TARTU ÜLIKOOL
Arvutiteaduse instituut
Informaatika õppekava

Allan Alikas

Privaatsust tagavad anonüümimistarkvarad

Bakalaureusetöö (9 EAP)

Juhendaja: Sulev Reisberg

Tartu 2022

Privaatsust tagavad anonüümimistarkvarad

Lühikokkuvõte:

Lõputöö eesmärk on võrrelda üldistamisel põhinevaid anonüümimistarkvarasid, et võtta neist sobivaim kasutusele Health Sense projektis. Võrdlusse kaasati kolm tarkvara: ARX, Amnesia ja Anonimatron. Võrdlus viidi läbi 10 hindamiskriteeriumi alusel, millest osad pärinevad ISO-9421 standardist ja osad tulenevad Health Sense projekti vajadustest. Võrdluse läbiviimiseks loodi 3 komplekti testandmeid, mida kasutati anonüümimisel sisendina. Sobivaimaks tarkvaraks osutus ARX, mis integreeriti Health Sense projektis valmivasse tarkvarakomplekti.

Võtmesõnad:

Java, anonüümimine, avaandmed

CERCS: P170 Arvutiteadus, süsteemid

Privacy preserving anonymization software

Abstract:

The aim of this thesis is to compare generalization-based anonymization software in order to implement the most suitable one in the Health Sense project. Three software were included in the comparison: ARX, Amnesia, and Anonimatron. The comparison was based on 10 evaluation criteria, some of which were derived from the ISO-9421 standard, and others were based on the requirements of the Health Sense project. For the comparison, three sets of test data were compiled and used in the anonymization process. The most suitable software was ARX which was integrated into the software suite that is being developed in the Health Sense project.

Keywords:

Java, anonymization, open data

CERCS: P170 Computer science, systems

Sisukord

Sissejuhatus

Andmekogudesse talletatakse suures mahus ja väga erisuguseid andmeid. Andmete tekke ning kogumise ajal tihtipeale seostatakse need isikuandmetega.

Isikuandmed on kõik andmed, mis väljendavad informatsiooni tuvastatava füüsilise isiku kohta [1]. Isikuandmeid liigitatakse mitmesse rühma ning kõik on reguleeritud mitmete seaduste ja määrustega. Antud lõputöös käsitletakse eelkõige terviseandmeid, mis on eriliiki isikuandmed. Eriliiki isikuandmed on andmed, mille avalikustamine võib seada ohtu inimese tervise või tekitada varalist kahju [2].

Riigiasutused koguvad palju erinevaid isikuandmeid. Statistika tegemise eesmärgil võib nende andmete põhjal teha andmeväljastusi, kuid peavad olema täidetud inimeste anonüümsuse nõuded, et nendele inimestele ei oleks võimalik andmetes leiduva informatsiooniga kahju tekitada.

Andmeväljastuse andmetes esinevate inimeste privaatsuse kaitsmiseks on võimalik rakendada erinevaid meetmeid. Andmete anonüümimine on üheks selliseks meetmeks [3]. Andmete anonüümimiseks on loodud mitmeid meetodeid ning tarkvarasid. Samas puudub hea praktika, milliseid neist kasutada ja milliste parameetritega andmeid anonüümida. Anonüümimisel kaitstakse andmetes olevate inimeste privaatsust sellega, et andmeid muudetakse ja üldistatakse. Antud lõputöös käsitletakse andmete anonüümimist üldistamise baasil. Näiteks võib üldistamine tähendada, et täpne arvuline vanus 25 muudetakse vahemikuks 20-29. Andmete üldistamisel säilitatakse kasutuskõlblikkus statistika arvutamiseks.

Eesti riik soovib muuta riiklikes andmekogudes olevaid andmeid avaandmeteks [4]. Avaandmed on andmed, mis on tehtud kõigile avalikult kättesaadavaks, et vabalt kasutada ja jagada eesmärgist olenemata [5]. Tervisega seotud avaandmete loomise ning haldamise lihtsustamiseks loodi Health Sense projekt.

Tartu Ülikool arendab Health Sense projekti raames Tervise ja Heaolu Infosüsteemide Keskusele (edaspidi TEHIK) tarkvara andmeväljastuste anonüümimiseks. Üheks ülesandeks on tarkvarasse integreerida komponent, mis kasutaks anonüümimiseks andmete üldistamist.

Antud töö eesmärgiks oli analüüsida kolme erinevat privaatsust tagavat anonüümimistarkvara, mis kasutavad anonüümimisel andmete üldistamist, valida võrdleva

metoodika abil Health Sense projekti jaoks sobivaim ning integreerida see projektis arendatavasse tarkvarakomplekti. Töö jaguneb neljaks peatükiks. Esimeses peatükis selgitatakse teemaga seonduvaid mõisteid. Teises peatükis antakse taustainfot nii probleemiga kaasnevatele teemadele kui ka Health Sense projektile. Kolmandas peatükis on kirjeldatud tarkvarade võrdluse metoodikat, testandmeid ning tarkvara integratsiooni metoodikat. Neljandas peatükis on toodud võrdluse tulemused ja arutelu. Tööle on lisatud GitHubi hoidla viide, kus on võimalik näha integreeritud tarkvara ning testandmeid.

1. Mõisted ja terminid

Olem (ingl. *entity*) on miski, millel on omadusi ja seoseid.[1] Näiteks inimene, andmebaasi rida jne.

Atribuut (ingl. *attribute*) on nimeline olemi karakteristik või omadus, millega saab kirjeldada ta olekut, ilmet või muid aspekte.[2]

Klassifikatsioon (ingl. *classification*) on oluliste atribuutide alusel loodud olemite jaotussüsteem.[3]

RHK-10 ehk ICD-10 on rahvusvahelise haiguste ja nendega seotud terviseprobleemide ning kaebuste statistilise klassifikatsiooni kümnes versioon, mida haldab Maailma Terviseorganisatsioon.[4]

ATC on rahvusvaheline raviainete toime klassifikatsioon, mida haldab Maailma Terviseorganisatsioon.[5]

EHAK on Eesti haldus- ja asutusüksuste klassifikatsioon, mida haldab Statistikaamet.[6]

Pseudonüümitud andmed (ingl. *pseudonymized data*) on andmed, millele on rakendatud identiteedi maskeerimise protsessi.[7]

Kvaasi-identifikaator (ingl. *quasi-identifier*) on andmeelement, mis eraldiseisvalt ei tuvasta isikut üheselt, kuid koos teiste selliste andmeelementidega võimaldab isiku tuvastamist.[8]

---

[1] https://akit.cyber.ee/term/422-olem

[2] https://akit.cyber.ee/term/1833-atribuut

[3] https://sonaveeb.ee/search/unif/dlall/dsall/klassifikatsioon/1

[4] https://www.who.int/standards/classifications/classification-of-diseases

[5] https://www.riigiteataja.ee/akt/120112018007

[6] https://geoportaal.maaamet.ee/est/Andmed-ja-kaardid/Haldus-ja-asustusjaotus-p119.html

[7] https://akit.cyber.ee/term/914-pseudonuumimine

[8] https://akit.cyber.ee/term/3535-kvaasi-identifikaator

Üldistushierarhia (ingl. *generalization hierarchy*) on ühiste atribuutidega olemite struktureeritud rühmitus. Üldistushierarhia on astmestatud nii, et kõrgema taseme komponent kirjeldab üldisemat olemit kui madalama taseme komponent.[9]

Ekvivalentsiklass (ingl. *equivalence class*) on ühesuguste anonüümitud andmetega kirjete kogum.[10]

[9] https://faculty.kfupm.edu.sa/ICS/mwaslam/ICS014/generalization_hierarchies.htm

[10] https://akit.cyber.ee/term/3552

2. Taustainfo

Selles peatükis selgitatakse anonüümimise ning Health Sense projektiga seonduvaid teemasid ning kirjeldatakse, miks need on tähtsad projekti kontekstis ja ka laiemalt.

2.1 Isikuandmed

Isikuandmed on kõik andmed, mille abil on võimalik tuvastada füüsilist isikut ning selle isiku füüsilisi, psüühilisi, füsioloogilisi, majanduslikke, kultuurilisi või sotsiaalseid omadusi, suhteid ja kuuluvust [1].

Isikuandmed klassifitseeritakse kolme rühma: tavalised, tundlikud ja eriliiki isikuandmed. Tavalisteks isikuandmeteks loetakse kõiki selliseid andmeid, mille alusel on võimalik füüsilist isikut otse või kaudselt tuvastada. Eriliiki isikuandmete alla kuuluvad enamus selliseid andmeid, millest ilmneb isiku päritolu, vaated, veendumused ja kuuluvused gruppidesse. Samuti kuuluvad eriliiki isikuandmete alla biomeetrilised ja terviseandmed. Eriliiki isikuandmete avalikustamisel võib sattuda ohtu vastava inimese tervis ja elu. Tundlike isikuandmete all mõeldakse selliseid andmeid, mis ei kuulu eriliiki isikuandmete loetellu, kuid mis samuti kujutavad suuremat ohtu isiku privaatelule. Tundlikeks loetakse andmeid, mille avaldamisega kaasneb oht elule ja tervisele, identiteedivargusele, varalisele ja mainekahjule jms. [2]

Inimeste kohta kogutud isikuandmed on tihtipeale kombinatsioonid eelmainitud kolmest rühmast. Füüsiliste isikute privaatsuse ja turvalisuse tagamiseks on kehtestatud erinevad seadused ja määrused, mis piiravad isikuandmete töötlemist ning avalikustamist. Neist üheks olulisemaks on Euroopa Liidu isikuandmete kaitse üldmäärus.

2.2 Isikuandmete kaitse üldmäärus

Isikuandmete kaitse üldmäärus (General Data Protection Regulation, GDPR) [6] on 2016. aastal Euroopa Liidus kehtestatud määrus, mis sätestab reeglid Euroopa Liidu elanike isikuandmete kogumise ning töötlemise jaoks. GDPR seab piirangud tegevustele, mida on lubatud andmetega teha ning isegi asukoha, kus andmete töötlemine on lubatud. Määrust peavad järgima kõik ettevõtted ja asutused, mis baseeruvad Euroopa Liidus või töötlevad Euroopa Liidus elavate inimeste andmeid [7].

Üheks hiljutiseks näiteks ettevõttest, mis ei tahtnud täita vastavaid määrusi, on Meta (varasemalt Facebook). 2022. aastal ähvardas Meta lõpetada Euroopa Liidus oma teenuste pakkumise, kuna Euroopa Liit ei lubanud elanike andmeid liigutada USAsse, kus andmete nõuetele vastav töötlemine ei ole garanteeritud [8].

