Euroopa Liit
Euroopa Sotsiaalfond

Eesti tuleviku heaks

Uurimisprojekti „Andmeaitade (teiseste andmekogude) loomise põhimõtete väljatöötamine“ lõpparuanne

Uuringu tellija: Riigikantselei, tarkade otsuste fond

Uuringu partnerid: Riigi Infosüsteemi Amet ja Sotsiaalministeerium

TTÜ Küberneetika Instituut

Vahur Kotkas
Hele-Mai Haav

Tallinna Tehnikaülikool

Jaak Tepandi
Enn Õunapuu

Data Capital OÜ

Jaanus Grauberg

TTÜ Küberneetika Instituut

Tallinn 2013

Soovitav viide: V. Kotkas, H-M. Haav, J. Tepandi, E. Õunapuu, J. Grauberg, Uurimisprojekti "Andmeaitade (teiseste andmekogude) loomise põhimõtete väljatöötamine" lõpparuanne, TTÜ Küberneetika Instituut, 2013

Tellija: Riigikantselei ja Riigikantselei tarkade otsuste fond
Partnerid: Riigi Infosüsteemide Amet ja Sotsiaalministeerium

ISBN: 978-9949-430-65-9 (võrguväljaanne, pdf)

TTÜ Küberneetika Instituut
Akadeemia tee 21
12618 Tallinn
Tel. 6204150
Fax 6204151
E-post dir@ioc.ee
www.ioc.ee

Käesoleva töö valmimisse on andnud olulise panuse kaTellijapoolne töörühm, kuhu kuulusid

Ahto Kalja, Uuno Vallner, Anti Urm, Heiko Vainsalu, Rein Murakas, Siim Sikkut ja Kaisa-Maarja Jagula.

Uuringut rahastasid Euroopa Liidu

Sotsiaalfond ja Riigikantselei läbi Riigikantselei tarkade

otsuste fondi.

SISUKORD

1 SISSEJUHATUS 4

2 PROBLEEMI KIRJELDUS JA UURINGU EESMÄRK 6

3 METOODIKA 9

3.1. METOODIKA ÜLEVAADE 9

3.1 INTERVJUUDE JA ANKEETKÜSITLUSTE LÄBIVIIMINE 11

4 INTERVJUUDE JA ANKEETIDE ANALÜÜSI TULEMUSED 17

4.1 KOKKULEPPED (SEADUSED, STANDARDID, LEPINGUD) 17

4.2 KINDLUS (INFOTURVE, USALDUSVÄÄRSUS, IDENTIFITSEERIMINE) 26

4.3 MÕTTEVIIS (TEADMISED, MOTIVATSIOON, KOOSTÖÖ) 32

4.4 VÕIMALDAJAD (IT TARISTU, KVALIFITSEERITUD INIMRESSURSID, OLEMASOLEVAD ANDMEAIDAD) 38

4.5 LIIKUMAPANEVAD JÕUD (VÕIMALUSED JA OHUD, TEHNOLOOGIA, KONKURENTS, ÜHISKOND) 41

5 TEABEMATERJALIDE ANALÜÜSI TULEMUSED 50

5.1 TEADUS- JA TEHNOLOOGIATRENDIDE ANALÜÜS 50

5.2 RAHVUSVAHELISE PRAKTIKA ANALÜÜS 53

6 PAKUTAVAD LAHENDUSED JA ETTEPANEKUD 68

6.1 ANDMEAITADE SEADUSANDLUS JA STANDARDID 70

6.2 ANDMELAONDUSE ARENDAMISE ORGANISATSIOONILISED ASPEKTID 82

6.3 ANDMEAITADE KASUTAMINE JA TEHNOLOOGIA 85

7 KOKKUVÕTE 93

8 LÜHENDID JA MATERJALIDE LOETELU 99

8.1 LÜHENDID JA SÕNASELETUSED 99

8.2 KASUTATUD MATERJALID 99

9 LISAD 105

9.1 INTERVJUUDE ANALÜÜS 105

9.2 ANKETEERIMISE TULEMUSED 105

9.3 ANKEETIDE KÜSIMUSTIK SIHTRÜHMADE LÕIKES 105

9.4 VAHESEMINARI ETTEKANNE 12.12.2012 105

9.5 LÕPPSEMINARI ETTEKANNE 11.04.2013 105

1 SISSEJUHATUS

Andmeait on kindlale valdkonnale (või probleemile) orienteeritud, teisene, integreeritud, ajast sõltuv, püsiv või loogiliselt integreeritud andmekogum, mille eesmärgiks on toetada otsuste tegemist. Andmeaidad koostatakse juba olemasolevatest andmekogudest spetsiifiliste andmetöötlusülesannete ja aruannete koostamiseks. Tänapäeval kasutatakse üha enam ka laiemat määratlust, mille kohaselt andmeait on andmete kasutamise meetodite, tehnoloogiate ja praktikate kompleks, mille eesmärk on teha paremaid otsustusi ning pakkuda paremaid teenuseid ja mida võib realiseerida väga mitmesuguste vahenditega. Eesti riiklikus sektoris loodavad andmeaidad aitaksid toetada poliitikakujundajate ja elluviijate otsuste kvaliteeti, kuna otsuste aluseks saaks võtta erinevatest andmekogudest agregeeritud informatsiooni.

Uuring analüüsib andmeaitade poliitilisi, sotsiaalseid, organisatoorseid, infotehnoloogilisi, metodoloogilisi ja juriidilisi aspekte ning pakub välja kasutatava tehnoloogia ja reeglistiku ühtlustatud andmeaitade loomiseks riigis. Uuringu põhitulemuseks on metoodilised soovitused ja juhised andmeaitade (ka andmeanalüüsi) abil lahendatavate riiklike probleemide (sh poliitiliste, aruandluse, teaduslike jms) ja tehtavate otsuste osas, nende probleemide lahendamiseks andmeaitades ladustatavate andmete kogumise, haldamise, privaatsuse, koosvõime ja analüüsi läbiviimiseks ning kolmandatele osapooltele kättesaadavaks tegemiseks.

Uuringu läbiviimisel on lähtutud „Eesti infoühiskonna arengukava 2013“ suunistest, nende hulgas eriti p. 4.3.1 „Avaliku sektori toimimise tõhustamine“ toodud ülesanded tegevussuundades „Avaliku sektori haldustoimingute ja menetlusloogikate ümberkujundamine vastavalt IKT rakendamisest tulenevate eeliste ja võimaluste ärakasutamiseks“ ja „Poliitikakujundamise tõhustamine parema andmekasutuse ning infoühiskonna mõju ja väljakutseid käsitlevate uuringute läbiviimise kaudu”.

Käesoleva lõpparuande teises osas täpsustatakse probleemi kirjeldust, uuringu eesmärke ning uurimisküsimusi. Kolmandas osas kirjeldatakse kasutatavat metoodikat. Neljas osa sisaldab intervjueerimise ja anketeerimise tulemusi, mida on täiendatud kirjanduse ja muude allikate tulemustega vastavalt vajadusele. Viiendas osas analüüsitakse teabematerjalide põhjal teadus- ja tehnoloogiatrende ning välismaa praktikaid. Kuues osa sisaldab pakutavaid lahendusi ning ettepanekuid. Seitsemendas ja kaheksandas osas on kokkuvõte ning

lühendite ja kirjanduse loetelud. Lisad sisaldavad täpsustavaid detailseid materjale – intervjuude analüüsi ja anketeerimise tulemusi ning küsimustikke, samuti vahe- ja lõppseminari ettekannete slaide.

2 PROBLEEMI KIRJELDUS JA UURINGU EESMÄRK

Praeguseks on mitmed ministeeriumid ja nende haldusalade asutused (Sotsiaalministeerium, Siseministeerium jt) oma tööülesannetest lähtuvalt asunud andmeaitasid välja töötama. Tihti ei tehta seda koordineeritult, kuna riigis puudub üldtunnustatud andmeaitade loomise ja kasutamise poliitika. Ka valdkonnasiseselt lahendab iga töörühm oma probleeme iseseisvalt. Samas näiteks on Sotsiaalministeeriumi haldusalasse planeeritavate andmeaitade eeldatav kasutusala seotud äärmiselt laia otsuste tegemise tasandiga ja ulatub väljapoole sotsiaalvaldkonna piire (teadusuuringud, statistika, finantsvaldkonna planeerimine jne).

Käeasoleva uuringu põhieesmärgiks on pakkuda välja lahendusi andmeaitade haldamise protsessi korrastamiseks ja harmoneerimiseks riigi infosüsteemis tervikuna, aidates lahendada sellega ka erinevate ametkondade andmete ladustamise ja analüüsi probleeme. Uuringul on nii poliitiline, sotsiaalne, organisatoorne, infotehnoloogiline, metodoloogiline kui ka juriidiline aspekt.

Lähtudes nimetatud põhieesmärgist on käesoleva uuringu alameesmärgid järgmised:

1. Riigi infosüsteemis rakendatavate andmeaitade tehnoloogia ja kogutavate andmete analüüs, selgitamaks välja riigis kujunenud andmete ladustamise ja analüüsi hetkeolukord.

2. Uurida välja, millised infotehnoloogilised, organisatoorsed, poliitilised, seadusandlikud ja sotsiaalsed probleemid on tekkinud andmeaitade rakendamisel riigi erinevates haldusalades.

3. Uurida maailmas levinud uusimaid suundi andmete ladustamise ja suurte andmekogumite analüüsi tehnoloogias; teha kindlaks, kuidas andmeaitade tehnoloogiat on kasutatud teiste riikide infosüsteemides ning analüüsida ka Eesti erasektori praktikaid nimetud valdkondades.

4. Et erinevates haldusalades kasutatavaid andmeaitu ei saa vaadelda isoleeritult teistest andmekogudest, siis tuleb uurida andmeaitade võimalikku (semantilist) koosvõimet teiste andmekogudega (ka andmeaitadega).

Ülalpool nimetud uurivate ülesannete tulemite alusel pakutakse uuringus välja lahendusi ja metoodilisi soovitusi Eesti riigi infosüsteemi andmeaitade arendamise, haldamise ja andmeanalüüsi parendamiseks ning harmoniseerimiseks selleks, et riigile ja selle kodanikele tähtsate otsuste langetajad saaksid võtta vastu teadmispõhiseid otsuseid. Seega on projekti lõppeesmärgiks tagada otsuste vastuvõtmise kvaliteedi parendamine.

Lisaks vaadeldakse käesolevas uuringus ka kõigi ülalpool mainitud probleemiga kaasnevaid juriidilisi aspekte, eriti seost pakutud lahenduste sobivusega meie riigi seadusandlusega; vajadusel pakutakse vastavaid seadusandluse muudatusi. Juriidilised aspektid omavad erilist tähtsust isikuandmete privaatsuse säilitamisel ja seega pööratakse erilist tähelepanu andmete anonüümimise või pseudonüümimise probleemidele andmete ladustamisel ja hilisemal andmete analüüsil.

Uuringu raames otsiti muuhulgas vastuseid järgmistele uuringuküsimustele:

1. Kui palju on tänase seisuga avalikus sektoris andmeaitasid olemas, kui palju on loomisel ning planeerimisel?

2. Milliseid andmeid, millistest valdkondadest ja milliste otsustustasandite jaoks üldse andmeaitadesse kogutakse ja mida perspektiivis (nii olemasolevate konkreetsete plaanide järgi kui ka kaugemas tulevikus) koguda tahetakse?

3. Millised andmed on vaja andmeaita sisestamiseks kodeerida (eelkõige isikuandmete kaitsest lähtuvalt) ja millised mitte?

4. Kuidas on korraldatud eri ministeeriumide ja ametkondade koostöö andmeaitadesse kogutavate andmete kogumisel, laadimisel ja kasutamisel? Kuidas on korraldatud asutuste vahel andmeaida andmekooseisu muudatuste teostamise protsess?

5. Millised on andmeaitade loomise ja haldusega seotud seadusandlikud probleemid? Milliseid regulatsioone oleks vaja muuta, täiendada või lisada?

6. Kuidas kujunevad praktikas välja vajadused andmete hoidmiseks ja töötlemiseks andmeaitades?

7. Kuivõrd on andmeaida andmete koosseis ajutine ja peale töötlusi kustutatav ning kuivõrd on vaja aasta aastalt üha uusi andmeid andmeaita lisada, et tekiksid võimalused andmete kumuleerimiseks ja aegridade analüüsiks?

8. Kellele ja mis alusel andmeaitadest andmeid väljastatakse?

9. Kuidas jaguneb vastutus andmeaidas olevate andmete kvaliteedi osas? Kuidas andmeaidad omavahel edaspidi suhtlevad? Kas selleks on seadusandlusse vaja lisada uusi regulatsioone?

10. Millised on lahendused ja regulatsioonid avaliku sektori andmeaitade osas teistes riikides? Kuhu suurte andmehulkade töötlus tänapäeva parimaid praktikaid silmas pidades suundub?

11. Mida on andmeaitade loomisel ja halduses õppida Eesti erasektori praktikatest?

12. Mida peaks eelnevast lähtudes Eestis andmeaitade tehnoloogia, seadusandluse ja kasutamise alal edasi tegema?

Uuringu läbiviimisel lähtuti Eesti riigi seadusandlusest ja infoühiskonna arengukavadest; eriti „Eesti infoühiskonna arengukava 2013"[1] vastavatest punktidest. Näiteks paragrahvis 4.3.1 „Avaliku sektori toimimise tõhustamine" toodud suunised, mis käsitlevad avaliku sektori haldustoimingute ja menetlusloogikate ümberkujundamist vastavalt IKT rakendamisele ning poliitikakujundamise tõhustamist parema andmekasutuse kaudu on otseselt seotud käesoleva uuringu teemade ringiga. Lisaks võib nimetada veel riigi infosüsteemide koosvõime raamistikke, so veebide, infoturbe, tarkvara ja Eesti IT koosvõime raamistikke[2], milledest viimane näeb ette semantikatehnoloogiate arendamise ja kasutuselevõtu lähimate aastate jooksul, olles seotud andmeaitade kui andmekogumite koosvõime probleemidega. Peale nende üldiste suuniste lähtuti kehtivast Eesti ja Euroopa Liidu andmekogusid puudutavast seadusandlusest, uuriti ka väljaspool EL olevate riikide vastavat seadusandlust.

Uuringuküsimustele vastamiseks kasutati andmeaitade valdkonna tehnoloogilist arengut käsitlevaid teabematerjale (näiteks analoogiliste rahvusvaheliste uuringute tulemused, tehnoloogiatrendide uuringute tulemused jms), teiste riikide ja Eesti erasektori praktikat andmeaitade rakendamisel otsustusprotsesside toetuseks tehtavaks andmeanalüüsiks ning andmeaitade kasutamise, loomise, haldamise ja õigusliku reguleerimisega seotud huvigruppide hinnanguid andmeaitadega seotud erinevatele aspektidele (intervjuude ja ankeetide põhjal).

