

TARTU ÜLIKOOL  
Loodus- ja täppisteaduste valdkond  
Arvutiteaduse instituut  
Andmeteaduse õppekava

Rain Eichhorn

# Eesti rahvastikuregistrile sarnase sünteetilise andmestiku reeglipõhine genereerimine

Magistritöö (15 EAP)

Juhendaja(d): Sven Laur, DSc (Tech)  
Liina Kamm, PhD

Tartu 2025

# Eesti rahvastikuregistrile sarnase sünteetilise andmestiku reeglipõhine genereerimine

## Lühikokkuvõte:

Infosüsteemide arendamine ja testimine nõuab realistlikke testandmeid, eriti olukordades, kus pärisandmete kasutamine on privaatsuspiirangute tõttu keelatud. Reeglipõhine sünteetiliste andmete genereerimine võimaldab luua testandmeid, mis jäljendavad pärisandmete struktuuri ja loogikat ilma isikuandmeid töötlemata. Selles töös näitan, kuidas luua Eesti rahvastikuregistrile sobivat reeglipõhist sünteetiliste andmete generaatorit, kasutades ainult avalikke lähteandmeid ja registri äriloogikat. Loodud Pythonil põhinev tööriist võimaldab automaatselt genereerida kasutuslugusid, toetades tõhusat ja turvalist süsteemide testimist. Tulemused pakuvad praktilist alust edasisteks uuringuteks ja turvalist alternatiivi pärisandmete kasutamisele.

## Visuaalne kokkuvõte:



## Võttesõnad:

sünteetilised andmed, reeglipõhine andmete genereerimine, testandmed, rahvastikuregister, andmekaitse, infosüsteemide testimine, GDPR vastavus

## CERCS:

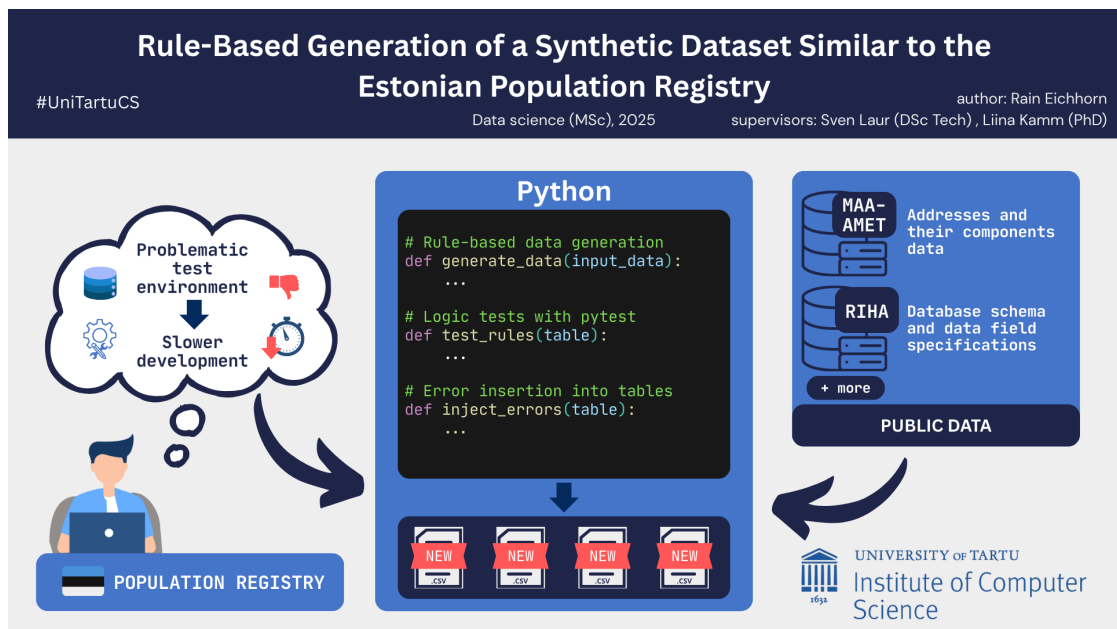
P170 Arvutiteadus, arvanalüüs, süsteemid, kontroll

# Rule-Based Generation of Synthetic Data Similar to the Estonian Population Registry

## Abstract:

The development and testing of information systems require realistic test data, especially in cases where privacy restrictions prohibit the use of real personal data. Rule-based synthetic data generation provides a way to create datasets that mimic the structure and logic of real data without processing personal information. In this thesis, I present the creation of a rule-based synthetic data generator for the Estonian Population Registry, relying solely on public input data and system business logic. The developed Python-based tool automatically generates use cases, enabling efficient and secure system testing. The results offer a practical foundation for future research and a privacy-preserving alternative to real data.

## Graphical abstract:



## Keywords:

synthetic data, rule-based generation, test data, population registry, data privacy, information systems testing, GDPR compliance

## CERCS:

P170 Computer science, numerical analysis, systems, control

# Sisukord

|          |                                                                  |           |
|----------|------------------------------------------------------------------|-----------|
| <b>1</b> | <b>Sissejuhatus</b>                                              | <b>6</b>  |
| 1.1      | Probleem ja taust . . . . .                                      | 6         |
| 1.2      | Eesmärk ja tähtsus . . . . .                                     | 7         |
| 1.3      | Varasemad tööd . . . . .                                         | 7         |
| 1.4      | Töö struktuur . . . . .                                          | 8         |
| <b>2</b> | <b>Süntetilised andmed</b>                                       | <b>9</b>  |
| 2.1      | Ülevaade ja õiguslik raamistik . . . . .                         | 9         |
| 2.2      | Süntetiliste andmete taksonoomia . . . . .                       | 10        |
| 2.3      | Süntetiliste andmete eelised ja piirangud testimisel . . . . .   | 12        |
| <b>3</b> | <b>Metoodika</b>                                                 | <b>13</b> |
| 3.1      | Käsitlusala . . . . .                                            | 13        |
| 3.2      | Sünteesimeetodi valimine ja implementatsioon . . . . .           | 14        |
| 3.2.1    | Kasutuslugude põhine andmekooskõla tingimuste rikkumine . .      | 15        |
| 3.2.2    | Testimine . . . . .                                              | 15        |
| 3.3      | Võrdlus masinõppemeetoditega . . . . .                           | 15        |
| <b>4</b> | <b>Andmegeneraator</b>                                           | <b>17</b> |
| 4.1      | Andmete genereerimise protsess . . . . .                         | 17        |
| 4.1.1    | Sisendandmete lugemine . . . . .                                 | 17        |
| 4.1.2    | Tabelite genereerimine . . . . .                                 | 17        |
| 4.1.3    | Järeltegevused . . . . .                                         | 24        |
| 4.2      | Mõju privaatsusele . . . . .                                     | 24        |
| 4.2.1    | Juhuslik kattuvus . . . . .                                      | 25        |
| 4.2.2    | Läbipaistvus ja privaatsusauditi võimalikkus . . . . .           | 25        |
| 4.2.3    | Teadmuspõhise sünteesi eelised privaatsuse seisukohalt . . . . . | 26        |
| 4.3      | Valideerimine ja kvaliteedikontroll . . . . .                    | 27        |
| 4.4      | Andmete muutmine kasutuslugude jaoks . . . . .                   | 29        |
| <b>5</b> | <b>Tulemused</b>                                                 | <b>30</b> |
| 5.1      | Andmegeneraatori ülesehitus ja töömaht . . . . .                 | 30        |
| 5.2      | Jõudlustulemused . . . . .                                       | 30        |
| 5.3      | Metoodika praktiline valideerimine . . . . .                     | 31        |
| <b>6</b> | <b>Kokkuvõte</b>                                                 | <b>34</b> |
|          | <b>Viidatud kirjandus</b>                                        | <b>38</b> |

|                            |           |
|----------------------------|-----------|
| <b>Lisad</b>               | <b>39</b> |
| I. Repositoorium . . . . . | 39        |
| II. Tabelid . . . . .      | 40        |
| III. Litsents . . . . .    | 50        |

# 1 Sissejuhatus

## 1.1 Probleem ja taust

Tarkvaraarendus ja infosüsteemide testimine sõltuvad kvaliteetsetest ja realistlikest testandmetest. Andmete kättesaadavus ja realistlikkus mõjutavad otseselt süsteemide töökindlust, turvalisust ning võimet täita neile seatud funktsionaalseid nõudeid. Eriti oluline on testandmete kättesaadavus ja realistlikkus tundlike andmete puhul, mida kasutatakse rahvastiku-, sotsiaal- või tervishoiusüsteemides. Sellistes valdkondades on andmete kasutamine piiratud seaduste ja eetikanõuetega ning testandmete loomine on keerukas [5, 35].

Euroopa Liidu isikuandmete kaitse üldmäärus (GDPR) kehtestab ranged reeglid, mis piiravad isikuandmete töötlemist, sh ka nende kasutamist test- ja arenduskeskkondades [5]. Praktikast tähendab see, et tarkvaraarendajad ja testijad ei saa süsteemi testimiseks kasutada pärisandmeid, mistõttu võivad olulised funktsionaalsused jääda testimata. Samas võib ebarealistlike või käsitsi koostatud testandmete kasutamine viia eksitavate tulemusteni – kas ei ilmne päriselus esinevad probleemid või pööratakse tähelepanu olematutele olukordadele. Mõlemad stsenaariumid võivad tuua kaasa olulisi kulusid, süsteemivigu ja turvariske.

Tänapäeva infosüsteemide keerukus, sealhulgas seosed erinevate andmeobjektide vahel, ajalisus, sõltuvused ja muutuv äri loogika, eeldavad testandmeid, mis peegeldavad tegelikku kasutuskonteksti. Privaatsusnõuete ja andmekaitsepiirangute tõttu on juurdepääs sellistele andmetele sageli piiratud või sootuks keelatud [35]. See loob olukorra, kus testimiskvaliteet sõltub otseselt arendajate võimest simuleerida realistlikke andmeid ilma pärisandmeid kasutamata.

Sünteeetilised andmed pakuvad selles olukorras mõjusat lahendust. Sünteeetilised andmed säilitavad tegelike andmete struktuurid, statistilised ja loogilised omadused, kuid ei sisalda isikuandmeid ega konkreetseid pärisandmetel põhinevaid kirjeid. Selline lähenemine võimaldab arendada, testida ja valideerida infosüsteeme oluliselt turvalisemalt ning kooskõlas kehtivate õigusaktidega [16, 34, 35].

Siseministeeriumi infotehnoloogia- ja arenduskeskus (SMIT) haldab ja arendab Eesti siseturvalisuse ning elanikkonna toimimise seisukohalt olulisi infosüsteeme, sealhulgas rahvastikuregistrit. SMIT-i roll on tagada nende süsteemide töökindlus, turvalisus ja vastavus seadusandlikele nõuetele. See töö on loodud tuginedes SMIT-i vajadustele, et pakkuda reeglipõhist sünteeetiliste andmete lahendust Eesti rahvastikuregistri arenduse ja testimise jaoks [6].

Rahvastikuregister on Eesti riiklik alusandmekogu, mis sisaldab põhjalikku teavet kõikide Eesti elanike kohta, nagu nimed, isikukoodid, sünni- ja surmaandmed, elukohtaandmed ning peresündmused. Seda registrit kasutatakse alusena paljudes avalikes teenustes, haldusotsustes ning statistiliste analüüside tegemisel. Registri ulatuslik kasutusvaldkond eeldab, et testimiseks kasutatavad andmed on võimalikult realistlikud ja kajastavad tegelikku olukorda täpselt [33]. Reeglipõhine sünteeetiliste andmete gene-

reerimine, mis tugineb äriloogikal, pakub tugevat kompromissi Eesti rahvastikuregistri testandmete loomisel [32]. Erinevalt keerukatest generatiivsetest mudelitest on reeglipõhine lähenemine läbipaistvam, paremini kontrollitav ning hõlpsam kohandada konkreetse registriloogikaga [21].

## 1.2 Eesmärk ja tähtsus

Selle töö eesmärgiks on arendada välja reeglipõhise andmegeneraatori prototüüp, mis võimaldab luua Eesti rahvastikuregistri andmetega sarnaseid testandmeid. Selle jaoks analüüsisin nii teoreetilisi lähtepunkte kui ka praktilisi lahendusi. Töö käigus arendasin välja tarkvara, mis genereerib erinevaid isikuandmeid (nagu näiteks aadressid, dokumendid ja suhted) nii, et need peegeldaksid võimalikult hästi tegeliku registri loogikat. Privaatsusriskide vältimiseks kasutasin generaatori loomisel ainult avalikke andmeid.

Töö praktiline tähtsus seisneb kolmes aspektis.

- **Testimisvõimekuse parandamine:** Süsteemide arendajad ja testijad saavad töötada realistlikes tingimustes ka ilma juurdepääsuta tundlikele andmetele.
- **Privaatsuse tagamine:** Lahendus viib isikuandmete lekked või väärkasutuse miinimumi, võimaldades samal ajal usaldusväärset arendustegevust.
- **Skaleeritavus ja korduvkasutatavus:** Väljatöötatud generaatorit saab kohandada ka teiste registrite või andmestike jaoks, sh teistes riikides või domeenides.

Oluline osa töö rakenduslikust väärtusest seisneb selles, et loodud generaator laiendab arendajate, testijate ja kvaliteedispetsialistide võimalusi koostada sisulisi ja mitmekesiseid kasutuslugusid. Reeglipõhised sünteetilised andmed võimaldavad luua stsenaariume, mis kajastavad erinevaid päriselulisi olukordi, katavad keerukamaid andmestruktuure ja võimaldavad kontrollida süsteemide käitumist äriloogika piires.

Töö tulemusena arendasin välja tööriista prototüübi, mida kasutasin testandmete loomiseks ning mida saab siduda ka erinevate testkeskkondadega. Lisaks loob see aluse edasisteks uurimistöödeks sünteetiliste andmete genereerimise teemal, pakkudes rakenduslikku ja teoreetilist väärtust.

## 1.3 Varasemad tööd

Sünteetiliste andmete kasutamine tarkvaraarenduses ja andmekaitstes on Eestis viimastel aastatel järjest enam tähelepanu saanud. Näiteks on Eesti Andmekaitse Inspektsioon rõhutanud sünteetiliste andmete olulisust olukordades, kus tegelike isikuandmete kasutamine ei ole lubatud või on piiratud. Sünteetilised andmed võimaldavad luua testandmestikke, mis jäljendavad tegelike andmete struktuuri ja loogikat, kuid ei sisalda päris isikuandmeid, aidates seeläbi tagada privaatsusnõuete täitmise tarkvaraarenduse ja testimise

protsessides. Sellised tööriistad on eriti olulised avaliku sektori arenduskeskkondades, kus andmekaitsevenõuded on ranged [20].

Süntetilisi andmeid rakendatud keele- ja kõnetehnoloogias: näiteks eesti keele grammatikakontrolli arenduses loodi vigadega sünteetilised tekstikorpused, mis võimaldasid treenida korrektuurimudeleid [26]. Samuti on Tartu Ülikooli ja Eesti Keele Instituudi koostöös välja töötatud närvivõrkudel põhinev kõnesünteesi süsteem, mille treenimisel kasutati sünteetilist kõneandmestikku, tagamaks paremat kõnekvaliteeti ja mitmekesisust [8]. Rahvusvaheliselt on turule tulnud ka platvormid nagu Syntho, mis võimaldavad andmete struktuurilisi ja statistilisi omadusi säilitades genereerida testandmestikke, ilma et need sisaldaksid tegelikke isikuandmeid [7].

Põhjalikult on uuritud sünteetiliste andmete kasutusvõimalusi õiguskaitseasutustes, kus tundlike isikuandmete töötlemine seab ranged juriidilised ja tehnilised piirangud. Magistr töö [13] raames töötati välja raamistik sünteetiliste andmete rakendamiseks Eesti Politsei- ja Piirivalveametis, mille eesmärk oli parandada andmekaitset ja infoturvet ning võimaldada andmete jagamist teadlastele ning välistele arendajatele. Uuring tõi esile sünteetiliste andmete potentsiaali PET-tehnoloogiana (ingl k *privacy enhancing technology*) ning rõhutas vajadust tasakaalu leidmise järele privaatsuse ja andmekasutuse vahel, eriti siseturvalisuse valdkonnas.

Rahvastikuregistri kontekstis pakub asjakohast vaatenurka uuring, kus mobiilpositsioneerimise andmeid kasutati elukohaandmete täpsustamiseks, viidates sellele, et ametlikud registriandmed ei pruugi alati olla asjakohased ning andmete usaldusväärsuse parandamiseks on vaja uusi lähenemisi [25]. See töö jätkab samas suunas, pakkudes reeglipõhist sünteetiliste andmete genereerimise mudelit, mis sobitub Eesti avaliku sektori vajadustega ja toetab testkeskkondade arengut.

## 1.4 Töö struktuur

Töö struktuur on korraldatud järgmiselt. Teises peatükis selgitatakse sünteetiliste andmete mõistet, õiguslikku raamistikku ja taksonoomiat. Kolmandas peatükis kirjeldatakse meetodikat: käsitlusala piiramist, reeglipõhist sünteesimeetodit ning andmete valideerimist. Neljandas peatükis tutvustatakse loodud andmegeneraatorit koos töövooga, privaatsusmõjude ja kvaliteedikontrolliga. Viiendas peatükis käsitletakse töö tulemusi, sealhulgas generaatori ülesehitust, jõudlust ning meetodika praktilist rakendamist. Kuuendas peatükis esitatakse kokkuvõte ning tuuakse välja võimalikud edasiarenduste suunad.

