks. Sõltuvalt avastatud probleemide keerukusest ja kiireloomulisusest võib olla vajalik meetodite omavaheline kombineerimine või eelistatav hoopis ühe konkreetse meetodi valik. Näiteks võib kiireloomulise probleemi esmaseks analüüsiks sobida lihtne viis korda „Miks?“ meetod.

Kolm meetodit andmekvaliteedi juurpõhjuste analüüsimiseks on järgmised:

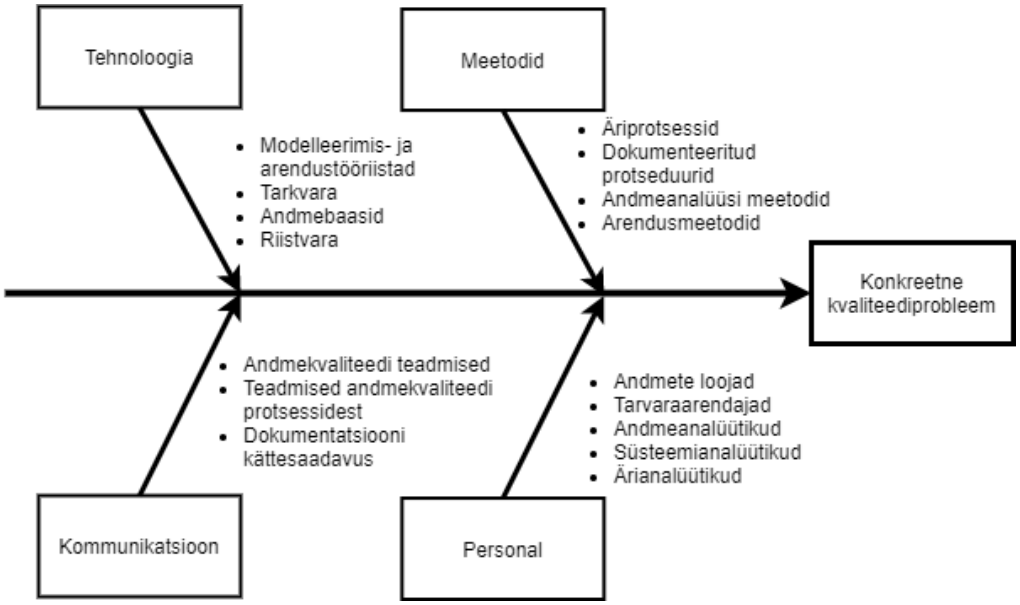
- Küsi 5 korda „Miks?“ (*Five „Whys“*)
- Jälita (*Track and Trace*)
- Põhjus ja tagajärg diagramm (*Cause-and-Effect/ Fishbone Diagram*)

Esimene meetod on küsida **5 korda „Miks?“** (*Five „Whys“*). Meetodit kasutatakse laialdaselt tööstuses probleemide juurpõhjuste leidmiseks. Meetodit rakendades tuleb kõigepealt selgelt sõnastada tuvastatud andmekvaliteedi probleem. Mida selgemalt probleem on sõnastatud, seda lihtsam on leida tuvastada probleemi juurpõhjust. Järgmiseks tuleb viis korda küsida „Miks?“. Alustada tuleks sõnastatud kvaliteediprobleemist küsimusega „Miks selline tulemus saadi?“ või „Miks selline 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 soovituste prioriseerimist. Viimaks tuleb meetodi rakendamise tulemid dokumenteerida. Dokumentatsioon peaks kindlasti kirjeldama juurpõhjused, soovitusi nende lahendamiseks, meetodi käigus läbitud järeldusteni jõudmise protsessi ning tuvastatud otseseid ja kaudseid ärilisi mõjusid.