[1] Eesti infoühiskonna arengukava 2013, p.4.3.1, Majandus- ja Kommunikatsiooniministeerium, Tallinn 2009
[2] http://www.riso.ee/et/koosvoime/raamistik

3 METOODIKA

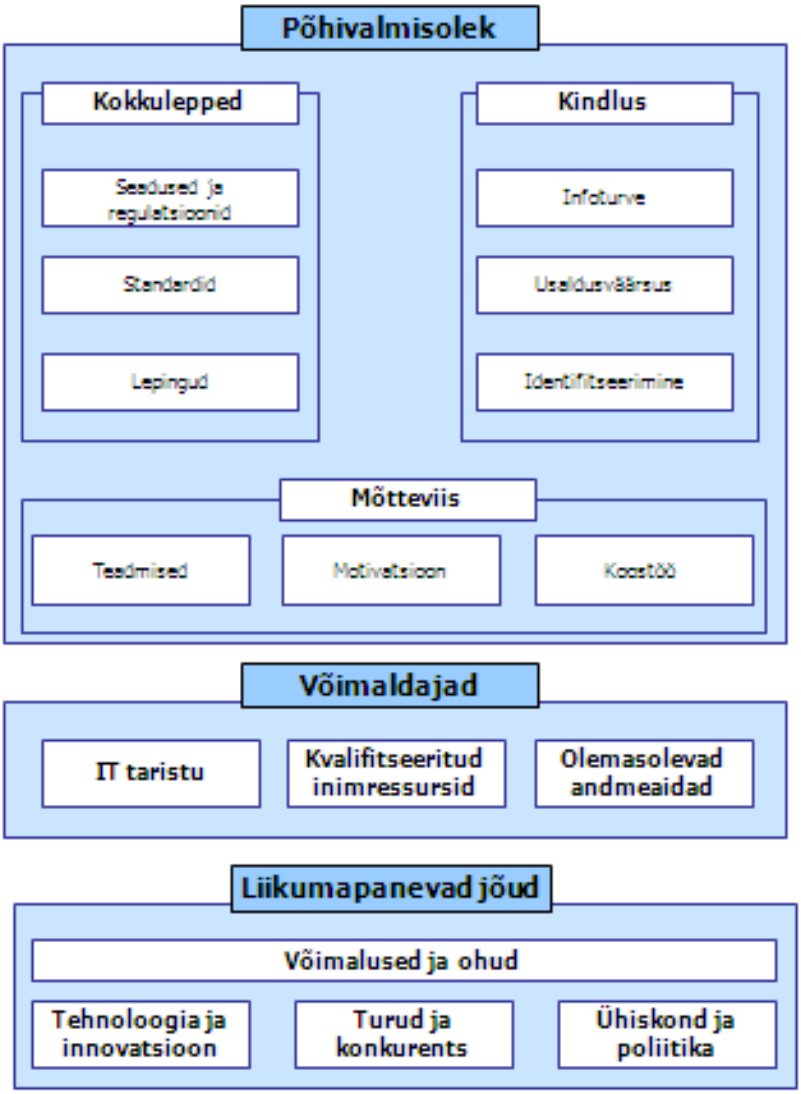
3.1. METOODIKA ÜLEVAADE

Uuringu läbiviimise metoodika tuleneb uuringu eesmärgist, ülesannetest ja saadavatest lähteandmetest. Vastavalt sellele oli uuringu metoodika kavandatud järgmiselt:

- **Olemasoleva teabe kogumine ja analüüs**, mis põhineb analoogiliste rahvusvaheliste uuringute tulemustel, andmeaitade ja andmeanalüüsi alastel tehnilistel artiklitel ja raamatutel, vastaval seadusandlusel, tooteuuringutel jms.
- **Teabe ja hinnangute kogumine intervjuude ja ankeetküsitluste abil,** selgitamaks välja andmeaitade kasutamise, loomise, haldamise ja õigusliku reguleerimisega seotud huvigruppide arvamusi, kogemusi, vajadusi, ettepanekuid ja tulevikuvisioone.
- **Vaheseminar** arutamaks diskussiooni vormis vastavate huvigruppidega uurimuse vahetulemusi.
- **Töökoosolekud** Tellija-poolsete partnerite esindajatega ja uuringu juhtkomitee liikmetega uurimise käigu täpsustamiseks ning vahe- ja lõpptulemuste arutamiseks.
- **Uuringu tulemuste analüüs ja esitamine.** Vahe- ja lõpparuande koostamine uuringu tulemuste esitamiseks.
- **Uurimuse lõpptulemusi tutvustav seminar** uuringu tulemuste tutvustamiseks kõigile huvigruppidele.

Uuringu metoodika **teabe ja hinnangute kogumiseks intervjuude ja ankeetküsitluste abil** nägi ette, et kõigepealt viiakse läbi intervjuud ettenähtud sihtgruppides selleks, et kaardistada andmeaitade probleemistikku, katsetada küsimustikku ja häälestada valikvastuste skaalasid ning vajadusel püstitada täiendavaid hüpoteese. Anketeerimine oli ette nähtud analüüsi läbiviimiseks ja hüpoteeside kontrollimiseks ning vastavate järelduste tegemiseks. Nii intervjuude kui ankeetide küsimustike struktureerimisel kasutati metoodikat, mis põhineb eBCM mudeli (The E-Business Community Model, e-äri kogukonna mudel, vt Joonis 1) lähenemisviisil. See võimaldab jagada andmeaitade valdkonna teemablokkideks vastavalt uuringu eesmärgile, hüpoteesidele ja uurimisülesannetele. See mudel on välja

töötatud projektis eBCM-RAP (finantseeritud Norra Innovatsioonikeskuse poolt) ja seda on kohandatud vastavalt käesoleva projekti vajadustele.



Joonis 1 Uurimuses kasutatud intervjuude ja ankeetide küsimustike koostamise metoodika mudel (kohandatud eBCM mudelist vastavalt käesoleva projekti vajadustele)

Käesoleva projekti jaoks kohandatud eBCM mudeli teemablokid on järgmised.

1. Põhivalmisolek iseloomustab seda, kuivõrd ühiskond on valmis andmeaitade rakendamiseks. Andmeaitade rakendamine eeldab kokkuleppeid - seadusi ja regulatsioone, standardeid ning lepinguid, mis toetavad andmeaitade juurutamist ja

kasutamist. Samuti on hädavajalik kindlustunne suuremate riskide suhtes, mille annab infoturbe tagamine, usaldusväärsete andmete kindlustamine ning kasutajate identifitseerimine. Lõpuks on oluline vastav mõtteviis - andmeaitade kohta käivad teadmised, inimeste motivatsioon ning asutuste vaheline koostöö.

2. Ainult põhimõttelisest valmisolekust ei piisa, on vaja ka võimaldajaid - taristut, eriti IT osas, aga samuti kvalifitseeritud inimressurssi ning olemasolevate andmeaitade realiseerimise kogemusi.

3. Valmisolek ja võimaldajad pannakse tööle liikumapanevate jõudude poolt. Eelkõige loovad andmeaidad uusi võimalusi info kättesaamiseks ning motiveerivad maandama kaasnevaid ohte. Kogu andmeaitade suuna on muutnud võimalikuks tehnoloogia ja innovatsioon. Turud ja firmadevaheline konkurents võimendab tehnoloogilist arengut. Ühiskond ja poliitilised jõud tunnetavad vajadust agregeeritud informatsiooni põhjal tehtavate otsuste järele, mis motiveerib andmeaitade arengut.

Käesoleva uuringu soovituste koostamisel lähtuti anketeerimise tulemustest, intervjuudel räägitust, teabematerjalide analüüsist, välismaisest kogemusest (ka välisintervjuude tulemustest) ja isiklikest ekspertteadmistest ning kogemustest.

3.1 INTERVJUUDE JA ANKEETKÜSITLUSTE LÄBIVIIMINE

3.1.1 Intervjuude läbiviimise metoodika

Uurimismetoodika nägi ette intervjuude läbiviimise järgmistes sihtrühmades:

- IT-ala koordineerijad riigis,
- IT-alase seadusandluse spetsialistid,
- ministeeriumide ja ametkondade andmeaitade spetsialistid,
- andmeaitade spetsialistid erasektoris,
- andmete kasutajad (ametnikud, poliitikud, teadlased, statistikud jt.).

Intervjuude küsimused olid jagatud järgmistesse gruppidesse: üldinfo, andmeaitadega seonduv infopoliitika ja seadusandlus, andmeaitade organisatoorsed küsimused, andmeaitade infotehnoloogiline aspekt ja intervjueeritava taustainfo.

Intervjuude küsimustik koostati lähtudes järgmisest hüpoteesist: riigi infosüsteemi andmeaitade valdkonnas on vaja teha nii poliitilisi, seadusandlikke, organisatoorseid kui tehnoloogilisi muutusi. Intervjuude küsimustikud on toodud käesoleva aruande Lisades.

Intervjuusid oli kavas läbi viia vähemalt üks iga sihtgrupi kohta, va riigisektori andmeaitade spetsialistid, kus planeeriti viia läbi 3 intervjuud. Intervjuud olid ette nähtud viia läbi poolstruktureeritud intervjuudena ja salvestada. Intervjuude analüüsi tulemusi planeeriti tutvustada huvigruppidele projekti vaheseminaril, et saada tagasidet ja sisendit ankeetküsitluse läbiviimiseks.

Peale selle oli kavas teha 3 lühiintervjuud e-posti vahendusel 3 välisriigi andmeaitade spetsialistiga, et koguda rahvusvahelist kogemust ja head tava. Nende intervjuude küsimused on toodud selle aruande ptk 5.2.2.

3.1.2 Anketeerimise metoodika

Anketeerimise üldine metoodika jälgib järgmisi Massachusetts Institute of Technology (MIT) anketeerimise põhimõtteid (http://web.mit.edu/surveys/survey-guidelines.pdf). Ankeetide küsimustiku koostamiseks on kasutatud eBCM mudeli (The E-Business Community Model, e-äri kogukonna mudel, vt Joonis 1) lähenemisviisi, mille järgi jagati ankeetküsimused järgmisteks üldisteks teemablokkideks: kokkulepped, kindlus, mõtteviis, võimaldajad ja liikumapanevad jõud. Ankeedi struktuur vastavalt nendele teemadele on täpsemalt kirjeldatud järgmises osas ja ankeedi küsimustik on toodud käesoleva uuringu lisas.

Ankeetide küsimustiku koostamise aluseks olid intervjuude küsimustikud ja intervjuude läbiviimisel saadud tagasiside. Näiteks osa intervjuude küsimusi ei töötanud, osade küsimuste skaalad olid kas ebatäpsed või vajasid laiendamist jms. Samuti võeti arvesse juhtgrupi töökoosolekul ja vaheseminaril kõlanud arvamused ja ettepanekud. Lisaks arvestati, et ankeetide küsimustik ei oleks ülepaisutatud. Et kasutati veebipõhist anketeerimast, siis arvestati ka anketeerimise vahendi Google Drive Form võimalusi ja piiranguid.

3.1.3 Ankeedi struktuur

Ankeedi küsimustiku struktuur põhineb järgneval:

- eBCM mudeli lähenemisviis määrab teemablokid.

- Teemablokid sisaldavad küsimusi, mis määravad lepingus ja pakkumuskutses nõutud tulemused. Teemablokid sisaldavad ka muid küsimusi, mis tulenevad eespool kirjeldatud metoodikast.
- Iga sihtrühm vastab temale sobivatele teemablokkidele.

Vastavalt anketeerimise üldistele põhimõtetele ja eBCM mudelile on allpool käsitletud järgmisi teemablokke:

- Üldinfo vastajate kohta.
- Kokkulepped (seadused, standardid, lepingud).
- Kindlus (infoturve, usaldusväärsus, identifitseerimine).
- Mõtteviis (teadmised, motivatsioon, koostöö).
- Võimaldajad (IT taristu, kvalifitseeritud inimressursid, olemasolevad andmeaidad).
- Liikumapanevad jõud (võimalused ja ohud, tehnoloogia ja innovatsioon, turud ja konkurents, ühiskond ja poliitika).

Teemablokid ja uurimusteemad ning teemablokkide jaotus sihtrühmade kaupa on toodud järgnevas tabelis (Tabel 1). Märkus: kaks tabelit on ühendatud, et tagada ankeedi struktuuri parem modifitseeritavus.

Tabel 1 Uurimisküsimuste ja sihtrühmade jaotus teemablokkide kaupa

Sihtrühm / uurimisküsimus / uuringu tulem / teemablokk	Üld-info	Kokku-lepped	Kindlus	Mõtte-viis	Võimal-dajad	Liik-p jõud
Pakkumuskutse. Uurimisküsimused ja teemablokid						
1. Kui palju on tänase seisuga avalikus sektoris andmeaitasid olemas, kui palju on loomisel ning planeerimisel?					+	
2. Milliseid andmeid, millistest valdkondadest ja milliste otsustustasandite jaoks üldse andmeaitadesse kogutakse ja mida perspektiivis (nii olemasolevate konkreetsete plaanide järgi kui ka kaugemas tulevikus) koguda tahetakse?						+
3. Millised andmed on vaja andmeaita sisestamiseks kodeerida (eelkõige isikuandmete kaitsest lähtuvalt) ja millised mitte?			+			

Sihtrühm / uurimisküsimus / uuringu tulem / teemablokk	Üld-info	Kokku-lepped	Kindlus	Mõtte-viis	Võimal-dajad	Liik-p jõud
4. Kuidas on korraldatud eri ministeeriumide ja ametkondade koostöö andmeaitadesse kogutavate andmete kogumisel, laadimisel ja kasutamisel? Kuidas on korraldatud asutuste vahel andmeaida andmekooseisu muudatuste teostamise protsess?				+		
5. Millised on andmeaitade loomise ja haldusega seotud seadusandlikud probleemid? Milliseid regulatsioone oleks vaja muuta, täiendada või lisada?		+				
6. Kuidas kujunevad praktikas välja vajadused andmete hoidmiseks ja töötlemiseks andmeaitades?						+
7. Kuivõrd on andmeaida andmete koosseis ajutine ja peale töötlusi kustutatav ning kuivõrd on vaja aasta aastalt üha uusi andmeid andmeaita lisada, et tekiksid võimalused andmete kumuleerimiseks ja aegridade analüüsiks?		+				
8. Kellele ja mis alusel andmeaitadest andmeid väljastatakse?		+				
9. Kuidas jaguneb vastutus andmeaidas olevate andmete kvaliteedi osas? Kuidas andmeaidad omavahel edaspidi suhtlevad? Kas selleks on seadusandlusse vaja lisada uusi regulatsioone?		+		+		
10. Millised on lahendused ja regulatsioonid avaliku sektori andmeaitade osas teistes riikides? Kuhu suurte andmehulkade töötlus tänapäeva parimaid praktikaid silmas pidades suundub?		+				
11. Mida on andmeaitade loomisel ja halduses õppida Eesti erasektori praktikatest?						+
12. Mida peaks eelnevast lähtudes Eestis andmeaitade tehnoloogia, seadusandluse ja kasutamise alal edasi tegema?		+	+	+	+	+
Pakkumuskutse. Oodatavad uuringu tulemid ja teemablokid						
Uus teadmine andmeaitade tehnoloogiast sh milliseid infotehnoloogilisi vahendeid ja keskkondi peaks kasutama.					+	
Metoodilised soovitused andmete kogumiseks, töötlemiseks ja väljastamiseks avaliku sektori andmeaitadest.				+		
Ettepanekud andmeid ja nende kvaliteeti puudutava vastutuse reguleerimise osas andmeaida tasandil.		+				
Ettepanekud andmeaitade vahelise andmekasutuse kohta ehk kuidas peaksid andmeaidad omavahel suhtlema.			+	+		
Ettepanekud andmetöötluse paremaks korraldamiseks riigi infosüsteemis (sh tehnoloogilised, organisatsioonilised ja juriidilised aspektid).		+		+		+