## 2 Sünteetilised andmed

### 2.1 Ülevaade ja õiguslik raamistik

Sünteetilised andmed on kunstlikult loodud andmed, mis jäljendavad pärisandmete statistilisi ja struktuurseid omadusi, kuid ei sisalda tegelike isikute kirjeid ega kaudselt tuvastatavaid isikuandmeid. Selliseid andmeid saab genereerida olemasolevate andmekogumite põhjal või luua täielikult nullist, tuginedes konkreetsetele reeglitele või masinõppe mudelitele [16, 35]. Eesmärgiks on pakkuda realistlikke alternatiive tegelikele andmetele, võimaldades andmeid turvaliselt jagada, testida, arendada ning analüüsida ilma reaalsete isikute privaatsust ohustamata. Viimasel ajal on sünteetiliste andmete loomisel hakatud üha enam kasutama masinõppe ja generatiivsete tehisintellekti mudelite abi, näiteks generatiivseid vastandvõrke (GAN; ingl k *generative adversarial network*), mis võimaldavad luua keerukaid ja struktureeritult realistlikke andmekogumeid [15, 37].

Sünteetiliste andmete vajadus kasvab pidevalt, kuna andmete töötlemisele ja jagamisele seatakse järjest karmimaid piiranguid, eriti tundlikes valdkondades nagu tervishoid ja finantsteenused, samuti riiklikes andmekogudes. Nendes valdkondades seab isikuandmete konfidentsiaalsuskohustus ranged piirangud andmete vabale liikumisele ja jagamisele. Traditsioonilised de-identifitseerimise tehnikad, nagu pseudonüümimine ja agregatsioon, ei pruugi alati pakkuda piisavat privaatsuskaitset, eriti kui on võimalik andmekogusid kombineerides kaudselt isikuid tuvastada [11, 14]. Lisaks privaatsusprobleemidele on praktikas ilmnenud ka väljakutseid sünteetiliste andmete kvaliteedis – näiteks võib kaduma minna semantiline tähendus või jääda märkamata valdkonnaspetsiifiline keerukus [12].

Euroopa Liidus reguleerib andmetöötlust peamiselt isikuandmete kaitse üldmäärus (GDPR), mis kehtestab selged ja ranged raamid isikuandmete töötlemiseks, kogumiseks ja jagamiseks. GDPR-i eesmärgiks on tagada isikuandmete töötlemisel kõrge privaatsustase, nõudes andmete anonüümsuse ja pseudonüümsuse tagamist. Selleks, et andmeid saaks pidada anonüümseks, tuleb arvestada kõigi võimalike meetoditega, mille abil võiks andmesubjekti otseselt või kaudselt tuvastada [9, 11]. Just siin tuleb selgelt esile sünteetiliste andmete eelis – kuna need ei sisalda päris isikuandmeid, on identifitseerimise risk minimaalne. Siiski on oluline mõista, et sünteetiliste andmete staatus GDPR-i mõistes sõltub konkreetsetest asjaoludest ning andmete loomise protsessi läbipaistvus ja dokumenteeritus on võtmetähtsusega [31].

Samas ei ole sünteetilised andmed alati täiesti vabastatud GDPR-i kohaldamisest. Kui sünteetilised andmed on loodud tegelike andmete põhjal ja nende loomise protsess pole piisavalt range, võib tekkida kaudse re-identifitseerimise oht. Sellest tulenevalt on valitsusasutused, teadusasutused ja andmeanalüüsi platvormid hakanud kasutama täiendavaid privaatsuskaitse tehnoloogiaid, nagu diferentsiaalprivaatsus ja reeglipõhised genereerimismeetodid, et vähendada võimalikke tagasituvastuse riske ning tagada andmete turvalisus [9, 10]. Sünteetiliste andmete kasutamise põhimõtted erinevad rahvusvaheli-

selt, sõltudes iga riigi seadustest ja eetikanõuetest. See toob esile vajaduse kujundada andmelahendusi vastavauses kohaliku regulatiivse ja kultuurilise raamistikuga [17].

Reeglipõhine sünteetiliste andmete genereerimine, mida selles töös rakendan, loob andmeid selgete ja rangete eelmääratletud reeglite järgi, mitte ei kopeeri statistilisi mustreid olemasolevatest andmetest. Selline meetod tagab, et loodavad andmed vastavad täpselt seaduses ja äriloogikas sätestatud nõuetele, näiteks elukohtade kehtivus, dokumentide vorming, isikuandmete seosed ja ajalised piirangud [3]. Reeglipõhine genereerimine aitab tagada, et loodavad sünteetilised andmed on realistlikud ja sobivad tarkvaraarenduse ning süsteemide testimise jaoks, vältides samal ajal tegelike isikute privaatsusriske. Samuti võimaldab see andmestike kontrollitavust ja korduvkasutatavust, mis on oluline suuremahuliste infosüsteemide arendamisel [27].

Lisaks GDPR-ile võivad sünteetiliste andmete loomist ja kasutamist mõjutada ka spetsiifilised valdkondlikud regulatsioonid, näiteks tervishoiusektoris meditsiiniliste andmete kaitse või finantssektoris finantsandmete töötlemise nõuded. Selliste valdkondlike regulatsioonide tõttu tuleb sünteetiliste andmete loomine korraldada koostöös domeeniekspertidega, tagamaks vastavus kõigile kehtivatele õigusaktidele ja kvaliteedistandarditele. Samuti on oluline kaasata andmeanalüütikuid ja IT-spetsialiste, et tagada andmete tehniline sobivus ja usaldusväärsus erinevates kasutusstsenaariumites [15].

## 2.2 Sünteetiliste andmete taksonoomia

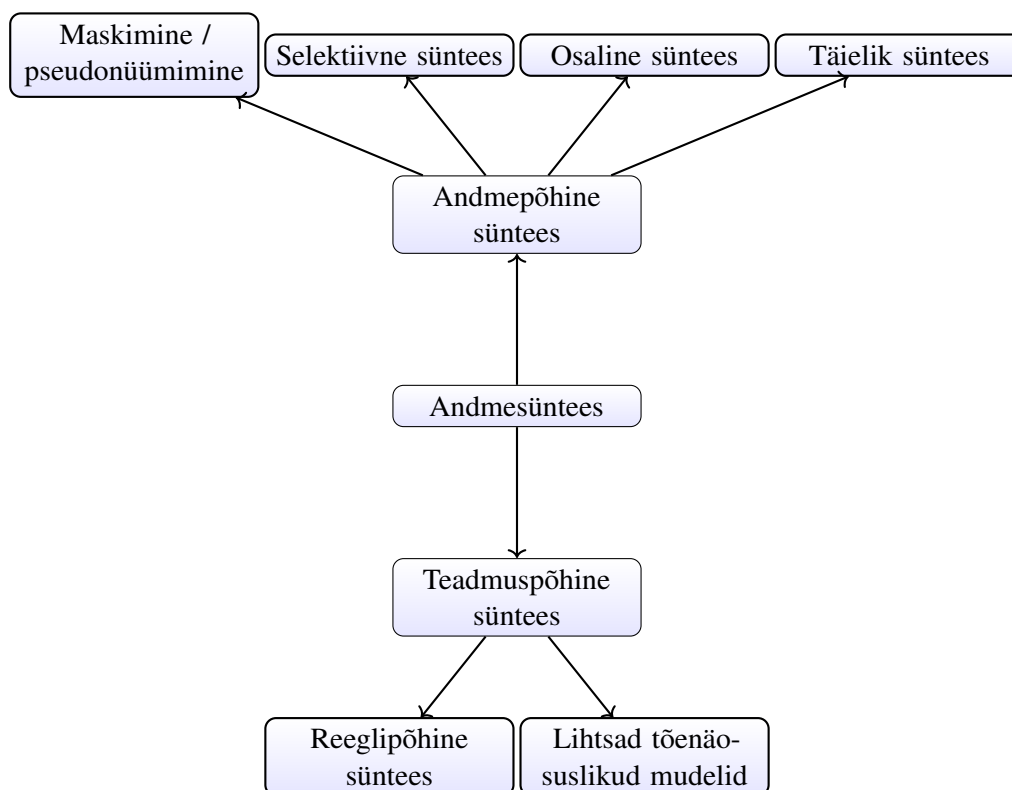
Sünteetiliste andmete genereerimise meetodid saab liigitada erinevate kriteeriumide alusel, sõltuvalt nii kasutatavast meetodikast kui ka loodud andmete eesmärgist. Meetodite eristamine aitab selgitada, millal on sobilik kasutada andmepõhiseid või teadmuspõhiseid lähenemisi ning milliseid privaatsustagatise igauks võimaldab.

Sünteetiliste andmete genereerimise meetodiline taksonoomia keskendub sellele, kuidas andmeid sünteesitakse, ning eristab kahte peamist lähenemisviisi: andmepõhist ja teadmuspõhist sünteesi (joonis 1). Andmepõhisel sünteesil põhineb sünteetiliste andmete loomine olemasolevate andmekogumite statistiliste omaduste jäljendamisel. Selle raames eristatakse mitmeid alamviise.

- **Täielik süntees** – kogu andmestik sünteesitakse mudelite abil. Näiteks Urban Institute lõi täielikult sünteetilise sotsiaalteenuste andmestiku, mis sisaldas muutujaid nagu sissetulekutoetused, rass, sugu ja haridustase [19].
- **Osaline süntees** – sünteesitakse ainult tundlikud või valitud väljad. Näiteks kasutati osalist sünteesi USA tööjõuandmetes, kus tundlikud väljad asendati sünteetiliste väärtustega, säilitades samas ülejäänud andmed [22].
- **Selektiivne süntees** – sihitud süntees konkreetsetele populatsioonidele või mustritele. Näiteks MOSTLY AI võimaldab tingimuslikku andmete sünteesimist, kus

kasutaja saab määrata kindlad tingimused, mille alusel sünteetilised andmed genereeritakse [24].

- **Maskimine ja pseudonüümimine** – pärisandmed asendatakse muudetud väärtustega, mis säilitavad algsete andmete struktuuri või vormingu, kuid ei võimalda otsest tuvastamist. Näiteks krediitkaardi numbrid asendatakse vormingusarnaste, kuid mitteeksisteerivate numbritega, et säilitada andmete struktuur testimise eesmärgil [18].



Joonis 1. Sünteetiliste andmete metoodiline taksonoomia.

Teadmuspõhine süntees tugineb eeldefineeritud reeglitele ja domeenipõhisele loogikale, mitte olemasolevate andmete statistilistele mustritele. Selle lähenemise raames eristatakse kahte peamist suunda.

- **Reeglipõhine süntees** – andmed luuakse vastavalt äriloogikale ja normidele. Näiteks Syntho AI pakub reeglipõhist sünteetiliste andmete generaatorit, mis võimaldab luua andmeid nullist vastavalt määratud reeglitele [36].

- **Auditeeritav tõenäosuslik mudeldamine** – lihtsustatud, kontrollitavate parameetritega mudelid. Näiteks kasutatakse sünteetiliste andmete genereerimiseks Bayesi võrku, mille abil saadakse andmed teadaolevate parameetrite alusel [28].

## 2.3 Sünteetiliste andmete eelised ja piirangud testimisel

Sünteetilised andmed on muutunud oluliseks tööriistaks nii teadusuuringutes kui ka tehismintellekti arenduses, kus nende eesmärk on tagada juurdepääs kvaliteetsetele andmestikele ilma isikute privaatsust ohustamata [21]. Enamasti hinnatakse sünteetiliste andmete kvaliteeti statistilise kasutuse ja masinõppe kontekstis, kus peamine eesmärk on säilitada võimalikult täpselt pärisandmete jaotused, korrelatsioonid ja muud statistilised omadused. Sünteetilised andmed peavad olema piisavalt realistlikud, et masinõppemudelid, mille treenimiseks neid kasutatakse, suudaksid üldistada pärisandmete käitumist [30]. Samuti on oluline, et sünteetilisest andmetest saadud järeldused oleksid usaldusväärsed ja võrreldavad pärisandmete põhjal tehtud järeldusega [16]. Ebatäpsed või kallutatud sünteetilised andmestikud võivad tugevdada juba olemasolevaid eelarvamusi ja kahjustada mudelite tulemuslikkust või otsustusprotsesse [14].

Statistilisele täpsusele keskenduv lähenemine on süsteemide testimise kontekstis ebaoluline või isegi eksitav. Testimisel on sünteetiliste andmete eesmärk süsteemi funktsionaalsuse, äriloogika ja kasutusjuhtumite realistlik katmine, aga mitte statistiline vastavus. Andmestik peab sisaldama tüüpilisi juhtumeid, mis kajastavad süsteemi tavapärasest kasutusest, ning olulisi erijuhte. Näiteks osaliselt täidetud vormid, piiripealsed väärtused või loogiliselt keerulised sisendid. Samas ei ole vaja säilitada haruldasi statistilisi ekstreemsusi, mille esinemise tõenäosus on marginaalne või mille mõju süsteemile on aktsepteeritav jääkriski tasemel. Siin ei ole esmane mure mitte järelduste usaldusväärsus populatsiooni kohta, vaid süsteemi käitumise põhjalik ja realistlik testimine erinevates tingimustes.

Testimiseks sobivate sünteetiliste andmete genereerimine nõuab seetõttu reeglipõhist ja eesmärgipärast lähenemist. Reeglid peavad tagama andmete sisemise loogika vastavuse süsteemis kehtivatele ärireeglitele, näiteks väljade seosed ja dokumentide tüübid. Oluline on vältida nii liigset rangust, mis võib kõrvale jätta olulised piirsituatsioonid, kui ka liigset täpsust, mis matkib pärisandmeid nii üksikasjalikult, et see võib suurendada privaatsusriske [11]. Testandmete kvaliteeti ei saa hinnata samade põhimõtete järgi nagu masinõppes või statistikaanalüüsis. Siin on määrav see, kuivõrd andmestik toetab süsteemi käitumise testimist eriolukordades ja katab vajalikud äriloogilised stsenaariumid, mitte see, kui täpselt see kajastab populatsiooni statistilisi omadusi.

## 3 Metoodika

### 3.1 Käsitlusala

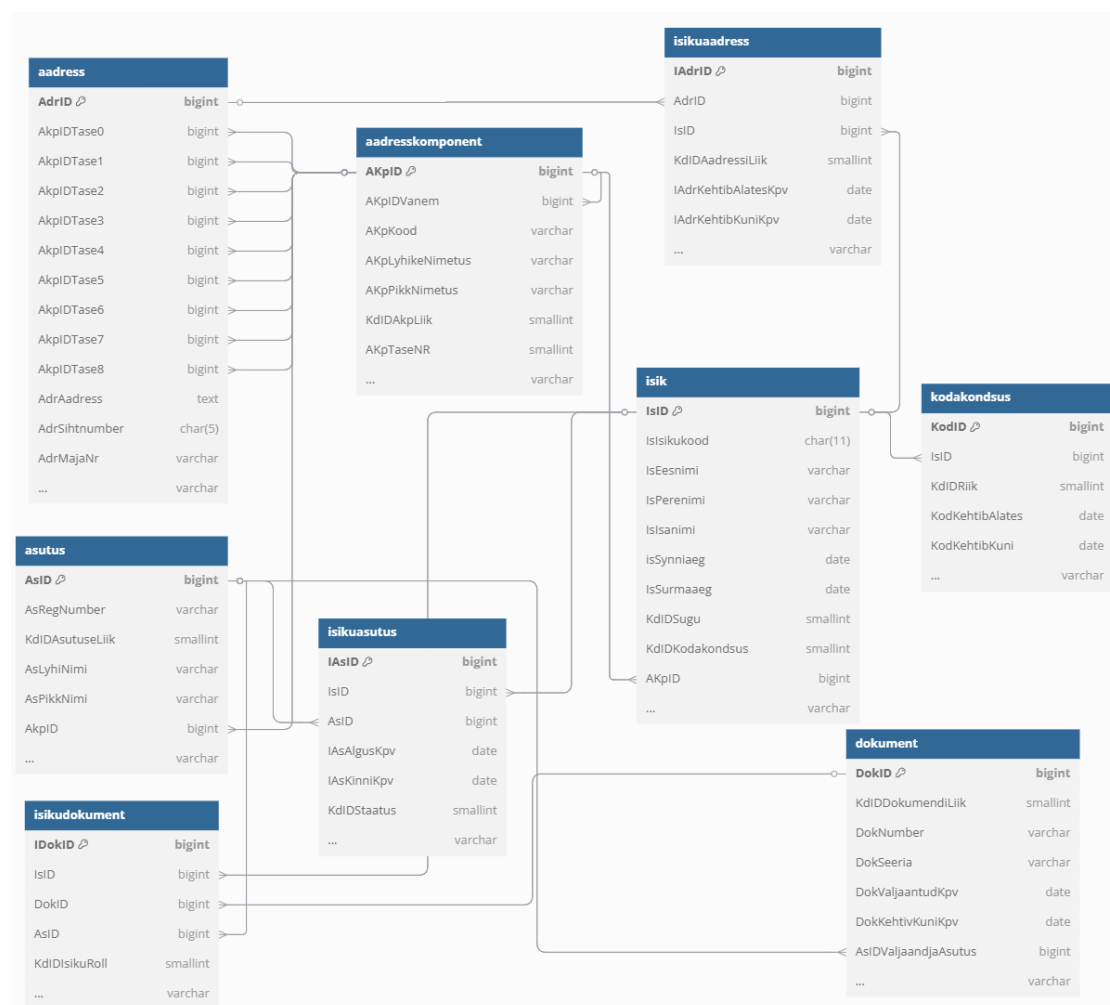
Määrasin töö käsitlusala, lähtudes SMIT-i poolt kirjeldatud neljast kasutusloost. Kasutuslood on ärioloogikal põhinevad stsenaariumid, mis kirjeldavad süsteemi tüüpilisi või kriitilisi kasutusolukordi. Sellised olukorrad on testimisel kriitilise tähtsusega, kuna need nõuavad realistlikke ja keerukate seostega andmekombinatsioone. Valitud juhtumite hulka kuuluvad olukorrad:

1. isikul on samaaegselt mitu kehtivat aadressi,
2. dokumendi kehtivusaeg on lõppenud,
3. elukoht algab tulevikus,
4. isikule on määratud vastuolulised kodakondsused.