Jälitamise (*Track and Trace*) meetodit saab kasutata tuvastamaks probleemi esmakordset ilmnemist informatsiooni elutsüklis. Nagu ka eelnevalt kirjeldatud meetodi puhul on esimeseks sammuks tuvastatud kvaliteediprobleemi selge sõnastamine. Järgmiseks tuleb kindlaks määrata infomatsiooni elutsükkel ning informatsiooni jälitamise marsruut. Kui informatsiooni elutsüklit on juba varasemalt kirjeldatud saab kasutada olemasolevat informatsiooni ning seda vastavalt vajadusele täiendada. Probleemi esmakordse esinemise tuvastamiseks tuleb järgnevalt võrrelda iga elutsükli sammu puhul sisendandmeid ning sammu teostamise järgset väljundit. Üks võimalus kirjeldatud võrdluse teostamiseks on sisend- ja väljundandmete profileerimine. Võrdluse tulemusena tuvastatakse samm protsessis, kus sisendandmed on õiged, kuid väljundandmed valed. Järgmiseks tuleb välja selgitada vajaminevad muudatused, et andmed oleksid õiged ka pärast probleemse sammu läbimist. Siinkohal võib täiendavalt rakendada teisi meetodeid (Küsi 5 korda „Miks?“ või Põhjus ja tagajärg diagramm), et tuvastada probleemses protsessis tehtavad andmekvaliteeti mõjutavad tegevused. Viimaks tuleb tulemused dokumenteerida. Dokumentatsioon peaks kindlasti kirjeldama juurpõhjused, soovitusi nende lahendamiseks, meetodi käigus läbitud 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 enamat 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 põhjuseid ehk kategooriad. Selleks võib diagrammile lisada enimlevinud üldised põhjused, kasutada varasema analüüsi käigus tuvastatud põhjuseid (näiteks jälitamise meetodi rakendamisel kogutud info), korraldada koosoleku liimetega ajurünnak või kasutada kombinatsiooni mitmest nimetatud viisist. Koosoleku liikmete küsitlemist tuleb jätkata kuni on

jõutud probleemi juurpõhjuseni. Koosolekul leitud juurpõhjused tuleb dokumenteerida nende edasiseks kasutamiseks. Tuvastatud juurpõhjuste põhjal tuleb seejärel koostada soovitused kvaliteediprobleemi juurpõhjuse lahendamiseks. Viimaks tuleb taas kõik tulemused dokumenteerida. Dokumentatsioon peaks kindlasti kirjeldama juurpõhjused, soovitusi nende lahendamiseks, meetodi käigus läbitud järeldusteni jõudmise protsessi ning tuvastatud otseseid ja kaudseid ärilisi mõjusid.



Joonis 7:Põhjus ja tagajärg diagrammi näidis

5.2 Kvaliteediprobleemide mõju hindamine

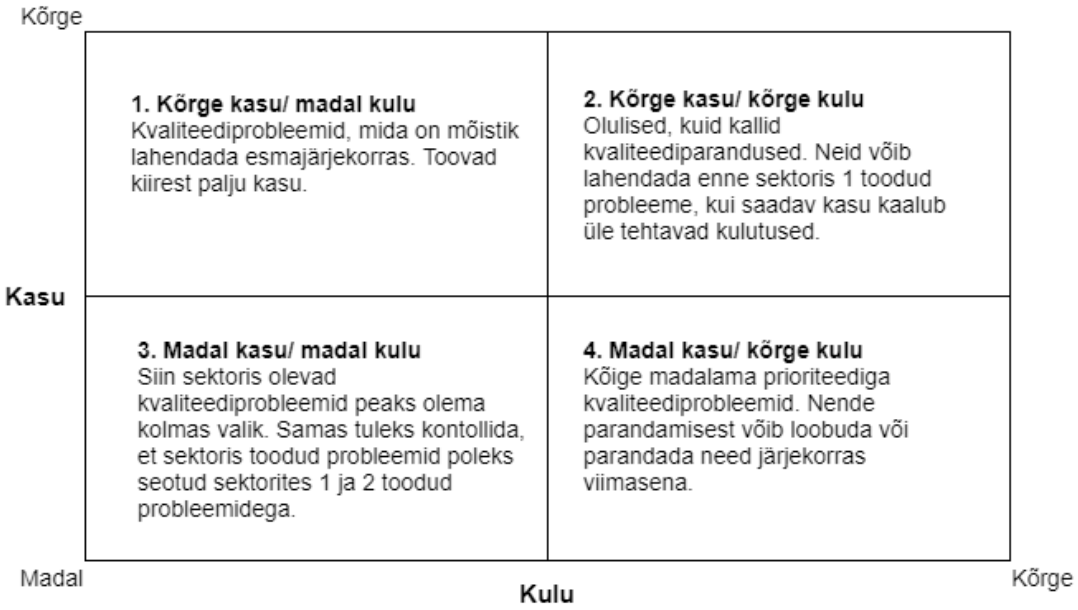
Peale andmekvaliteedi probleemide tuvastamist ning probleemide juurpõhjuste analüüsi tuleb identifitseerida ja prioriseerida probleemide lahendamiseks tehtavad muudatused. Selleks teostatakse esmalt kvaliteediprobleemide ärilise mõju hindamine. Teostatud analüüsi tulemused tuleb edastada andmeomanikule andmekvaliteedi 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 soovitused 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 ärile, mis kaasneb konkreetse kvaliteediprobleemi lahendamisega. Kasu võib olla ka kvaliteediprobleemi lahendamisega kaasnev produktiivsuse tõus või mis tahes muu konkreetse asutuse jaoks oluline muutus. Kulu võib olla kvaliteediprobleemi lahendamise maksumus rahas, selleks tehtava töö maht (ajakulu) või muu konkreetse asutuse jaoks oluline kulutus. Arutelu alguses tuleks tuua mõni näide, et meetodi kasutamine oleks kõigile arusaadav. Järgmiseks tuleb iga prioriseeritava kvaliteediprobleemi jaoks määrata hindamiskriteerium. Hindamiskriteerium võib olla kvalitatiivne (näiteks kliendimugavus) või kvantitatiivne (näiteks ajakulu). Näiteks kui kliendimugavus on oluline kasu tuleks hindamisel küsida: „Mis on arutatavate kvaliteediprobleemide nimekirjast probleemi number 1 mõju kliendimugavusele (skaalal madal kuni kõrge)?“. Kui kvaliteediprobleemi lahendamiseks kuluv aeg on oluline kulu tuleks küsida: „Kui palju aega kulub arutatavate kvaliteediprobleemide nimekirjast probleemi number 1 lahendamiseks (skaalal madal kuni kõrge)?“. Hindamisel on võimalik arvestada ka mitut kriteeriumit, kuid valitud kriteeriumite arv peaks olema piisavalt väike, et hindamisprotsess oleks hallatav.