Sihtrühm / uurimisküsimus / uuringu tulem / teemablokk	Üld-info	Kokku-lepped	Kindlus	Mõtte-viis	Võimal-dajad	Liik-p jõud
Ettepanekud andmeaitade andmehõive laiendamiseks ja uute andmeaitade loomiseks.						+
Pakkumus (vajadusel, lisaks eelnevale)						
2. Uurida välja millised nii infotehnoloogilised, organisatoorsed, poliitilised, seadusandlikud kui sotsiaalsed probleemid on tekkinud andmeaitade rakendamisel riigi erinevates haldusalades.			+	+	+	+
4. Tuleb uurida andmeaitade võimalikku (semantilist) koosvõimet teiste andmekogudega (ka andmeaitadega).					+	
Tagada otsuste vastuvõtmise kvaliteedi parendamine.						+
Võetakse arvesse ka vastava valdkonna arenguid 5-10 a. perspektiivis.						+
Eraldi tähelepanu pööratakse soovitustes andmete kvaliteedile ja selle tagamise regulatiivsetele mehhanismidele.		+	+			
Uuringu tulemuseks on ka andmeaitade loomise, haldamise ja hallatavate andmete privaatsusega seotud juriidilised ja organisatoorsed ettepanekud.		+		+		
Intervjuudest, teemablokkidest jne lähtuvad küsimused (vajadusel, lisaks eelnevale)						
Seadused, standardid, lepingud. RIHA ja andmeaida vahekord tuleks selgitada. Kas kasutatakse standardeid selles valdkonnas?		+				
Infoturve, usaldusväärsus, identifitseerimine. Üldised suundumused ühiskonnas, kas soodustavad või ei andmeanalüüsi? Kuidas suhtutakse privaatsusega seoses?			+			
Teadmised, motivatsioon, koostöö. Motivatsioon ja valmisolek koostööks. Inimeste andmeaitade alased teadmised (haridus?), kvalifikatsioon ja teadlikkuse tase. Kas on vajadus andmeaitade vahelise koosvõime järele ja millises vormis?				+		
IT taristu, kvalifitseeritud inimressursid, olemasolevad andmeaidad. Ressursside (inimesed, raha, riist-ja tarkvara jms) olemasolu (või hankimise võimalus) andmelaonduse arendamiseks.					+	
Võimalused ja ohud, tehnoloogia ja innovatsioon, turud ja konkurents, ühiskond ja poliitika. Kuidas on teistes riikides andmelaondus jm selline tehnoloogia arenenud? Kas see sunnib meie ministeeriumides ka asjaga tegelema?						+
Kas andmeaitades võiks info olla mingil määral agregeeritud või peaksid andmed olema objektidena?						
Sihtrühmade teemablokid						
Ministeeriumide ja ametkondade andmeaitade spetsialistid	+	+	+	+	+	+
IT-ala koordineerijad riigis	+	+	+	+	+	+
IT-alase seadusandluse spetsialistid	+	+	+	+		

Sihtrühm / uurimisküsimus / uuringu tulem / teemablokk	Üld-info	Kokku-lepped	Kindlus	Mõtte-viis	Võimal-dajad	Liik-p jõud
Andmeaitade spetsialistid erasektoris	+	+	+	+	+	+
Andmete kasutajad – eri tasandite ametnikud, poliitikud, teadlased, statistikud jne	+	+	+	+	+	+

3.1.4 Intervjuude ja ankeetide läbiviimine

Põhiliste välitööde käigus viidi läbi 8 intervjuud järgmistes sihtrühmades:

- IT-ala koordineerijad riigis (RISO - 1 intervjuu),
- IT-alase seadusandluse spetsialistid (AKI - 1 intervjuu),
- ministeeriumide ja ametkondade andmeaitade spetsialistid (Sotsiaalministeerium, Statistikaamet, Eesti Haigekassa, Politsei ja Piirivalveamet - kokku 4 intervjuud)
- andmeaitade spetsialistid erasektoris (1 intervjuu),
- andmete kasutajad (erasektor- 1 intervjuu).

Intervjuud salvestati ja nende kirjalikud kokkuvõtted esitati koos uuringu vahearuandega.

Peale selle viid läbi 3 intervjuud vastavalt Ameerika Ühendriikide, Suurbritannia ja Hollandi andmeaitade ekspertidega. Need intervjuud tehti e-posti vahendusel.

Anketeerimine viidi läbi elektroonselt veebiküsimustiku abil. Anketeeriti spetsialiste ning kasutajaid, kusjuures anketeeriti ka juba intervjueeritud inimesi. Anketeerimise sihtrühmad olid järgmised: andmeaitade riigisektori spetsialistid, andmeaitade erasektori spetsialistid ja andmete kasutajad.

Sihtrühmade anketeerimiseks koostati eraldi ankeedid spetsialistide ja kasutajate anketeerimiseks. Saadeti laiali 105 ankeeti, neist 11 korduvalt. Laekus 27 vastust spetsialisti ankeedile (sellest 17 riigi- ja 10 erasektorist) ja 10 vastust kasutaja ankeedile.

Ankeetide tulemused on esitatud käesoleva aruande lisas 9.2.

4 INTERVJUUDE JA ANKEETIDE ANALÜÜSI TULEMUSED

Ankeetidele vastajatest üle kolmandiku olid tipp- ja keskastme juhid, ligi pooled spetsialistidest juhid. Kõigi töö oli seotud andmeaitade loomise, haldamise, rakendamise või kasutamisega ning valdav enamus (ligi 90% vastanutest) omab kõrgharidust.

Arvestades vastajate positsioone ja nende seoseid andmeaitade loomise, haldamise, rakendamise ja kasutamisega, peaksid vastused adekvaatselt katma kõiki soovitud küsimuste teemasid: andmeaitadega seonduv infopoliitika ja seadusandlus, andmeaitade organisatoorsed küsimused, andmeaitade infotehnoloogiline aspekt alates andmeaitade loomisest ja lõpetades nende tööshoidmisega.

Sihtgruppide lõikes kattuvate küsimuste puhul on vastused üldjuhul summeeritud. Sihtgruppide lõiked on eraldi välja toodud vaid siis, kui see annab lisainfot uuritavate küsimuste kohta.

NB! Diagrammidel toodud vastajate arvud ei pruugi summeeruda vastajate koguarvuks kui mõni vastanutest jättis antud küsimusele vastamata. Seda võib ka tõlgendada „Ei oska öelda“ vastusena.

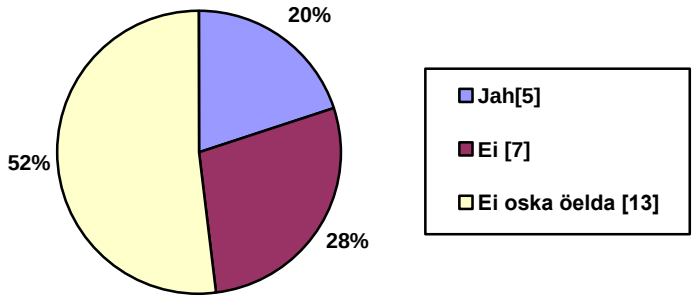
4.1 KOKKULEPPED (SEADUSED, STANDARDID, LEPINGUD)

Andmeaitade rakendamine eeldab seadusi ja regulatsioone, standardeid ning lepinguid, mis soodustavad andmeaitade juurutamist ja kasutamist. Näiteks pakuvad andmekaitseseadused (isikuandmete kaitse seadus, avaliku teabe seadus, riikliku statistika seadus jm) küll teatud regulatsiooni andmeaitadega tööks, kuid valdkondlikud seadused võivad käia nende andmete hoidmise ja töötlusega halvasti kokku või olla kohati vastuolulised.

Käesolevas jaotises analüüsitakse, millised on andmeaitade loomise ja haldusega seotud seadusandlikud probleemid ning milliseid regulatsioone oleks vaja muuta, täiendada või lisada.

4.1.1 Rahulolu riigi infopoliitikaga andmeaitade valdkonnas

Küsimusele "Kas olete üldiselt rahul meie riigi infopoliitikaga andmeaitade valdkonnas?" laekunud vastuste jaotus on toodud järgneval joonisel (Joonis 2).



Joonis 2 Rahulolu riigi infopoliitikaga (spetsialistid)

Märgitud probleemide klassid on järgmised.

- Õigusaktid pole piisavad, eriti seoses isikuandmete kaitsega.
- Riiklik koordineerimine pole piisav.
- Andmekogu ja andmeaida vahelised seosed pole selged. Andmeaida staatus ja registreerimine RIHAs pole selge.
- Riigisektoris on vähene teadlikkus andmeaitade valdkonna kohta - mis on andmeladu, kuidas seda ehitatakse, kuidas projekti läbi viia, kuidas vastavaid hankeid korraldada.

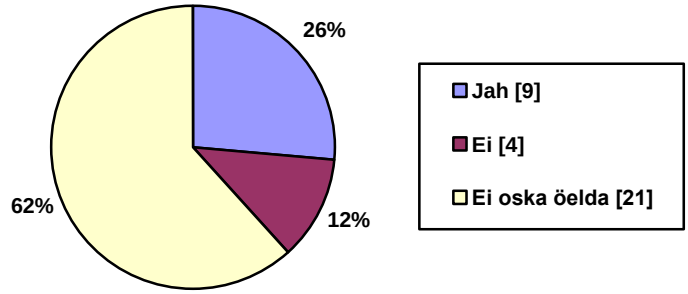
Kuna enamik vastustest käis pigem regulatsioonide kohta, vaatame neid allpool koos regulatsioonide kohta käiva küsimusega. Konkreetselt infopoliitika kohta võib märkida järgmisi kommentaare.

- Ei. Tundub, et riigil pole ühtset infopoliitikat andmeaitade osas. Kui infopoliitika ongi olemas, siis pole ta allpool tunnetatav. Üldiselt infopoliitika on pigem ministeeriumide keskne – igaüks teeb seda mida vajalikuks peab.
- Ei. Poliitikat õieti polegi. RIHA tasemel on sellest korduvalt juttu olnud. Andmeaitade kooskõlastused saabuvad osakonda, kuid RIHA ei toeta andmeaitade dokumenteerimist. Riigis pole paigas, milliseid andmeaitasid riigil on vaja ja mismoodi neid tuleks luua ning seostada. Kohapealne teadmine on asutustes tendentslik – igal pool leiutatakse jalgratast. Puudub ühtne poliitika.

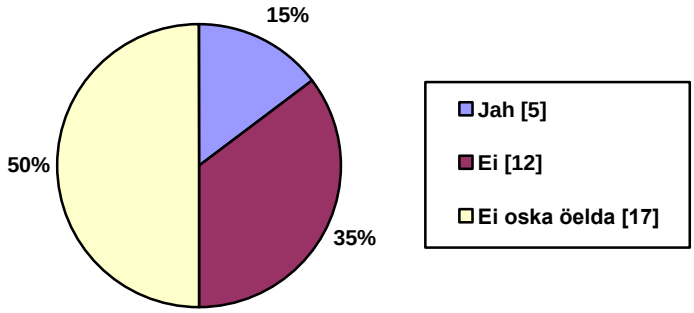
Kokkuvõttes, teadlikkus riigi infopoliitikast andmeaitade valdkonnas ja sellega rahulolu ei ole kõrge.

4.1.2 Seadusandlike regulatsioonide muutmise vajadus andmeaitade valdkonnas

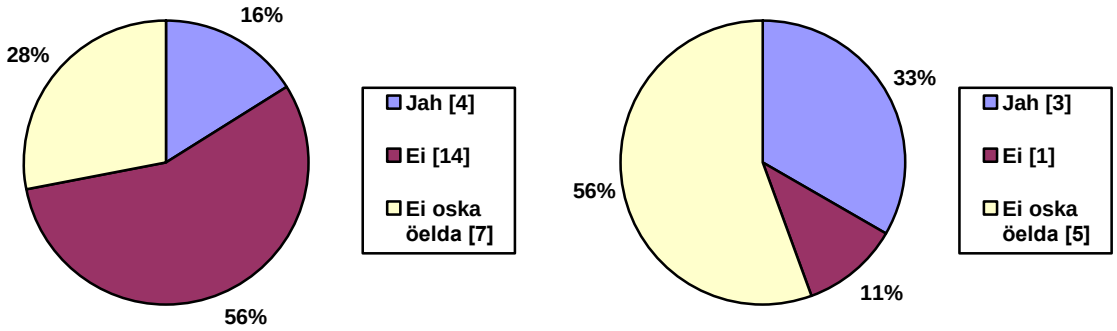
Seadusandlikke regulatsioone puudutavatele küsimustele antud vastuste jagunemine ankeetides on illustreeritud järgnevate joonistega.



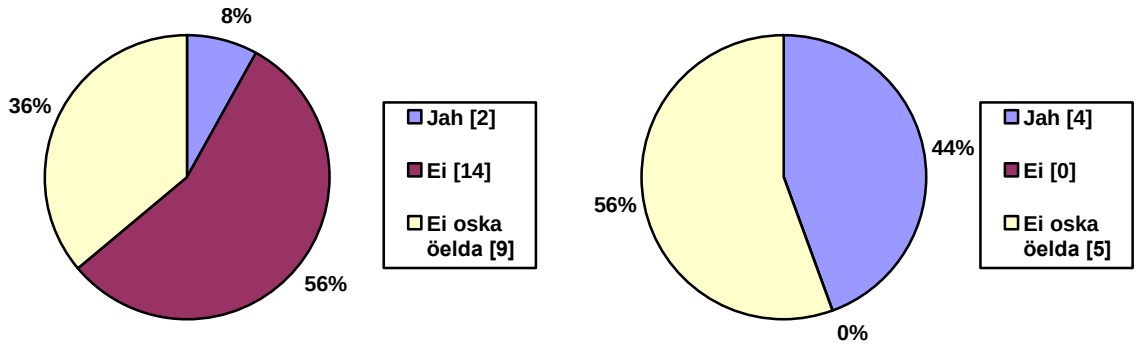
Joonis 3 Riigi andmeaitade loomise ja haldusega seotud seadusandlike regulatsioonide muutmise vajadus



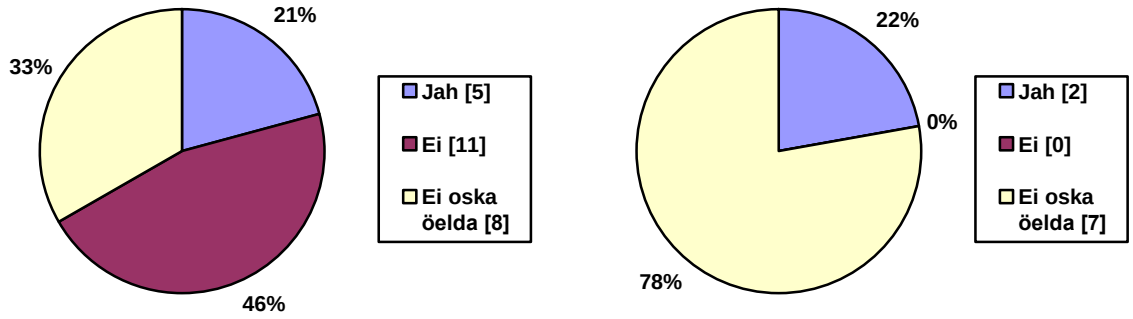
Joonis 4 Seadusandlikud regulatsioonid takistavad andmeaitade koosvõimet



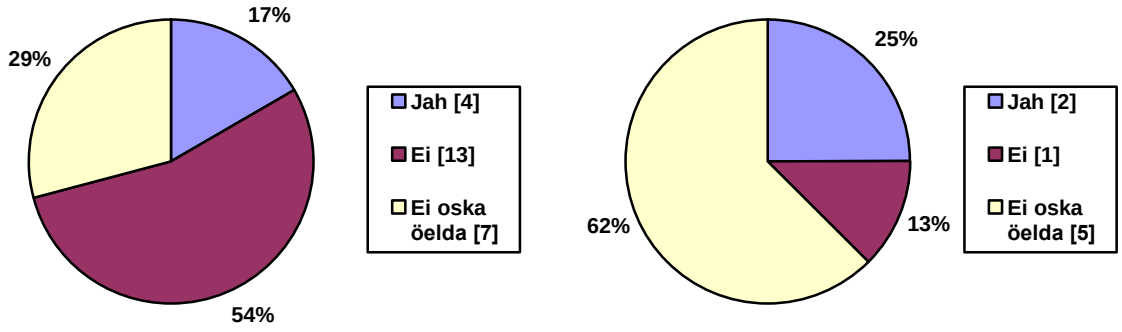
Joonis 5 Erinevate andmekogude ühendamine ühte andmeaita vajab eraldi seadusandlikku regulatsiooni (spetsialistid - kasutajad)



Joonis 6 Andmeaitade omavaheliseks suhtlemiseks on seadusandlusse vaja lisada uusi regulatsioone (spetsialistid - kasutajad)



Joonis 7 On vaja uusi regulatsioone andmeid ja nende kvaliteeti puudutava vastutuse sätestamise osas andmeaida tasandil (spetsialistid - kasutajad)



Joonis 8 On vaja uusi regulatsioone seoses andmeaitades hallatavate andmete privaatsusega (spetsialistid - kasutajad)

Andmeaitade loomise ja haldusega seotud seadusandlike regulatsioonide kooskõla hinnangut riigi üldiste andmeaitade loomist ja haldust puudutavate regulatsioonidega kajastab Tabel 2.