Kuigi rahvastikuregistri RIHA kirjelduses on kokku 32 andmetabelit [1], kesken-  
dusin töös piiratud andmevalikule, mille sobitasin nimetatud kasutuslugudega. Valisin  
käsitlusala selliselt, et keskenduda prototüübi süsteemsele ja põhjalikule arendusele.  
Valisin need tabelid, mis on kasutuslugude katmiseks kõige olulisemad ning mille vahel  
esineb piisavalt andmestruktuurilisi sõltuvusi. See võimaldas hinnata loodava sünteesi  
keerukust ning andmestiku loogilist järjepidevust. Valitud tabelid, millele töö tugineb, on  
järgmised:

- aadress – Eestis kehtivad aadressiobjektid;
- aadresskomponent – kehtivad ja endised hierarhilised aadressielemendid;
- isik – isikute põhiandmed;
- dokument – isikutega seotud dokumendid;
- asutus – asutuste metaandmed;
- isikuasutus – seosed isikute ja asutuste vahel;
- isikuaadress – seosed isikute ja nende aadresside vahel;
- isikudokument – seosed isikute ja dokumentide vahel;
- kodakondsus – isikute kodakondsuse andmed.

Valitud tabelitevahelised sõltuvused ja relatsioonid on toodud joonisel 2. Auditväljad, na-  
gu IASIDLooja ja IASIDMuutja, mis viitavad tabelile isikuasutus, on joonise selguse  
huvides välja jäetud, kuid sisalduvad kõigis vastavates tabelites.



Joonis 2. Genereeritud tabelite relatsioonid ilma auditväljadeta.

## 3.2 Sünteesimeetodi valimine ja implementatsioon

Arvestades töö eesmärgi ja privaatsusnõudeid, valisin sünteetiliste andmete loomiseks reeglipõhise lähenemise. See võimaldas koostada andmestiku, mis järgib rahvastiku-registri ärioloogikat, ilma et kasutataks päris isikuandmeid. Reeglipõhine süntees sobib eriti hästi tundlike registrite testimiseks, kuna see on läbipaistev, auditeeritav ja kergesti kohandatav [21, 29].

Arendasin andmegeneraatori Pythonis, kuna see pakkus häid tööriistu andmetööt-luseks (Pandas) ja testimiseks (pytest). Ehituse jagasin loogilisteks komponentideks: lähteandmete laadimine, ärioloogikal põhinev andmete genereerimine, kasutuslugudest lähtuvate andmevigade lisamine (valikuline etapp) ning loodud andmestiku valideerimine.

Selline arhitektuur võimaldab süsteemi laiendada ja kohandada vastavalt vajadusele.

### **3.2.1 Kasutuslugude põhine andmekooskõla tingimuste rikkumine**

Lisaks korrektsele andmete sünteesile töötasin välja eraldi komponendi, mis võimaldab sihipäraselt luua vigaseid olukordi vastavalt määratletud kasutuslugudele. Selle komponendi abil muutsin olemasolevat korrektset andmestikku kontrollitud viisil. Näiteks saab lisada isikule mitu samaaegselt kehtivat elukohta või määrata dokumendile aegunud kehtivusaja, säilitades siiski selle staatuse „KEHTIV“. Selline lahendus võimaldab testida süsteemi käitumist äriloogikat rikkuvates olukordades, säilitades samal ajal baasandmete loogilise järjepidevuse.

### **3.2.2 Testimine**

Genereeritud andmete korrektsuse tagamiseks koostas automaattestid, kasutades Pythonis pytest raamistikku. Testide koostamisel lähtusin põhimõttest, et need seaksid edaspidiste arendajate jaoks selged kontrollid andmete struktuuri ja äriloogika kohta. Samuti on testidega võimalik kontrollida kasutuslugude andmevigade rakendumist. Kontrollisin testidega iga tabeli puhul, et kõik kohustuslikud väljad oleksid täidetud, primaarvõtmed oleksid unikaalsed ning tabelitevahelised välisvõtmed viitaksid olemasolevatele ridadele.

## **3.3 Võrdlus masinõppemeetoditega**

Masinõppe- ja generatiivsetel mudelitel põhinevad lahendused, nagu generatiivsed vastandvõrgud (GAN) ja variatsioonilised autoenkoodrid (VAE), võimaldavad sünteesida realistlike andmeid, tuginedes olemasolevate andmekogumite mustritele. Sellised mudelid on eriti sobivad olukordades, kus on vaja säilitada keerulised seosed ja kõrged dimensioonid, näiteks pildi- või kõneandmetes. Nende suurimaks puuduseks on aga vajadus pärisandmete järele: mudelite treenimiseks tuleb omada juurdepääsu andmestikele, mis võivad sisaldada tundlikku teavet. GDPR-i kontekstis tähendab see, et loodud sünteetilise andmestiku anonüümsus peab olema tõendatavalt tagatud, vastasel juhul võib kaasneda privaatsusnõuete rikkumise oht [14, 31].

Reeglipõhine süntees ei vaja treeningandmeid. Selle asemel luuakse andmed äri- loogika alusel, kasutades spetsifikatsioone ja määratletud struktuurireegleid. Selline lähenemine muudab sünteetilise andmete genereerimise protsessi täielikult auditeeritavaks: iga loodud väärtust või seost saab põhjendada reegluga ning on võimalik kontrollida, et ei esine varjatud või juhuslikke seoseid, mis viitaksid tegelikele isikutele. Lisaks tagab reeglipõhine süntees, et loodud andmestik on kooskõlas testkeskkonnas rakenduva loogikaga ning võimaldab sihipäraselt modelleerida just neid erijuhte, mis on süsteemi töökindluse seisukohalt kõige olulisemad [21].

Testimise kontekstis on andmepõhistel mudelitel üks nõrkus: need kipuvad reprodutseerima keskmisi või tüüpilisi juhtumeid ning ei kata hästi haruldasi, kuid kriitilisi piirsituatsioone. Testandmestiku kvaliteedi mõttes ei ole alati soovitatav peegeldada reaalandmete sagedusjaotusi – sageli tuleb just haruldasi olukordi esile kutsuda, et süsteemi käitumist nende suhtes hinnata. Reeglipõhine süntees võimaldab seda otse: näiteks saab teadlikult luua olukorra, kus isikul on samaaegselt mitu kehtivat aadressi, aegunud dokument või vastuoluline kodakondsus. Andmepõhised mudelid seevastu vajavad väga täpselt kureeritud treeningandmeid, et selliseid juhtumeid esile tuua [23].

Privaatsuse aspektist pakub reeglipõhine lähenemine selget eelist: kuna see ei kasuta pärisandmeid ega statistiliselt tuletatud mudeleid, puudub oht, et sünteesitud andmestik sarnaneb tahtmatult mõne reaalse isiku andmetega. Reeglid ise on dokumenteeritavad ja auditeeritavad ning neid saab luua täielikult spetsifikatsioonide alusel, ilma et peaks analüüsima olemasolevaid andmeid. Kui reeglid siiski tuletatakse olemasolevate andmete analüüsi teel, tuleb neid hinnata eraldi privaatsusriskide osas – näiteks kas mõni invariant andmestikus võib viia isiku kaudse tuvastamiseni. Sellise riski vähendamine on oluline, et reeglite loomine oleks jälgitav ja auditeeritav.

## 4 Andmegeneraator

### 4.1 Andmete genereerimise protsess

Andmete genereerimise üldine protsess on automatiseeritud ning üles ehitatud sammude-na, mis tagavad loogiliselt kooskõlalised ja õiged seosed erinevate tabelite vahel. Koodi kirjutamisel kasutasin Pythonit ning andmetetöötlemiseks Pandas teeki. Genereerimis-protsess pseudokoodina on joonisel 3 ning genereerimise töövoo visuaal joonisel 4.

#### 4.1.1 Sisendandmete lugemine

Skriptid alustavad eeltöötlust vajalike lähtefailide (CSV ja Excel) lugemisega. Need failid sisaldavad pärisandmeid, mis on avalikult kättesaadavad. Kasutasin andmeid aadressi-andmete süsteemi (ADS) avalikust teenusest [2] ning äriregistri asutuste loetelust [4]. ADS-ist saadud failid sisaldavad aadressiobjektide andmeid ilma isikuseosteta, samas kui äriregistri andmestik hõlmab üksnes asutustega seotud teavet.

Rahvastikuregistri andmekoosseisu kirjeldusest selgub, et seal kasutatakse kodifikaatorit. See on loetelupõhine klassifikaator, mille abil tähistatakse standardiseeritud väärtusi (näiteks isikusuhte tüüp, aadressi staatus, kodakondsus, dokumendi liik jm). Siiski ei ole registris kasutatava kodifikaatori täielik loend koos väärtustega avalikult kättesaadav. Seetõttu koostasın vajaliku kodifikaatori ise, lähtudes RIHAs dokumenteeritud registri-struktuurist ning X-tee teenuste kirjeldustest, mis määratlevad erinevad andmepäringud ja nende väljundid [1]. Loodud tabelid koondasin ühte Pythoni sõnastikku, et neid oleks edasistes sünteesietappides hõlpsam käsitleda.

#### 4.1.2 Tabelite genereerimine

Kõik ülesande lahendamiseks vajalike tabelite genereerimisloogika arendasin ise, tagades loogiliste seoste ja välisvõtmete järjepidevuse. Protsess algab aadressi- ja aadressikomponentide tabelite loomisega, jätkub ajutise seostetabeli genereerimisega, seejärel asutuste ning isikutega seotud tabelite täitmisega ning lõpeb dokumentide ja sidetabelite koostamisega. Viidatud tabelid on kirjeldatud lisa II.

Iga tabeli täitmisel kasutasin erinevaid reeglipõhiseid ja juhuslikke otsuseid, mis määravad näiteks staatuse, kuupäevade olemasolu või seoste loomise. Tõenäosused on kirjas selgelt ja koondatult funktsioonide ülaosas, et võimaldada nende kiiret kohandamist arenduse või testide käigus. Neid saab lähemalt uurida koodis (vaata lisa I).

---

```

# 1. SISENDANDMETE LUGEMINE
load_data():
    - Lae kõik sisendfailid (kodifikaator.csv, aadress.csv,
      aadressskomponent.csv, asutus.xlsx)

# 2. TABELITE GENEREERIMINE
generate_tables(data, num_records, seed):
    - Genereeri aadressitabelid (generate_address,
      generate_address_komponent)
    - Moodusta ajutised isiku seosed (
      generate_temp_relationships)
    - Seo isikud asutustega (build_isik_asutus)
    - Genereeri asutuste tabel (generate_asutus)
    - Lisa iga isiku 1-N aadressi (generate_isikuaadress)
    - Koosta isikuandmed ja kodakondsus (generate_isik,
      generate_kodakondsus)
    - Genereeri igale dokumendile kirjed (
      generate_person_document, dokumendid koos ISID)
    - Seo dokumendid isikutega (generate_isikudokument)
    - Pärast seostamist salvesta ka 09_dokument.csv (dokumendid
      ilma IsID-ta)
    - Tagasta sõnastik: {"01_aadress.csv": df_adr, ...}

# 3. ANDMETE SALVESTAMINE
save_to_csv(dataframes):
    - Salvesta iga tabel CSV-ks

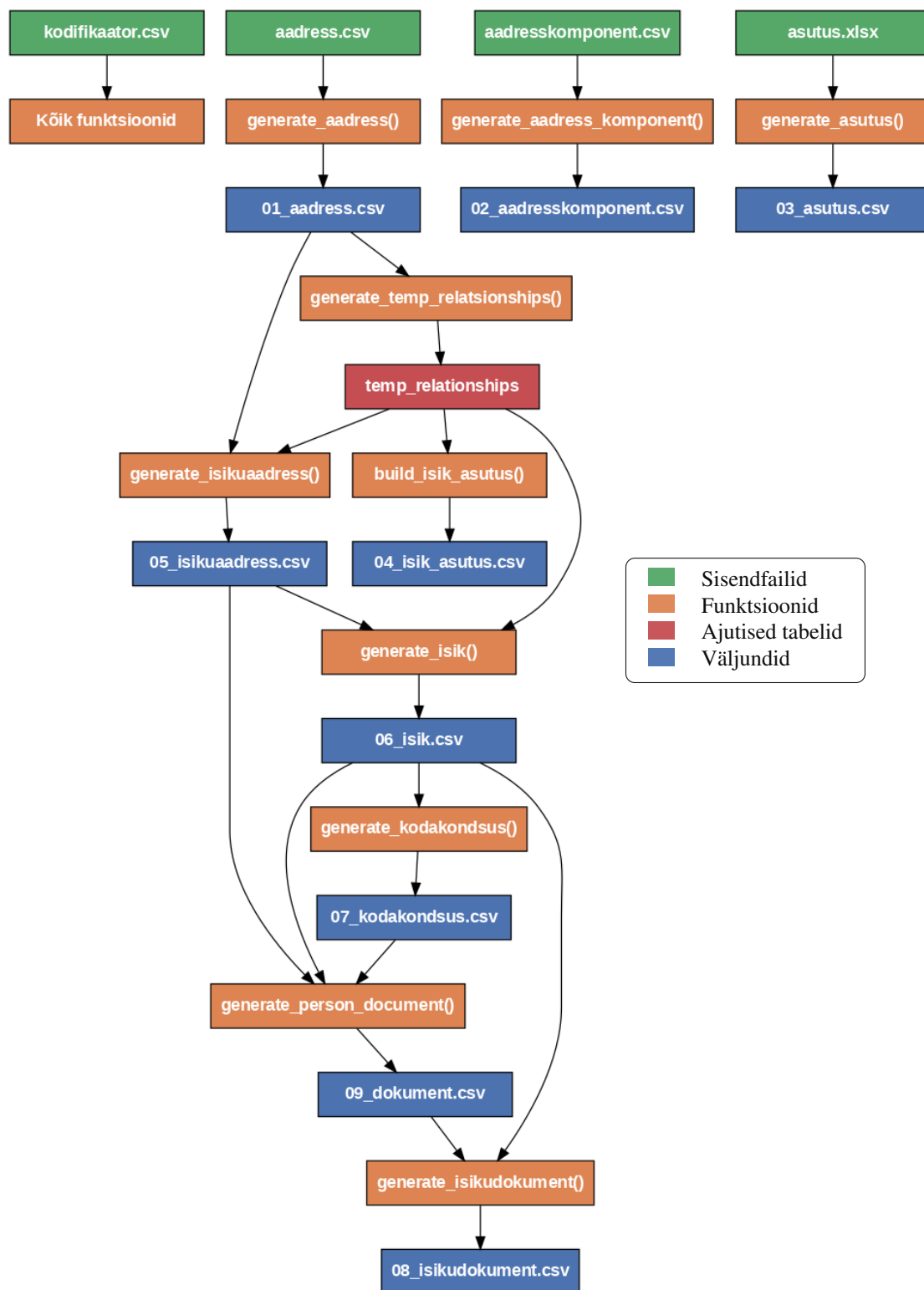
# 4. IAs-METAANDMETE LISAMINE
inject_ias_metadata_to_csv_folder(df_isik_asutus):
    - Ava iga väljundkaustas olev CSV
    - Lisa veerud IAsIDLooja, IAsIDMuutja, AsID
    - Vali kordumatu ajatempliga sobiv IAsID (lru_cache)
    - Uuenda read ja salvesta CSV uuesti

# 5. PÕHIPROTSESS
main(num_records, output_folder, seed):
    - Määra väljundkaust
    - Lae andmed (load_data)
    - Genereeri andmestikud (generate_tables)
    - Salvesta CSV-d (save_to_csv)
    - Lisa IAs metaandmed (inject_ias_metadata_to_csv_folder)

```

---

Joonis 3. Andmete genereerimise kirjeldus pseudokoodina.



Joonis 4. Andmegeneraatori genereerimisprotsessi töövoog.

**Tabel address.** Tabeli väljade kirjeldus on toodud tabelis 2. Tabeli täitmiseks kasutatakse funktsiooni `generate_aadress`, mis võtab sisendiks juhuslikult valitud read Eesti aadressiregistri failist. Valitud kirjete arv on võrdne isikute kirjete arvuga. Iga kirje kohta määratakse seisundikood `KdIDAdressiStaat` veerus `AADR_OLEK` sisalduva väärtuse alusel, kasutades klassifikaatorite tabelit. Igale kirjele genereeritakse unikaalne viiekohaline postiindeks ning määratakse kindlaks kõrgeim ruumitaseme komponent `AkpIDTase0`, mille väärtus fikseeritakse Eesti koodile (EE). Tulemuseks on tabel, mille igas kirjes on mitmetasemeline aadresside hierarhia (maakond → vald → asula jne), klassifikaatoripõhine aadressi staatus ja vajalikud metaandmed. Veerud, mida lähteandmetes pole, täidetakse vaikimisi tühiväärtustega. Aadressid ei sisalda isikukode ning seetõttu ei kaasne nende kasutamisega privaatsusriske, kuid need võimaldavad modelleerida erinevaid aadresside kehtivusstsenaariume kasutusloodes. Postiindeksid genereeritakse, kuna algses andmestikus neid ei sisaldu ning nende vastavus tegelikele aadressidele ei ole testimise seisukohalt funktsionaalselt oluline; vajadusel saab selle täpsustamist käsitleda tulevastes töödes. Kuna kasutuslugudes ei olnud postiindeksi täpsus nõutud, aitab see lahendus ühtlasi vähendada andmestiku privaatsusriske.

**Tabel aadresskomponent.** Tabeli väljade kirjeldus on toodud tabelis 3. Funktsioon `generate_aadress_komponent` loob ruumilise hierarhia üksikute aadressi komponentide lõikes. Algandmete põhjal teisendatakse kehtivuskuupäevad masinloetavaks ning igale kirjele määratakse loomise aeg, mis on kuni kümme aastat enne kehtivuse algust. Staatus määramiseks kasutatakse kahte klassifikaatorikoodi: "KEHTIV", kui puudub kehtivuse lõpp, ning "KEHTETU", kui lõppkuupäev on antud. Iga kirje saab unikaalse identifikaatori `AKpID`, mis põhineb rea indeksil. Puuduvad veerud täidetakse vaikimisi tühjade väärtustega, säilitades sellega struktuurse sidususe hilisemates andmetöötlusetappides. Tabeli struktuur järgib rahvastikuregistri loogikat ja võimaldab hiljem siduda elukohaandmeid ruumiliste tasemete kaupa.