Lisaks eraettevõtetele kehtivad GDPRi reeglid ka avalikele asutustele. Seetõttu peavad kõik isikuandmeid sisaldavad andmeväljastused ja ülekandmised olema GDPRiga kooskõlas [7]. Sealhulgas peavad reegleid järgima kõik avaandmete väljastused.

2.3 Avaandmed

Avaandmed on andmed, mis on tehtud kõigile avalikult kättesaadavaks, et vabalt kasutada ja jagada eesmärgist olenemata. Neid võib kasutada äriliste ja mittetulunduslike ettevõtmiste käivitamiseks, uuringute läbiviimiseks ning andmepõhiste otsuste tegemiseks [5].

Eesti riik soovib liikuda avaandmete suunas, muutes riiklikes andmekogudes olevaid andmeid avaandmeteks. Samuti luuakse keskkonda, kus inimestel oleks võimalik avaandmeid vaadelda ning alla laadida [4]. Terviseandmetest avaandmete loomise protsessi lihtsustamisega tegeleb Health Sense projekt, mida on täpsemalt kirjeldatud peatükis “Health Sense projekt” [9].

Andmete avalikustamisega kaasneb kohustus tagada inimeste privaatsus. Privaatsuse tagamiseks kasutatakse andmete anonüümimist, millest räägib lähemalt järgmine alapeatükk.

2.4 Anonüümimine

Anonüümimine on protsess, mis püüab isikuandmeid muuta nii, et andmete põhjal ei oleks võimalik üheselt tuvastada vastavat isikut [3]. Anonüümimiseks on enamasti kasutusel viit tüüpi operatsioonid: üldistamine, supressioon ehk pärssimine, anatomisatsioon ehk lahkamine, permutatsioon ja perturbatsioon ehk häirimine. Selles töös käsitletakse anonüümimist üldistamise baasil.

Üldistamise baasil anonüümimiseks kasutatakse tavaliselt üldistushierarhiaid, mis kirjeldavad, kuidas andmed peaksid muutuma, et tagada andmetes olevate isikute privaatsus. Näiteks Eesti aadressiandmed on hierarhilised ning on seotud Statistikaameti poolt kehtestatud Eesti haldus- ja asutusjaotuse klassifikaatoriga (EHAK). Kui kasutada EHAKit

üldistushierarhiana, siis on näiteks võimalik andmekogu andmetes Põhja-Tallinna linnaosa asendada hierarhias kõrgemal tasemel oleva Tallinna haldusüksusega. See tagab, et asendus on korrektne ning andmetest ei kaotata liigselt informatsiooni.

Anonüümimisalgoritm on andmetöötlusmeetod, mis muudab isikut tuvastavaid andmeid, et ennetada isikutuvastust. Üldistamise baasil töötavatest anonüümimisalgoritmidest on rohkem informatsiooni järgmistes töödes: „An Extensive Study on Statistical Data Anonymization Algorithms" [10] ja "Data privacy preservation algorithm with k-anonymity" [11].

Algoritmi tööks on vaja teada andmete klassifikatsioone, mis määravad, milliste reeglite põhjal vastavaid andmevälju muuta. Klassifikatsioonid on järgmised:

1. identifikaatorid on andmeväljad, mis üheselt tuvastavad isiku;
    - Näiteks isikukood on seotud ühe kindla inimesega.
2. kvaasi-identifikaatorid on andmeväljad, mis üksi ei identifitseeri isikut üheselt, kuid mitme sellise välja kombinatsioon võimaldab isiku tuvastada;
    - Näiteks ainult aadressi alusel ei ole sageli võimalik isikut üheselt tuvastada. Aga mõningase lisainfoga, nagu näiteks sünnikuupäev, saab üldjuhul tuvastada üksikisiku.
3. sensitiivsed (tundlikud) andmeväljad sisaldavad andmeid, mis on privaatsed, kuid millega ei ole võimalik isikut üheselt tuvastada;
    - Näiteks haigusdiagnoosi põhjal ei ole võimalik isikut tuvastada. Kuid juba tuvastatud isikule võib diagnoosi lekkimine tekitada kahju.
4. mittesensitiivsed andmed on andmed, mis ei kuulu ühtegi eelmainitud klassifikatsiooni.
    - Näiteks inimese lemmikvärv, sest enamasti pole selle põhjal võimalik isikut tuvastada ega ole ka võimalik tuvastatud isikule sellega kahju tekitada.

Anonüümimisel identifikaatorid enamasti lihtsalt eemaldatakse. Kvaasi-identifikaatoreid ja sensitiivseid andmeid muudetakse anonüümimise käigus vastavalt kasutatavale anonüümsuse mudelile. Mittesensitiivsed andmeväljad jäetakse esialgsele kujule.

Kahest tuntumast anonüümsuse mudelist on lähemalt juttu järgmistes alapeatükkides.

2.4.1 *k*-anonüümsus

k-anonüümsus on üks populaarsemaid anonüümsuse mõõdikuid, mida kasutatakse anonüümsuse kaitse formaalsetes mudelites ning anonüümimisalgoritmides. Andmed on *k*-anonüümsed, kui iga olemi andmed ei ole eristatavad vähemalt *k*-1 muu olemi andmetest samas andmekogus [12]. Näide *k*-anonüümsuse baasil anonüümimisest on toodud joonisel 1.

Enne:

id	gender	age	drug_name	ATC	date	doctor_registry_code
1	M	56	LORAMYC	A01AB09	14.03.2022	D0003
2	M	53	KALMENTE	R01AD09	13.03.2022	D0006
3	N	45	ZANTAC 75	A02BA02	14.03.2022	D0002
4	N	47	NEXMEZOL 20 MG	A02BC05	14.03.2022	D0004
5	M	34	NEXMEZOL 20 MG	A02BC05	15.03.2022	D0005
6	N	42	FORTRANS	A06AD81	15.03.2022	D0005
7	M	65	LORAMYC	A01AB09	15.03.2022	D0003
8	N	25	KALMENTE	R01AD09	14.03.2022	D0005
9	M	31	FORTRANS	A06AD81	14.03.2022	D0005
10	N	52	LECROLYN	S01GX01	13.03.2022	D0004
11	N	66	KALMENTE	R01AD09	13.03.2022	D0001
12	N	53	KALMENTE	R01AD09	13.03.2022	D0004
13	N	41	CERUCAL	A03FA01	15.03.2022	D0002
14	M	51	FORTRANS	A06AD81	13.03.2022	D0005
15	N	52	FORTRANS	A06AD81	14.03.2022	D0001
16	N	22	CERUCAL	A03FA01	14.03.2022	D0005
17	M	18	FORTRANS	A06AD81	15.03.2022	D0004
18	N	45	FORTRANS	A06AD81	14.03.2022	D0005
19	N	22	LECROLYN	S01GX01	13.03.2022	D0003
20	M	64	ZANTAC 75	A02BA02	13.03.2022	D0005
21	M	31	ZANTAC 75	A02BA02	13.03.2022	D0005
22	M	23	LECROLYN	S01GX01	15.03.2022	D0004
23	M	20	KALMENTE	R01AD09	14.03.2022	D0005
24	M	62	KALMENTE	R01AD09	15.03.2022	D0003
25	N	71	LORAMYC	A01AB09	14.03.2022	D0002

Pärast:

id	gender	age	drug_name	ATC	date	doctor_registry_code
1	*	*	LORAMYC	A01AB09	3.2022	D0003
2	M	*	KALMENTE	R01AD09	3.2022	D0006
3	*	*	ZANTAC 75	A02BA02	3.2022	D0002
4	N	*	NEXMEZOL 20 MG	*	3.2022	D0004
5	*	*	NEXMEZOL 20 MG	*	*	D0005
6	N	*	FORTRANS	A06AD81	3.2022	D0005
7	*	*	LORAMYC	A01AB09	3.2022	D0003
8	N	*	KALMENTE	R01AD09	3.2022	D0005
9	M	*	FORTRANS	A06AD81	3.2022	D0005
10	*	*	LECROLYN	S01GX01	3.2022	D0004
11	N	*	KALMENTE	R01AD09	3.2022	D0001
12	N	*	KALMENTE	R01AD09	3.2022	D0004
13	N	*	CERUCAL	*	3.2022	D0002
14	M	*	FORTRANS	A06AD81	3.2022	D0005
15	N	*	FORTRANS	A06AD81	3.2022	D0001
16	N	*	CERUCAL	*	3.2022	D0005
17	M	*	FORTRANS	A06AD81	3.2022	D0004
18	N	*	FORTRANS	A06AD81	3.2022	D0005
19	*	*	LECROLYN	S01GX01	3.2022	D0003
20	*	*	ZANTAC 75	A02BA02	3.2022	D0005
21	*	*	ZANTAC 75	A02BA02	3.2022	D0005
22	*	*	LECROLYN	S01GX01	3.2022	D0004
23	M	*	KALMENTE	R01AD09	3.2022	D0005
24	M	*	KALMENTE	R01AD09	3.2022	D0003
25	*	*	LORAMYC	A01AB09	3.2022	D0002

Joonis 1: Anonüümimine *k*-anonüümsuse baasil. Ülemisel pildil on andmed enne ja alumisel pärast anonüümimist.

Joonisel 1 on esitatud andmekogu enne ja pärast anonüümimist, kui on tagatud *k*-anonüümsus = 3. Kvaasi-identifikaatoriteks on määratud väljad „gender“, „age“, „ATC“ ja „date“. Vaatleme näiteks esiletõstetud ridasid 1, 5 ja 7. Enne anonüümimist on võimalik

eristada kolme erinevat isikut. Pärast anonüümimist pole võimalik kvaasi-identifikaatorite alusel neid ridasid üksteisest enam eristada, sest ridade vastavad väärtused ühtivad täpselt või on mõni väärtustest asendatud üldistatud väärtusega. Kui korrata sama analüüsi terves anonüümitud andmekogus, on võimalik näha, et iga rea kohta leidub veel vähemalt 2 rida, mida pole võimalik valitud reast eristada. Seega on pärast anonüümimist tagatud *k*-anonüümsus = 3.

2.4.2 *l*-hajutus

l-hajutus on *k*-anonüümsuse edasiarendus, mille eesmärk on paremini kaitsta tundlikke atribuute. Andmed on *l*-hajutusega, kui igal tundlikul atribuudil on igas ekvivalentsiklassis vähemat *l* erinevat ja ühtlaselt jaotatud väärtust [13]. Näide *k*-anonüümsuse ja *l*-hajutuse baasil anonüümimisest on toodud joonisel 2.