Tabel 2 Seadusandlike regulatsioonide kooskõla (spetsialistid)

Väga hea	2	7%
Hea	5	19%
Rahuldav	13	48%
Mitterahuldav	1	4%
Kooskõla puudub	1	4%
Ei oska öelda	4	15%

Intervjuudes leiavad riigi esindajad, et seadusandlike regulatsioonide muudatused on vajalikud. Erasektori esindajad seevastu ei näe vajadust muudatusteks. See võib tähendada järgmist.

- Riigiasutustes on andmete ühiskasutuse vajadus suurem, andmeaitu vajatakse rohkem ja seepärast on regulatsioonide ebapiisavus suurem probleem.
- Riigiasutustes on tegemist spetsiifiliste probleemidega (RIHAsse kandmine), mis ei puuduta otseselt erasektorit.
- Riigiasutused on põhilised andmete andjad ja seepärast huvitavad regulatsioonid neid rohkem.
- Erasektoris lähenetakse igale andmeaitade loomise juhtumile konkreetselt ja leitakse lahendused juhtumipõhiselt, puudub vajadus lahendusi üldistada.
- Muud erinevused riikliku ja erasektori vahel.

Muudatuste vajadused intervjuudes ja ankeetides jagunevad järgnevatesse põhigruppidesse.

- Andmeaidad tuleks määratleda regulatsioonide tasemel.
- Tuleks sätestada andmeaidas paiknevate andmete privaatsuse nõuded.
- Tuleks sätestada andmeaitade registreerimine ja haldamine (vt ka järgmine alateema).

Konkreetselt võib intervjuudest ja ankeetidest tulenevalt ära tuua järgmisi kommentaare (osa vastuseid oli seotud infopoliitika küsimusega), mis soovitavad andmeaitade määratlemist.

- Tuleks luua normid andmeaitade loomiseks ja halduseks. Kui andmed on kaugelt tuvastatavad, siis tuleb praegu järgida isikuandmeseadust. Selles osas oleks vaja uut regulatsiooni. Võimaluste laiendamiseks peaks andmeaitade jaoks olema norm.

- Regulatsioonides tuleks eraldi välja tuua andmelaondus kui eraldi eesmärkidega ja ülesehitusega andmekogundus. Peaks olema eraldi regulatsioon andmeaitade ja nende koosvõime kohta. Siis oleks lihtsam andmeaitu luua ja see peaks lahendama ka andmeaitade koosvõime küsimused.

- Selles valdkonnas oleks vajalik luua selgus ja põhimõttelised reeglid, kas (kuidas?) peab olema võimalik andmeid jagada ja kasutada.

- Avaliku teabe seaduse järgi on andmeait ja andmekogu võrdsed mõisted, seal ei räägita eraldi andmeaitadest. Kuigi seaduse tasemel ei ole ka praegu probleeme andmeaitade loomisel, võiks siiski igal andmeaidal olla seaduslik alus, mingi õigusakt (nt asutusesisene määrus vms).

- Kõige üldisemalt peaks andmeaidad seaduses välja tooma kui teatud eri klass andmekogusid, millel on rida spetsiifilisi omadusi (nad ei sisalda tavaliselt esmatasandi andmeid vaid on mõeldud ühekordseks või perioodiliseks töötluseks).

- Riigisektoris pole andmelaonduse seadusandlus paika pandud, puudub ühtne arusaam, et millistel tingimustel on andmeladu eraldi andmekogu. Andmelao eesmärk on siiski teine kui andmekogul kuigi ta kasutab operatiivsüsteemist (operatiivandmekogust) kopeeritud andmeid.

Järgmised kommentaarid käsitlevad eelkõige privaatsusega seotud regulatsioone.

- Andmeaidas paiknevate andmete privaatsus võiks olla selgemalt reguleeritud. Riikliku statistika seaduse $34 lõige 3 vajaks täpsustamist.

- Isikuandmete kaitse vajadus toob kaasa IS-i täiendava keerukuse, kuid alati pole seaduse järgi selge, kus on isikuandmete kaitse piir (nt millal kaitse on piisav). Seaduses on liiga palju tõlgendatavat.

- Tervise infoga seotud andmete osas on kindlasti vaja kindlustunnet selles, kuivõrd eri andmeaitade andmeid võiks/peaks/tohiks siduda.

- Tuleks muuta avaliku teabe seadust ja RIHA määrust. Võibolla ka andmekaitseseadust, kuigi see muutus piirdub ainult tagasikodeerimise osaga. Seda

üldiselt praegu ei lubata kasutada. Statistikaks tohib saada identifitseerivate väljadega andmeid.

- Seadused peaksid võimaldama automaatsete rakenduste loomist, mis võimaldab erinevate andmebaasides olevate isikuandmete ühist analüüsimist selliselt, et ei oleks tuvastatavad konkreetsed isikud (programm teostab andmete agregeerimise vajalikule tasemele).
- Tervise infoga seotud andmete puhul on vaja kindlustunnet selles, kas eri andmeaitade andmeid võib siduda.

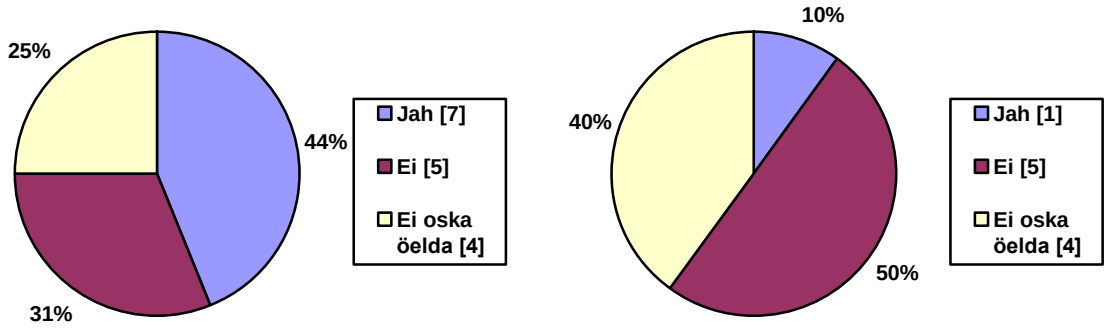
Järgmised kommentaarid on seotud eelkõige andmeaitade loomise ja haldamisega.

- Pole kindel, kas küsimus on ainult andmeaitades, pigem algab kõik infosüsteemidest ja sellest, kuidas sealt andmed edasi kanduvad.
- Andmeait võiks isikuandmeid kajastada nagu operatiivsüsteem (so originaalkujul). Andmekaitse peaks tulema rakenduste kihi poolt. Hea oleks omada valdkonnaüleseid andmeaitu (nn superandmeaitu).
- Ühekordsete päringute tegemine pole õige asi - andmeait pole ajutine. Määruse järgi tohib andmeid hoida andmeaidas 1 päev aruannete genereerimiseks. Praegu seda soovitust ignoreeritakse - hoitakse kõiki andmeid, kogu ajalugu (7 a). Andmeait on andmekogu sisemine asi, jagatud mitme andmekogu vahel. Andmeait on infosüsteemi (andmekogu) üks komponent ning ei peaks olema eraldi RIHA's.

Kokkuvõttes võib öelda, et küsimustele uute regulatsioonide vajalikkuse kohta vastas jaatavalt üldjuhul vähem ning eitavalt - rohkem vastajaid. Muutmise peamised valdkonnad on seotud andmeaitade määratlemisega, andmete privaatsusega andmeaitades ning andmeaitade loomise ja haldamisega.

4.1.3 Andmeaitade registreerimine RIHAs

Andmeaitade registreerimist RIHAs pooldavad enam riigisektori esindajad (vt Joonis 9).



Joonis 9 Andmeaitade registreerimise vajalikkus RIHAs (riigi- ja erasektori spetsialistid)

Registreerimise toetajad leiavad, et see on vajalik selguse ja vastutuse tagamise huvides ning märgivad: "Kui andmeaita kogutakse põhiandmekogust erinevaid andmeid, siis oleks kindlasti vajalik ka andmeaida ja selle mujalt kogutavate andmete kirjeldus". Intervjuude vastustest: "RIHA määrus võiks defineerida andmeaida kui ühe andmekogu liigi ja vastavalt sellele kehtestada RIHA eeskirjad andmeaitadele". Veel kommentaare:

- Kui andmeaidal on teine eesmärk kui operatiivsüsteemil või kui andmeait luuakse mitme andmekogu baasil, siis ta tuleks registreerida RIHAs.
- Kui andmeait on operatiivandmebaasi koopia, siis pole vaja.

Registreerimisega mitte nõus olevad vastajad leiavad, et sellest ei oleks andmeaida omanikul kasu, sellisel juhul kaob andmeaitade loomise ja muutmise paindlikkus ("andmeaida loomiseks piisab, kui see on vajalik avaliku sektori asutuse põhitegevuse täitmiseks, st selle loomine ja muutmine on puhtalt organisatsiooni enda otsustada") ning taoliste andmekogude aktuaalsena hoidmine RIHAs on väga töömahukas.

Intervjuudes avaldati ka seisukohta, et enne tuleks täpsustada RIHAs registreerimine: "Praegu RIHAs registreerimisel toimub kontseptuaalne segadus: mõisted andmekogu, andmeait ja infosüsteem on erinevalt käsitletud seaduse poole pealt. RIHAs on praegu registreeritud infosüsteemid, kuid RIHA määruse järgi peaksid olema registreeritud andmekogud".

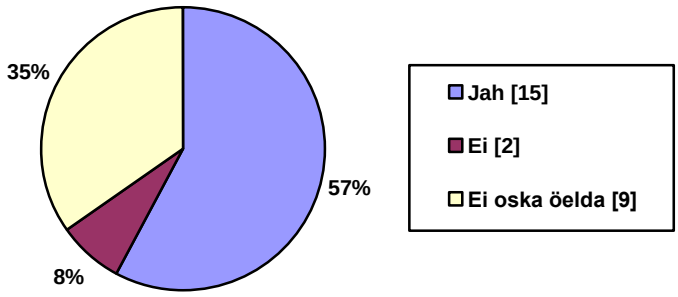
Kindla seisukohata vastajatest hindas üks vastaja, et registreerimise vajadus sõltub andmeaidast.

Kokkuvõttes, vastajad on erineval arvamusel selle kohta, kas andmeaitu tuleks RIHAs registreerida. Vastus võib sõltuda järgnevast.

- Andmeaida suhtest alussüsteemidega (nt kas andmeait on aluseks oleva andmekogu koopia, kas andmeait luuakse mitme andmekogu baasil).
- Sellest, kas andmeait luuakse vaid sisemisteks vajadusteks.
- Andmeaidas olevate andmete säilitamise kestvusest.
- Enne andmeaitade registreerimise küsimust tuleks täpsustada, mida RIHAs registreeritakse - infosüsteeme, andmekogusid vms.

4.1.4 Standardite kasutamine

Küsimus "Kas Eesti peaks kasutama rahvusvahelisi standardeid avaliku sektori andmeaitade osas?" esitati spetsialistidele Küsimusele vastati jaatavalt enam kui pooltel juhtudel (vt Joonis 10).



Joonis 10 Rahvusvaheliste standardite kasutamise vajalikkus andmeaitade osas (spetsialistid)

Enamus vastajatest leidis, et Eesti peaks kasutama rahvusvahelisi standardeid avaliku sektori andmeaitade osas. Põhjenduseks toodi, et rahvusvaheliste kokkulepete järgimine tuleb üldiselt kasuks, et saab üle võtta häid praktikaid, et saab vältida liigset tööd ("jalgratta leiutamist") ning et standardite kasutamine loob aluse andmete ristkasutuseks. Üks vastaja hindas standardite kasutamise vajalikuks teatud ulatuses ("ainult siis ja selles osas, mis tagab erinevate andmekogude ühis- ja ristkasutuse").

Kokkuvõttes võib järeldada, et need vastajad, kes pidasid ennast pädevaks sellele küsimusele vastama, hindasid rahvusvaheliste standardite kasutamist valdavalt vajalikuks.

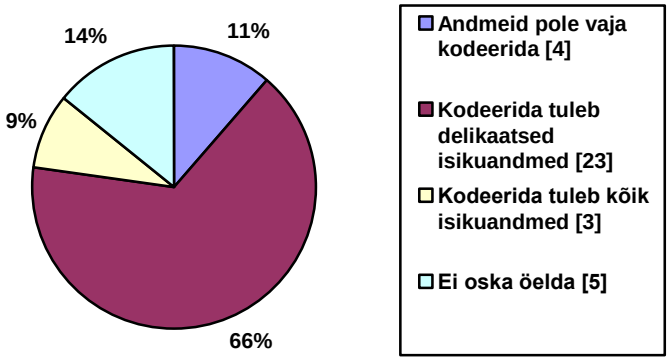
4.2 KINDLUS (INFOTURVE, USALDUSVÄÄRSUS, IDENTIFITSEERIMINE)

Infoturbe ja identifitseerimise küsimusi vaadatakse allpool esmajoones seoses personaalsete andmete töötlemisega, andmete usaldusväärsust - seoses nende kvaliteediga.

4.2.1 Isikuandmete kaitse, andmete kodeerimine ja identifitseerimine

Jaotises vastatakse küsimusele, millised andmed on vaja andmeaita sisestamiseks kodeerida (eelkõige isikuandmete kaitsest lähtuvalt) ja millised mitte. Samuti analüüsitakse spetsiaalsete kodeerimise tehnoloogiate kasutamist ja selle edasist vajadust.

Küsimusele "Milliseid andmeid on vaja andmeaita sisestamiseks kodeerida (eelkõige isikuandmete kaitsest lähtuvalt) ja millised mitte?" vastas 66% (23 küsitletut 35-st), et on vaja kodeerida delikaatseid isikuandmeid. Kolm vastajat pakkusid kõigi isikuandmete kodeerimist, neli - et andmeid pole vaja kodeerida (vt Joonis 11).