Joonis 8: Kasu ja kulu näidismaatriks

Järgmise sammuna tuleb eelnevalt seatud kriteeriumite põhjal määrata kvaliteediprobleemidele hinnangud. Üks võimalus selle teostamiseks on lasta kõigil arutelul osalejatel panna oma hinnang kirja ning seejärel neid grupis arutada. Teine võimalus on asetada eri võimalused otse maatriksile nii, et need oleksid kõigile nähtavad ning seeläbi jõuda arutelu käigus ühise kokkuleppeni kvaliteediprobleemi lõpliku asukoha osas maatriksil. Kui kõik arutatavate kvaliteediprobleemide nimekirjas olnud probleemid on hinnatud ning maatriksile

kantud, tuleb antud hinnangud veelkord üle vaadata ning kontrollida, kas eri kvaliteediprobleemide 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).

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 elutsükkel planeerimisest, hankimisest, talletamisest ja jagamisest, säilitamisest, rakendamisest ning kustutamisest või arhiveerimisest. Hindamise ja prioriseerimise tehnika keskendub andmete rakendamise faasile, kus toimub andmeressursside kasutamine eesmärkide saavutamiseks. Andmeid võib järjestada ja hinnata nii konkreetsete kirjete põhjal kui ka andmete gruppidena, 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 seadmisesse 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 andmete antud hinnang. Näiteks kui protsessis X oli andmetele antud hindeks C, kuid protsessis Y oli andmetele antud hinnang A siis on andmete lõplik üldine hinnang A. Viimaks

tuleb analüüsida kogu kogutud informatsiooni ning antud hinnangute järgi tuvastada äriliselt kõige olulisemad andmed. Lõpuks tuleb taas tulemused dokumenteerida. Alljärgnevalt on esitatud näidistabel illustreerimaks tehnika rakendamist (Tabel 8)

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

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

Tasuvusanalüüsi teostamiseks tuleb esmalt identifitseerida kõik kvaliteediprobleemi parandamiseks tehtavad kulutused (ehk tuvastada alginvesteering). Näiteks kulutused tööjõule, väljaõppele, riistvarale ja tarkvarale. Järgmiseks tuleb identifitseerida kvaliteediprobleemi parandamisega kaasnev potentsiaalne rahaline kasu (ehk investeeringu tulu). Seejärel tuleb leida kvaliteedi parandamisest tulenev kokkuhoid, mille arvutamiseks tuleb leida kasu ja kulu vahe. Lisaks tuleb 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). 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.