Enne:

id	gender	age	drug_name	ATC	date	doctor_registry_code
1	M	56	LORAMYC	A01AB09	14.03.2022	D0003
2	M	53	KALMENTE	R01AD09	13.03.2022	D0006
3	N	45	ZANTAC 75	A02BA02	14.03.2022	D0002
4	N	47	NEXMEZOL 20 MG	A02BC05	14.03.2022	D0004
5	M	34	NEXMEZOL 20 MG	A02BC05	15.03.2022	D0005
6	N	42	FORTRANS	A06AD81	15.03.2022	D0005
7	M	65	LORAMYC	A01AB09	15.03.2022	D0003
8	N	25	KALMENTE	R01AD09	14.03.2022	D0005
9	M	31	FORTRANS	A06AD81	14.03.2022	D0005
10	N	52	LECROLYN	S01GX01	13.03.2022	D0004
11	N	66	KALMENTE	R01AD09	13.03.2022	D0001
12	N	53	KALMENTE	R01AD09	13.03.2022	D0004
13	N	41	CERUCAL	A03FA01	15.03.2022	D0002
14	M	51	FORTRANS	A06AD81	13.03.2022	D0005
15	N	52	FORTRANS	A06AD81	14.03.2022	D0001
16	N	22	CERUCAL	A03FA01	14.03.2022	D0005
17	M	18	FORTRANS	A06AD81	15.03.2022	D0004
18	N	45	FORTRANS	A06AD81	14.03.2022	D0005
19	N	22	LECROLYN	S01GX01	13.03.2022	D0003
20	M	64	ZANTAC 75	A02BA02	13.03.2022	D0005
21	M	31	ZANTAC 75	A02BA02	13.03.2022	D0005
22	M	23	LECROLYN	S01GX01	15.03.2022	D0004
23	M	20	KALMENTE	R01AD09	14.03.2022	D0005
24	M	62	KALMENTE	R01AD09	15.03.2022	D0003
25	N	71	LORAMYC	A01AB09	14.03.2022	D0002

Pärast:

id	gender	age	drug_name	ATC	date	doctor_registry_code
1	*	*	LORAMYC	*	*	D0003
2	M	*	KALMENTE	R01AD09	3.2022	D0006
3	N	*	ZANTAC 75	*	14.03.2022	D0002
4	N	*	NEXMEZOL 20 MG	*	14.03.2022	D0004
5	M	*	NEXMEZOL 20 MG	*	15.03.2022	D0005
6	*	*	FORTRANS	A06AD81	3.2022	D0005
7	M	*	LORAMYC	*	15.03.2022	D0003
8	N	*	KALMENTE	R01AD09	3.2022	D0005
9	*	*	FORTRANS	A06AD81	3.2022	D0005
10	N	*	LECROLYN	*	3.2022	D0004
11	N	*	KALMENTE	R01AD09	3.2022	D0001
12	N	*	KALMENTE	R01AD09	3.2022	D0004
13	N	*	CERUCAL	*	3.2022	D0002
14	*	*	FORTRANS	A06AD81	3.2022	D0005
15	*	*	FORTRANS	A06AD81	3.2022	D0001
16	N	*	CERUCAL	*	14.03.2022	D0005
17	*	*	FORTRANS	A06AD81	3.2022	D0004
18	*	*	FORTRANS	A06AD81	3.2022	D0005
19	N	*	LECROLYN	*	3.2022	D0003
20	*	*	ZANTAC 75	*	*	D0005
21	*	*	ZANTAC 75	*	*	D0005
22	M	*	LECROLYN	*	15.03.2022	D0004
23	M	*	KALMENTE	R01AD09	3.2022	D0005
24	M	*	KALMENTE	R01AD09	3.2022	D0003
25	N	*	LORAMYC	*	14.03.2022	D0002

Joonis 2: Anonüümimine *k*-anonüümsuse ja *l*-hajutuse baasil. Ülemisel pildil on andmed enne ja alumisel pärast anonüümimist.

Joonisel 2 on kujutatud andmekogu enne ja pärast anonüümimist, kui on tagatud *k*-anonüümsus = 3 ja *l*-hajutus = 3. Kvaasi-identifikaatoriteks on määratud väljad „gender“, „age“, „ATC“ ja „date“. Tundlikuks väljaks on määratud „doctor_registry_code“.

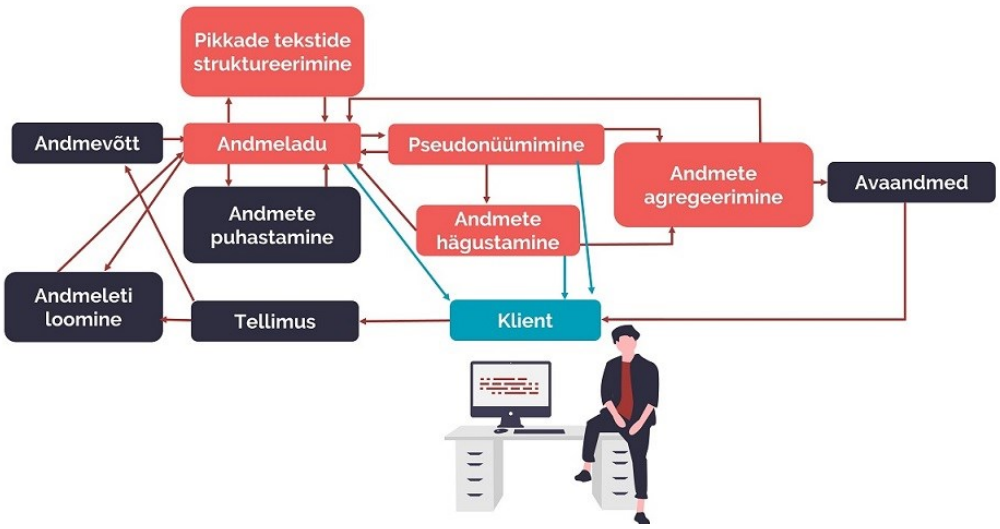
Vaatleme esiletõstetud ridasid 8, 11 ja 12. Enne anonüümimist on võimalik eristada kolme erinevat isikut. Pärast anonüümimist ei ole võimalik kvaasi-identifikaatorite alusel ridasid üksteisest eristada ehk on täidetud *k*-anonüümsus = 3. Kvaasi-identifikaatorite nelik (gender = N, age = *, ATC = R01AD09, date = 3.2022) moodustab ühe ekvivalentsiklassi. Selles ekvivalentsiklassis on kolm erinevat tundliku välja „doctor_registry_code“ väärtust: D0001, D0004, D0005. Seega on tagatud ka *l*-hajutus = 3.

Kui korrata sama analüüsi kõigi ridade kohta, on võimalik näha, et iga rida kuulub ühte ekvivalentsiklassi vähemalt 2 teise olemiga ning igas ekvivalentsiklassis on vähemalt 3 erinevat tundliku välja väärtust. Seega on terves anonüümitud andmekogus tagatud *k*-anonüümsus = 3 ja *l*-hajutus = 3.

2.5 Health Sense projekt

Health Sense on projekt, mida rahastab Norway Grants Programme „Green ICT“. Projekti eesmärgiks on luua keskkond, mille kaudu oleks võimalik väljastada Tervise ja Heaolu Infosüsteemide Keskuse (TEHIK) poolt hallatavate andmekogude andmeid anonüümitud kujul. Keskkonna arhitektuur on toodud joonisel 3.

Projektis osalevad mitmed organisatsioonid nii Eestist kui ka väljastpoolt. Nendeks on Sotsiaalministeerium, TEHIK, Tartu Ülikool, Tallinna Tehnikaülikool, Western Norway University of Applied Sciences ja Helse Wekt IKT.



Joonis 3: Health Sense arhitektuur. Käesolev töö käsitleb andmete hägustamise komponenti [9].

Projekt koosneb tervest hulgast komponentidest, mida valmistavad erinevad osapooled.

Pikkade tekstide struktureerija eesmärk on leida tekstidest sensitiivseid terviseandmeid ning tõsta need eraldi andmeväljadesse.

Andmeladu luuakse TEHIKu ning Tallinna Tehnikaülikooli koostööna. Ülikool koostab andmelao jaoks andmemudelid ning andmelao ja andmeanalüüsikeskkonna loob TEHIK.

Pseudonüümija koosneb kahest eraldisesisvast osast: mootor ja administreerimisliides. Pseudonüümija ülesanne on asendada sensitiivsed andmed, mida ei ole võimalik üldistada, pseudonüümidega.

Andmete hägustaja teostab Tartu Ülikool. Hägustaja eesmärk on anonüümida andmeid üldistamise baasil. Käesolev lõputöö on läbi viidud selle projekti raames.

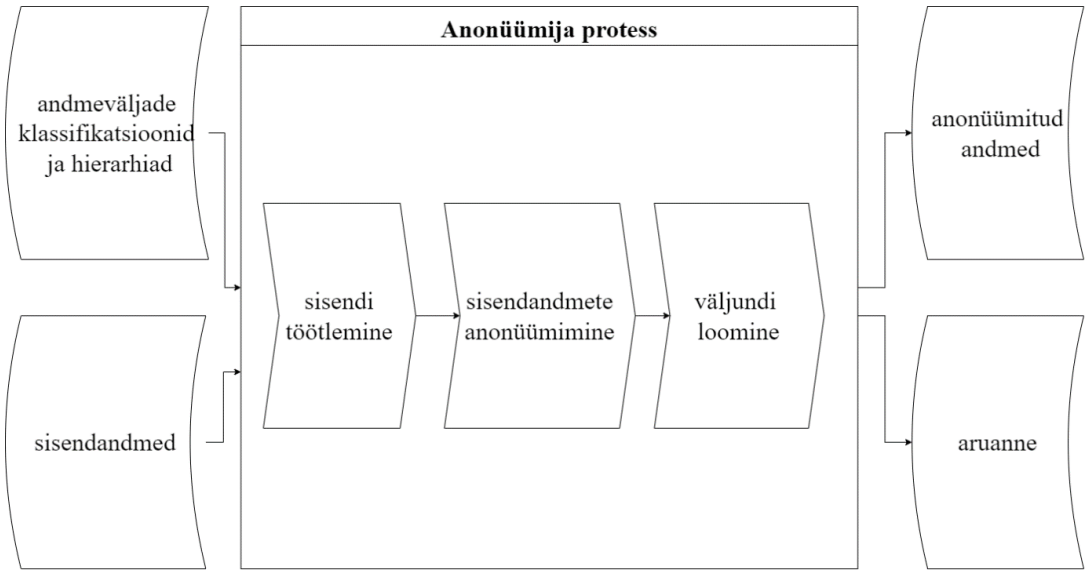
2.6 Anonüümimistarkvara arendus Health Sense projektis

Health Sense projekti raames arendab Tartu Ülikool andmete hägustajat ehk anonüümijat, mille ülesandeks on muuta Tervise ja Heaolu Infosüsteemide Keskuse (TEHIK) poolt väljastatavaid andmeid nii, et neid võib käsitleda avaandmetena. Vastavalt vajadusele

väljastab TEHIK andmeid Eesti avaandmete portaalis või otse teadlastele. Projekt kestab 2021-2023.