Joonis 11 Andmeaita sisestatavate andmete kodeerimise vajadus

Samast grupist hindasid 17 vastajat, et nad ei kasuta spetsiaalseid kodeerimise või krüpteerimise tehnoloogiaid ning 11 - et nad kasutavad selliseid tehnoloogiaid. Spetsiaalsete kodeerimise või krüpteerimise tehnoloogiate vajadust toetas 11 vastajat, eitavalt suhtus sellesse 10 vastajat.

Täpsustused spetsiaalsete kodeerimise või krüpteerimise tehnoloogiate kasutamise kohta võib jagada kolme rühma: kodeeringu vajalikkus; kodeeringu meetodid; täpsustavad kommentaarid.

Märgitakse, et kodeerimise või krüpteerimise vajalikkus sõltub andmeaida eesmärgist. Näiteks, kui andmeait toetab operatiivrakenduse aruandlust, ei saa isikuandmeid kodeerida.

Spetsiaalse kodeerimise vajalikkus sõltub ka andmeaida arhitektuurist; pigem on oluline teema, kuidas vältida isikute kaudset tuvastamist. Pakutakse võimalust, et andmed võiksid süsteemis olla objektidena ning kasutaja saaks sellises detailsuses teostada üldanalüüse, kuid tulemusi näeks kasutaja üksnes agregeeritud kujul. Andmeaidad andmeaida funktsionaalsuse mõttes peaks sisaldama kasutajate jaoks ainult agregeeritud infot, sestap ei tohiks seal kodeerimise vajadust (andmeaitade agregeeritus pole mitte võimalus, vaid kasutaja aspektist pigem kohustus, just see defineerikski andmeaida). Märgitakse, et kodeeritakse vaid delikaatseid isikuandmeid ning et spetsiaalsete tehnoloogiate vajalikkus sõltub andmeaida andmetest ja riskidest.

Kodeeringu võimaluste ja meetoditena tuuakse ära järgmised.

- Isikukoodid on asendatud juhuarvuna genereeritud koodiga. Tagasi teisendust saab vajadusel teha põhiandmekogus. Põhiandmekogu tabeliruum on krüpteeritud.
- Kasutame X-tee turvaserveri kodeerimisteenust.
- Eemaldatakse isikut identifitseerivad tunnused.
- Isikuandmed tulevad läbi kodeerimiskeskuse, et vältida isiku tuvastamise võimalust.
- Kasutame statistilise andmeaida loomisel ilmselt BI tarkvara enda kodeerimisvõimekust.
- Turvalisus tuleb tagada teisel tasandil. Isikuandmeid võib hoida eraldi krüpteeritult, kuid ülesande täitmiseks peavad need olema igal ajahetkel lahti krüpteeritavad ning lisanduvate andmetega seostatavad. Igasugune agregeeritus välistab võimaluse uute vajaduste ilmnemisel kasutada uut laadi agregeeringuid, seetõttu peavad olema kõik objektidena, kuid varustatuna ajas ühilduvate klassifikaatoritega, mis võimaldab iga ajapunkti kohta taasesitada sellel ajahetkel kehtinud infot, kuid olla samas ka uues ajapunktis võrreldav ülejäänud maailmaga.
- Meie poolt pakutava andmeaida lahendus võimaldab andmete krüpteerimist ning maskeerimist (vajalik arenduste puhul) ning samuti rollide lahusust. See on hädavajalik andmeturbe tagamiseks. Samuti sinna juurde kuuluv auditeerimise tugi.

Küsimusele "Kas andmeaitades võiks info olla mingil määral agregeeritud või peaksid andmed olema objektidena?" vastas 8 vastajat, et andmed peavad olema objektidena, 4

vastajat - et andmed võivad olla agregeeritud ning 16 vastajat, et see sõltub ülesandest või andmete laadist. Ülejäänud küsitletud ei osanud vastata või ei vastanud sellele küsimusele.

Kokkuvõttes, prevaleerib seisukoht, et kodeerida tuleks delikaatseid isikuandmeid; sellest tugevamaid (kodeerida tuleks kõiki andmeid) ja nõrgemaid (kodeerida pole vaja) lahendusi toetati ligikaudu võrdselt väikese arvu vaatajate poolt. Spetsiaalsete meetodite kasutamist ja vajadust nii kinnitati kui ka eitati. Märgiti, et kodeerimise vajalikkus sõltub mitmetest teguritest ja seda ei saa kõigi andmeaitade puhul ühtemoodi hinnata ning et kodeerimiseks on ka praegu olemas mitmesugused võimalused ja meetodid. Ka agregeerimise vajaduse küsimuses ei esitatud ühest valikut, vaid toetati pigem seisukohta, et seda tuleb otsustada eraldi sõltuvalt ülesandest ja andmetest.

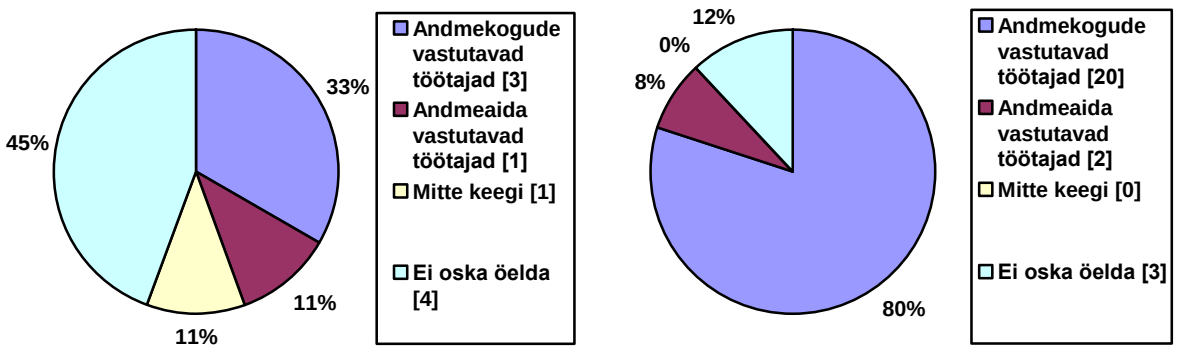
Toodud vastused ja hinnangud on kasulikud mitte vaid siis, kui andmed sisestatakse andmeaita ja nende peal tehakse analüüsi ("*taking the data to the analysis*"), vaid ka siis, kui analüüs tehakse mitmesugusest allikatest tulevate andmete põhjal, mida andmeaita ei koondata ("*taking the analysis to the data*"). Sellega seoses tekkivaid võimalikke lahendusi käsitletakse järgmistes peatükkides.

4.2.2 Andmete kvaliteet

Andmete kvaliteedi eest vastutavad andmekogude vastutavad ja volitatud töötlejad. Kuna andmeait liidab kokku erinevate andmekogude andmed, pole selge, kuidas jaguneb vastutus andmeaidas olevate andmete kvaliteedi osas. Samuti pole teada, kuidas ollakse rahul andmete kvaliteediga olemasolevates andmeaitades. On vaja anda nii tehnoloogiline kui ka seadusandlik hinnang sellisel teel tekkinud uuele olukorrale.

Andmete kvaliteeti puudutas mitu küsimust. Küsimusele "Kui andmeait ühendab mitme andmekogu andmeid, siis kes Teie arvates peaks vastutama andmete kvaliteedi eest?" vastati valdavalt, et vastutama peaksid vastavate andmekogude vastutavad ja volitatud töötajad (vt. Joonis 12). Kasutajatest üks vastaja hindas, et mitte keegi ei vastuta. Jooniselt on ka näha, et spetsialistid olid vastutuse jagunemise osas märgatavalt kindlamal seisukohal.

Ankeetide küsimusele "Kas on vaja uusi regulatsioone andmeid ja nende kvaliteeti puudutava vastutuse sätestamise osas andmeaida tasandil?" vastas 7 küsitletut jaatavalt, 11 - eitavalt, 15 ei osanud vastata.

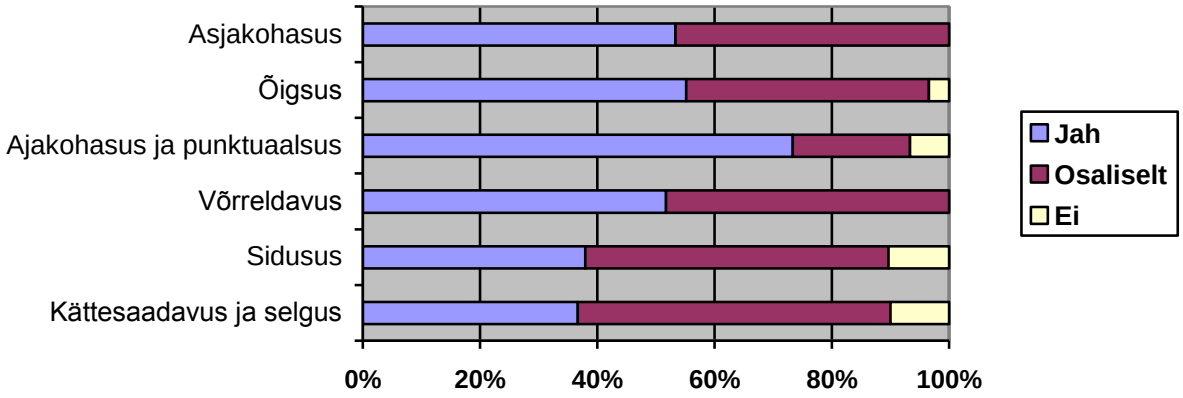


Joonis 12 Andmete kvaliteedi eest (kasutajad - spetsialistid)

Küsimusele "Kas väliste andmekogude kvaliteet on üldiselt rahuldav?" vastas 16 anketeeritut jaatavalt ("Jah" või "Pigem jah"), eitavaid vastuseid oli 11.

Rahulolu andmeaita laaditavate andmete kvaliteediga käsitles täiendav küsimus, mis põhines järgmistel EUROSTAT andmekvaliteedi parameetritel (Handbook on Data Quality Assessment Methods and Tools. European Commission, 2007). Järgnevas loetelus on toodud küsimused koos jaatavate, osaliselt jaatavate ja eitavate vastuste arvudega, Joonis 13 esitab sama graafiliselt.

- Asjakohasus (relevance) - kas andmed vastavad kasutaja vajadustele? ("Jah" - 16, "Osaliselt" - 14, "Ei" - 0).
- Õigsus (accuracy) - kas andmed on lähedased tegelikele väärtustele? ("Jah" - 16, "Osaliselt" - 12, "Ei" - 1).
- Ajakohasus ja punktuaalsus (timeliness and punctuality) - kas andmed on ajakohased ja õigeaegselt väljastatud? ("Jah" - 22, "Osaliselt" - 6, "Ei" - 2).
- Võrreldavus (comparability) - kas erinevad (nt erinevate ajaperioodide, geograafiliste asukohtade jne) andmed on omavahel võrreldavad? ("Jah" - 15, "Osaliselt" - 14, "Ei" - 0).
- Sidusus (coherence) - kas andmeid saab usaldusväärselt kombineerida erineval viisil ja erinevate kasutusviiside jaoks? ("Jah" - 11, "Osaliselt" - 15, "Ei" - 3).
- Kättesaadavus ja selgus (accessibility and clarity) - kas andmed on kättesaadavad ja arusaadavalt esitatud (sh koos metaandmetega ja kasutamise kitsendustega)? ("Jah" - 11, "Osaliselt" - 16, "Ei" - 3).



Joonis 13 Rahulolu andmeaita laaditavate andmete kvaliteediga

Intervjuudes toodi järgmisi täpsustusi seoses kõikuva või mittevastava andmete kvaliteediga andmeallikates.

- Ei kasutata õigeid klassifikaatoreid.
- Allikad ise ei analüüsi oma andmekogude andmekvaliteeti.
- Andmeaida pidaja annab allikatele tagasisidet andmekvaliteedi osas, kuid ta ei tohi anda sisulist tagasisidet probleemide kohta. Nt. ei tohi öelda, et inimene on saanud palka, aga on registreeritud töötuks jms. Esineb sisulisi vastuolusid ühe objekti kohta käivates andmetes, mis asuvad erinevates andmekogudes. Kui andmeaida pidaja selle avastab, puudub tal seaduslik õigus seda vastavate andmekogude töötlejatele teatada.
- Ei kontrollita sisestusi (nt. et aastaarv oleks 4-kohaline) see on peamiselt vanade andmeallikate probleem. Kui ise infot ei kasutata, siis pole ka kontrollitud.
- On põhimääruse ja praktilise elu erinevused. Aadresside identifikaatoreid ei rakendata. Nt. KMA ja ADS-i aadressid hästi ei sobi kokku. Kõik riigi IS-d pole liidestatud aadressiandmete süsteemiga (ADS). Kui ADS-s puudub vajalik aadress, siis peaks selle sinna kõigepealt lisama ja siis alles mingis infosüsteemis seda kasutama (nt sisestama) hakkama. Ruumistatistika jaoks on vajalik aadressi sidumine objektiga.
- Puudulik sisestuskontroll.

- Ajaloos on olnud kasutusel mitmed operatiivsüsteemid ja vanast uude süsteemi andmete ülekandmine tekitab vigu. Tekivad andmete ühilduvuse probleemid (sh andmete semantika probleemid).

- Registrid on huvitatud andmekvaliteedist, aga puudub ressurss kvaliteedi parandamiseks.

- Vead tulenevad põhiliselt inimese tegevusest. Kõik algab ka vastutusest – kes vastutab, kas andmete koguja või sisestaja või keegi kolmas. Võib tekkida vääriti arusaamu, sest menetlusprotsessi käigus andmed muutuvad. On tööl andmekvaliteedijuht, kes tegeleb andmete kvaliteedi monitooringuga. On seatud kriteeriumid, mille baasil kvaliteeti hinnata. Loomisel on andmete kvaliteedi monitooringu keskkond. Viiakse läbi andmekvaliteedi siseauditeid. Andmelao andmekvaliteedi kontrolli tulemusena antakse tagasisidet lähtesüsteemidele.

- Andmeaitade arendajatel on tekkinud juba kogemus näha lähteallikate andmete kvaliteedi erinevusi. Andmeait aitab tuvastada lähteallikate andmekvaliteedi probleeme ja lõpuks ühildub lähteallikate andmekvaliteet. On olemas andmemonitooringu süsteemid, mis teavitavad vigastest andmetest. Mitmestest allikatest valitakse allikad, milles on kvaliteetsemad andmed.

Kokkuvõttes näitavad saadud vastused, et rahulolu andmete kvaliteediga andmeaitades hinnatakse paremaks, kui seda enne uuringut oleks võinud oletada. Suhteliselt väiksem oli seejuures rahulolu andmete sidususe, kättesaadavuse ja selgusega.

Vastajad pakkusid valdavalt, et kui andmeait ühendab mitme andmekogu andmeid, siis andmete kvaliteedi eest peaks vastutama vastavate andmekogude vastutavad ja volitatud töötajad (kuigi hetkeseisuga see alati nii ei ole). Selline vastutus võib kaasa tuua järgmised probleemid.

- Kui andmete andmine välisesse andmeaita eeldab vastavate andmekogude vastutavate ja volitatud töötajate lisategevusi andmete kvaliteedi tagamisel, ei pruugi viimased olla andmete väljastamisest huvitatud.

- Kui andmete andmine välisesse andmeaita tekitab vastavate andmekogude vastutavatele ja volitatud töötajatele vastutuse andmeaida andmete põhjal tehtavate otsuste eest (nt riigi tasandil tehtavad strateegilised otsused), siis võivad vastutavad ja volitatud töötajad olla sellise vastutuse suhtes eitaval seisukohal.

Küsitlus ei näidanud selget uute regulatsioonide vajadust andmeid ja nende kvaliteeti puudutava vastutuse sätestamise osas andmeaida tasandil.