Hetkel ei võta funktsioon arvesse kõiki Eesti aadressikomponente. Kuigi täieliku hierarhia laadimine oleks vajalik suuremahuliste katsete või laadimiste jaoks, ei pidanud ma otstarbekaks kogu Eesti komponentide andmestiku korduvat ja mahukat töötlemist iga käivituse korral. Sellest tulenevalt keskendusin väiksemahulise proovivalimi töötlemisele, mis on piisav funktsionaalse loogika ja testide käivitamiseks. Vajadusel saab selle otsuse ümber teha, võimaldades laadida kõik komponendid.

**Ajutine tabel isikuandmetest ja erinevatest seostest (temp\_relationships).** Tabeli väljade kirjeldus on toodud tabelis 4. Ajutine tabel lahendab probleemi, et andmebaasi teiste tabelite (näiteks `isik`, `isikuaadress` ja `isikuasutus`) täitmiseks on vaja genereerida omavahel seotud isikuid koos järjepidevate perekondlike seostega. Ilma sellise vahetabelita oleks keeruline säilitada realistlikke seoseid, nagu vanemad ja lapsed või abikaasad ning neile vastavad ühised atribuudid (nt aadressid, kodakondsus, saabumis- ja

lahkumiskuupäevad).

Funktsioon `generate_temp_relationships` loob ajutise tabeli, mis simuleerib inimeste perekondlikke seoseid ning varustab edasisi tabeleid (nagu `isik`, `isikuaadress` ja `isikuasutus`) vajalike lähteandmetega. Skript genereerib täiskasvanute ja laste kirjeid, määrates juhuslikult nende vanuse, soo ja suhtestaatuse. Täiskasvanutele määratakse perekonnaseis ning osa neist paigutatakse abielupaaridesse. Lastele omistatakse juhuslikult vanemad (kas üks või kaks vanemat). Arvutatakse sünnikuupäevad ning teatud tõenäosusega ka surmakuupäevad. Haridustase määratakse sünnikuupäeva põhjal, kasutades lihtsat funktsiooni `get_education_for_age` ja kodifikaatorit. Kodakondsus valitakse tõenäosuslikult – valdavalt Eesti kodakondsus, muidu lisatakse saabumis- ja teatud tõenäosusega lahkumiskuupäev.

Tabelisse arvutatakse laiutiotsingu abil perekonna identifikaator, mis ühendab partnerid ja nende ühised lapsed üheks terviklikuks perekonnaks. Aadressid määratakse kindlaksmääratud tõenäosusega abielupaaridele, ülejäänutele määratakse aadress juhuslikult. Lastele antakse üldjuhul sama aadress, mis vanemal – täisealiseks saades jäävad nad kas samale aadressile või kolivad teatud tõenäosusega.

Kui aadresside tabel sisaldab `AkpIDTase2` väärtusi, määratakse isikutele ka kohalik omavalitsus. Saabumiskuupäevad joondatakse nii, et kui vanem on Eestisse saabunud pärast lapse sündi, määratakse ka lapsele saabumiskuupäev. Lõpuks määratakse täisealistele juhusliku tõenäosusega ka seos asutusega (`AsID`), mida kasutatakse hiljem auditiveergude (näiteks `IAsIDLooja` või `IAsIDMuutja`) täitmisel. Sisuliselt genereerime siin ametnike andmeid.

**Tabel asutus.** Tabeli väljade kirjeldus on toodud tabelis 5. Asutustest koosnev tabel luuakse funktsiooniga `generate_asutus`, mis võtab sisendiks äriregistrist eksporditud nimekirja ja rikastab iga kirjet juhuslikult genereeritud metaandmetega. Määratakse asutuse algus- ja teatud tõenäosusega lõppkuupäev, kirjete loomise- ja muutmise ajad ning juhuslik aadress. Staatuse määramine põhineb veerul `Staatus`: kui väärtuseks on „Registrisse kantud“, märgitakse asutus „KEHTIVAKS“, muidu „KEHTETUKS“. Igale asutusele lisatakse unikaalne asutusekood, määratakse aadressi viide (`AkpID`) ja valitakse juhuslik asutuse liik klassifikaatori põhjal. Valminud tabelit kasutatakse isikute ja asutuste vaheliste seoste loomiseks.

**Tabel isikasutus.** Tabeli väljade kirjeldus on toodud tabelis 6. Funktsioon `generate_isik_asutus` loob ühe seosekirje, mis seob konkreetse isiku (`IsID`) kindla asutusega (`AsID`). See kirje modelleerib töösuhet või muud institutsionaalset seost ning sisaldab loogiliselt määratletud kuupäevavälju, staatust ja metaandmeid. Alguskuupäev (`IAsAlgusKpv`) valitakse juhuslikult määratud ajavahemikust, millele järgneb teatud tõenäosusega lühiajaline (kuni 7 päeva) kinnitamiskuupäev (`IAsKinniKpv`).

Staatuse väljal `KdIDStaatus` määratakse teatud protsendiga juhtudest väärtus "KEH-

TIV", ülejäänud juhtudel "KEHTETU". Teatud tõenäosusega juhtudest määratakse ka kustutamiskuupäeva, mis muudab staatuse automaatselt "KEHTETUKS". Kirjele lisatakse kasutajanimi, salasõna ja kontaktandmed, mis võimaldavad modelleerida juurdepääsuloogikat ja auditeerimisvõimalusi. Loomise ja muutmise ajad genereeritakse loogilise järjestusega. Tabel on kasutusel isiku-asutuse seoste loomiseks ja auditväljade täitmiseks väljundtabelites.

**Tabel isikuaadress.** Tabeli väljade kirjeldus on esitatud tabelis 7. Funktsioon `generate_isikuaadress` koostab iga isiku või pere elukoha ajaloo, jagades selle järjestikusteks ajaperioodideks. Funktsioon kasutab sisendiks tabelit `temp_relationships`. Täiskasvanule genereeritakse üks kuni kolm elukohaperioodi. Iga periood sisaldab aluskuupäeva `IAdrKehtibAlatesKpv`, lõppkuupäeva `IAdrKehtibKuniKpv`, staatuse koodi `KdIDAadressiStaat` ning liigi `KdIDAadressiLiik`. Viimane periood jääb avatuks ("KEHTIV"), varasemad lõppevad ja märgitakse staatusega "KEHTETU".

Lastele luuakse elukoha ajalugu, kopeerides see ühe juhuslikult valitud vanema ajaloost. Kui lapsel on mitu vanemat, eelistatakse vanemat, kelle kohta on olemas aadressiperioodid. Sellise lähenemisega säilitatakse pere loogiline struktuur ja realistlik aadressi ajalugu.

Kui isikul on registreeritud surmakuupäev `Surmaaeg`, siis kõrbitakse kõik kehtivad aadressiperioodid selle kuupäevani. Lisaks arvestatakse perekonna liikmete saabumise (`IsSaabusEesti`) ja lahkumiskuupäevi (`IsLahkusEestist`), kohandades vajadusel elukohaajaloo perioode vastavalt. Saabumis- ja lahkumiskuupäevad määratakse perekonna kõigi liikmete varaseima saabumise ja hiliseima lahkumise alusel. Näiteks, kui emal on saabumiskuupäev 2012-05-01 ja isal ning lastel see puudub, siis määratakse sama kuupäev kogu perele. Samamoodi, kui ühel lapsel on lahkumiskuupäev 2021-09-15, laieneb see kogu perele juhul, kui teistel see puudub.

Kõik järjestikused perioodid ühendatakse omavahel veergude `IAdrIDJargmine` ja `DokIDLopuAlus` abil. Kui ajalugu sisaldab vigaseid kirjeid (nt kui `IAdrKehtibAlatesKpv` on hilisem kui `IAdrKehtibKuniKpv`), eemaldatakse need ja allesjäänud perioodid ühendatakse uuesti. Kõik kirjed saavad ka auditväljad: `LoodiKpv` määratakse enne perioodi algust; `MuudetiKpv` ja `KustutatiKpv` määratakse kindla tõenäosusega ja sõltuvalt sellest, kas periood on lõppenud või kehtiv.

**Tabel isik.** Tabeli väljade kirjeldus on esitatud tabelis 8. Funktsioon `generate_isik` koostab lõplikud isik tabeli kirjed. Sisendiks kasutab tabeli `temp_relationships` ja tabelit `isikuaadress`. Iga inimese kohta koostatakse kirje, mis sisaldab erinevaid isikuandmeid nagu näiteks (`KdIDSugu`), sünniaeg (`isSynniaeg`), surmaaeg (`IsSurmaaeg`) ning isiku staatused (`KdIDIsikuStaat`, `KdIDKirjeStaat`) olenevalt sellest, kas tegemist on elava või surnud isikuga.

Nimed genereeritakse Eesti andmestikele sobiva Faker teegi abil. Samuti arvutatakse

isikukood (IsIsikukood) vastavalt sünniajale ja soole, kuid kontrollnumbri (isikukoodi viimane number, mis tagab koodi õigsuse) kehtivust ei kontrollita ega arvutata. Selline lähenemine tagab, et suurem osa loodud isikukoodidest ei lange kokku päris isikute tegelike koodidega. Rahvuse (KdIDRahvus), emakeele (KdIDEmakeel) ja kodakondsuse (KdIDKodakondsus) väljad seotakse vaikumisi Eesti koodiga, kuid kui isikul esineb saabumiskuupäev Eestisse (IsSaabusEesti), määratakse talle juhuslikult mõni muu rahvuskood.

Inimese kohaliku omavalitsuse saabumiskuupäev (IsSaabusKOVikpv) täidetakse kontrollides isiku ja tema kehtiva aadressi seost isikuaadress tabelist – kui isik on seotud aadressiga, leitakse tema elukoha alguskuupäev. Samuti tuletatakse kohaliku omavalitsuse aadressikomponendi identifikaator (AKpID) olemasolevast ajutisest tabelist. Lõpus lisatakse igale kirjele ka auditiveerud: LoodiKpv, MuudetiKpv ja KustutatiKpv, mis määratakse kindlaks juhusliku loogika ja kindlate ajavahemike põhjal.

**Tabel kodakondsus.** Tabeli väljade kirjeldus on esitatud tabelis 9. Funktsioon `generate_kodakondsus` koostab iga inimese kohta ühe kirje kodakondsuse tabelisse, lähtudes väljadest `IsID`, `KdIDKodakondsus` ja `isSunniaeg`. Funktsioon määrab iga kodakondsuse kirje alguskuupäeva `KodKehtibAlates`, mis valitakse juhuslikult inimese sünniaja või määratud minimaalse kuupäeva ja tänase päeva vahel. Teatud tõenäosusega määratakse ka lõppkuupäev `KodKehtibKuni`, mille korral muutub staatuse väärtus "KEHTETU"; muul juhul jääb staatuseks "KEHTIV"(väli `KdIDStaatus`).

Iga kirje seotakse dokumentidega, mille identifikaatorid määratakse kasvava loenduri alusel väljade `DokIDAlguseAlus` ja `DokIDLopuAlus` kaudu. Lõppdokument lisatakse ainult juhul, kui lõppkuupäev on määratud ja juhuslik tingimus seda lubab. Määratakse ka auditiväljad: `LoodiKpv` on alati varasem kui `KodKehtibAlates`, ning `MuudetiKpv` ja `KustutatiKpv` lisatakse sõltuvalt tõenäosuslikest tingimustest. Kui kirje on kustutatud, muutub ka tema lõplik staatus automaatselt "KEHTETU", tagades andmesisese järjepidevuse.

**Tabel isikudokument.** Tabeli väljade kirjeldus on esitatud tabelis 10. Funktsioon `generate_isikudokument` loob isikute ja dokumentide vahele sidetabeli, milles iga rida vastab konkreetse dokumendi seosele konkreetse inimesega. Iga kirje sisaldab välja nagu näiteks `IDokIsikukood`, `IDokEesnimi`, `IDokID`, `DokID` jne. Lisaks luuakse teatud tõenäosusega ka varasemad andmeväljad, näiteks `IDokVanaIsikukood`, mis võimaldab simuleerida juhtumeid, kus isikuandmeid on dokumendil hiljem muudetud.

Kirjele lisatakse `AsID`, mis tuletatakse väljaandjast (`AsIDValjaandjaAsutus`), ja metaandmed, sealhulgas `LoodiKpv`, `MuudetiKpv` ja `KustutatiKpv`. Loomise kuupäev `LoodiKpv` genereeritakse dokumendi loomise ja jooksva aja vahelise vahemiku põhjal. Teatud juhtudel määratakse ka muudatus- või kustutamiskuupäev, et simuleerida dokumentide elutsükli erinevaid etappe.

**Tabel dokument.** Tabeli väljade kirjeldus on esitatud tabelis 11. Funktsioon `generate_person_document` võimaldab genereerida dokumente kahel erineval viisil: olemasoleva andmestiku põhjal või täielikult juhuslikult. Mõlemal juhul luuakse kirjed tabelisse Dokument, kus igal real on unikaalne identifikaator DokID, koos muu olulise teabega nagu näiteks DokNumber, DokSeeria, DokValjaantudKpv, DokKehtivKuniKpv, KdIDDokumendiLiik jne. Iga kirje on seotud konkreetse isikuga läbi välja IsID.

Kui funktsioon töötab režiimis `source`, kasutatakse sisendtabelis määratud veerge (nt `IAdrKehtibAlatesKpv`, `IAdrKehtibKuniKpv`), et määrata dokumendi kehtivusaeg ning staatus "KEHTIV" või "KEHTETU". Režiimis `random` genereeritakse kõik väljad juhuslikult. Tõenäosuse alusel määratakse, kas dokument on kehtetu või jääb aktiivseks.

Funktsioon määrab ka kirjete kuupäevad nagu `LoodiKpv`, `MuudetiKpv` ja `KustutatiKpv`. Dokumendi tüüp (nt "ELUKOHATEADE", "JUHILUBA") määratakse argumendiga `doc_type`, mille kaudu leitakse sobiv `KdIDDokumendiLiik` väärtus tabelist kodifikaa-  
tor. See võimaldab dokumente liigitada ja ka hiljem siduda näiteks elukohaandmetega või isikuga seotud sündmustega.

#### 4.1.3 Järeltegevused

Kui kõik vajalikud tabelid on genereeritud, siis skript salvestab need CSV-failidena (komadega eraldatud väärtused) väljundkausta. Kuna tabelite genereerimise käigus tekitatakse ametnike andmeid (isikasutus), siis veergude `IAsIDLooja`, `IAsIDMuutja` ja `AsID` lisamine toimub alles pärast tabelite genereerimist. Sellise töövooga on kogu andmete genereerimise protsess paindlik ning hõlpsasti laiendatav: võimalik on muuta geneeritavate kirjete arvu, täiendada või kohandada funktsioone uute teststsenaariumide jaoks või lisada valideerimismehhanisme.

## 4.2 Mõju privaatsusele

Süntetiliste andmete genereerimise peamine eelis on võimalus testida süsteeme ilma päris isikuandmeid kasutamata, kuid sellega kaasnevad olulised mõjud privaatsusele. Töös loodud andmed on täielikult sünteesitud või tuginevad avalikule teabele. Ka sellisel juhul tuleb arvestada mitme ohuga, mis võivad viia juhusliku isikutuvastuseni või infoturvariskideni. Genereeritud andmestik tugineb kahte liiki allikatele.

1. Avalikud ressursid (Eesti aadressiandmed ja äriregister), mis ei sisalda isikukode ega muid otseseid tunnuseid.
2. Täielikult sünteesitud väärtused nagu nimed ja isikukoodid, mille loob Faker-teek või ise arendatud abifunktsioonid failis `utils.py`.

Allpool analüüsin, kui tõenäoline on, et sünteesitud kirje võiks siiski juhuslikult kattuda päris inimese andmetega, ja millised praktilised sammud seda riski leevendavad.

#### 4.2.1 Juhuslik kattuvus

Süntheetiliste andmete loomisel võib teoreetiliselt tekkida olukord, kus mõne isiku kohta genereeritud andmekombinatsioon – näiteks eesnimi, sugu, sünniaeg ja elukoht – langeb juhuslikult kokku mõne reaalse inimese andmetega. Sellist olukorda käsitletakse töös kui juhuslikku kattuvust (ingl k *coincidental reidentification*). Tegemist ei ole taastuvastamisega, kuna andmed ei pärine pärisandmestikust ega ole loodud seda matkima. Kattuvus tekib üksnes siis, kui sõltumatult sünteesitud väärtused moodustavad kombinatsiooni, mis võib meenutada päriselt eksisteeriva isiku kirjet.

Käesoleva töö kontekstis on juhuslik kattuvus privaatsuse mõttes ebaoluline risk, eeldusel et ei teki andmepaare, mis võimaldaksid üheselt tuvastada konkreetset isikut. Näiteks ei peeta nime ja vale isikukoodi koos esinemist probleemseks, samuti mitte kehtiva isikukoodi olemasolu juhul, kui nimi, sünniaeg ja elukoht on kõik sõltumatult genereeritud. Sünteetiline nimi koos sünnikuupäevaga ei ole taastuvastamise alus, kuna selliseid kombinatsioone esineb Eesti rahvastikus sadu või tuhandeid. Kriitiliseks tuleks lugeda olukord, kus sünteetiline kirje koosneb andmetest, mis pärinevad avalikest allikatest ja mille kombineerimisel tekib tahtmatult reaalne faktistruktuur, näiteks täpselt samad nimed ja seosed, mis eksisteerivad avalikus registris.