Projekti eesmärk ei ole välja arendada uut anonüümimise metoodikat. Samuti ei ole eesmärk taasluua või imiteerida juba olemasolevaid anonüümimistarkvarasid. Mõistlik on kasutada olemasolevaid tarkvarakomponente, kuid need tuleb kohandada terviklikuks ning Eesti terviseandmete ja TEHIKu jaoks sobivaks tarkvarakomplektiks.

Anonüümija protsessi skeem on toodud joonisel 4. See koosneb kolmest peamisest sammust: sisendi töötlemine, anonüümimine ning väljundi väljastamine koos aruandega. Aruandes on sisendi ja väljundi võrdlusnäitajad ning riskianalüüs. Käesolev töö keskendub anonüümimise sammule, kus töö alustamise hetkel oli kasutusel Mondriani anonüümsusalgoritm. Mondriani algoritm on üks kiiremaid anonüümimisalgoritme, kuid sellega tihtipeale üldistatakse andmeid rohkem kui teiste algoritmidega [14], mis ei lähe kokku Health Sense projekti eesmärkidega. Käesoleva töö eesmärgiks oli välja selgitada ning seejärel integreerida sobivaim olemasolev avatud lähtekoodiga anonüümimistarkvara.



Joonis 4: Health Sense anonüümija protsess

Anonüümimisel oli seatud eesmärgiks saavutada *k*-anonüümsus väärtusega 5 minimaalse infokaoga, välja arvatud juhud, kui andmete mahust tulenevalt sobib paremini väiksem *k*-anonüümsuse väärtus. Anonüümimismeetodiks valiti üldistamine hierarhiate baasil, kuna terviseandmetes on kasutusel mitmeid hierarhilisi klassifikatsioone.

Samuti oli seatud eesmärgiks tarkvara töösuutlikkus vigaste sisendandmetega, mis ei ole kaetud hierarhiatega. Näiteks võib vanuse üldistushierarhia defineerida tasemed kuni 120 eluaastani, kuid andmestikus võib esineda isik, kes on 122 aastat vana.

3. Metoodika

Käesolevas töös võrreldakse anonüümimistarkvarasid, mis on avatud lähtekoodiga ning ei nõua kasutamiseks tasulise litsentsi hankimist. Võrdlusest jäeti välja tarkvarad, mis ei toeta sisend- ja väljundandmete talletamist CSV-failis ega PostgreSQL andmebaasis, sest need on Health Sense skoobis kõige laialdasemalt kasutatavad tehnoloogiad [15].

Autor otsis Internetist erinevaid tarkvarasid, millega oleks võimalik anonüümida andmeid hierarhilise üldistamise baasil. Kasutati Google otsingumootorit ning otsingusõnu „anonymization tool”**.** Kuigi vasteid leiti rohkem, olid kõige populaarsemad järgnevad kolm anonüümimistarkvara: ARX [16] (versioon 3.9.0), Amnesia [17] (versioon 1.2.9) ja Anonimatron [18] (versioon 1.4). Nende tutvustus on toodud järgmises alapeatükis.

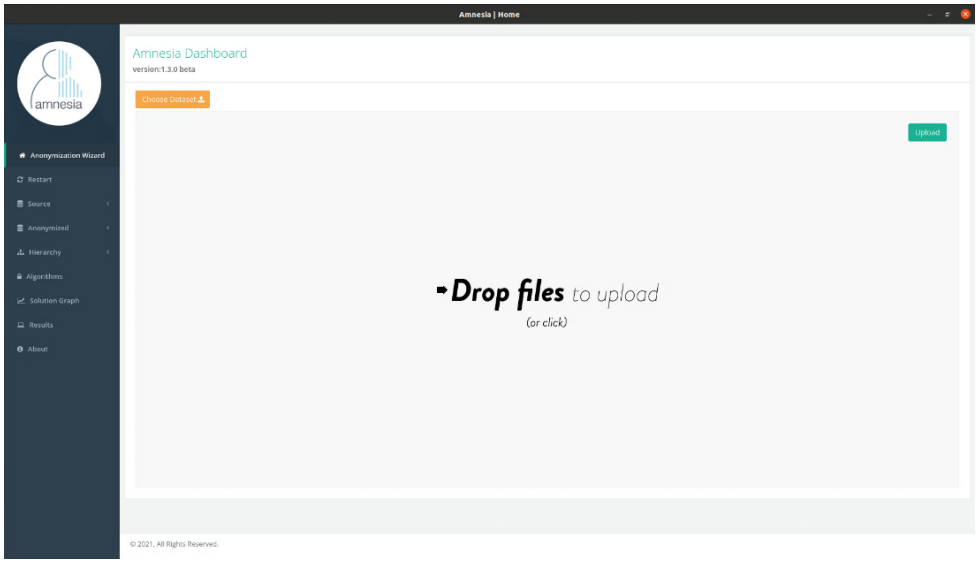
Seejärel hinnati iga tarkvara vastavalt kriteeriumitele, mida on kirjeldatud peatükis 3.2.1 „Tarkvarade hindamiskriteeriumid”. Võrdluse tulemusena loodud hindamismaatriksi põhjal valiti parimate tulemustega anonüümimistarkvara ning integreeriti see Health Sense projekti.

3.1 Võrreldavad tarkvarad

Selles alapeatükis antakse ülevaade kolmest vabavaralisest anonüümimistarkvarast, mida kaaluti integreerimiseks Health Sense projektis arendatavasse tarkvarakomplekti.

3.1.1 Amnesia

Amnesia on anonüümimistarkvara, mis sai alguse 2015. aastal Kreeka tarkvaraarendaja Dimitris Tsitsigkos poolt. Joonisel 5 on kujutatud selle tarkvara graafilise kasutajaliidese põhiakent.



Joonis 5: Amnesia põhiaken

Tarkvara loomist rahastas OpenAIRE projekt. Amnesia loojate eesmärgiks oli arendada terviseandmete jaoks anonüümimistarkvara, mis tagaks GDPRi nõuete täitmise. Anonüümimiseks on võimalik kasutada *k*-anonüümsusel põhinevaid anonüümsuse mudeleid.

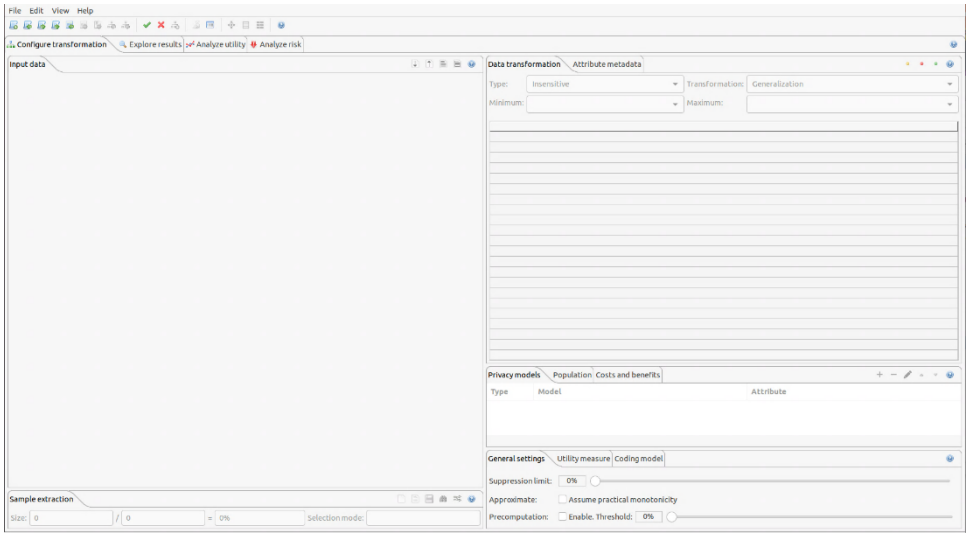
3.1.2 Anonimatron

Anonimatron on 2010. aastal alguse saanud vabavaraline projekt, mille arendus jätkub tänaseni. Anonimatron oskab anonüümida privaatseid andmeid või luua privaatsete andmete alusel sünteetilisi andmeid.

Tarkvaraarenduses on tihti vajalik testida uut funktsionaalsust võimalikult erinevate testandmetega, kuid sellel eesmärgil ei tohi kasutada päris isikuandmeid. Seetõttu tarkvaraarendajad sageli loovad käsitsi piiratud hulgal sünteetilisi andmeid, mis enamasti on palju pealiskaudsemad kui reaalsuses kasutatavad andmed. Ning see põhjustab defekte arendatavas tarkvaras, mille lahendamiseks tuleb kulutada täiendavat aega ja raha. Anonimatroni põhieesmärgiks ongi kvaliteetsete testandmete loomine eelmainitud viisidel.

3.1.3 ARX

ARX - Data Anonymization Tool on 2013. aastal alguse saanud vabavaraline tööriist, mida on enamasti arendanud Fabian Prasser ja Florian Kohlmayer. Joonisel 6 on kujutatud ARXi graafilise kasutajaliidese põhiakent.

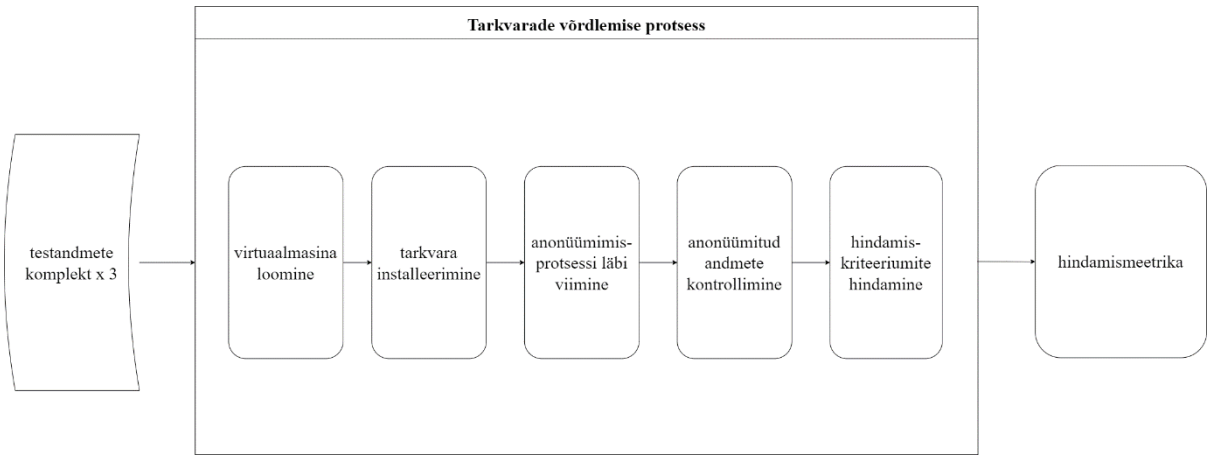


Joonis 6: ARXi põhiaken

ARX on hetkel üks kõige laialdasemate võimalustega vabavarasid, mis tegeleb andmete anonüümimisega. ARX pakub kõige suuremat valikut anonüümimismeetoditest, mis toetavad kvaasi-identifikaatorite ja sensitiivsete andmete anonüümimist.