Andmeaitade andmekvaliteedi probleemide põhilised allikad on kvaliteediprobleemid andmeaitade aluseks olevates andmekogudes. Muuhulgas võivad probleemid tuleneda sellest, et andmete kvaliteet vastab küll algse andmekogu vajadustele, kuid seda on raske liidestada teiste andmekogude andmetega, tekivad semantilised probleemid jne. Seega võib andmelattu laadimine kaasa tuua lisanõuded andmete kvaliteedile, mille rahuldamine nõuab andmekogu töötlejalt lisaressurssi ja mille jälgimisest algse andmekogu töötlejad ei ole otseselt huvitatud.

Arvestades andmeallikate töötlejate võimalikku vähest motivatsiooni, peaksid andmeaitade pidajad ka ise aitama kaasa andmete kvaliteedi parandamisele andmeaidas, näiteks valides sobivaid andmekogusid andmeallikatena, rakendades tööle andmekvaliteedijuhi, seades kvaliteedi hindamise kriteeriumid, luues andmete kvaliteedi monitooringu keskkonna, viies läbi andmekvaliteedi siseauditeid ja andes tagasisidet lähtesüsteemidele.

4.3 MÕTTEVIIS (TEADMISED, MOTIVATSIOON, KOOSTÖÖ)

4.3.1 Teadmised

Intervjueeritavatest ja ankeetidele vastanutest olid 92% kõrgharidusega. Seega on käesolevat valimit arvestades andmeaitadega tegelejatel olemas üldine laiem vaade probleemide lahendamisele, sealhulgas andmetöötlusele ja infotehnoloogia vahendite kasutamisele.

Vastused ankeetides esitatud tehnilist laadi küsimustele olid enamasti konkreetsed ("Ei oska öelda" tüüpi vastused olid vähemuses), mis näitab vastajate orienteerumist andmeaitade praegustes tehnilistes aspektides.

Küsimusele "Kas olete tutvunud (kursis) väga suurte, variatiivsete andmetüüpidega ja erinevatest allikatest kogutud andmete töötlemise ning analüüsi (nn big data analytics) tehnoloogiatega?" vastas jaatavalt alla 11% küsitletuist. See näitab, et andmeaitade edasiste perspektiivide osas on teadlikkus väike.

Vastates küsimusele rahulolust riigi infopoliitikaga andmeaitade valdkonnas märgiti ühes intervjuus, et riigisektoris on vähene teadlikkus andmeaitade valdkonna kohta. Ei ole

reeglina selget ettekujutust sellest, et mis on andmeladu, kuidas seda ehitatakse, kuidas projekti läbi viia, kuidas vastavaid hankeid korraldada. Andmelao lahenduste hanked peaksid olema mõnevõrra spetsiifilised. Hangetel kohtab pealisehitusena tihti nõudeid nt objektorienteerituse vmt enda arendatud tarkvaralahenduste kohta. Tundub, et hanke tegijad ei tunne andmeaitade tehnoloogiat. Kohati näib, et kasutatakse vanu hankepõhjasid ja sobimatud nõuded on jäänud sisse. Tihti pannakse operatiivsüsteemi ja andmelao hanked kokku – ei anta endale aru, et nende vahel on tugev ajaline seos.

Andmeaitade alast koolitust pidas vajalikuks 83% ankeetidele vastanuist.

Kokkuvõttes võib järeldada, et vastajate üldine IT-alane teadlikkus ning tehnilised teadmised andmeaitade valdkonnast on paremad kui teadlikkus selle valdkonna suundumustest. Kohati võib olla probleeme andmelao projektide hankimise ja läbiviimisega. Vastanutel on soov täiendada oma andmeaitade alaseid teadmisi ja oskusi.

4.3.2 Motivatsioon

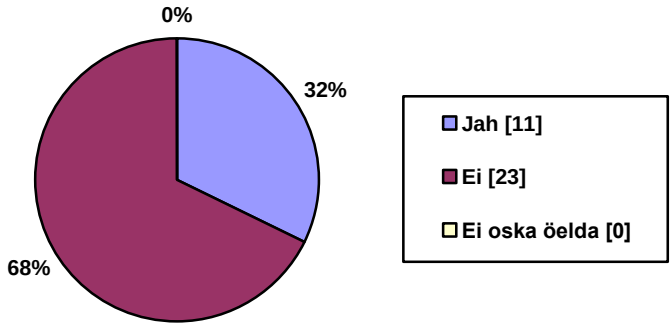
Küsimusele "Kas andmeaidad võiksid aidata kaasa otsuste vastuvõtmise kvaliteedi parendamisele?" vastas 92% anketeerituist jaatavalt. Märgiti muuhulgas, et andmeaidad võimaldavad välja tuua suuremat pilti ja seoseid valdkondade vahel, kus ilma andmekaeveta see praegu võimalik ei ole; et suur osa äriotsustest tehakse seni tunnetuse, mitte faktide baasil; et tervisepoliitika otsused võiksid olla tehtud Tervise Infosüsteemi andmete pealt; et andmeaitade kasutamine võimaldaks liikuda rohkem reaalajas info põhjal tehtavate ja terviklikumate otsusteni. Andmeaita konsolideeritud andmetel põhineval otsusel on kindlasti suurem väärtus, kui ainult ühe andmebaasi andmete põhjal tehtud otsustel - eeldusel, et andmeaita andmete laadimisel on eelnevalt kontrollitud ka andmete kvaliteeti. Lisaks peaksid andmed andmeaidas olema pikema perioodi vältel, mis võimaldab ajadünaamikat paremini jälgida ja selle põhjal vastavaid otsuseid teha. Pikad andmeread annavad aluse seaduspärasuste tundmaõppimisele. Neid tundes on võimalik prognoosida protsessi kulgu ning tänapäevaseid analüüsimeetodeid kasutades arvestada põhjus-tagajärg seoseid, mis on otsuste tegemise aluseks.

Küsimusele "Kuidas kujunevad praktikas välja vajadused andmete hoidmiseks ja töötlemiseks andmeaitades?" vastanuist hindas üle 86% anketeerituist, et vajadus tuleneb kasutajatelt, kes vajavad andmeid, või juhtkonnalt, kes püstitab eesmärgid. Vaid viis vastajat märkisid, et vajadus tuleneb seadusandlusest, IT spetsialistidelt või muudest allikatest.

Kokkuvõttes hindavad vastanud kõrgelt andmeaitade poolt pakutavaid uusi võimalusi. Algatus andmeaitadega tegelemiseks tuleneb enamasti kasutajate või ettevõtte vajadustest, luues seega tugeva motivatsiooni andmeaitadega tegelejate poolel. Nagu märgitud eespool, ei pruugi samasugune motivatsioon olla nende andmekogude töötlejatel, kust andmeaitade andmed pärinevad.

4.3.3 Koostöö ja muudatuste teostamine andmeaitade valdkonnas

Küsimusele "Kas Teie valdkonna andmeait/andmeaidad kasutavad (st koguvad, laadivad) teiste valdkondade andmekogude/andmeaitade andmeid?" vastas enamus eitavalt (vt Joonis 14). Täpsustavast küsimusest selgus, et teiste valdkondade andmeaitade andmeid ei kasuta keegi. Üle 30% kasutab andmeid muudest andmekogudest. Üks vastaja kasutab andmeid sama valdkonna andmeaitadest.



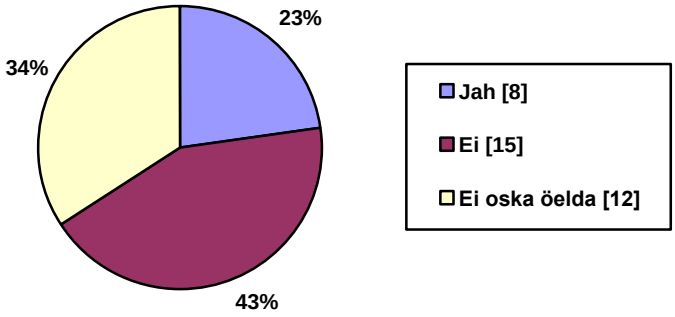
Joonis 14 Valdkondade andmekogude/andmeaitade andmete kasutamine

Küsimusele "Kuidas on korraldatud asutuste vahel andmekogu või andmeaida andmekooseisu muudatuste teostamise protsess?" vastasid 41% anketeeritud spetsialistidest, et enamasti muudatused kooskõlastatakse andmeid kasutava asutusega, neist teavitatakse kasutavat asutust vähemalt nädal ette või muudatused lepitakse eelnevalt kokku. Siiski leidsid 19% anketeeritud spetsialistidest, et muudatusi ei kooskõlastata ja neist ei teavitata. Muudatuste sünkroniseerimine ja teavitamine toimub mõnel juhul X-tee või RIHA vahendusel.

Muuhulgas märgitakse, et alati ei taha andmete töötlejad teha pingutusi andmete andmiseks, et andmekogud ei teavita oma andmestruktuuri muutustest kuigi peaks, et ebamäärane juriidiline regulatsioon takistab teiste andmeaitade andmete

laadimist/kasutamist ning et muudatuste tegemisel on probleemiks andmeaitades olevate andmete isikustamatus.

Küsimusele "Kas Teil on olnud probleeme oma valdkonna andmeaitade semantilise koosvõimega teiste andmekogudega (ka andmeaitadega)?" vastasid jaatavalt 23% anketeeritutest ja 43% hindasid, et probleeme pole selles valdkonnas olnud (vt Joonis 15).



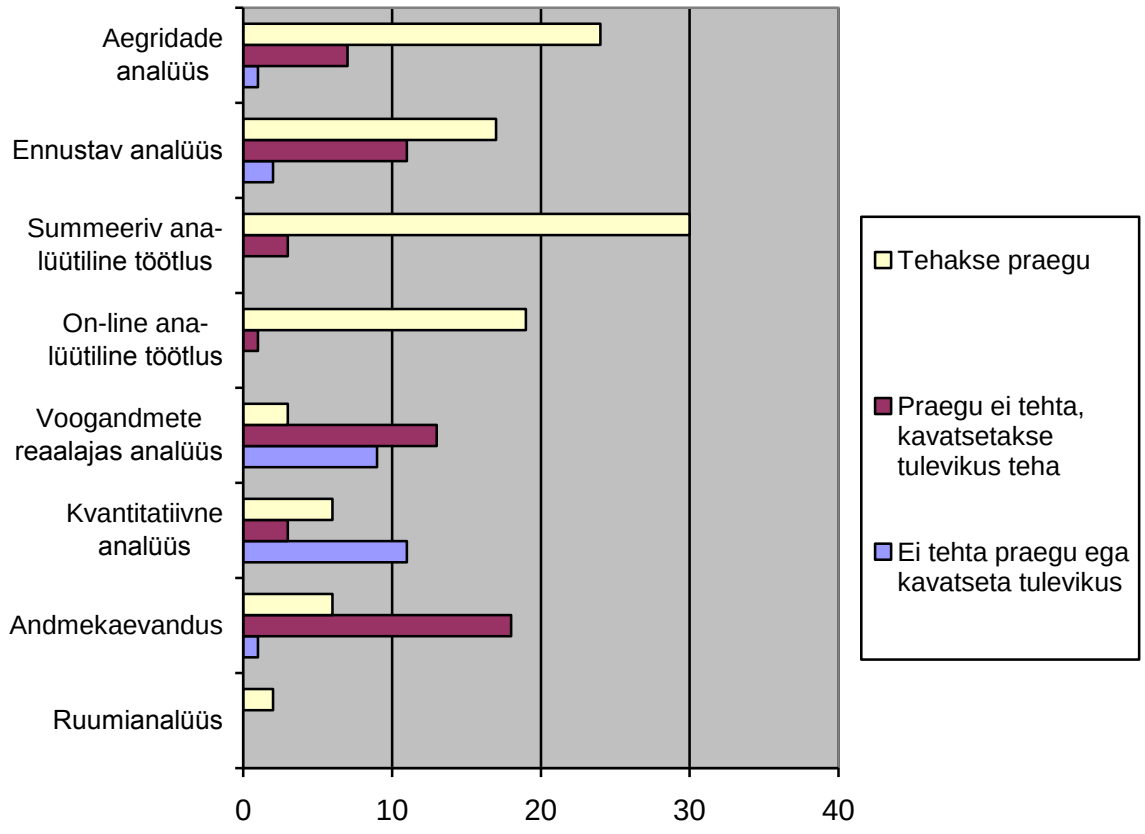
Joonis 15 Probleemid semantilise koosvõimega

Märgitakse, et kui andmeaida koosseis ja semantika on ette antud, siis on suur töö leida semantiline kooskõla andmeallikaks olevate andmekogude andmetega. Statistikud ja operatiivinfosüsteemi kasutajad võivad kasutada sama mõistet laiemas või kitsamas tähenduses. Samuti ei ole statistikute jaoks vajalikke mõisteid operatiivbaasides olemas, vaid need tuleb andmeaidas tuletada. Tuletamiseks peab andmeaida arendaja teadma väga põhjalikule andmete loogikat operatiivsüsteemis. Probleemid oma valdkonna andmeaitade semantilise koosvõimega võivad olla tingitud õigusaktides kasutatavast erinevast terminoloogiast. Iga andmekogu on arendatud kui eraldiseisvat, semantilisest koosvõimest alguses ei räägitud.

Üks teadus- ja tehnikavaldkonnas tegutsev anketeeritav märgib: "Kuna meie tegeleme andmete teatud laadi ühildamisega rahvusvahelistest nõuetest lähtuvalt, siis aluseks olevate andmekogude kirjeldused ei ole olnud piisavad või annavad vale ettekujutuse, mida tegelikkuses tähistatakse teatud andmeväljadel. Siiski, piisavalt detailsete kirjelduste olemasolul, on võimalik detaile tuvastada, kuid kindlasti ei tohiks usaldada vaid semantikat, see saab olema eksitav, kuni pole kõikide andmekogude üleselt vastavates terminites kokkulepitud. On kahetsusväärne, kui andmekogud loovad oma ettekujutusest lähtuvalt semantilisi kirjeldusi, mis ei ole koostatud koostöös valdkonna sisuinimestega ning pole läbinud keskset sisukoordinatsiooni."

Andmeid väljastatakse andmeaitadest selleks, et võtta vastu otsuseid, teha andmeanalüüsi, anda kasutajatele võimalus päringute tegemiseks või anda võimalus nende kasutamiseks laiemale üldsusele (avaandmed). Küsimusele "Mis tasemel otsuseid Teie arvates võetakse vastu Teie valdkonnas hallatavate andmeaitade andmete baasil?" vastas üle 80% anketeerituist, et lahendatakse valdkondlikke probleeme; üle 70% anketeeritutest vastas ka, et võetakse vastu rahvusvahelise, üldriikliku ja/või valdkonnaülese taseme otsuseid (võis valida mitu varianti).

Küsimusele "Kas Teie valdkonnas kasutusel oleva andmeaida andmete põhjal tehakse praegu või võiks teha tulevikus andmeanalüüsi?" vastas ligi viiendik anketeeritute koguarvust, et ühtegi liiki andmeanalüüsi praegu ei tehta; ülejäänud tegid juba praegu mingit liiki andmeanalüüsi (nt aegridade analüüs, ennustav analüüs, summeeriv analüüs jne, vt Joonis 16).