Selle vältimiseks ei looda sünteetiliste isikute ja avalike asutuste vahel seoseid tegelike rollide või struktuuride alusel. Asutuste andmed on pärit äriregistrist, kuid kõik seosed isikutega määratakse juhuslikult, vältimaks kokkulangemist näiteks tegeliku juhatuse liikmega. Samamoodi pärinevad aadressid küll avalikust aadressiregistrist, ent neid ei seostata kunagi konkreetse sünteetilise isikuga mingi loogika alusel – aadressid määratakse juhuslikult.

Isikukoodid genereeritakse kontrollnumbriga küll formaalselt sarnasel kujul, ent nende arvutuses ei rakendata valideerimisalgoritmi. Seetõttu ei vasta suurem osa isikukoodide tegelikule isikukoodistruktuurile ega oma tuvastamisväärtust. Nimed, rahvused ja muud demograafilised andmed valitakse laiaast tõenäosuslikust jaotusest ning ei tugine ühelegi olemasolevale andmekogumile.

Arvestades, et loodud andmestik on mõeldud kasutamiseks sisemistes testkeskkondades, on juhuslikud kokkulangevused teoreetiliselt võimalikud, kuid nende tähenduslikkus praktilise privaatsusohu seisukohast on minimaalne. Oluline on tagada, et sünteesi käigus ei moodustu tahtmatult reaalseid fakte ega olukordi, kus kombinatsioonid muutuksid sisuliselt tuvastatavateks. Sellise ohu vältimiseks on töö käigus järgitud kindlaid sünteesireegleid, mis välistavad avalike andmete mehaanilise kokkupaneku viisidel, mis võiksid põhjustada ebasoovitavaid kattuvusi.

#### 4.2.2 Läbipaistvus ja privaatsusauditi võimalikkus

Reeglipõhise andmesünteesi tugevuseks on lihtne auditeeritavus. Iga reegel, mille abil andmeid sünteesitakse, on inimloetav ja tehniliselt jälgitav. Selline auditeeritavus või-

maldab läbi viia formaalse privaatsusauditi, mille käigus saab veenduda järgnevas.

- Ükski reegel ei põhine tegelikest andmetest tuletatud invariantidel. Reeglid, mille alusel sünteetilisi andmeid luuakse, põhinevad dokumenteeritud struktuuril, ametlikel kirjeldustel ja üldistel loogilistel eeldustel, mitte olemasolevate andmete analüüsil ega andmetest tuletatud seaduspäradel.
- Loodud andmestikus puuduvad statistilised või semantilised mustrid, mis viitaksid otseselt pärisandmestikele. Kuigi sünteetilisest andmestikus võib mõnikord taotleda reaalsusele sarnaseid jaotusi (nt sagedusi või korrelatsioone), tuleb vältida olukorda, kus need mustrid jäljendavad pärisandmeid liiga täpselt või sobituvad andmestikuga liigselt. Selline ülejäljendamine (ingl k *overfitting*) võib suurendada riski, et väljundandmestik sarnaneb liialt originaalandmetega ja võimaldab teha ebatäpseid järeldusi konkreetsete isikute kohta.

Kirjeldatud auditeeritavus annab olulise eelise võrreldes musta kasti tüüpi generatiivsete mudelitega, nagu GAN või VAE, mille sisemist toimimist on keeruline jälgida ning mille väljund võib tahtmatult peegeldada treeningandmestiku struktuuri.

#### **4.2.3 Teadmuspõhise sünteesi eelised privaatsuse seisukohalt**

Teadmuspõhine (ehk spetsifikatsioonil põhinev) süntees tagab privaatsuse, lähtudes domeeniteadmistest ja äriloomikast, mitte olemasolevatest andmetest. See tähendab, et isegi juhul kui mõni väljund näib sarnane pärisandmetele, pole sellel põhjuslikku ega tuvastatavat sidet ühegi konkreetse isikuga. Selline lähenemine on kooskõlas diferentsiaalprivaatsuse filosoofiaga: isegi kui keegi sooviks kindlaks teha, kas mõni konkreetne isik oli algsetes andmetes olemas, ei võimalda sünteetiline andmestik sellist järeldust teha, sest selles puudub isikuga otseselt seostatav info. Kuigi siin ei rakendatud diferentsiaalprivaatsust selle tehnilises tähenduses, on saavutatud sarnane privaatsusefekt tänu sellele, et andmed sünteesiti loogiliselt isoleeritud reeglite ja spetsifikatsioonipõhise kontrolli kaudu [10].

### 4.3 Valideerimine ja kvaliteedikontroll

Lisaks andmete genereerimisele on oluline kontrollida loodud andmestiku vastavust defineeritud reeglitele ning äriloogikale. Töös valideerisin andmeid põhiliselt automaatsete testide abil, kasutades Pythonis pytest raamistikku. Testidega veendusin, et iga tabeli sees ja tabelite vahel kehtivad nii struktuursed kui ka sisulised seosed (näiteks viited välisvõtmetele ja tabelite vahelised seosed, kuupäevavahemike kattuvused, staatusemuutuste loogika). Need testid kehtestavad selged kontrollid andmete struktuuri ja äriloogika kohta ka edaspidiste arenduste jaoks. Teste tegin kokku 59. Alljärgnevalt kirjeldan valideerimise teste (pseukoodiline kirjeldus joonisel 5).

**Kuupäevade ja staatuste loogika.** Kontrollitakse tabelites kas iga kirje kehtivuse algus- ja lõppkuupäev on omavahel loogilises suhtes (nt  $I\text{AdrKehtibAlatesKpv} \leq I\text{AdrKehtibKuniKpv}$ ). Samuti veendutakse, et staatus muutuks vastavalt kuupäevadele: kui lõppkuupäev on olemas, siis peab staatuseväli olema "KEHTETU". Lisaks kontrollitakse, et kuupäevad ei ulatuks ebaloogiliselt tulevikku ega oleks varasemad loogilisel järjestikuses ajaskaalal.

**Viidete kehtivus.** Veendutakse, et kõigil välisvõtmetel on vasted olemas. Näiteks  $IsID$ , mida kasutatakse isikudokument tabelis, peab viitama reaalselt eksisteerivale isikule isik tabelis. Samuti kontrollitakse, et  $AsID$  väärtused isikasutus tabelis kattuksid asutus tabeli ridadega.

**Järjestikused ajaperioodid.** Tabel isikuaadress sisaldab perioodipõhiseid kirjeid, mis kirjeldavad järjestikuseid muutusi. Kontrollitakse, et iga uus periood algaks eelmisest loogiliselt – näiteks, et  $DokIDLopuAlus$  ja  $I\text{AdrIDJargmine}$  viitaksid õigetele dokumentidele ning ajavahemikud ei kattuks valesi ega jääks katki. See aitab kiiresti tuvastada puuduvad või valesid üleminekud.

**Staatuste ja sündmuste kooskõla.** Testitakse ka sündmuste ja staatuste vastavust. Näiteks kui kirje sisaldab kustutamiskuupäeva ( $KustutatiKpv$ ), siis peab selle staatus olema "KEHTETU". Samamoodi peab surnud isikutel olema  $IsSurmaaeg$  täidetud ja  $KdIDIsikuStaatust$  väärtuseks "SURNUD".

**Vormingu- ja sisuvalideerimine.** Mitmes testis kontrollitakse automaatselt genereeritud väärtuste vormingut. Näiteks vaadatakse, et väljad nagu  $IsIsikukood$  sisaldaksid ainult numbreid ning oleksid õige pikkusega. Samuti kontrollin, et väljad nagu "Vananimi" või "OLD-xxxx" vastaksid eelnevalt määratud struktuurile ja ei sisaldaks keelatud sümboleid.

Valideerimise koodi puhul tegin enamus tabelite jaoks eraldi testfailid. Seal kontrollisin üksikasjalikult kõiki olulisi seoseid ja äriloogikast tulenevaid eeldusi. Tänu sellisele ülesehitusele oli võimalik kiiresti tuvastada loogikavigu, mis manuaalsel ülevaatusel oleks jäänud märkamatuks.

---

```
# TESTIDE KESKKONNA SEADISTAMINE
- Impordi pytest ja pane paika kataloogid, kus paiknevad
  genereeritud CSV-failid (nt "output/")

# STRUKTUURIKONTROLL
- Veendu, et igas loodud failis on eeldatavad veerunimed (
  nt "IsID", "DokID").

- Kontrolli, et primaarvõtmed (nt "IDokID", "KodID") on
  unikaalsed ja mitte tühjad.

# LOOGILISED SEOSED JA AJALISED REEGLID
- Kontrolli, et "LoodiKpv" <= "MuudetiKpv" <= "KustutatiKpv" (
  kui need pole null).

- Kui "KustutatiKpv" != null, peab staatus olema "KEHTETU".

- Perioodide (nt "IAAdrKehtibAlatesKpv".."IAAdrKehtibKuniKpv"
  ) korral ei tohi esineda vigast kattuvust ja lõppkuupä
  evad peavad olema hilisemad alguskuupäevadest.

# VIIDETE KEHTIVUS
- Kontrolli FK seoseid: kas "isikudokument.DokID"
  eksisteerib tegelikult "dokument.DokID" hulgas?

- Kas "isikuaadress.IsID" kattub "isik.IsID" väärtustega?

# ERIOLUKORDAD JA LAIENDUSED
- Testi, et delete-juhtudel (nt x% kirjetest on juhuslikult
  kustutatud) pole vastuolusid (näiteks "MuudetiKpv">"
  KustutatiKpv").

- Lisa täiendavaid teste (nt nimelised kontrollid "Vana-"
  prefiksiga andmetele vms).
```

---

Joonis 5. Andmete valideerimise kirjeldus pseudokoodina.

## 4.4 Andmete muutmine kasutuslugude jaoks

Valideerimisele on lisatud ka võimalus luua vigaseid andmeid, mis lubab simuleerida äärmuslikke või ebatavalisi olukordi, mis aitaksid süsteeme põhjalikumalt arendada. Kasutuslugude loomine algab olemasolevatest, andmegeneraatori poolt loodud CSV-failidest, kust valitakse teatud hulk kirjeid. Seejärel muudetakse teadlikult konkreetseid välju, et tekiks vastuolu: näiteks seatakse kehtivusaeg minevikku või lisatakse uus elukoht isikule. SMIT pakkus välja järgmised kasutuslood.

**Kasutuslugu I.** Mitme samaaegselt kehtiva aadressi loomine ühele isikule. See kasutuslugu simuleerib olukorda, kus ühele isikule luuakse rohkem kui üks kehtiv elukoha aadress sama ajaperioodi vältel. Üldjuhul on reegel, et isikul võib olla vaid üks kehtiv elukoht korraga. Juhuslikult valitud isikutele genereeritakse täiendavad elukohad koos vastava "ELUKOHATEADE" dokumendiga.

**Kasutuslugu II.** Juhiloa märkimine kehtivaks, kuigi selle DokKehtivKuniKpv on möödunud. See juhtum keskendub olukorrale, kus dokument (näiteks juhiluba) märgitakse süsteemis kehtivaks, kuigi dokumendi "kehtiv kuni" kuupäev on minevikus. Kasutusloo loomiseks valitakse juhuslik arv kehtivaid juhilubasid ning muudetakse nende "DokKehtivKuniKpv" kuupäev minevikku, jättes dokumendi staatuse endiselt kehtivaks.

**Kasutuslugu III.** Elukoha IAdrKehtibAlatesKpv nihutamine tulevikku, et inimesel tekiks „tulevikuaadress”. Selles kasutusloos lükatakse elukoha aadresside ja vastavate "ELUKOHATEADE" dokumentide kehtivuse alguskuupäev määratud arvu päevade võrra tulevikku.

**Kasutuslugu IV.** Olukord, kus ühel isikul on korraga mitu vastuolulist kodakondsust, millest üks on "MÄÄRATLEMATA". Tavaliselt on isikul kindel riiklik kodakondsus ning "MÄÄRATLEMATA" staatuse ei tohiks eksisteerida paralleelselt konkreetse riigi kodakondsusega. Kasutusloo puhul luuakse valitud isikutele uus kodakondsuskirje, mille staatuseks määratakse sihilikult "MÄÄRATLEMATA", lisaks luuakse vastav dokumendi kirje.

Sellised sihilikult vigased kasutuslood aitavad arendamisel lihtsalt tuvastada puudusi või vigu valideerimisreeglites ja päringutes. Lõplik valideerimine generaatori sobivuse osas SMIT-i arendus- ja testikeskkonnas on järgmine loogiline samm. See sisaldaks endas generaatori integreerimist SMIT-i testiprotsessidesse ning hinnangute kogumist reaalses kasutuses tekkivate kasutuslugude põhjal. Selline valideerimine võimaldaks hinnata, kuivõrd generaator katab vajalikkude äri loogikat, kuivõrd see suudab tekitada realistlikke ja mitmekesiseid olukordi, ning kas toodetud andmestik toetab tõhusalt süsteemide töökindluse ja andmekvaliteedi kontrolli. Tagasiside põhjal oleks võimalik täiustada nii andmevaliku loogikat kui ka kasutuslugude katvust, et suurendada generaatori väärtust praktilises arendustöös.

## 5 Tulemused

### 5.1 Andmegeneraatori ülesehitus ja töömaht

Töö üks olulisi väljundeid on täielikult funktsionaalne andmegeneraatori prototüüp koos andmevigade lisamise ja valideerimise moodulitega. Kogu lahendus on realiseeritud Pythonis ja koosneb 32 Python-failist, kogumahus 4342 koodirida (koos kommentaaridega). Automaatsete testide kogumaht on 59 testi, mis jagunevad 9 eraldiseisvasse testifaili ja katavad kõik väljundtabelid. Koodi on võimalik näha Github repositooriumis (vaata lisa I). Projekt koosneb kolmest peamisest moodulist (vaata joonis 6):

- **generation/** – sünteetiliste tabelite loomine (nt isik, aadress, dokument);
- **error\_generation/** – nelja kasutusloo alusel juhuslike, kuid kontrollitud andmevigade lisamine;
- **validation/** – struktureeritud pytest-põhine valideerimissüsteem koodi kontrollimiseks.

Kogu süsteemi realiseerimiseks kulus hinnanguliselt 200–250 tundi, mis sisaldas endas koodi kirjutamist, testimist ja valideerimist. Arvestades tehtud töö mahtu ja tehnilist keerukust (ajaliselt sõltuvad väljad, relatsioonilised sõltuvused, privaatsusnõuded), on lahenduse loomine ühe arendaja poolt realistlik, ent nõuab head äriloogika mõistmist ja süsteemiarhitektuurset lähenemist.

### 5.2 Jõudlustulemused

Tabelis 1 on esitatud nelja järjestikuse eksperimendi tulemused, kus sünteetiliste andmete generaatorile anti sisendparameetrina ette soovitud isik-kirjete arv (vastavalt 100, 1000, 10 000 ja 100 000). Tabelis on toodud iga tabeli kohta genereeritud kirjete arv ning viimases reas kogu andmestiku loomise käitusaeg sekundites. Kõik katsed viidi läbi samas arvutuskeskkonnas (32 GB RAM, AMD Ryzen 7 5800X 8-core protsessor, NVIDIA GeForce RTX 4070 S), tagades tulemuste võrreldavuse. Lisaks mõõdeti pärast andmete genereerimist ka andmestiku valideerimiseks koostatud pytest-põhiste testide käitusajad, mis olid vastavalt 0,4; 0,9; 5,3 ja 50,9 sekundit. Tulemused kinnitavad, et generaator suudab efektiivselt käsitleda ka suuremahulisi andmekogumeid ning on hästi skaleeritav vastavalt vajadusele.

Sünteetiliste andmete generaatori väljundiks on komplekt CSV-vormingus tabeleid, mille struktuur ja sisu vastavad rahvastikuregistri loogikale ja andmekirjeldustele. Kokku loodi üheksa tabelit: aadress, aadresskomponent, isik, dokument, asutus, isikuasutus, isikuaadress, isikudokument ja kodakondsus. Tabelid on omavahel seotud viisil, mis tagab struktuurse ja semantilise kooskõla SMIT-i määratud kasutusluge ja rahvastikuregistri andmete kirjeldustega.

```

RRgen/
├── README.md
├── generate.py
├── generate_errors.py
├── run_validation_tests.py
├── data/
│   ├── README.md
│   ├── aadress.csv
│   ├── kodifikaator.csv
│   └── ...
├── output/
│   ├── 01_aadress.csv
│   ├── 02_aadresskomponent.csv
│   └── ...
└── src/
    ├── error_generation/
    │   ├── README.md
    │   ├── main.py
    │   ├── test_case_01.py
    │   └── ...
    ├── generation/
    │   ├── README.md
    │   ├── utils.py
    │   ├── main.py
    │   ├── isik.py
    │   └── ...
    └── validation/
        ├── README.md
        ├── helpers.py
        ├── test_aadress.py
        └── ...

```

Joonis 6. Koodi struktuur: põhimoodulid andmete genereerimiseks, kasutusloo vigade lisamiseks ja valideerimiseks.

### 5.3 Metoodika praktiline valideerimine

Töö käigus viisin läbi mitmeid praktilisi katseid, keskendudes andmete sisemisele loogikale ja võimekusele realiseerida määratletud kasutuslugusid. Ligikaudu kolme arendajakuu jooksul arendasin välja lahenduse, mis võimaldab luua GDPR-iga kooskõlas olevat testandmestikku koos automaatsete valideerimismehhanismidega.