3.2. Tarkvarade võrdlemise metoodika

Tarkvarade võrdlemise metoodika seab raamistiku vaadeldavate tarkvarade objektiivseks võrdlemiseks. Iga tarkvara hinnati vastavalt kriteeriumitele, mis on kirjeldatud järgmises alapeatükis. Iga tarkvara jaoks loodi virtuaalmasinad hindamiskriteeriumites loetletud operatsioonisüsteemidega ning viidi läbi anonüümimisprotsess nagu on näidatud joonisel 7.



Joonis 7: Tarkvarade võrdlemise protsess

Anonüümimisprotsessis kasutati kolme komplekti testandmeid. Igast testandmete komplektist oli omakorda kaks versiooni: korrektne ja vigane. Eesmärk oli võimalikult täpselt simuleerida andmeid, mis võivad esineda Health Sense projekti skoobis.

3.2.1 Tarkvarade hindamiskriteeriumid

Kõigi võrreldavate tarkvarade puhul võeti võrdluse aluseks ISO-9421 standard [19], mis kirjeldab inimese ja arvuti interaktsiooni põhimõtteid. Need moodustasid hindamiskriteeriumid 1 kuni 6. Kriteeriumite hulgast jäeti välja ISO-9421 standardi punkt „Sobivus kasutaja ülesannete täitmiseks", kuna tarkvarad valiti juba eelnevalt välja selle põhjal, et tegemist on anonüümimistarkvaraga. Kriteeriumid 7 kuni 10 hindasid täiendavaid Health Sense projekti jaoks olulisi omadusi, milleks olid integreeritavus, hierarhiate veakindlus, operatsioonisüsteemide tugi ja tarkvara tugi. Hindamiskriteeriumeid hinnati 0-10 punkti skaalal, kus 0 punkti oli halvim võimalik tulemus ning 10 punkti parim tulemus.

Hindamiskriteeriumid olid järgmised:

1. **Kasutajale kuvatava info asjakohasus.** Hinnang põhines sellel, kui hästi tegi tarkvara kasutajale selgeks, milliseid tegevusi on võimalik teha, ning kas eksisteeris abi juhul, kui kasutaja ei osanud protsessis edasi liikuda. Autor hindas järgmisi alapunkte:
    a. Kas tarkvaras on näha, milliseid faile see suudab töödelda?
    b. Kas tarkvaras on nupp „Abi", kust kasutajad saaksid lisainformatsiooni tarkvara kasutamise kohta?
    c. Kas tarkvara kuvab ajakulukate protsesside ajal kasutajale edenemisriba?

    Iga alapunkti eest oli võimalik saada 1/3 võimalikest punktidest.

2. **Kasutaja ootustele vastavus.** Hinnang moodustus selle põhjal, kas tarkvaraga oli võimalik sooritada tegevusi nagu tarkvara sulgemine ning tulemuste alla laadimine vastavalt levinud tavadele. Kriteeriumi hindamiseks vastas autor järgmistele küsimustele:
    a. Kas tarkvara on võimalik käivitada ühe nupuvajutusega?
    b. Kas tarkvara on võimalik sulgeda ühe nupuvajutusega?
    c. Kas anonüümitud andmeid on võimalik tarkvara vahendusel failina alla laadida?

Alapunktid a ja b andsid kumbki 1/5 võimalikest punktidest hindamiskriteeriumis ning alapunkt c andis 3/5 võimalikest punktidest.

3. **Õpitavus.** Tarkvara õpitavuse hindamiseks vaatles autor, kas tarkvara oli suuteline abistama kasutajat anonüümimisprotsessi käigus. Selleks hindas autor järgmisi punkte:
    a. Kas tarkvara pakub vaikeväärtusi seadistustes olevate väärtuste jaoks?
    b. Kas tarkvara selgitab, kuidas toetatud anonüümsuse mudelid töötavad?
    c. Kas tarkvara juhib anonüümimisprotsessi käigus kasutajat järgmise vajaliku sammu juurde? Näiteks anonüümimisprotsessi alustamise nupp muutub aktiivseks alles pärast faili üles laadimise lõppu vms.

    Iga alapunkt andis 1/3 võimalikest punktidest.

4. **Juhitavus.** Hinnati, kas oli võimalik pikaajalisi tegevusi katkestada ning vahetulemusi salvestada, et neid oleks võimalik hiljem taaskasutada. Hinnangu moodustamiseks vastas autor järgmistele küsimustele:
    a. Kas on võimalik katkestada anonüümimisprotsessi?
    b. Kas on võimalik salvestada vahetulemusi/seadistusi?

    Alapunktid andsid mõlemad 1/2 võimalikest punktidest.

5. **Veakindlus.** Veakindluse kontrollimiseks kasutas autor anonüümimisprotsessis ebakorrektseid seadistusi. Hinnati, kas tarkvara andis kasutajale vea olemuse kohta selgesti mõistetavat tagasisidet. Hinnang anti järgmiste küsimuste põhjal:
    a. Kas tarkvara keelab tühja faili anonüümimist?
    b. Kas tarkvara keelab anonüümimist ilma seadistuseta?
    c. Kas tarkvara kontrollib ja piirab seadistustesse sisestavaid väärtusi? Näiteks, kas tarkvara takistab arvulise väärtuse asemel sõne sisestamist.

    Alapunkt a andis 1/5 võimalikest punktidest hindamiskriteeriumis ning alapunktid b ja c andsid kumbki 2/5 võimalikest punktidest.
6. **Kasutaja kaasamine.** Antud kriteeriumi jaoks hindas autor, kui hästi suutis tarkvara edasi anda positiivseid töövooge, kus kõik sammud õnnestusid, ning kas kasutajat kaasati tarkvara edasisse arendusse. Selleks hindas autor järgnevaid punkte:
    a. Kas tarkvara teavitab kasutajat anonüümimisprotessi õnnestumisest?

b. Kas tarkvaras eksisteerib viis, mis võimaldab kasutajal kergesti edastada võimaliku tarkvaradefekti detaile?

Iga alapunkt andis 1/2 hindamiskriteeriumi võimalikest punktidest.

7. **Integreeritavus.** Integreeritavuse hinnangu eesmärgiks oli välja selgitada, kas on võimalik automatiseerida anonüümimisprotsessi käsurea skriptiga. Selleks hindas autor järgnevaid punkte:
    a. Kas tarkvara on võimalik välja kutsuda käsurealt?
    b. Kas tarkvara on võimalik seadistada käsurealt?
    c. Kas tarkvara on võimalik seadistada tekstifaili abil?
    d. Kas tarkvara võimaldab seadistada relatiivse või absoluutse failitee tulemuse salvestamiseks?

    Kõik alapunktid andsid 1/4 võimalikest punktidest.

8. **Operatsioonisüsteemide tugi.** Antud kriteerium on oluline, kuna Health Sense projektis ei ole seni määratud kindlaid operatsioonisüsteeme, mille toetamine on nõutud. Seetõttu valis autor ise mõned enamlevinud operatsioonisüsteemid ning kontrollis virtuaalmasinate abil, kas tarkvara ja selle olulisemad funktsionaalsused töötavad valitud operatsioonisüsteemidel ilma tõrgeteta. Seejärel vastas autor järgnevatele küsimustele:
    a. Kas tarkvara toetab operatsioonisüsteemi Windows 10?
    b. Kas tarkvara toetab operatsioonisüsteemi macOS 12?
    c. Kas tarkvara toetab GNU/Linux distributsiooni Ubuntu 20.04?

    Iga toetatud operatsioonisüsteem andis 1/3 võimalikest punktidest.
9. **Tarkvara tugi.** Selle kriteeriumiga hinnati tarkvara dokumentatsiooni kvaliteeti ning alternatiivsete abikanalite olemasolu. Selleks hindas autor järgmisi punkte:
    a. Kas tarkvaral on olemas dokumentatsioon?
    b. Kas tarkvara loojate poolt on loodud samm-sammulisi juhendeid tarkvara kasutamiseks?
    c. Kas eksisteerib foorum, kus kasutajal oleks võimalik abi küsida?
    d. Kas tarkvara veebisaidil või dokumentatsioonis on kirjeldatud milliseid anonüümsuse mudeleid tarkvara toetab?

Alapunktid a ja b andsid 2/5 võimalikest punktidest ning alapunktid c ja d andsid 1/10 võimalikest punktidest.

10. **Hierarhiate veakindlus.** Selles kriteeriumis hinnati, kuidas tarkvara käsitles olukordi, kui anonüümimist vajavate veergude jaoks pole lisatud üldistushierarhiat või etteantud hierarhiates on puuduvaid väärtusi. Autor hindas järgmisi punkte:
    a. Kas tarkvara suudab anonüümimist vajavale veerule automaatselt genereerida hierarhia, juhul kui see ei ole ette antud?
    b. Kas tarkvara pakub võimalust salvestada automaatselt genereeritud hierarhia hilisemaks taaskasutamiseks?
    c. Kas tarkvara suudab anonüümida, kui hierarhias on puuduvaid väärtusi?
    d. Kas tarkvara pakub võimalust automaatselt täiendada olemasolevat hierarhiat puuduvate väärtustega, mis esinesid andmestikus?

    Kõik alapunktid andsid 1/4 võimalikest punktidest.

Kriteeriumite põhjal moodustus hindamisraamistik, mida rakendati valitud tarkvarade võrdlemiseks ning parima valimiseks. Lisaks hinnati väljaspool raamistikku anonüümimise tulemuse korrektsust. Korrektsete sisendandmete puhul oli korrektne väljund rangelt nõutud ning tarkvara, mis seda nõuet rikkus, diskvalifitseeriti koheselt. Sel põhjusel ei antud tarkvaradele korrektse anonüümimise tulemuse eest eraldi punkte.