Joonis 16 Andmeanalüüsi tüübid praegu ja tulevikus

Võimalust, et konkreetse valdkonna andmed võiksid olla avaandmed (küsimus "Kas teie valdkonna andmeaida andmed võiksid olla avaandmed või lingitud avaandmed?"), toetas

täiel või osalisel määral 63% küsimusele vastanutest. Eitavalt vastanud (17% küsimusele vastanutest) põhjendasid otsust sellega, et tegu on äriettevõtete andmete, delikaatsete isikuandmetega või teaduslikel eesmärkidel kasutatavate andmetega, mis ei ole mõeldud üldkasutuseks. Samuti toodi esile seda, et andmeaida andmete avaandmeteks tegemisel saab ühest kohast ühe isiku või ettevõtte kohta ohtlikult palju infot teatavaks, mis tänases ühiskonnas ei ole kindlasti veel aktsepteeritav. Märgiti ka, et avaandmete eesmärk ja olemus vajab veel paljus täpsustamist; idee tundub toores, vaatamata selle vajalikkusele. Andmete kättesaadavaks tegemisel üldsusele oleks vaja riiklikul tasandil läbi mõelda ja tellida ühtne andmete kaudse tuvastamist mittevõimaldav töövahend kõigile riigiasutustele. See vajadus puudutab mitte ainult andmeaitade, vaid mistahes andmekogude väljundeid, mida soovitakse avalikustada eesmärgiga vähendada spetsialistidele tulevat päringute hulka.

Andmeaita päringuid tegevate kasutajate arv kõigub suurtes piirides (küsimus "Kui palju on Teie poolt loodud andmeaitadel tavaliselt kasutajaid, kes läbi mingite vahendite/rakenduste teevad andmeaita päringuid?"). 35% vastustest on kasutajate arv 1...10, 39% vastustest on kasutajate arv vahemikus 11...75.

Küsimusele "Kas avaliku sektori asutused võiksid andmeaitade loomisel ja halduses õppida Eesti erasektori praktikatest?" vastas kaks kolmandikku anketeerituist positiivselt, eitava vastuse andis vaid üks vastaja. Intervjuudes hinnati muuhulgas, et suurematelt erafirmadelt saaks õppida hangete korraldamist - näiteks, analüüsi ja realisatsiooni vahele peaks jääma otsustamise koht, kus tellija saab otsustada, mis läheb realiseerimisele; andmeaida lahendusi oleks mõistlikum teha avatud eelarvega, aga see on vastuolus riigihangete seadusega. Pakuti ka, et mitmes erasektori valdkonnas (nt telekom, energia) on oma andmeaidad, aga andmeid pole sealt võimalik kätte saada. Märgiti, et pigem on erasektoril õppida riigilt, sest riigil on ehk pikemaajalisem kogemus. Konstateeriti, et andmeaitade loomise ja halduse konkreetsed probleemid on iseloomulikud rohkem avaliku sektori asutustele - erasektori andmeaitadega on erinev olukord, kuna need ei allu samadele õigusregulatsioonidele ja neid ei tule registreerida RIHAs.

Kokkuvõttes võib koostöö ja muudatuste teostamise intervjueerimise ja anketeerimise tulemused võtta kokku järgmiselt.

Teiste valdkondade andmeaitade andmeid praegu ei kasutata, küll aga kasutatakse umbes kolmandiku andmeaitade puhul andmeid teiste valdkondade andmekogudest.

Väliste andmekogude andmete kasutamisel toimub üle 40% juhtudel muudatuste kooskõlastamine või nendest teavitamine, kuid umbes viiendik vastajatest märgib, et

muudatustest ei teavitata ning toob esile järgmisi probleeme teiste valdkondade andmete kasutamisega: andmestruktuuri muudatuste teavitamine ja nende muudatuste haldamine siht-andmeaidas, kõik andmete omanikud ei ole huvitatud oma andmete andmisest, andmete isikustamatus.

Semantilise koosvõime probleeme tõid esile alla veerandi vastajatest, märkides seejuures, et semantilise koosvõime probleemid võivad tuleneda erinevatest allikatest - näiteks, mõistete erinev kasutamine, õigusaktide erinev terminoloogia, mõistete puudumine, andmete esituse loogika teadmise vajadus jne.

Valdav enamus anketeerituist märgib, et andmeid väljastatakse andmeaitadest nii valdkondliku kui ka rahvusvahelise, üldriikliku ja/või valdkonnaülese taseme otsuste tegemiseks. Neli viiendikku vastajatest teevad väljastatavate andmete põhjal mingit liiki andmeanalüüse.

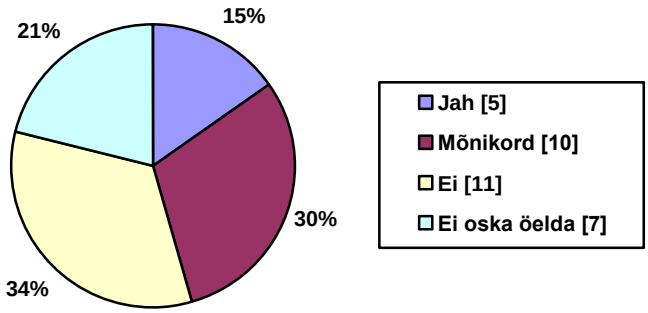
Oma valdkonna andmete kasutatavaks tegemist avaandmetena toetas täiel või osalisel määral ligi kaks kolmandikku anketeeritutest. Eitavalt vastanud põhjendasid otsust sellega, et tegu on äriettevõtete andmete, delikaatsete isikuandmetega või teaduslikel eesmärkidel kasutatavate andmetega, mis ei ole mõeldud üldkasutuseks. Andmekasutajate arv jääb 74% vastanute puhul vahemikku 1...75.

Kaks kolmandikku anketeerituist hindas, et avaliku sektori asutused võiksid andmeaitade loomisel ja halduses õppida Eesti erasektori praktikatest. Muuhulgas, suurematelt erafirmadelt saaks õppida hangete korraldamist avatud eelarvega, mille puhul analüüsi ja realisatsiooni vahele peaks jääma edasisi tegevusi määrav otsustamise koht. Siiski märgiti ka, et pigem on erasektoril õppida riigilt, sest riigil on ehk pikemaajalisem kogemus.

4.4 VÕIMALDAJAD (IT TARISTU, KVALIFITSEERITUD INIMRESSURSID, OLEMASOLEVAD ANDMEAIDAD)

Küsimusele "Kas Teie asutusel on piisavalt ressursse (inimesed, raha, riist-ja tarkvara) andmeaitade arendamiseks?" vastati enam kui kolmel neljandikul juhtudest positiivselt - ressursid on kas olemas või neid saab vajadusel hankida. Vaid kaks anketeeritut (üks kasutaja ja üks riigisektori spetsialist) vastasid, et ressursse ei ole ja neid pole ka võimalik hankida. Seega võib järeldada, et ressursipuudus inimeste, raha või tehnoloogia osas ei ole andmeaitade arenduse põhiprobleem.

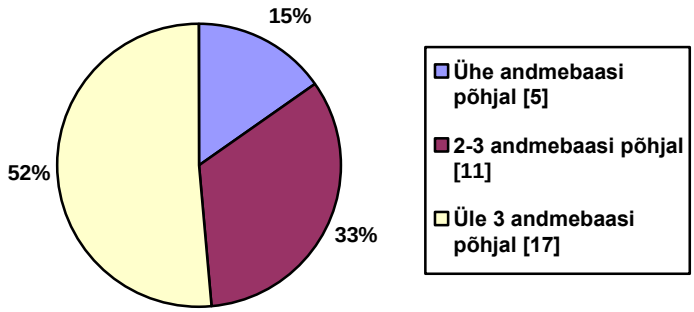
Küsimusele "Kas erinevate andmekogude andmete ühendamiseks ühte andmeaita on vaja eraldi tehnoloogilisi lahendusi?" vastati kolmandikul juhtudel eitavalt ning ligi pooltel juhtudel positiivselt (jah või mõnikord, vt. Joonis 17). Seejuures märgitakse, et lähteandmebaasid on erinevate andmebaasimootorite peale ehitatud ning on vaja andmete erinevaid struktuure ühendavaid/tõlgendavaid liideseid. Kolmel korral mainitakse X-tee lahendusi, kahel korral - Extract, Transform and Load (ETL) tööriistu / raamistikke. Võib seega järeldada, et andmeaitade loomine nõuab paljudel juhtudel eraldi tehnoloogilisi lahendusi. Erasektori spetsialistidest vaid 10% märgivad sellist vajadust, millest võib järeldada, et erasektoris on sellisele vajadusele mõeldud aegsasti ja kogu andmemajandus on ehitatud ühilduvatele platvormidele.



Joonis 17 Eraldi tehnoloogiliste lahenduste vajalikkus andmeaitade puhul

Kolm ankeetküsimust käsitles olemasolevate, loodavate ja planeeritavate andmeaitade arvu. Valdav enamus anketeerituist hindasid nende organisatsioonis olemasolevate andmeaitade arvuks 1-3; kahe vastaja hinnangul on andmeaitu organisatsioonis 3-6, kahel juhul üle 10. Kaks kolmandikku vastajatest pakkusid loomisel olevate andmeaitade arvuks 1-3 ja kaks viiendikku - planeerimisel olevate andmeaitade arvuks 1-3; ülejäänud vastasid, et uusi andmeaitu loomisel/planeerimisel ei ole või ei vastanud. Võib eeldada seega, et enamikus organisatsioonides jääb andmeaitade arv ka lähitulevikus põhiliselt vahemikku 1-3, üksikjuhtudel rohkem.

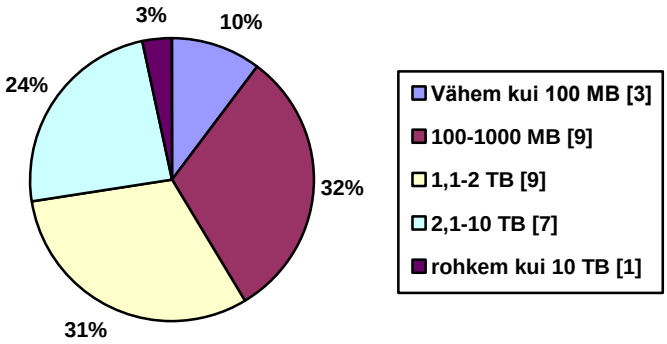
Küsimusele, mitme andmebaasi põhjal on vastaja valdkonnas kasutusel olev andmeait koostatud, vastasid enam kui pooled, et andmeait on koostatud enam kui 3-st andmebaasist (vt Joonis 18).



Joonis 18 Andmeaida aluseks olevate andmebaaside arv

Kui seostada see järeldus ühe intervjuudes pakutud kriteeriumiga RIHAs registreerimiseks (kui andmeait luuakse mitme andmekogu baasil, siis ta tuleks registreerida RIHAs), siis tuleks vaadeldud andmeaidad RIHAs registreerida.

Küsimusele algandmete mahu kohta valdkonna andmeaidas vastas vaid üks kasutaja, et andmete maht on suurem kui 10 TB (vt joonis). Seega võib järeldada, et Eestis praegu kasutatavad andmeaidad ei ole veel suure mahuga (*large data sets*, mahuga 10 TB ja enam, vt *Mark Scott, The Shortcut Guide to Large Scale Data Warehousing and Advanced Analytics*).



Joonis 19 Algandmete maht valdkonna andmeaidas

Intervjuudes ja ankeetides uuriti, kuivõrd on andmeaida andmete koosseis ajutine ja peale töötlusi kustutatav ning kuivõrd on vaja aasta aastalt üha uusi andmeid andmeaita lisada, et tekiksid võimalused andmete kumuleerimiseks ja aegridade analüüsiks. Vastavale küsimusele pakkusid kasutajad, IT spetsialistid kui ka seadusandluse spetsialistid 79% juhtudel, et andmeaitu luuakse andmete pikaajaliseks kogumiseks ja agregeerimiseks (aastaid). 15% vastanutest hindas, et kasutamise ajahorisont sõltub ülesandest ja olukorrast, üks kasutaja - et andmeait luuakse vahetuks tulemuste saamiseks (luuakse, kasutatakse ja

kustutatakse). Märgitakse ka, et lühiajalist andmeaita võiks asendada päringute mehhanism ning et vaja on mõlemat lähenemist. Võib seega järeldada, et Eestis on praegu valdav pikaajaliste andmeaitade loomise lähenemisviis ("*taking the data to the analysis*") ning et operatiivne analüüs operatiivsete allikate otsese kasutamisega ("*taking the analysis to the data*") ei ole veel läbi löönud. Tõenäoliselt on küsimus ka andmeaitade loomise efektiivsuses - lühiajaliseks agregeerimiseks on efektiivsem kasutada päringute mehhanismi.

4.5 LIIKUMAPANEVAD JÕUD (VÕIMALUSED JA OHUD, TEHNOLOOGIA, KONKURENTS, ÜHISKOND)

Motivatsiooni, ohte, tehnoloogiat ja muid liikumapanevatesse jõududesse puutuvaid teemasid on vaadeldud ka eespool. Selles jaotises viidatakse lühidalt juba saadud tulemustele ning analüüsitakse mõningaid lisanduvaid aspekte.

4.5.1 Võimalused ja ohud

Andmeaitade poolt pakutavaid uusi võimalusi on analüüsitud eespool motivatsiooni käsitlevas jaotises 4.3.2. Üle 90% anketeerituist hindab, et andmeaidad võiksid aidata kaasa otsuste vastuvõtmise kvaliteedi parendamisele. Praktikas tulenevad vajadused andmete hoidmiseks ja töötlemiseks andmeaitades enamasti kasutajatelt, kes vajavad andmeid, või juhtkonnalt, kes püstitab eesmärgid. See loob tugeva motivatsiooni andmeaitadega tegelemiseks. Vaid riigi spetsialistid märgivad seadusandlusest tulenevaid vajadusi.

Intervjuudes esitatud küsimusele "Kas Teie valdkonnas töötav andmeait loodi mingi probleemi või otsustuse tegemise vajadustest lähtuvalt?" vastajad on enamuses seisukohal, et andmeait loodi mingi probleemi või otsustuse tegemise vajadustest lähtuvalt. Muuhulgas mainitakse vajadust juhtimisotsuste vastuvõtmiseks, ärianalüüsiks ning selleks, et andmeid ühtemoodi hoida, formaate ühtlustada ja andmeid mitte dubleerida. Eitava vastuse andnud erasektori spetsialisti kommentaar oli järgmine: "Tavaliselt algab andmeaida loomine probleemist ja kliendiga suhtlemise käigus probleemide valdkond laieneb ja määravaks saab otsustuste tegemise vajadus. Tänapäeval pigem ei, sest määravaks on otsustamise vajadus ja andmed, mille alusel otsustatakse kujunevad välja töö käigus. Metoodika toetab pigem ettepoole vaatamist. Elu sunnib vaatama, mis andmed on ja mis analüüsi nendest annab välja pigistada – eesmärk peab olema piisavalt lai".

Küsimusele, millistest vajadustest lähtudes on andmeait loodud, vastas kaks kolmandikku anketeerituist, et neid motiveeris vajadus saada andmeid kiiresti ning neid integreerida (oli võimalik valida mitu vastust). Ligi pooled vastajatest märkisid vajadust korrektsete andmete järele, kolm vastajat soovisid saada andmeid odavalt. Märgiti ka aruannete ja statistika vajadusi, vajadust vältida ajaloolisi päringuid operatiivbaasist jne.