Realiseerisin neli SMIT-i poolt kirjeldatud kasutuslugu. Selleks lisasin sünteesitud andmestikku sihilikult loogikareeglitega vastuolus olevaid olukordi. Näiteks modelleerisin olukordi, kus isikul on mitu samaaegselt kehtivat elukohta, kehtiv dokument mille

Tabel 1. Genereeritud kirjete arv erinevates tabelites vastavalt määratud sisendparameetritele.

| Tabel                    | Eksperiment I | Eksperiment II | Eksperiment III | Eksperiment IV   |
|--------------------------|---------------|----------------|-----------------|------------------|
| isik                     | 100           | 1000           | 10 000          | 100 000          |
| aadress                  | 100           | 1000           | 10 000          | 100 000          |
| aadresskomponent         | 100           | 1000           | 10 000          | 100 000          |
| kodakondsus              | 100           | 1000           | 10 000          | 100 000          |
| asutus                   | 121           | 121            | 121             | 121              |
| isikuasutus              | 8             | 68             | 713             | 7295             |
| isikuaadress             | 173           | 1637           | 16 279          | 164 039          |
| isikudokument            | 398           | 3855           | 38 290          | 384 180          |
| dokument                 | 398           | 3855           | 38 290          | 384 180          |
| <b>Kirjete arv kokku</b> | <b>1498</b>   | <b>13 536</b>  | <b>133 693</b>  | <b>1 324 815</b> |
| <b>Käitusaeg</b>         | <b>20 sek</b> | <b>30 sek</b>  | <b>111 sek</b>  | <b>1340 sek</b>  |
| <b>Kogumaht</b>          | <b>0,3 MB</b> | <b>2,0 MB</b>  | <b>19,5 MB</b>  | <b>197,0 MB</b>  |

kehtivusaeg on möödunud, tulevikus algav elukoht või vastuolulised kodakondsused (nt konkreetne riik koos määratlemata staatusega). Pärast andmevigade lisamist käivitasin valideerimiskripti, et veenduda testide toimimises.

Testide tulemused kinnitasid, et loodud kontrollmehhanismid töötavad ootuspäraselt. Esimese kasutusloo puhul tuvastati mitmel isikul (IsID) rohkem kui ühe samaaegselt kehtiva aadressikirje. Joonisel 7 on testide üldine käivituseväljund. Rohelised punktid näitavad edukaid teste, punased F-märgid osutavad äriloogikat rikkuvatele olukordadele, mis tekitati sihilikult. Joonisel 8 on näha vea detailne väljund, kus funktsioon `assert_single_active()` tuvastas, et sama isiku kohta esines mitu kehtivat aadressikirjet.

```
platform win32 -- Python 3.8.15, pytest-7.1.2, pluggy-1.0.0
rootdir: C:\Users\Rain\Desktop\MAGISTRITÖÖ\RRgen
plugins: anyio-3.5.0, Faker-35.2.0
collected 59 items

src\validation\test_aadress.py .....
src\validation\test_asutus.py .....
src\validation\test_isik.py .....
src\validation\test_isikuaadress.py ....F....
src\validation\test_isikudokument.py .....
src\validation\test_kodakondsus.py .....
src\validation\test_temp_person_document.py .....
src\validation\test_temp_relationships.py .....
```

Joonis 7. Valideerimiskoodi üldväljund pärast kasutusloo I andmevigade lisamist.

```

def assert_single_active(
    df: pd.DataFrame,
    id_col: str,
    *,
    status_col: str | None = None,
    active_code: int | str | None = None,
):
    """Ensure at most one active row per entity.

    If status_col is None, df is assumed to be pre-filtered to active rows.
    """
    if status_col is not None and active_code is not None:
        active_df = df[df[status_col] == active_code]
    else:
        active_df = df
    counts = active_df.groupby(id_col).size()
    multiples = counts[counts > 1]
    > assert multiples.empty, f"Multiple active rows per {id_col}: {multiples}"
E   AssertionError: Multiple active rows per IsID: IsID
E   2872    2
E   2269    2
E   4780    2
E   9477    2
E   9761    2
E   dtype: int64

src\validation\helpers.py:77: AssertionError

```

Joonis 8. Valideerimisvea detailne vaade (Kasutuslugu I: mitu kehtivat elukohta sama-aegselt).

Valideerimise põhjal saab järeldada, et loodud sünteetiline andmestik sobib hästi arendus- ja testkeskkondades kasutamiseks. Usaldusväärselt tuvastati kõik kasutuslugudest tuletatud äriloogika rikkumised. Kuna lahendus jagati loogilisteks mooduliteks ja dokumenteeriti süsteemi käitumist, saab generaatorit tulevikus hõlpsasti laiendada uute kasutuslugude, tabelite või domeenide jaoks.

## 6 Kokkuvõte

Käesoleva töö eesmärgiks oli valmistada reeglipõhine sünteetiliste andmete generaatori prototüüp, mis võimaldaks luua Eesti rahvastikuregistri loogikaga sarnaseid testandmeid. Töö keskendus praktilisele lahendusele, mis ühendab andmete genereerimise, valideerimise ning ärilooikat rikkuvate kasutuslugude modelleerimise. Lahendus on realiseeritud Pythonis ning koosneb selgelt eristatud moodulitest, mis hõlmavad andmete loomist, kontrollitud vigade lisamist ja automaatset testimist.

Kokku loodi üheksa tabelit, mis katavad olulisemad registriandmete liigid: aadressid, isikuandmed, kodakondsus, dokumendid ja seosed asutustega. Lahendus suudab genereerida nii väikesemahulisi kui ka suurehulgalisi andmestikke, säilitades samal ajal loogilise järjepidevuse ja andmesisese kooskõla. Arendustöö käigus realiseeriti neli SMIT-i poolt määratletud kasutuslugu ning valideeriti, et sünteetiliselt tekitatud andmevead tuvastatakse usaldusväärselt automaatsete testidega.

Privaatsuse tagamiseks kasutati üksnes avalikke andmeid ja sünteetiliselt genereeritud väärtusi, välistades kokkulangemised reaalse isikutega. Rakendatud reeglipõhine lähenemine võimaldab läbipaistvat, auditeeritavat ja kergesti laiendatavat andmete sünteesi, mis sobib hästi arendus- ja testkeskkondadesse, kus tegelike isikuandmete kasutamine on piiratud või keelatud.

Töö tulemusena loodud generaatoril on suur potentsiaal SMIT-is ja teistes Eesti riigiasutustes rakendamiseks. Edasises arenduses võiks keskenduda olemasolevate reeglite täiendamisele ja laiendamisele, hõlmates ka teisi andmetabeleid ning äriprotsesse. Samuti oleks otstarbekas tegeleda loodud andmete integreerimisega olemasolevate testimisplatvormidega, mis võimaldaks automatiseeritud ja kiiret andmestike uuendamist. Lisaks võiks kasutuslugude valikut laiendada, et toetada veelgi keerukamate ja realistlikumate testolukordade loomist.

## Viidatud kirjandus

- [1] Infosüsteem Eesti rahvastikuregister. <https://www.riha.ee/Infos%C3%BCsteemid/Vaata/rr>. Viimati külastati 2025-05-15.
- [2] Maa-ameti aadressiandmete infosüsteemi avalik teenus. <https://xgis.maaamet.ee/adsavalik/extracts>. Viimati külastati 2025-05-15.
- [3] Rahvastikuregistri seadus. <https://www.riigiteataja.ee/akt/118102024004>. Viimati külastati 2025-05-15.
- [4] Äriregister. [https://ariregister.rik.ee/est/company\\_search\\_result/6d1e3cd?page=1](https://ariregister.rik.ee/est/company_search_result/6d1e3cd?page=1). Viimati külastati 2025-05-15.
- [5] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation). <https://eur-lex.europa.eu/eli/reg/2016/679/oj>, 2016. Official Journal of the European Union, L119, pp. 1–88.
- [6] Siseministeriumi infotehnoloogia- ja arenduskeskus (SMIT). <https://www.smit.ee/et>, 2025. Viimati külastati 2025-05-15.
- [7] Syntho AI. Guide to Synthetic Data Generation: Definition, Types, Applications. <https://www.syntho.ai/et/guide-to-synthetic-data-generation-definition-types-applications/>, 2024. Viimati külastati 2025-05-15.
- [8] Tanel Alumäe, Koit Tätte, and Kairit Sirts. Neural Speech Synthesis for Estonian. *arXiv preprint arXiv:2010.02636*, 2020.
- [9] Dan Bogdanov, Eduardo Brito, Paula Etti, Liina Kamm, Peeter Laud, Tanel Mällo, Andre Ostrak, Kati Sein, Riivo Talviste, and Maria Toomsalu. Privaatsuskaitse tehnoloogiate Eestis rakendamise teekaart. Technical Report ID D-16-215, Cybernetica AS, Majandus- ja Kommunikatsiooniministeerium, Tallinn, Estonia, March 2023. Avalik aruanne. Projektijuhid: Liina Kamm, Nele Nisu.
- [10] Cynthia Dwork. Differential privacy: A survey of results. pages 1–19, 2008.
- [11] Darren Edge, Weiwei Yang, Kate Lytvynets, Harry Cook, Claire Galez-Davis, Hannah Darnton, and Christopher M White. Design of a privacy-preserving data platform for collaboration against human trafficking. *arXiv preprint arXiv:2005.05688*, 2020.

- [12] Yanai Elazar et al. Measuring and mitigating unintended bias in synthetic data. *Transactions of the Association for Computational Linguistics*, 10:1–16, 2022.
- [13] Paula Etti. Sünteetiliste andmete kasutamise võimalused avalikus sektoris: raamistik ja juhtumiuuring Eesti Politsei- ja Piirivalveameti näitel. Magistritöö, Tallinna Tehnikaülikool, Tallinn, Juuni 2024. <https://digikogu.taltech.ee/et/item/9e453e2c-f342-442c-a8ad-2ea0a695b909>.
- [14] MichalŠ Gal and Orla Lynskey. Synthetic data: legal implications of the data-generation revolution. *Iowa L. Rev.*, 109:1087, 2023.
- [15] Alexandre Goncalves et al. Generation and evaluation of synthetic healthcare data. *BMC Medical Research Methodology*, 20(1):1–15, 2020.
- [16] Shuang Hao, Wenfeng Han, Tao Jiang, Yiping Li, Haonan Wu, Chunlin Zhong, Zhangjun Zhou, and He Tang. Synthetic data in AI: Challenges, applications, and ethical implications. *arXiv preprint arXiv:2401.01629*, 2024.
- [17] Andreas Holzinger et al. What do we need to build explainable AI systems for the medical domain? *arXiv preprint arXiv:1712.09923*, 2017.
- [18] Imperva. What is Pseudonymization | Safeguarding Data with Fictional IDs. <https://www.imperva.com/learn/data-security/pseudonymization/>, 2023.
- [19] Urban Institute. Generating a Fully Synthetic Human Services Dataset. Technical report, Urban Institute, 2023.
- [20] Marten Juurik. Teaduseetika ja andmekaitse uurimistöodes. [https://etag.ee/wp-content/uploads/2022/11/Teaduseetika-ja-andmekaitse-uurimistoodes\\_Juurik\\_Marten.pdf](https://etag.ee/wp-content/uploads/2022/11/Teaduseetika-ja-andmekaitse-uurimistoodes_Juurik_Marten.pdf), 2022. Eesti Teadusagentuur.
- [21] Marta Lenatti, Alessia Paglialonga, Vanessa Orani, Melissa Ferretti, and Maurizio Mongelli. Characterization of synthetic health data using rule-based artificial intelligence models. *IEEE Journal of Biomedical and Health Informatics*, 27(8):3760–3769, 2023.
- [22] Roderick J.A. Little et al. Disclosure Limitation Approaches: Synthetic Data. In *Protecting Individual Privacy in the Struggle Against Terrorists: A Framework for Program Assessment*. National Academies Press, 2004.
- [23] Edlira Martiri. Synthetic Data Generation: Methods, Applications, and Multidisciplinary Use Cases. In *Data Envelopment Analysis (DEA) Methods for Maximizing Efficiency*, pages 118–138. IGI Global Scientific Publishing, 2024.

- [24] MOSTLY AI. Perform Conditional Data Generation. <https://mostly.ai/blog/conditional-data-generation>, 2023.
- [25] Sulev Mägi and Mari-Liis Jakobson. Rahvastikuregistri elukoha andmete täpsustamine mobiilpositsioneerimise andmetel. [https://sisu.ut.ee/wp-content/uploads/sites/377/tp2\\_rahvastikuregistri\\_elukoha\\_andmete\\_tapsustamine\\_mobiilpositsioneerimise\\_andmetel.tp2\\_aruanne.pdf](https://sisu.ut.ee/wp-content/uploads/sites/377/tp2_rahvastikuregistri_elukoha_andmete_tapsustamine_mobiilpositsioneerimise_andmetel.tp2_aruanne.pdf), 2021. Viimati külastati 2025-05-15.
- [26] Kaili Müürisep, Kairit Sirts, Andriela Rääbis, and Maarja-Liisa Pilvik. Eesti-keelse teksti automaatkorrekatuur: projekti ektb25 lõpparuanne. *arXiv preprint arXiv:2402.11671*, 2024.
- [27] Neha Patki, Roy Wedge, and Kalyan Veeramachaneni. The synthetic data vault. *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 399–410, 2016.
- [28] Ann Maria Philip. Synthetic Data Generation Using Bayesian Networks: A Step-by-Step Guide. *Medium*, 2023.
- [29] Michael Platzer and Ivona Krchova. Rule-adhering synthetic data—the lingua franca of learning. *arXiv preprint arXiv:2209.06679*, 2022.
- [30] Trivellore E Raghunathan. Synthetic data. *Annual review of statistics and its application*, 8(1):129–140, 2021.
- [31] Luc Rocher, Julien Hendrickx, and Yves-Alexandre De Montjoye. Estimating the success of re-identifications in incomplete datasets using generative models. *Nature Communications*, 10(1):3069, 2019.
- [32] Philipp Röchner. On the Fidelity-Privacy Tradeoff of Synthetic Cancer Registry Data. *Studies in Health Technology and Informatics*, 316:621–625, 2024.
- [33] Siseministeerium. Rahvastikuregister. <https://www.siseministeerium.ee/tegevusvaldkonnad/rahvastikutoimingud/rahvastikuregister>, 2025. Viimati külastati 2025-05-15.
- [34] Aiden Smith, Paul C Lambert, and Mark J Rutherford. Generating high-fidelity synthetic time-to-event datasets to improve data transparency and accessibility. *BMC Medical Research Methodology*, 22(1):176, 2022.
- [35] Zoey Sun. Query Aware Synthetic Data Generation. Master’s thesis, EECS Department, University of California, Berkeley, May 2023.

- [36] Syntho AI. Rule-Based Synthetic Test Data Generator. <https://www.syntho.ai/rule-based-synthetic-data/>, 2023.
- [37] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling Tabular data using Conditional GAN. *arXiv preprint arXiv:1907.00503*, 2019.

## **Lisad**

### **I. Repositoorium**

Andmegeneraator, automaatteestide ja andmevigade lisamise lähtekood, mis selle magistritöö raames realiseeriti on avalikud ja kättesaadavad GitHub-i repositooriumist: <https://github.com/koldsworth/RRgen>

## II. Tabelid

Tabel 2. Aadresside (aadress) tabeli väljade täitmise kirjeldus.

| Veeru nimetus      | Kommentaar                          | Täidetud | Allikas      | Genereerimise loogika                                     |
|--------------------|-------------------------------------|----------|--------------|-----------------------------------------------------------|
| AdrID              | Tabeli unikaalne võti               | Jah      | Lähteandmed  | ADR_ID                                                    |
| AkpIDTase0         | Viide riigi aadresskomponendile     | Jah      | Genereeritud | Fikseeritud Eesti riigi kood                              |
| AkpIDTase1         | Viide maakonna taseme komponendile  | Jah      | Lähteandmed  | TASE1_KOOD                                                |
| AkpIDTase2         | Viide valla taseme komponendile     | Jah      | Lähteandmed  | TASE2_KOOD                                                |
| AkpIDTase3         | Viide asula taseme komponendile     | Jah/Ei   | Lähteandmed  | TASE3_KOOD                                                |
| AkpIDTase4         | Viide väikekoha taseme komponendile | Jah/Ei   | Lähteandmed  | TASE4_KOOD                                                |
| AkpIDTase5         | Liikluspinna komponendi viide       | Jah/Ei   | Lähteandmed  | TASE5_KOOD                                                |
| AkpIDTase6         | Objekti taseme komponendi viide     | Jah/Ei   | Lähteandmed  | TASE6_KOOD                                                |
| AkpIDTase7         | Maja/krundi komponendi viide        | Jah/Ei   | Lähteandmed  | TASE7_KOOD                                                |
| AkpIDTase8         | Hoone osa komponendi viide          | Jah/Ei   | Lähteandmed  | TASE8_KOOD                                                |
| AdrMajaNr          | Maja/krundi number tekstina         | Jah/Ei   | Lähteandmed  | TASE7_NIMETUS                                             |
| AdrAadress         | Täisaadress tekstina                | Jah      | Lähteandmed  | TAISAADDRESS                                              |
| KdIDAdressiStaatus | Aadressi staatus                    | Jah      | Genereeritud | Staatus kaardistus AADR_OLEK                              |
| AdrSihtnumber      | Sihtnumber                          | Jah      | Genereeritud | Juhuslik vahemikus 10000–99999                            |
| LoodiKpv           | Kirje loomise aeg                   | Jah      | Lähteandmed  | ADS_KEHTIV                                                |
| ADS_ADS_OID        | ADS süsteemi OID                    | Jah/Ei   | Lähteandmed  | ADS_OID                                                   |
| ADS_ADOB_ID        | ADS süsteemi ADOB ID                | Jah/Ei   | Lähteandmed  | ADOB_ID                                                   |
| IAsIDLooja         | Kirje looja viide                   | Jah      | Genereeritud | Valitakse juhuslikult isikuasutus tabelist ametniku ID    |
| IAsIDMuutja        | Kirje muutja viide                  | Jah      | Genereeritud | Kui on Muudetikpv, siis valitakse juhuslikult ametniku ID |
| Muudetikpv         | Kirje muutmise aeg                  | Ei       | -            | Vaikimisi None                                            |
| KustutatiKpv       | Kirje kustutamise aeg               | Ei       | -            | Vaikimisi None                                            |
| AdrOige            | Töötuse tunnus                      | Ei       | -            | Vaikimisi None                                            |
| ADS_ADR_ID         | ADS süsteemi ADR ID                 | Ei       | -            | Vaikimisi None                                            |
| ADS_ADR TEKST      | ADS süsteemi ADR tekst              | Ei       | -            | Vaikimisi None                                            |
| ADS_ADR_MUDETI     | ADS ADR muutmise aeg                | Ei       | -            | Vaikimisi None                                            |
| ADS_ORIG_TUNNUS    | ADS originaaltunnus                 | Ei       | -            | Vaikimisi None                                            |
| ADS_MUUD_OBJEKTID  | ADS muud objektid                   | Ei       | -            | Vaikimisi None                                            |
| ADS_KOODAADDRESS   | ADS koodaadress                     | Ei       | -            | Vaikimisi None                                            |
| ADSILID            | ADS logi viide                      | Ei       | -            | Vaikimisi None                                            |