3.2.2 Anonüümimisprotsessi kirjeldus

Anonüümimisprotsess viidi läbi virtuaalmasinates, kuhu oli installeeritud eeltoodud operatsioonisüsteemid: Windows 10, macOS 12 ja GNU/Linux distributsioon Ubuntu 20.04. Võimaluse korral anti tarkvarale sisendandmed kasutades graafilist kasutajaliidest. Kui tarkvara seda võimalust ei pakkunud, anti sisend käsurea parameetrina. Anonüümitud andmed laaditi alla tarkvara vahendusel CSV-failina. Juhul, kui tarkvara toetas ainult PostgreSQL andmebaasi, siis anonüümitud andmed eksporditi andmebaasist CSV-faili. Anonüümimise tulemusi kontrolliti *k*-anonüümsuse mõõdiku abil, kui *k*-anonüümsus = 5. See tähendab, et tulemus oli korrektne, kui iga anonüümitud olemit ei olnud võimalik eristada 4 muust anonüümitud olemist. Tarkvara väljundi korrektsust kontrolliti käsitsi.

3.2.3. Testandmete kirjeldus

TEHIKu andmetes võib esineda vigasid, mis tekkisid inimliku eksituse tõttu andmete sisestamisel. Seetõttu oli anonüümimistarkvara kasutuskõlblikkuse hindamiseks vajalik kontrollida, kas tarkvara suudab anonüümida andmestikku, mille kõik väärtused pole kaetud üldistushierarhiate poolt. Hierarhiate veakindluse kriteeriumi hindamiseks koostati 3 komplekti testandmeid kahes versioonis:

1. korrektsed – kõik andmed järgivad üldistushierarhiaid;
2. vigased – andmetes esineb väärtuseid, mis ei ole kaetud üldistushierarhiaga.

Igas testandmete komplektis esineb üks või mitu andmevälja, mida peab kindlasti anonüümima, et tagada nõutud *k*-anonüümsus. Nendeks väljadeks on RHK-10, ATC, EHAK ja kuupäevad. Toodud väljade anonüümimine on vajalik, sest nende põhjal on kõige lihtsam viia läbi privaatsuse ründeid, kui ründaja on teadlik kindla isiku privaatelu aspektidest.

3.3 Tarkvara Health Sense projekti integreerimise metoodika

Health Sense projekti andmete anonüümija protsess on kujutatud joonisel 4. Käesolev töö käsitleb eelkõige „Sisendandmete anonüümimise" sammu, kus toimub andmete anonüümimine. Varasemalt kasutati projektis anonüümimiseks Mondriani algoritmi implementatsiooni. Antud töö tulemusena asendati Mondriani algoritm võimekama vabavaralise anonüümimistarkvaraga.

Sobivaima anonüümimistarkvara leidmiseks kasutati peatükis 3.2.1 „Tarkvarade hindamiskriteeriumid" kirjeldatud mõõdupuud. Pärast tarkvarade analüüsi valiti parim tarkvara ning integreeriti see Health Sense tarkvarakomplekti nii, et lahendus vastaks järgmistele nõutele:

1. lahendus sobib Health Sense projekti töövoogu;
2. integratsiooni on vajadusel lihtne asendada teistsuguse lahendusega;
3. integreeritud tarkvara on lihtne seadistada;
4. integratsiooni on lihtne tulevikus edasi arendada.

Autor kasutas lahenduse lähtekoodi talletamiseks hoidlat GitHubi keskkonnas. Lahendus koosneb Health Sense projekti integreerimiseks sobivast käsureaprogrammist ja kasutusjuhendist.

4. Tulemused ja arutelu

Järgnevates alapeatükkides kirjeldatakse tarkvarade võrdluse ja valitud tarkvara integratsiooni tulemusi. Samuti analüüsitakse võrdlusmetoodikat ning arutletakse, mida oleks saanud teha paremini.

4.1. Tarkvarade võrdluse tulemused

Käesoleva töö käigus analüüsiti ning võrreldi kolme anonüümimistarkvara: Amnesia, Anonimatron ja ARX. Võrdlus põhineb 10 kriteeriumil. Iga kriteeriumi hinnati skaalal 0-10 punkti. Kõik hinnangud on ümardatud täpsusega üks koht pärast koma. Lisaks kontrolliti anonüümimise tulemuste korrektsust, mis oli rangelt nõutud.

Tabel 1: Analüüsi tulemus punktidena

	Amnesia	**Anonimatron**	**ARX**
Korrektsed anonüümimise tulemused	✓	✓	✓
Kasutajale kuvatava info asjakohasus	6,7	3,3	10,0
Kasutaja ootustele vastavus	8,0	4,0	10,0
Õpitavus	6,7	3,3	6,7
Juhitavus	5,0	10,0	4,0
Veakindlus	6,0	4,0	10,0
Kasutaja kaasamine	0,0	5,0	0,0
Integreeritavus	7,5	7,5	10,0
Operatsioonisüsteemide tugi	6,7	10,0	10,0
Tarkvara tugi	8,0	5,0	9,0
Hierarhiate veakindlus	5,0	0,0	5,0
Summa	**59,6**	**52,1**	**74,7**

Kõik valitud tarkvarad anonüümisid testandmeid korrektselt ning neist ühtegi ei olnud vaja diskvalifitseerida.

Kõige asjakohasemat infot näitas kasutajale ARX, kuna vastas kõigile hindamiskriteeriumi alapunktidele ning sai seetõttu maksimaalse tulemuse. Amnesia kaotas punkte, kuna anonüümimise ajal ei näidatud edenemisriba. Kõige vähem punkte sai Anonimatron, sest tarkvaral puudub kasutajaliides, mistõttu ei olnud võimalik kuvada uuele kasutajale asjakohast infot. Anonimatron sai mõned punktid, sest tarkvaraga on kaasas tekstifail viidetega dokumentatsioonile ja teistele tugiressurssidele.

Kasutaja ootustele vastas kõige paremini ARX, mis jällegi vastas kõigile alapunktidele. Amnesia kaotas punkte, kuna seda polnud võimalik käivitada ühe nupuvajutusega, vaid pidi jooksutama tarkvara autorite poolt loodud skripti. Kõige vähem punkte sai Anonimatron, sest tarkvara käivitamiseks oli samuti vajalik skripti kasutamine ning anonüümimisprotsess jooksis koheselt algusest lõpuni ilma kasutaja sisendit ootamata. Anonimatroniga saab anonüümitud andmeid faili eksportida ainult andmebaasi vahendusel.

Õpitavuse kriteeriumist said ARX ja Amnesia võrdse arvu punkte. ARX ei tõsta esile anonüümimisprotsessi järgmisi samme, mistõttu võib kasutajal olla raske aru saada, mis peaks olema tema järgmine tegevus. Küll aga lihtsustab ARX seadistamist, sest piirab võimalike väärtuste lubatud vahemikke. Amnesia ei pakkunud seadistuste jaoks vaikeväärtuseid ega vahemikke, kuid tegi väga selgeks anonüümimisprotsessi sammud. Anonimatron ei pakkunud seadistuste jaoks vaikeväärtusi ega tõstnud anonüümimisprotsessi käigus järgmisi samme esile.

Parima juhitavusega tarkvaradeks osutusid ARX ja Anonimatron, sest mõlemas tarkvaras oli võimalik salvestada seadistusi ja katkestada anonüümimisprotsessi. Anonimatroni anonüümimisprotsess eeldab, et kõik vajalikud seadistused on tehtud konfiguratsioonifailis, mistõttu võib väita, et on võimalik seadistusi salvestada. Amnesia ei toetanud anonüümimisprotsessi katkestamist ning kaotas selle eest punkte.

Kõige veakindlam tarkvara oli ARX. Vigase faili korral tagastas ARX selge veateate ning seadistamisel piiras ebakorrektsete väärtuste sisestamist. Amnesia kaotas punkte, sest tühja seadistuse puhul kuvati küll veateade, kuid selle sisu ei sisaldanud tavakasutajale arusaadavat informatsiooni, mis aitaks viga lahendada. Anonimatron kaotas punkte, kuna tühja faili anonüümimisel ei tagastanud tarkvara veateadet ning tühja seadistuse puhul ei sisaldanud veateade vea lahendamiseks vajalikku informatsiooni. Ühtlasi ei kuvatud viga koheselt Anonimatroni käivitamisel, vaid alles anonüümimisprotsessi ajal, mis ajab kasutajat veelgi rohkem segadusse.

Kõige rohkem kasutajat kaasavaks tarkvaraks osutus Anonimatron. Anonimatron oli ainus, mis andis kasutajale selgelt teada, et anonüümimisprotsess lõpetati edukalt. Kuid Anonimatron ei paku lihtsasti kasutatavat võimalust vearaportite saatmiseks ja kaotas seetõttu punkte. Amnesia ei teavitanud kasutajat anonüümimise lõppemisest ega pakkunud võimalust tekkinud probleemide raporteerimiseks. Seetõttu kaotas Amnesia selles kriteeriumis kõik punktid. Samal põhjusel kaotas ka ARX kõik punktid. Kõigis tarkvarades on viide vastava tarkvara GitHub hoidlale, kuid ainult Anonimatron juhendas kasutajat, kuidas täpsemalt käsitsi viga raporteerida. Siiski ei olnud see autori arvates piisav täiendavate punktide teenimiseks.

Integreeritavuse poolest sai parima tulemuse ARX mõningate mööndustega. ARXi autorid toetasid varem käsureaversiooni, kuid tänapäeval on see asendatud Java teegiga. See tähendab, et vaikimisi pole käsurea kasutamine enam toetatud. Kuid autor otsustas siiski anda ARXile maksimaalsed punktid, kuna mõningase lisatööga on võimalik Java teegi põhjal käsurea kasutamise tugi taastada. Amnesia kaotas punkte, sest anonüümimisprotsess jookseb veebiserveri vahendusel ning seadistust on võimalik määrata ainult veebiserverisse päringut tehes, mitte tekstifaili abil. Anonimatron kaotas punkte, kuna väljundit ei ole võimalik suunata tekstifaili, vaid on nõutud andmebaasi kasutamine.

Parim operatsioonisüsteemide tugi oli Anonimatronil ja ARXil. Anonimatron töötas igal operatsioonisüsteemil, mille jaoks on olemas piisavalt uus Java virtuaalmasin (JVM), sealhulgas Windows, macOS ja GNU/Linux. ARXi jaoks oli võimalik alla laadida paigaldajad Windowsi, macOSi ning GNU/Linuxi jaoks. Amnesia kaotas punkte, kuna puudus macOS tugi.

Kõige paremini toetatud tarkvara oli ARX, millel oli laiaulatuslik dokumentatsioon tarkvara graafilise liidese, käsurealt väljakutsutava versiooni ning ka Java programmeerimiskeeles kirjutatud teegi jaoks. Amnesia kaotas punkte, kuna dokumentatsioonis ei selgitatud, kuidas toimub tarkvaras andmete anonüümimine. Anonimatroni dokumentatsioon käsitles rohkem probleemi olemust ning tarkvara kasutamise kohta oli oluliselt vähem informatsiooni, kui teistel vaadeldud tarkvaradel. Kõik tarkvarad kasutasid GitHub versioonihalduskeskkonna funktsionaalsusi foorumina, kus kasutajad saaksid küsida abi nii tarkvara loojatelt kui ka teistelt kasutajatelt.