Tabel 3. Aadressi komponentide (aadressskomponent) tabeli väljade täitmise kirjeldus.

| <b>Veeru nimetus</b> | <b>Kommentaar</b>         | <b>Täidetud</b> | <b>Allikas</b> | <b>Genereerimise loogika</b>                               |
|----------------------|---------------------------|-----------------|----------------|------------------------------------------------------------|
| AkpID                | Tabeli unikaalne võti     | Jah             | Genereeritud   | Genereeritakse järjestikuse indeksi põhjal                 |
| AkpKood              | Komponendi kood           | Jah             | Lähteandmed    | KOOD                                                       |
| AkpIDVanem           | Viide vanemkomponendile   | Jah             | Lähteandmed    | YLEMKOMP_KOOD                                              |
| AkpLyhikeNimetus     | Komponendi lühike nimetus | Jah             | Lähteandmed    | NIMETUS                                                    |
| AkpPikkNimetus       | Komponendi pikk nimetus   | Jah             | Lähteandmed    | NIMETUS_LIIIGIGA                                           |
| AkpTaseNR            | Komponendi tase           | Jah             | Lähteandmed    | TASE                                                       |
| AkpKehtivAlatesKpv   | Kehivuse algusaeg         | Jah             | Lähteandmed    | KEHTIV                                                     |
| AkpKehtivKuniKpv     | Kehivuse lõppaeg          | Jah/Ei          | Lähteandmed    | KEHTETU                                                    |
| LoodiKpv             | Kirje loomise aeg         | Jah             | Genereeritud   | Juhuslik kuupäev enne KEHTIV                               |
| KdIDStaatus          | Komponendi staatus        | Jah             | Genereeritud   | KEHTIV / KEHTETU järgi                                     |
| IASIDLooja           | Kirje looja viide         | Jah             | Genereeritud   | Valitakse juhuslikult isikuasutus tabelist ametniku ID     |
| IASIDMuutja          | Kirje muutja viide        | Jah             | Genereeritud   | Kui on Muudet iKpv, siis valitakse juhuslikult ametniku ID |
| Muudet iKpv          | Kirje muutmise aeg        | Ei              | -              | Vaikimisi None                                             |
| Kustutat iKpv        | Kirje kustutamise aeg     | Ei              | -              | Vaikimisi None                                             |
| KdIDAkpliik          | Komponendi liik           | Ei              | -              | Vaikimisi None                                             |
| AkpUnikaalsus        | Unikaalsuse tõeväärtus    | Ei              | -              | Vaikimisi None                                             |
| AkpIDAds             | Välise süsteemi ID        | Ei              | -              | Vaikimisi None                                             |
| TsID                 | Viide tööseansile         | Ei              | -              | Vaikimisi None                                             |
| AkpTase7NumberOsa    | Komponendi numbriline osa | Ei              | -              | Vaikimisi None                                             |
| AkpRoopNimi          | Komponendi rööpnimi       | Ei              | -              | Vaikimisi None                                             |
| ADS_KOMP_ID          | ADS komponendi ID         | Ei              | -              | Vaikimisi None                                             |
| ADS_TASE             | ADS tase                  | Ei              | -              | Vaikimisi None                                             |
| ADS_KOOD             | ADS kood                  | Ei              | -              | Vaikimisi None                                             |
| ADS_NIMETUS          | ADS nimetus               | Ei              | -              | Vaikimisi None                                             |
| ADS_NIMETUS_LIIIGIGA | ADS nimetus liigiga       | Ei              | -              | Vaikimisi None                                             |
| ADS_YLEKOMP_TASE     | ADS ülekomponendi tase    | Ei              | -              | Vaikimisi None                                             |
| ADS_YLEKOMP_KOOD     | ADS ülekomponendi kood    | Ei              | -              | Vaikimisi None                                             |
| ADS_KEHTIV           | ADS kehivuse algus        | Ei              | -              | Vaikimisi None                                             |
| ADS_KEHTETU          | ADS kehivuse lõpp         | Ei              | -              | Vaikimisi None                                             |
| ADS_MUDETI           | ADS muutmise aeg          | Ei              | -              | Vaikimisi None                                             |
| ADS_RR_NIMED         | RR nimetus                | Ei              | -              | Vaikimisi None                                             |
| ADSILID              | Viide logikirjele         | Ei              | -              | Vaikimisi None                                             |

Tabel 4. Ajutise tabeli (temp\_relationships) tabeli väljade täitmise kirjeldus.

| Veeru nimetus   | Kommentaar                            | Täidetud | Allikas      | Genereerimise loogika                                                                                                       |
|-----------------|---------------------------------------|----------|--------------|-----------------------------------------------------------------------------------------------------------------------------|
| IsID            | Isiku unikaalne identifikaator        | Jah      | Genereeritud | Järjestikune ID                                                                                                             |
| Vanuse staatus  | Kas isik on täiskasvanu või laps      | Jah      | Genereeritud | Juhuslikult valitud vastavalt vanusele (täiskasvanu alates 18 a)                                                            |
| Suhteseis       | Isiku perekonnanais                   | Jah      | Genereeritud | Täiskasvanutel juhuslikult „Vallaline”, „Abielus” või „Lahutatud”; lastel alati „Vallaline”                                 |
| Partneri ID     | Viide partnerile                      | Jah/Ei   | Genereeritud | Kui isik on „Abielus”, siis viide teisele täiskasvanule, kes on samuti abielus ja partnerita                                |
| Laps(ed)        | Viited lastele                        | Jah/Ei   | Genereeritud | Lastele määratakse juhuslikult üks või kaks täiskasvanut vanemateks; vanemate kirje sisaldab viiteid laste ID-dele          |
| Vanem(ad)       | Viited vanematele                     | Jah/Ei   | Genereeritud | Lastel 1–2 vanemat; täiskasvanutel tühi                                                                                     |
| Sugu            | Isiku sugu                            | Jah      | Genereeritud | Juhuslikult valitud „MEES” või „NAINE”                                                                                      |
| Sünniaeg        | Isiku sünnikuupäev                    | Jah      | Genereeritud | Täiskasvanutel vanus 18–100 aastat; lastel 0–17 aastat                                                                      |
| Surmaaeg        | Isiku surmakuupäev                    | Jah/Ei   | Genereeritud | Täiskasvanutel x% ja lastel y% tõenäosusega juhuslik surmakuupäev sünniaja ja tänase vahel                                  |
| Haridus         | Isiku haridustase                     | Jah      | Genereeritud | Määratakse vastavalt vanusele kasutades haridusklassifikaatorit                                                             |
| Kodakondsus     | Isiku kodakondsus                     | Jah      | Genereeritud | x% "EE" (Eesti), y% "MUU" (muu riik)                                                                                        |
| IsSaabusEesti   | Saandumise kuupäev Eestisse           | Jah/Ei   | Genereeritud | Kui kodakondsus on "MUU", siis juhuslik kuupäev sünniaja ja tänase vahel; lastel lisaks kontroll, et ei saabu enne vanemaid |
| IsLahkusEestist | Lahkumise kuupäev Eestist             | Jah/Ei   | Genereeritud | "MUU" kodakondsusega isikutel x% tõenäosusega juhuslik lahkumise kuupäev pärast saabumist                                   |
| Aadress         | Isiku elukoha aadress (aadress)       | Jah      | Lähteandmed  | Abielus paaridel x% sama aadress; lastel tavaliselt vanema aadress kuni 18-aastaseni                                        |
| Perekonna ID    | Viide perekonnale                     | Jah      | Genereeritud | Grupitakse BFS algoritmiga seotud isikud (partnerid, lapsed, vanemad) üheks perekonnaks                                     |
| KOV             | Kohaliku omavalitsuse viide (aadress) | Jah/Ei   | Lähteandmed  | Määratakse isiku aadressi alusel aadressi tabelist                                                                          |
| AsID            | Asutuse ID, kus isik töötab           | Jah/Ei   | Genereeritud | Täiskasvanutel x% tõenäosusega juhuslik asutuse ID, lastel puudub                                                           |
| vanus_aastates  | Vanus aastates                        | Jah      | Genereeritud | Arvutatakse sünnikuupäeva ja tänase kuupäeva järgi                                                                          |

Tabel 5. Asutuste tabeli (asutus) väljade täitmise kirjeldus.

| <b>Veeru nimetus</b> | <b>Kommentaar</b>          | <b>Täidetud</b> | <b>Allikas</b> | <b>Genereerimise loogika</b>                                                                                         |
|----------------------|----------------------------|-----------------|----------------|----------------------------------------------------------------------------------------------------------------------|
| AsID                 | Tabeli unikaalne vöti      | Jah             | Genereeritud   | Genereeritakse järjestikune ID                                                                                       |
| AsAlguseKpv          | Asutuse kehtivuse algus    | Jah             | Genereeritud   | Juhuslik kuupäev alates 1980 kuni praeguseni                                                                         |
| AsLopuKpv            | Asutuse kehtivuse lõpp     | Jah/Ei          | Genereeritud   | x% tõenäosusega juhuslik kuupäev alates algusest kuni praeguseni                                                     |
| AsKontakt            | Asutuse kontakt tekstina   | Jah             | Genereeritud   | Genereeritud kujul "info@asutus.ee"                                                                                  |
| AsRegNumber          | Registrikood               | Jah             | Lähteandmed    | Lähteandmetest                                                                                                       |
| KdIDAsutuseLiik      | Asutuse liik               | Jah             | Genereeritud   | Juhuslik väärtus klassifikaatorist (KdKodifikaatoriKood = 10)                                                        |
| AsLyhiNimi           | Asutuse lühike nimetus     | Jah             | Lähteandmed    | Lähteandmetest nimi                                                                                                  |
| AsPikkNimi           | Asutuse pikk nimetus       | Jah             | Lähteandmed    | Lähteandmetest täiendav nimi või nimi                                                                                |
| IASIDLooja           | Kirje looja viide          | Jah             | Genereeritud   | Valitakse juhuslikult isikuasutus tabelist ametniku ID                                                               |
| LoodiKpv             | Kirje loomise aeg          | Jah             | Genereeritud   | Juhuslik kuupäev enne asutuse alguse kuupäeva                                                                        |
| IASIDMuutja          | Kirje muutja viide         | Jah/Ei          | Genereeritud   | Kui on täidetud MuudetiKpv, siis valitakse juhuslikult isikuasutus tabelist ametniku ID                              |
| MuudetiKpv           | Kirje muutmise aeg         | Jah/Ei          | Genereeritud   | x% tõenäosusega juhuslik kuupäev pärast loomist kuni praeguseni                                                      |
| KustutatiKpv         | Kirje kustutamise aeg      | Jah/Ei          | Genereeritud   | x% tõenäosusega, ainult kui asutusel on lõpu kuupäev                                                                 |
| AsMuuKeelesNimi      | Asutuse võõrkeelne nimetus | Jah             | Genereeritud   | Vaikimisi tekst „Muu keeles nimi“                                                                                    |
| KdIDStaatus          | Asutuse staatus            | Jah             | Genereeritud   | Staatus määratakse vastavalt lähteandmete väljale staatus („Registrisse kantud“ siis "KEHTIV", muul juhul "KEHTETU") |
| AsIDYlemus           | Ülemasutus                 | Ei              | -              | Vaikimisi None                                                                                                       |
| KdIDRiik             | Viide riigile              | Jah             | Genereeritud   | Vaikimisi "EE" (Eesti) klassifikaatorist                                                                             |
| AsAsutuseAadress     | Asutuse aadress tekstina   | Jah             | Lähteandmed    | Aadress veerg                                                                                                        |
| AsInglise            | Nimetus inglise keeles     | Jah             | Genereeritud   | Vaikimisi tekst „Inglise keelne nimetus“                                                                             |
| AsSaksa              | Nimetus saksa keeles       | Jah             | Genereeritud   | Vaikimisi tekst „Saksa keelne nimetus“                                                                               |
| AsPrantsuse          | Nimetus prantsuse keeles   | Jah             | Genereeritud   | Vaikimisi tekst „Prantsuse keelne nimetus“                                                                           |
| AsAsutuseKood        | Asutuse kood               | Jah             | Genereeritud   | Genereeritud kujul "AS-<juhuslik arv vahemikus 1000–9999>"                                                           |
| AkpID                | Viide omavalitsusele       | Jah             | Genereeritud   | Juhuslik väärtus vahemikus 1–500                                                                                     |

Tabel 6. Isikute ja asutuste (isikasutus) tabeli väljade täitmise kirjeldus.

| Veeru nimetus           | Kommentaar                  | Täidetud | Allikas      | Genereerimise loogika                                                                      |
|-------------------------|-----------------------------|----------|--------------|--------------------------------------------------------------------------------------------|
| IAsID                   | Tabeli unikaalne võti       | Jah      | Genereeritud | Järjestikune ID                                                                            |
| IsID                    | Viide isikule (isik)        | Jah      | Genereeritud | Isiku ID temp_relationships tabelist, kui AsID on olemas                                   |
| AsID                    | Viide asutusele (asutus)    | Jah      | Genereeritud | temp_relationships tabelist                                                                |
| IAsAlgusKpv             | Seose kehtivuse algus       | Jah      | Genereeritud | Juhuslik kuupäev ajavahemikus 1900 kuni tänaseni                                           |
| IAsKinniKpv             | Seose kehtivuse lõpp        | Jah/Ei   | Genereeritud | x% töenäosusega kuupäev 0–7 päeva pärast algust (kuni tänaseni)                            |
| KdIDStaatus             | Seose staatus               | Jah      | Genereeritud | x% "KEHTIV", y% "KEHTETU"; kui KustutatiKpv on täidetud, siis alati "KEHTETU"              |
| IAsKasutajaNimi         | Süsteemne kasutajanimi      | Jah      | Genereeritud | Kujul user_IsID_AsID                                                                       |
| IAsKasutajaSalasona     | Süsteemne salasõna          | Jah      | Genereeritud | Kujul "pwdXXXX", kus XXXX on juhuslik neljakohaline arv                                    |
| IAsLisainfo             | Lisainfo tekstina           | Jah      | Genereeritud | Kujul „Lisainfo isikule X”                                                                 |
| IAsKirjeteArv           | Seotud kirjete arv          | Jah      | Genereeritud | Juhuslik täisarv vahemikus 0–200                                                           |
| AKpIDPadevusAla         | Pädevusala viide            | Ei       | -            | Vaikimisi None                                                                             |
| IAsMuudatused           | Muudatuste logi             | Ei       | -            | Vaikimisi None                                                                             |
| IAsAndmeväljad          | Kasutatavad andmeväljad     | Ei       | -            | Vaikimisi None                                                                             |
| IAsDokumendiLiigid      | Kasutatavad dokumendiliigid | Ei       | -            | Vaikimisi None                                                                             |
| IAsKontaktTeavitamiseks | Kontakt e-post              | Jah      | Genereeritud | Kujul kasutajanimi@domain.ee                                                               |
| IAsIDLooja              | Kirje looja viide           | Jah      | Genereeritud | Valitakse juhuslikult isikuasutus tabelist ametniku ID                                     |
| LoodiKpv                | Kirje loomise aeg           | Jah      | Genereeritud | Juhuslik kuupäev ajavahemikus 1900 kuni IAsAlgusKpv                                        |
| IAsIDMuutja             | Kirje muutja viide          | Jah/Ei   | Genereeritud | Kui on täidetud MuudetiKpv, siis valitakse juhuslikult isikuasutus tabelist ametniku ID    |
| MuudetiKpv              | Kirje muutmise aeg          | Jah/Ei   | Genereeritud | x% töenäosusega kuupäev loomisest kuni tänaseni                                            |
| KustutatiKpv            | Kirje kustutamise aeg       | Jah/Ei   | Genereeritud | x% töenäosusega kuupäev algusest kuni tänaseni; kui täidetud, muudetakse staatus "KEHTETU" |

Tabel 7. Isiku aadresside (isikuaadress) tabeli väljade täitmise kirjeldus.

| Veeru nimetus       | Kommentaar                  | Täidetud | Allikas      | Genereerimise loogika                                                                                                                                  |
|---------------------|-----------------------------|----------|--------------|--------------------------------------------------------------------------------------------------------------------------------------------------------|
| IAdrID              | Tabeli unikaalne võti       | Jah      | Genereeritud | Järjestikune ID algusega 1                                                                                                                             |
| AdrID               | Viide aadressile (aadress)  | Jah      | Genereeritud | Viide valitakse juhuslikult, viimane aadress vastavalt final_address_map-ile                                                                           |
| IsID                | Viide isikule (isik)        | Jah      | Genereeritud | Perekonna grupi isikute andmetest                                                                                                                      |
| IAdrKehtibAlatesKpv | Aadressi kehtivuse algus    | Jah      | Genereeritud | Alguskuupäev juhuslik varasema perioodi lõpust või alates varasemast lubatud kuupäevast, vajadusel kohandatakse pere saabumise järgi                   |
| IAdrKehtibKuniKpv   | Aadressi kehtivuse lõpp     | Jah/Ei   | Genereeritud | Viimane periood jäetakse avatuks ("KEHTIV"), eelnevatele määratakse juhuslik lõppkuupäev, vajadusel kohandatakse pere lahkumise ja surmakuupäeva järgi |
| KdIDAadressiLiik    | Aadressi liik               | Jah      | Genereeritud | Viimane aadress "ELUKOHT", varasemad "ENDINE ELUKOHT"                                                                                                  |
| KdIDAadressiStaatus | Aadressi staatus            | Jah      | Genereeritud | Viimane periood "KEHTIV", eelnevad "KEHTETU"                                                                                                           |
| DokIDAlus           | Alguse alusdokument         | Jah      | Genereeritud | Järjestikune dokumentide ID                                                                                                                            |
| IAdrIDJargmine      | Viide järgnevale aadressile | Jah/Ei   | Genereeritud | Viide järgmisele aadressile samal isikul, viimasel aadressil tühi                                                                                      |
| DokIDLopuAlus       | Lõpu alusdokument           | Jah/Ei   | Genereeritud | Viide järgmise aadressi DokIDAlusele, viimasel aadressil tühi                                                                                          |
| ADSILID             | Viide logikirjele           | Ei       | -            | Vaikimisi None                                                                                                                                         |
| IAdrPiirang         | Piirangu tõeväärtus         | Ei       | -            | Vaikimisi None                                                                                                                                         |
| IAsIDLooja          | Kirje looja viide           | Jah      | Genereeritud | Valitakse juhuslikult isikuasutus tabelist ametniku ID                                                                                                 |
| LoodiKpv            | Kirje loomise aeg           | Jah      | Genereeritud | Juhuslik kuupäev vahemikus earliest_date kuni aadressi kehtivuse alguskuupäevani                                                                       |
| IAsIDMuutja         | Kirje muutja viide          | Jah/Ei   | Genereeritud | Kui on täidetud Muudet iKpv, siis valitakse juhuslikult isikuasutus tabelist ametniku ID                                                               |
| MuudetiKpv          | Kirje muutmise aeg          | Jah/Ei   | Genereeritud | x% tõenäosusega pärast loomist                                                                                                                         |
| KustutatiKpv        | Kirje kustutamise aeg       | Jah/Ei   | Genereeritud | x% tõenäosusega pärast aadressi lõpu                                                                                                                   |

Tabel 8. Isiku (isik) tabeli väljade täitmise kirjeldus.

| Veeru nimetus     | Kommentaar                              | Täidetud | Allikas     | Geneereimise loogika                                   |
|-------------------|-----------------------------------------|----------|-------------|--------------------------------------------------------|
| IsID              | Isiku ID (temp_relationships)           | Jah      | Geneereitud | Võetakse temp_relationships.IsID                       |
| IsIsikukood       | Isikukood                               | Jah      | Geneereitud | Funktsioon generate_isikukood soo ja sünniaja põhjal   |
| IsEesnimi         | Eesnimi                                 | Jah      | Geneereitud | Faker                                                  |
| IsPerenimi        | Perenimi                                | Jah      | Geneereitud | Faker                                                  |
| IsEesnimiEri      | Eesnimi muus kirjaviisis                | Jah      | Geneereitud | Sama mis IsEesnimi                                     |
| IsPerenimiEri     | Perenimi muus kirjaviisis               | Jah      | Geneereitud | Sama mis IsPerenimi                                    |
| IsIsanimi         | Isanimi                                 | Jah      | Geneereitud | Faker juhuliku meenimi                                 |
| isSynniaeg        | Sünniaeg (temp_relationships)           | Jah      | Geneereitud | temp_relationships.Sünniaeg                            |
| IsSurmaaeg        | Surmaaeg (temp_relationships)           | Jah/Ei   | Geneereitud | temp_relationships.Surmaaeg                            |
| KdIDHaridus       | Hariduse kood (temp_relationships)      | Jah/Ei   | Geneereitud | temp_relationships.Haridus; puudumisel None            |
| KdIDEnakeel       | Emakeele kood (kodi fikaator)           | Jah      | Geneereitud | Vaikimisi eesti; välismaalastel juhulik mitte-EST kood |
| KdIDRahvus        | Rahvuse kood (kodi fikaator)            | Jah      | Geneereitud | Vaikimisi eesti; välismaalastel juhulik mitte-EST kood |
| KdIDKodakondsus   | Kodakondsuse kood (kodi fikaator)       | Jah      | Geneereitud | Vaikimisi eesti; välismaalastel juhulik mitte-EST kood |
| KdIDPerekonnaseis | Perekonnaseisu kood                     | Ei       | Geneereitud | Vaikimisi None                                         |
| KdIDIIsikuStaatus | Isiku olek (ELUS/SURNUD)                | Jah      | Geneereitud | Surmaaaja olemasolu alusel                             |
| KdIDKirjeStaatus  | Kirje olek (REGISTRIS/ARHIIVIS)         | Jah      | Geneereitud | Isiku staatuse põhjal                                  |
| KdIDPohjus        | Arhiiveerimise põhjus                   | Ei       | Geneereitud | Vaikimisi None                                         |
| KdIDSugu          | Sugu (temp_relationships)               | Jah      | Geneereitud | „MEES” / „NAINE”                                       |
| IsSaabusEesti     | Saabumine Eestisse (temp_relationships) | Jah/Ei   | Geneereitud | temp_relationships.IsSaabusEesti                       |
| IslahkusEestist   | Lahkumine Eestist (temp_relationships)  | Jah/Ei   | Geneereitud | temp_relationships.IsLahkusEestist                     |
| IsSaabusKOV-iKpv  | Saabumine KOV-i (isikuaadress)          | Jah/Ei   | Geneereitud | Varasem IAdrKehtibAlatesKpv isikuaadressidest          |
| AKpID             | Viide KOV-ile (temp_relationships)      | Jah/Ei   | Geneereitud | temp_relationships.KOV                                 |
| Idkiht            | Kiht ID                                 | Ei       | Geneereitud | Praegu None                                            |
| LoodiKpv          | Kirje loomise aeg                       | Jah      | Geneereitud | isSynniaeg või tänane kuupäev, kui sünniaeg puudub     |
| MuudetiKpv        | Kirje muutmise aeg                      | Jah/Ei   | Geneereitud | 50 % tõenäosusega [LoodiKpv...täna]                    |
| IasIDLooja        | Kirje looja viide (isikuasutus)         | Jah      | Geneereitud | Juhuliku ametniku ID                                   |
| IasIDMuutja       | Kirje muutja viide (isikuasutus)        | Jah/Ei   | Geneereitud | Täidetakse koos MuudetiKpv                             |
| KustutatiKpv      | Kirje kustutamise aeg                   | Ei       | Geneereitud | Vaikimisi None                                         |

Tabel 9. Kodakondsuse (kodakondsus) tabeli väljade täitmise kirjeldus.

| Veeru nimetus   | Kommentaar                       | Täidetud | Allikas      | Genereerimise loogika                                                                                     |
|-----------------|----------------------------------|----------|--------------|-----------------------------------------------------------------------------------------------------------|
| KodID           | Tabeli unikaalne võti            | Jah      | Genereeritud | Järjestikune täisarv alates 1                                                                             |
| IsID            | Viide isikule (isik)             | Jah      | Genereeritud | Kopeeritakse df_isik.IsID                                                                                 |
| KdIDRiik        | Riigi kood (kodifikaator)        | Jah      | Genereeritud | Võetakse veerust df_isik.KdIDKodakondsus                                                                  |
| KodKehtibAlates | Kehtivuse algus                  | Jah      | Genereeritud | Juhuslik kuupäev [saabumise/sünni kuupäev ... täna]                                                       |
| KodKehtibKuni   | Kehtivuse lõpp                   | Jah/Ei   | Genereeritud | 20 % tõenäosusega kuupäev [alates ... täna], muidu None                                                   |
| DokIDAlguseAlus | Algusdokumendi ID                | Jah      | Genereeritud | Järjestikune ID, alustades start_doc_id                                                                   |
| DokIDLopuAlus   | Lõppdokumendi ID                 | Jah/Ei   | Genereeritud | Määratakse ainult kui KodKehtibKuni pole None                                                             |
| KdIDStaatus     | Kodakondsuse staatus             | Jah      | Genereeritud | „KEHTIV” või „KEHTETU” vastavalt lõppkuupäevale                                                           |
| IAsIDLooja      | Kirje looja viide (isikuasutus)  | Jah      | Genereeritud | Juhuslik ametniku ID                                                                                      |
| IAsIDMuutja     | Kirje muutja viide (isikuasutus) | Jah/Ei   | Genereeritud | Täidetakse koos Muudetikpv                                                                                |
| LoodiKpv        | Kirje loomise aeg                | Jah      | Genereeritud | Juhuslik kuupäev [earliest_date ... KodKehtibAlates-1]                                                    |
| Muudetikpv      | Kirje muutmise aeg               | Jah/Ei   | Genereeritud | Kui staatus „KEHTIV” 50 % tõenäosus [LoodiKpv ... täna]; staatus „KEHTETU” = Kustutatikpv v KodKehtibKuni |
| Kustutatikpv    | Kirje kustutamise aeg            | Jah/Ei   | Genereeritud | määratud tõenäosus, ainult kui KodKehtibKuni määratud                                                     |

Tabel 10. Isiku dokumentide (isikudokument) tabeli väljade täitmise kirjeldus.

| <b>Veeru nimetus</b> | <b>Kommentaar</b>                      | <b>Täidetud</b> | <b>Allikas</b> | <b>Genereerimise loogika</b>                                                            |
|----------------------|----------------------------------------|-----------------|----------------|-----------------------------------------------------------------------------------------|
| IDokID               | Tabeli unikaalne võti                  | Jah             | Genereeritud   | Järjestikune ID                                                                         |
| IsID                 | Viide isikule (isik)                   | Jah             | Genereeritud   | Dokumendi ja isiku seosest                                                              |
| DokID                | Viide dokumendile (dokument)           | Jah             | Genereeritud   | Dokumendi tabelist                                                                      |
| IDokIsikukood        | Dokumendil olev isikukood (isik)       | Jah/Ei          | Genereeritud   | Isiku tabelist või None                                                                 |
| IDokEesnimi          | Dokumendil olev eesnimi (isik)         | Jah/Ei          | Genereeritud   | Isiku tabelist või None                                                                 |
| IDokPerenimi         | Dokumendil olev perenimi (isik)        | Jah/Ei          | Genereeritud   | Isiku tabelist või None                                                                 |
| IDokVanaEesnimi      | Dokumendil olev endine eesnimi         | Jah/Ei          | Genereeritud   | x% tõenäosusega eesliitega "Vana-" eesnimest                                            |
| IDokVanaPerenimi     | Dokumendil olev endine perenimi        | Jah/Ei          | Genereeritud   | x% tõenäosusega eesliitega "Vana-" perenimest                                           |
| IDokVanaIsikukood    | Dokumendil olev endine isikukood       | Jah/Ei          | Genereeritud   | x% tõenäosusega kujul "OLD-XXXX"                                                        |
| IDokIsanimi          | Dokumendil olev isanimi (isik)         | Jah/Ei          | Genereeritud   | Isiku tabelist või None                                                                 |
| KdIDIsikuRoll        | Isiku roll dokumendil                  | Jah             | Genereeritud   | Vaikimisi "Mingi roll"                                                                  |
| AsID                 | Dokumendi väljaandja asutus (dokument) | Jah/Ei          | Genereeritud   | Dokumendi tabelist (AsIDValjaandjaAsutus) või None                                      |
| IAsIDLooja           | Kirje looja viide (isikasutus)         | Jah             | Genereeritud   | Valitakse juhuslikult isikuasutus tabelist ametniku ID                                  |
| LoodiKpv             | Kirje loomise aeg                      | Jah             | Genereeritud   | Juhuslik kuupäev dokumendi loomise ja praeguse vahel                                    |
| IAsIDMuutja          | Kirje muutja viide (isikasutus)        | Jah/Ei          | Genereeritud   | Kui on täidetud MuudetiKpv, siis valitakse juhuslikult isikuasutus tabelist ametniku ID |
| MuudetiKpv           | Kirje muutmise aeg                     | Jah/Ei          | Genereeritud   | Määratud tõenäosusega pärast loomist kuni praeguseni                                    |
| KustutatiKpv         | Kirje kustutamise aeg                  | Jah/Ei          | Genereeritud   | Määratud tõenäosusega pärast loomist kuni praeguseni                                    |

Tabel 11. Dokumentide (dokument) tabeli väljade täitmise kirjeldus.

| Veeru nimetus          | Kommentaar                           | Täidetud | Allikas     | Geneerimise loogika                                                                                |
|------------------------|--------------------------------------|----------|-------------|----------------------------------------------------------------------------------------------------|
| IsID                   | Viide isikule (isik)                 | Jah/Ei   | Lähteandmed | Lähtetabelist või None, väli aitab isiku ja dokumendi seostamisel - peale geneerimist eemaldatakse |
| DokID                  | Tabeli unikaalne või                 | Jah      | Geneeritud  | Järjestikune ID või erandlikult lähtetabelist olemasolul                                           |
| KdIDDokumendiLiik      | Dokumendi liik                       | Jah      | Geneeritud  | Klassifikaatorist vastavalt dokumendi tüübile (nt "PASS", "ID-KAART", "MUU")                       |
| DokNumber              | Dokumendi number                     | Jah      | Geneeritud  | Formaadis "DOK-LIIK-0000001"                                                                       |
| DokSeeria              | Dokumendi seeria                     | Jah      | Geneeritud  | Formaadis "S-XXX", kus XXX on juhuslik arv                                                         |
| DokValjaantudKpv       | Dokumendi väljastamise kuupäev       | Jah      | Geneeritud  | Juhuslik kuupäev vahemikus loomise ja tänase vahel                                                 |
| DokKehtivKuniKpv       | Dokumendi kehtivuse lõpp             | Jah/Ei   | Geneeritud  | x% tööaastusega juhuslikult määratud kuupäev pärast väljastamist                                   |
| DokKehtetuAlatesKpv    | Dokumendi kehtetuks muutmise kuupäev | Jah/Ei   | Geneeritud  | Kui määratud kehtivuse lõpp, siis x% tööaastusega sama kuupäev või hilisem                         |
| AsIDValjaandjaAsutus   | Dokumendi väljastanud asutus         | Jah      | Geneeritud  | Juhuslik asutuse ID vahemikus 1–50                                                                 |
| KaIDKanne              | Viide kannele                        | Ei       | -           | Ei täideta (None)                                                                                  |
| IasIDLooja             | Kirje looja viide (isikasutus)       | Jah      | Geneeritud  | Valitakse juhuslikult isikuasutust tabelist ametniku ID                                            |
| LoodiKpv               | Kirje loomise aeg                    | Jah      | Geneeritud  | Juhuslik kuupäev alates 1980 kuni tänaseni                                                         |
| IasIDMuutja            | Kirje muutja viide (isikasutus)      | Jah/Ei   | Geneeritud  | Kui on täidetud MuudetiKpv, siis valitakse juhuslikult isikuasutust tabelist ametniku ID           |
| MuudetiKpv             | Kirje muutmise aeg                   | Jah/Ei   | Geneeritud  | Kui dokumentil on lõppkuupäev, siis muutmise aeg = lõppkuupäev                                     |
| KustutatiKpv           | Kirje kustutamise aeg                | Jah/Ei   | Geneeritud  | x% tööaastusega juhuslikult määratud kuupäev                                                       |
| KdIDDokumendiStaatus   | Dokumendi staatus                    | Jah      | Geneeritud  | KEHTIV, kui lõppkuupäeva pole, muidu "KEHTETU"                                                     |
| DokIDVanem             | Viide vanemdokumendile               | Ei       | -           | Ei täideta (None)                                                                                  |
| DokKehtivaLates        | Dokumendi kehtivuse algus            | Jah      | Geneeritud  | Tavaliselt sama mis väljastamise kuupäev                                                           |
| KdIDRIik               | Viide riigile                        | Jah      | Geneeritud  | Vahimisi "EE" (Eesti)                                                                              |
| DokAsutuseTekst        | Asutuse nimi tekstina                | Ei       | -           | Ei täideta (None)                                                                                  |
| AsIDHaldaJa            | Dokumendi haldaja asutus             | Jah      | Geneeritud  | Juhuslik asutuse ID vahemikus 1–50                                                                 |
| DokMarkus              | Märkus                               | Ei       | -           | Ei täideta (None)                                                                                  |
| DokValjastajaIsikukood | Väljastaja isikukood                 | Ei       | -           | Ei täideta (None)                                                                                  |
| DokValjastajaNimi      | Väljastaja nimi                      | Ei       | -           | Ei täideta (None)                                                                                  |
| MKaID                  | Viide menetlusele                    | Ei       | -           | Ei täideta (None)                                                                                  |
| DokSkaneeritud         | Kas dokument skaneeritud             | Ei       | -           | Ei täideta (None)                                                                                  |
| AsIDAsutusTellija      | Tellitud asutus                      | Ei       | -           | Ei täideta (None)                                                                                  |
| LRProtseduuriKirjeldus | Protseduuri kirjeldus                | Ei       | -           | Ei täideta (None)                                                                                  |
| DokSalastatud          | Kas dokument salastatud              | Ei       | -           | Ei täideta (None)                                                                                  |
| KdIDDokAlguseAlus      | Dokumendi alguse alus                | Ei       | -           | Ei täideta (None)                                                                                  |

### III. Litsents

#### **Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks**

Mina, **Rain Eichhorn**,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose  
**Eesti rahvastikuregistrile sarnase sünteetilise andmestiku reeglipõhine gene-reerimine**,  
mille juhendajad on Sven Laur ja Liina Kamm,  
reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.
2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commons'i litsentsiga CC BY NC ND 3.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.
3. Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.
4. Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Rain Eichhorn

**15.05.2025**