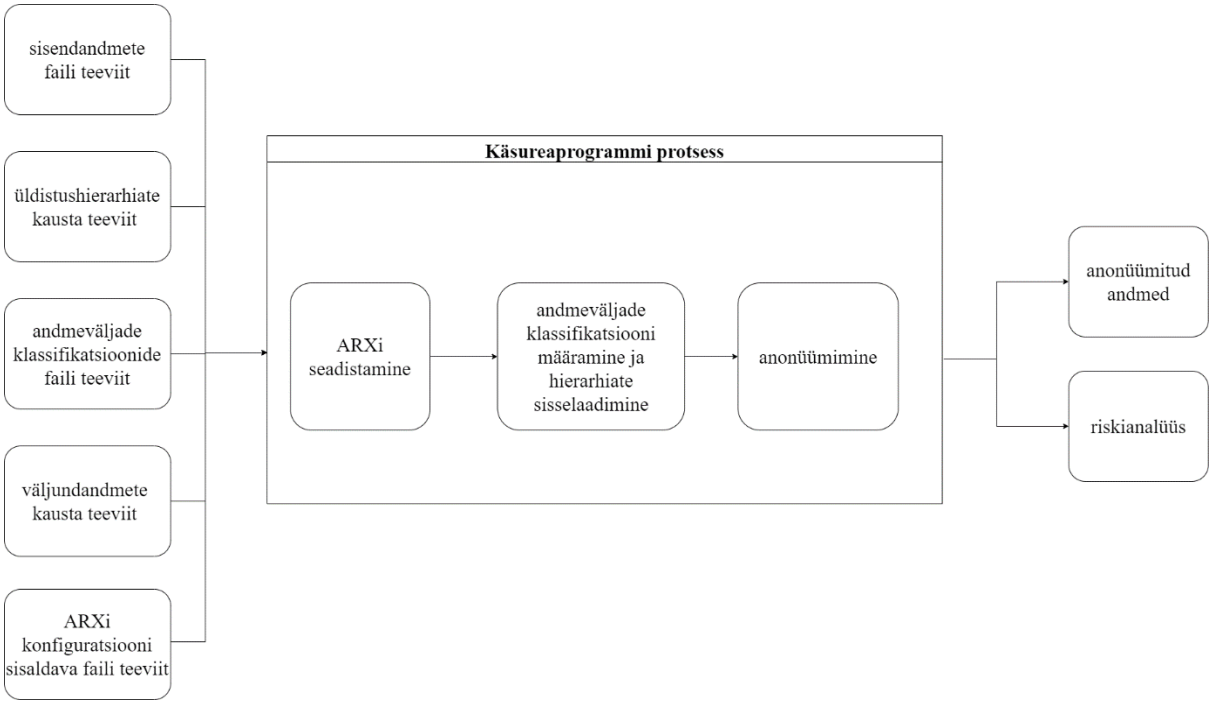
Hierarhiate toetus oli kõige parem ARXil ja Amnesial. Siiski esines mõlemal tarkvaral puudusi olukordades, kus hierarhia oli puudu või andmetes esines hierarhiast puuduvaid

väärtusi. Anonimatroni tugi hierarhiatele oli pealiskaudne ja seetõttu kaotas kõik punktid hindamiskriteeriumis.

Tabelis 1 on näidatud, et kõige kõrgema koondtulemuse sai ARX 74,7 punktiga. Järgnes Amnesia, millel oli 59,6 punkti, ning madalaima tulemuse sai Anonimatron 52,1 punktiga. ARXi palju kõrgemast koondtulemusest hoolimata tuleb nentida, et mõnedes hindamiskriteeriumites saavutasid nii Amnesia kui ka Anonimatron paremaid tulemusi kui ARX.

4.2. ARX integreerimine Health Sense projekti tarkvarakomplekti

Võrdluse tulemuste põhjal oli Health Sense projekti integreerimiseks sobivaim tarkvara ARX. Tarkvarade võrdlemise käigus oli selgunud, et ARX talitlusloogika on saadaval ka eraldiseisva Java teegina, ARX Java Library, mis on kasulik, sest võimaldab tulevikus integratsiooni lihtsamini edasi arendada vastavalt kasutajate vajadustele. ARX integreerimist alustatigi Health Sense projektis kasutamiseks kohandatud käsureaprogrammi loomisest. Programmi sisemine protsess on toodud joonisel 8.

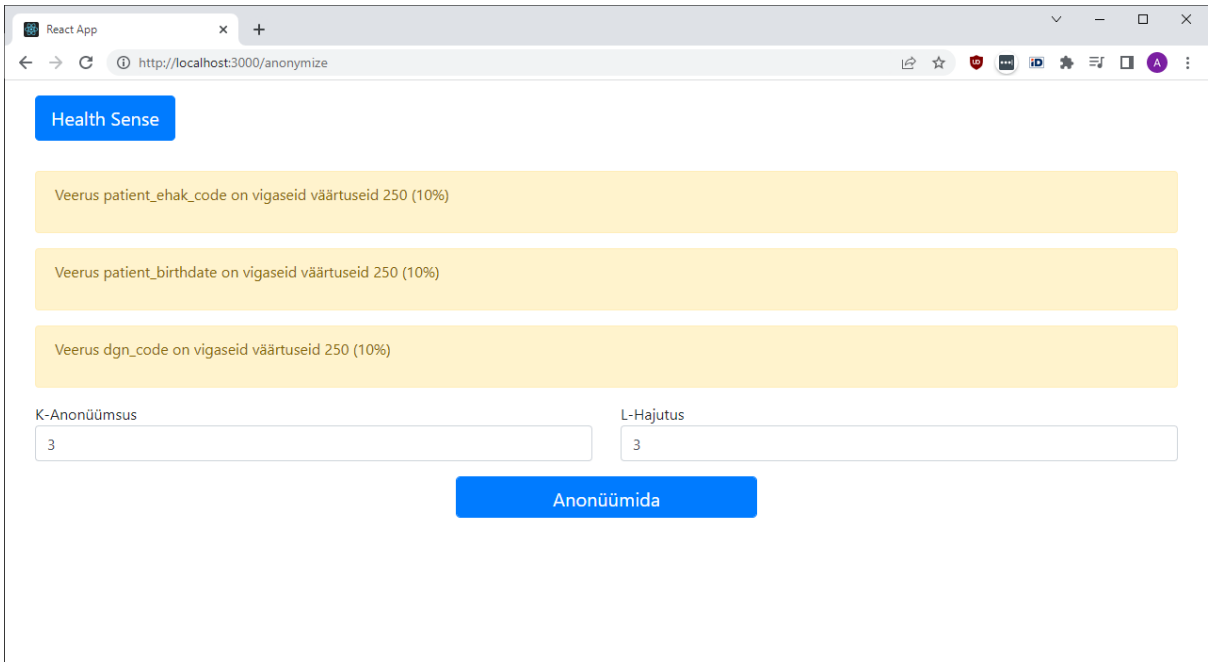


Joonis 8: Integreeritud käsureaprogrammi protsess

Programmi kirjutamiseks kasutati Java programmeerimiskeele versiooni 13. Loodud programmil on 5 kohustuslikku sisendparameetrit:

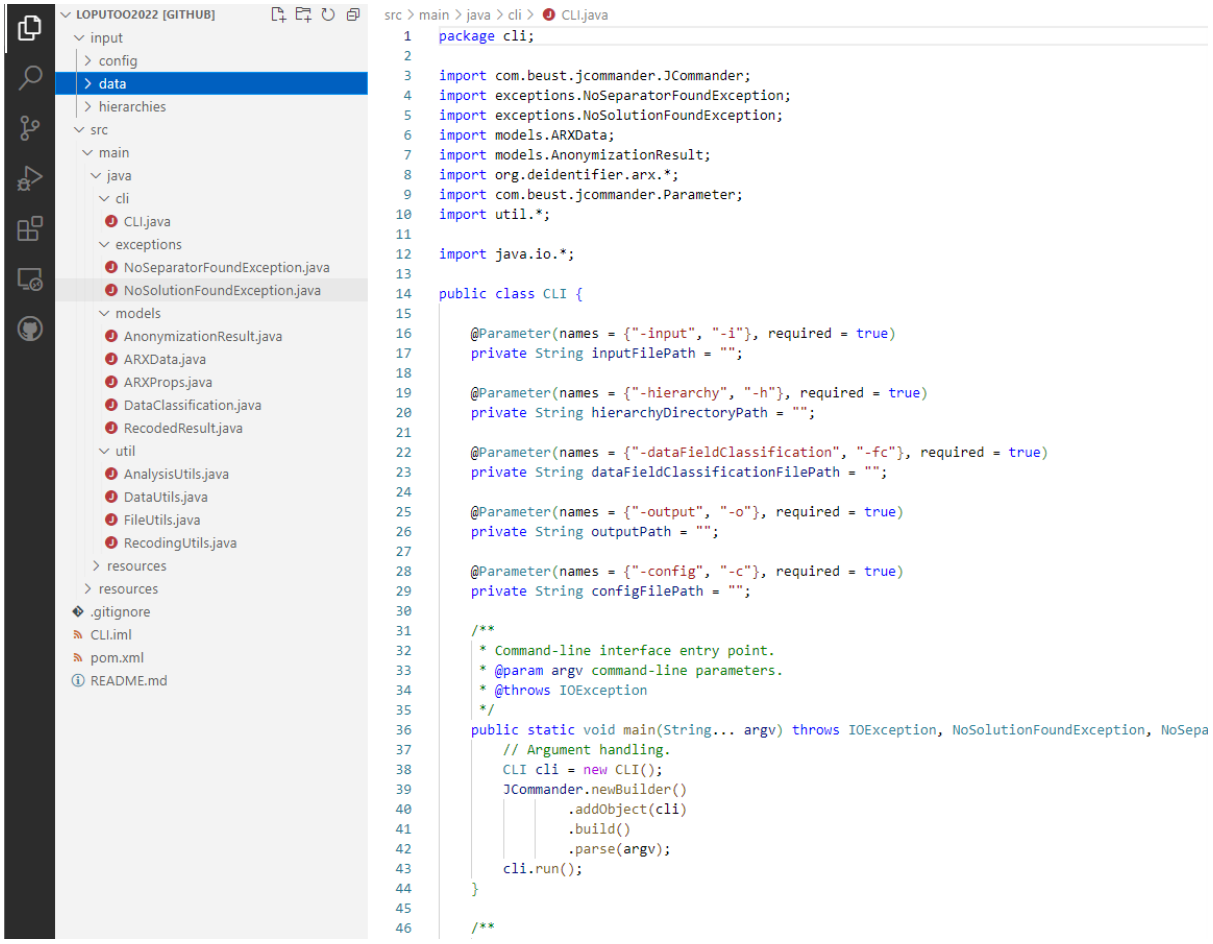
1. sisendandmete faili teeviit;
2. üldistushierarhiate kausta teeviit;
3. andmeväljade klassifikatsioonide faili teeviit;
4. väljundandmete kausta teeviit;
5. ARXi konfiguratsiooni sisaldava faili teeviit.

Programmi väljundiks on anonüümitud andmed ning võimalike privaatsust rikkuvate rünnete riskianalüüs. Integratsiooni testimiseks läbiti anonüümimisprotsess Health Sense tarkvarakomplekti kasutajaliideses, mis on toodud joonisel 9. Anonüümimistulemusi võrreldi tulemustega, mis saadi samade andmete anonüümimisel läbi ARXi graafilise kasutajaliidese.



Joonis 9: Health Sense kasutajaliides, kuhu on integreeritud käsureaprogramm

Lõputöö käigus loodud komponendi lähtekoodi on võimalik vaadelda Lisas I olevalt viitelt. Hoidlas on testandmed ning juhend, kuidas jooksutada loodud käsureaprogrammi. Joonisel 10 on toodud näide käsureaprogrammi lähtekoodist.



Joonis 10: Näide integreeritud käsureaprogrammi lähtekoodist

Loodud komponendi arendust jätkatakse Health Sense projekti raames ka väljaspool antud lõputööd. Esimesena on plaanis täiendada seadistusfaile, et täielikumalt ära kasutada ARXi võimalusi.

4.3. Võrdlusmetoodika analüüs

Lõputöö vältel oli kõige raskem anonüümimisega seotud domeeniteadmiste hankimine. Kuna anonüümimisel pole laialdaselt kasutusel olevaid häid tavasid, siis oli üllatavalt raske aru saada, mis on anonüümsuse mudel ning kuidas see seondub anonüümimisalgoritmiga. Ühtlasi ei ole kehtestatud kindlaid anonüümsuse mõõdikute väärtusi, mille korral saab kindlalt öelda, et andmed on anonüümsed. Seetõttu osutus sobivate väärtuste leidmine samuti keeruliseks. Teaduslike uuringute korral on näited tehtud 10-20 realiste näidisandmetega, mistõttu on raske öelda, millised anonüümsuse mõõdikute väärtused sobivad näiteks miljoni realise andmefaili korral, mis võivad esineda Health Sense projekti skoobis.

Tarkvarade võrdlusel tekkis probleeme, kuna Anonimatronil puudub graafiline kasutajaliides. Seetõttu oli vaja ISO-9421 standardis kirjeldatud kriteeriumeid käsitleda paindlikumalt, kui esialgu oli plaanitud. Graafiliste kasutajaliideste puhul on tarkvara sulgemise tüüplahenduseks ristiga nupp akna ülemises paremas nurgas. Anonimatroni puhul otsustati see võrdsustada CTRL + C klahvikombinatsiooniga käsureal, et teistel tarkvaradel ei oleks põhjendamatut eelist.

Ülejäänud kriteeriumite puhul ei olnud vajalik selliseid erisusi sisse viia, sest need kirjeldasid soovitud funktsionaalsust üldsõnalisemalt. Näiteks küsimusele „Kas anonüümitud andmeid on võimalik tarkvara vahendusel failina alla laadida?" positiivse vastuse andmiseks sobib graafilises liideses vastav nupp ning käsureaprogrammis viip, mis võimaldaks kasutajal valida, kuhu soovitakse tulemuste fail salvestada.

Tarkvarade võrdlemiseks ei olnud võimalik leida hindamisraamistikku, mis oleks sobinud käesoleva töö eesmärkidega. Esialgu leiti ISO-9421 standard tarkvarade hindamiseks, kuid see ei katnud kõiki Health Sense projekti jaoks olulisi omadusi. Selle tõttu lisati täiendavad hindamiskriteeriumid vastavate täpsustavate küsimustega.

Koostatud hindamisraamistikku on võimalik kasutada laiemalt erinevate tarkvarade võrdluse alguspunktina, kuna hinnatakse baasomadusi, mis võiks olla igal tarkvaral. Vastavalt olukorrale on siiski soovitatav raamistikku täiendada valdkonnapõhiste hindamiskriteeriumitega, et saavutada asjakohasem hinnang.

Kokkuvõte

Käesoleva lõputöö eesmärgiks oli süstemaatiliselt analüüsida üldistamisel põhinevaid anonüümimistarkvarasid ning integreerida parimate tulemustega tarkvara Health Sense projektis arendatavasse tarkvarakomplekti. Selleks koostati hindamisraamistik, mille alusel hinnati tarkvarade erinevaid omadusi. Hinnangud anti töö jaoks loodud testandmete anonüümimise abil.

Analüüsi tulemusena selgus, et töös võrreldud tarkvaradest sai hindamiskriteeriumite põhjal kõige paremad tulemused ARX. ARX on tarkvara, millega on võimalik andmestikke anonüümida ning koostada andmete jaoks riskianalüüsi. Antud töös väljendus see mugavas anonüümimisprotsessis, kus ARX pakkus lihtsasti mõistetavat tagasisidet ning lisainformatsiooni anonüümimisega seotud teemade kohta.

Paremuselt järgnes Amnesia tarkvara. Amnesia pakkus selget struktuuri anonüümimisprotsessis sooritatavate tegevuste jaoks ning pakkus veebiserveri vahendusel integreerimise võimalust. Amnesial esines puudusi tarkvara dokumentatsioonis ning lõppkasutaja kaasamisega tarkvara arendusse.

Kõige vähem punkte sai Anonimatron. Anonimatronil puudus graafiline kasutajaliides, mis tegi asjakohase informatsiooni kuvamise oluliselt raskemaks kui teistel tarkvaradel. Anonimatron saavutas kõrgeid tulemusi juhitavuse, kasutaja kaasamise ning operatsioonisüsteemide toe hindamiskriteeriumites. Anonimatroni veateated ei olnud informatiivsed ning ka dokumentatsioon oli liialt pealiskaudne.

ARXi integreerimisel Health Sense tarkvarakomplekti kasutati Java programmeerimiskeele versiooni 13, millega loodi käsureaprogramm kasutades ARXi loojate poolt pakutud teeki.

Töö käigus loodud hindamisraamistikku saab kasutada ka teiste samalaadsete võrdluste läbi viimiseks. Töö mahust tulenevalt on töös analüüsitud ainult kolme avatud lähtekoodiga anonüümimistarkvara. Kindlasti võiks viia läbi samalaadse analüüsi ka teiste tasuta ning tasuliste vahendite jaoks. Seeläbi saaks parema ülevaate turul leiduvatest anonüümimistarkvaradest ning kasutajatel oleks lihtsam valida sobivat tööriista oma probleemi

lahendamiseks.

Viidatud kirjandus

[1] "Andmekaitse Inspektsioon (AKI) | Isikuandmed." https://www.aki.ee/et/eraelu-kaitse/isikuandmed. (05.05.2022).

[2] "Andmekaitse Inspektsioon (AKI) | Isikuandmete liigitus." https://www.aki.ee/et/eraelu-kaitse/isikuandmed-ja-tootlemine/isikuandmete-liigitus. (05.05.2022).

[3] "Andmekaitse ja infoturbe leksikon | Anonüümimine." https://akit.cyber.ee/term/630-anonuumimine. (05.05.2022).

[4] "Eesti avaandmete portaal." https://avaandmed.eesti.ee/. (05.05.2022).

[5] "Mis on avaandmed?" https://data.europa.eu/et/trening/what-open-data. (05.05.2022).

[6] "General Data Protection Regulation." https://eur-lex.europa.eu/eli/reg/2016/679/oj. (05.05.2022).

[7] 2018 reform of EU data protection rules. European Comissioion. May 25, 2018. URL: https://european-union.europa.eu/privacy-policy_et. (05.05.2022)

[8] “Meta says it may shut down Facebook and Instagram in Europe over data-sharing dispute”, Feb 7, 2022. URL: https://www.cnbc.com/2022/02/07/meta-threatens-to-shut-down-facebook-and-instagram-in-europe.html. (05.05.2022)

[9] „Health Sense“ https://tehik.ee/projektid. (10.05.2022).

[10] Madan, S. and Goswami, D. P. (2018) ‘An Extensive Study on Statistical Data Anonymization Algorithms’, 2018 3rd International Conference and Workshops on Recent Advances and Innovations in Engineering (ICRAIE), Recent Advances and Innovations in Engineering (ICRAIE), 2018 3rd International Conference and Workshops on, pp. 1–5. doi: 10.1109/ICRAIE.2018.8710436.

[11] Mahanan, W., Chaovalitwongse, W. A. and Natwichai, J. (2021) ‘Data privacy preservation algorithm with k-anonymity’, WORLD WIDE WEB-INTERNET AND WEB INFORMATION SYSTEMS. doi: 10.1007/s11280-021-00922-2.

[12] Samarati, P. and Sweeney, L. (1998) ‘Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression’. doi: 10.1184/r1/6625469.

[13] Machanavajjhala, A. et al. (2006) ‘L-diversity: privacy beyond k-anonymity’, 22nd International Conference on Data Engineering (ICDE’06), Data Engineering, 2006. ICDE ’06. Proceedings of the 22nd International Conference on, p. 24. doi: 10.1109/ICDE.2006.1.

[14] Ayala-Rivera, V. (1) et al. (no date) ‘A Systematic comparison and evaluation of k-Anonymization algorithms for practitioners’, Transactions on Data Privacy, 7(3), pp. 337–370. (Accessed: 9 May 2022).

[15] PostgreSQL. https://www.postgresql.org/ (08.05.2022)

[16] Fabian Prasser, Florian Kohlmayer, “ARX – Data Anonymization Tool.” https://arx.deidentifier.org/. (05.05.2022).

[17] Dimitris Tsitsigkos “Amnesia.” https://amnesia.openaire.eu/. (05.05.2022).

[18] “Anonimatron.” https://realrolfje.github.io/anonimatron/. (05.05.2022).

[19] ISO 9241-11, “Ergonomic requirements for office work with visual display terminals (VDT)s- Part 11 Guidance on usability,” 1998.

Lisad

I. GitHubi repositoorium

Integreeritud käsureaprogrammi lähtekood asub viitel
https://github.com/allanalikas/Loputoo2022

II. Litsents

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Allan Alikas,

- annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose "Privaatsust tagavad anonüümimistarkvarad", mille juhendajaks on Sulev Reisberg, reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.
- Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commonsi 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.
- Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.
- Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Allan Alikas

10.05.2022