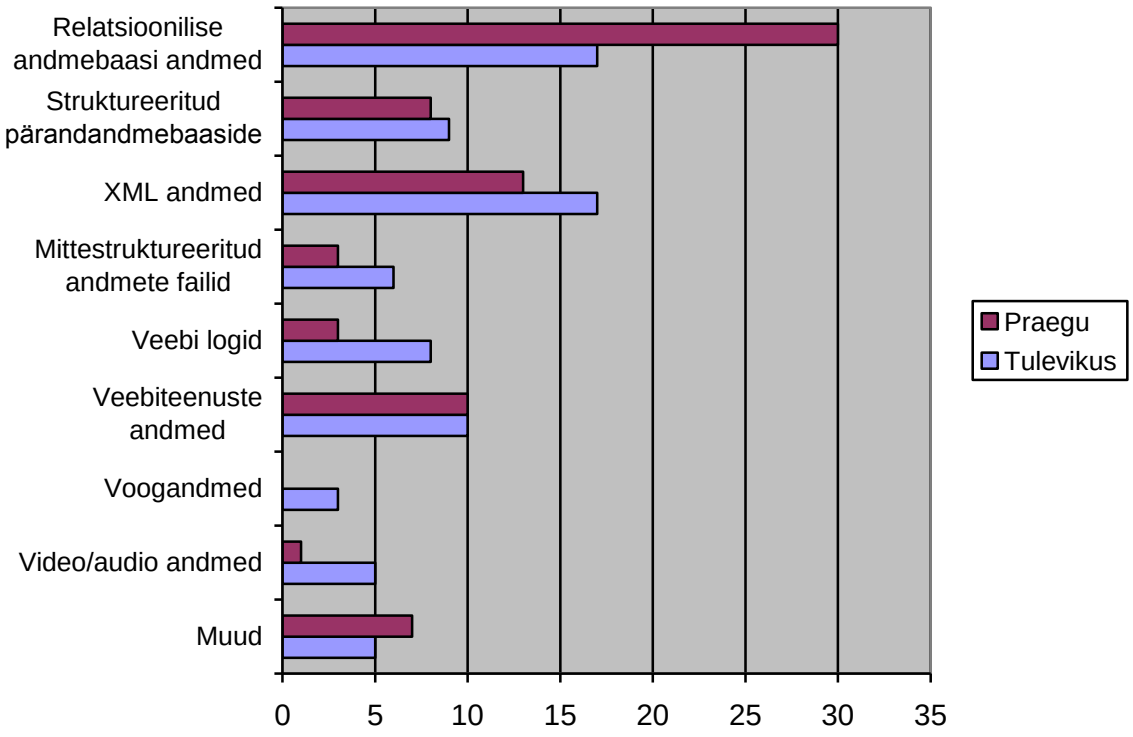
See, et ligi pooled vastajatest näevad vajadust korrektsete andmete järele kui andmeaitade loomise motivaatorit, näitab andmete kvaliteedi tähtsust. Samas ei looda vastajad üldiselt andmeaitade ökonoomsusele.

Küsimusele "Milliseid andmeallikaid/tüüpe kasutatakse Teie valdkonna andmeaida loomisel/laadimisel?" vastasid 81% anketeerituist, et nad kasutavad relatsioonilise andmebaasi andmeid (oli võimalik valida mitu vastust). XML andmeid kasutavad 35%, struktureeritud pärandandmebaaside andmeid - 32%, veebiteenuste andmeid - 27%, veebi logisid - 11%. Video/audio andmeid kasutab üks vastaja, voogandmeid ei kasutata.

Analoogilisele küsimusele tuleviku kohta ("Milliseid andmeallikaid/tüüpe Teie arvates võiks tulevikus kasutada Teie valdkonna andmeaidas/andmeaitades ärianalüüsi tegemiseks?") vastates nähakse relatsiooniliste andmebaasi andmete osatähtsuse vähenemist ja XML andmete ning veebi logide osatähtsuse suurenemist. Suureneb ka video/audio andmete kasutamine (5 vastajat). Kolm anketeeritut kavatsevad hakata kasutama voogandmeid (vt Joonis 20).

Seega on andmeaitade andmeallikate seisukohast jälgitav tendents andmebaasidest võetud andmete osatähtsuse vähenemisele ning erinevatest andmeallikatest pärinevate osaliselt mittestruktureeritud andmete suuremale kasutusele.

Ohud on seotud eelkõige infoturbe (sealhulgas privaatsuse tagamine ja identifitseerimine) ning andmete kvaliteediga (madala kvaliteediga andmed võivad kaasa tuua valed juhtimisotsused), mida on analüüsitud jaotises 4.2.



Joonis 20 Andmeallikate tüübid praegu ja tulevikus

4.5.2 Tehnoloogia

Tehnoloogilisi lahendusi käsitleti jaotises 4.4. Analüüsime siinkohal veel andmeaitade tarkvara kasutamist ja rahuolu sellega. Küsimusele "Millist andmeaida tarkvara kasutate?" vastasid ligi pooled küsimusele vastanuist, et kasutavad erinevaid Sybase tarkvara variante, järgnesid Oracle (19%), SAP tarkvara (10%) ja Microsofti (6%) tarkvara. Mitmesuguseid kombinatsioone eelpoolmainitud ja muudest tarkvaraplatvormidest kasutas 16% vastajatest.

Andmeaida halduseks kasutatava tarkvara platvormi omadustega (küsimus "Millisel määral olete rahul Teie andmeaida halduseks kasutatava tarkvara platvormi omadustega? ") ollakse valdavalt kas täiesti või osaliselt rahul - 98% vastustest üle kõigi kümne rahulolu valikute kategooria. Täiesti rahul ollakse 160 vastuses, osaliselt rahul - 91 vastuses. Kõige rohkem on täielikku rahulolu seoses andmete kiire laadimisega (54% vastajatest), integreeritavusega olemasolevasse IT keskkonda (57% vastajatest) ja ärianalüüsi süsteemide toetusega (51% vastajatest). Vaid neljal korral vastatakse "Pole rahul", kusjuures igast järgmisest valdkonnast on üks negatiivne arvamus: päringute jõudlus, tõrketaluvus, pilvearvutuse toetus, suurte andmemahtude toetus. Võib seega järeldada, et olemasolevate eesmärkide ja

andmetöötluse laadi raames on rahulolu andmeaida halduseks kasutatava tarkvara platvormi omadustega heal tasemel.

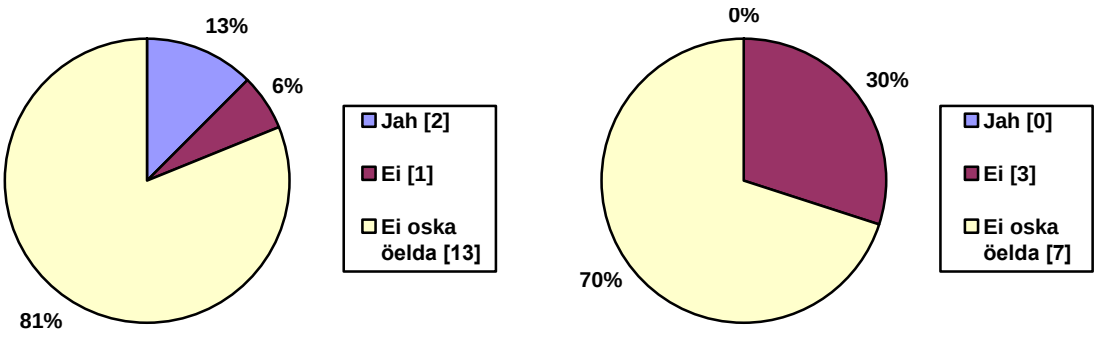
Samu rahulolu kategooriaid käsitles küsimus "Kui Teil oleks võimalus valida uus andmeaitade tarkvara platvorm, siis millised selle omadused oleksid olulised Teie valdkonna andmeaida jaoks?". Erinevate kategooriate peale kokku anti 264 vastust, millest 142 vastust oli "Väga tähtis", 92 vastust oli "Tähtis" ja 30 - "Mitte eriti tähtis". Kõige olulisemateks omadusteks (ükski vastaja ei hinnanud neid kui mitte eriti tähtsaid) osutusid päringute jõudlus, head administreerimisvahendid, tõrketaluvus, integreeruvus olemasolevasse IT keskkonda. Kõige sagedamini hinnati väga tähtsaks päringute jõudlust, kiiret laadimist, tõrketaluvust, integreeruvust olemasolevasse IT keskkonda ja ärianalüüsi süsteemide toetust. Kõige vähemtähtsamaks hinnati andmete kompressiooni (mitte eriti tähtsaks hindas 19% vastajatest), pilvearvutuse toetust (mitte eriti tähtsaks hindas 32% vastajatest) ja suurte andmemahtude toetust (mitte eriti tähtsaks hindas 19% vastajatest).

Kokkuvõttes võib järeldada, et ka perspektiivselt hinnatakse oluliseks pigem traditsioonilisi mittefunktsionaalseid omadusi - jõudlus, administreerimisvahendid, tõrketaluvus, integreeruvus, ärianalüüsi tugi; uusi võimalusi, nagu kompressiooni, pilvearvutuse toetust ja suurte andmemahtude toetust hinnatakse vähem oluliseks. See järeldus ei ühildu hästi ülaltoodud tendentsiga andmebaasidest võetud andmete osatähtsuse vähenemisele ning erinevatest andmeallikatest pärinevate osaliselt mittestruktureeritud andmete suuremale kasutusele. Erinevus võib olla tingitud osaliselt sellest, et teadlikkus vajadustest tekib kõigepealt, teadlikkus tehnoloogiatest võib võtta rohkem aega.

4.5.3 Konkurents

Konkurents on eriti oluline liikumapanev jõud ärimaailmas. Ankeetides uuriti, kas see toimib ka avalikus sektoris. Küsimusele "Kas andmeaitade tehnoloogia areng teistes riikides sunnib meie asutustes ka sellega tegelema?" vastas jaatavalt 27% anketeeritutest. Eitavalt vastajaid oli 38%, ülejäänud ei osanud öelda või ei vastanud.

Küsimusele "Kas Eesti peaks kasutama teiste riikide lahendusi ja regulatsioone avaliku sektori andmeaitade osas?" enamasti ei osatud vastata, vastati eitavalt või jäeti vastamata. Vaid kaks anketeeritut arvas otseselt, et teiste riikide lahendusi ja regulatsioone peaks kasutama (vt Joonis 21).



Joonis 21 Eesti peaks kasutama teiste riikide lahendusi ja regulatsioone avaliku sektori andmeaitade osas (riigi- ja erasektori spetsialistid)

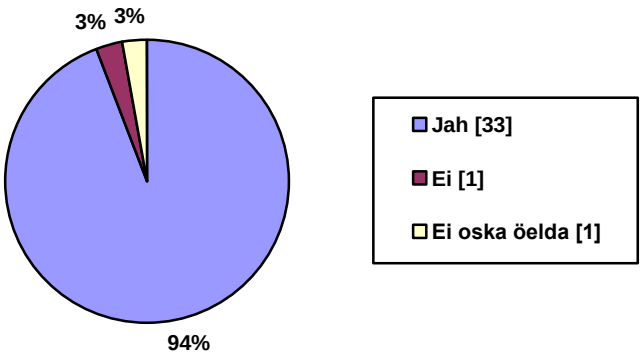
Märgiti, et Eestis on välja töötatud pädevad lahendused, et headest lahendustest võiks eeskuju võtta, et tehnoloogiad arenevad kiiresti ja kellegi vana süsteemi ülevõtmisesse tuleb suhtuda väga kriitiliselt ning et Eestis kehtivad regulatsioonid on võrreldes teiste riikidega eesrindlikud ja võimaldavad paremat integratsiooni kui mõne teise arenenud riigi seadusandlus - eeskujuks võib võtta parimaid praktikaid, aga mitte tervet seadusandlust tervikuna.

Kokkuvõttes võiks hinnata, et enamus vastanutest ei hinnanud teiste maade arenguid ja lahendusi andmeaitade valdkonnas olulise liikumapaneva jõuna.

4.5.4 Ühiskond

Ühiskonna mõju liikumapaneva jõuna iseloomustavad muuhulgas jaotises 4.3 vaadeldud teadmiste, motivatsiooni ja koostöö teemad. Näiteks, kuna andmeaidad võivad aidata kaasa otsuste vastuvõtmise kvaliteedi parendamisele, on ühiskonna toetus nende arendusele ja kasutamisele tugev.

Ankeetküsitluses uuriti veel, kas vastajate arvates soodustavad üldised suundumused ühiskonnas andmeaitade kasutamist ja andmeanalüüsi. (vt Joonis 22)

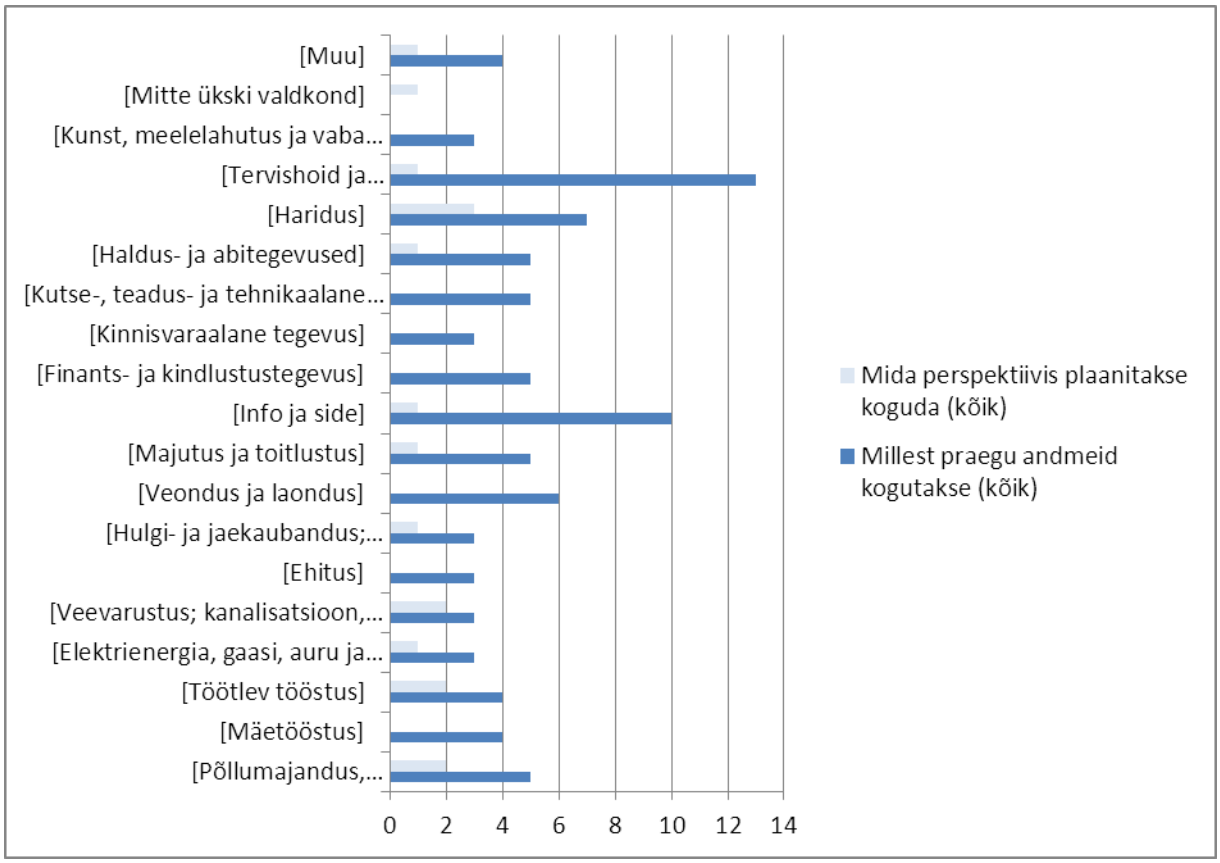


Joonis 22 Üldised suundumused ühiskonnas soodustavad andmeaitade kasutamist ja andmeanalüüsi

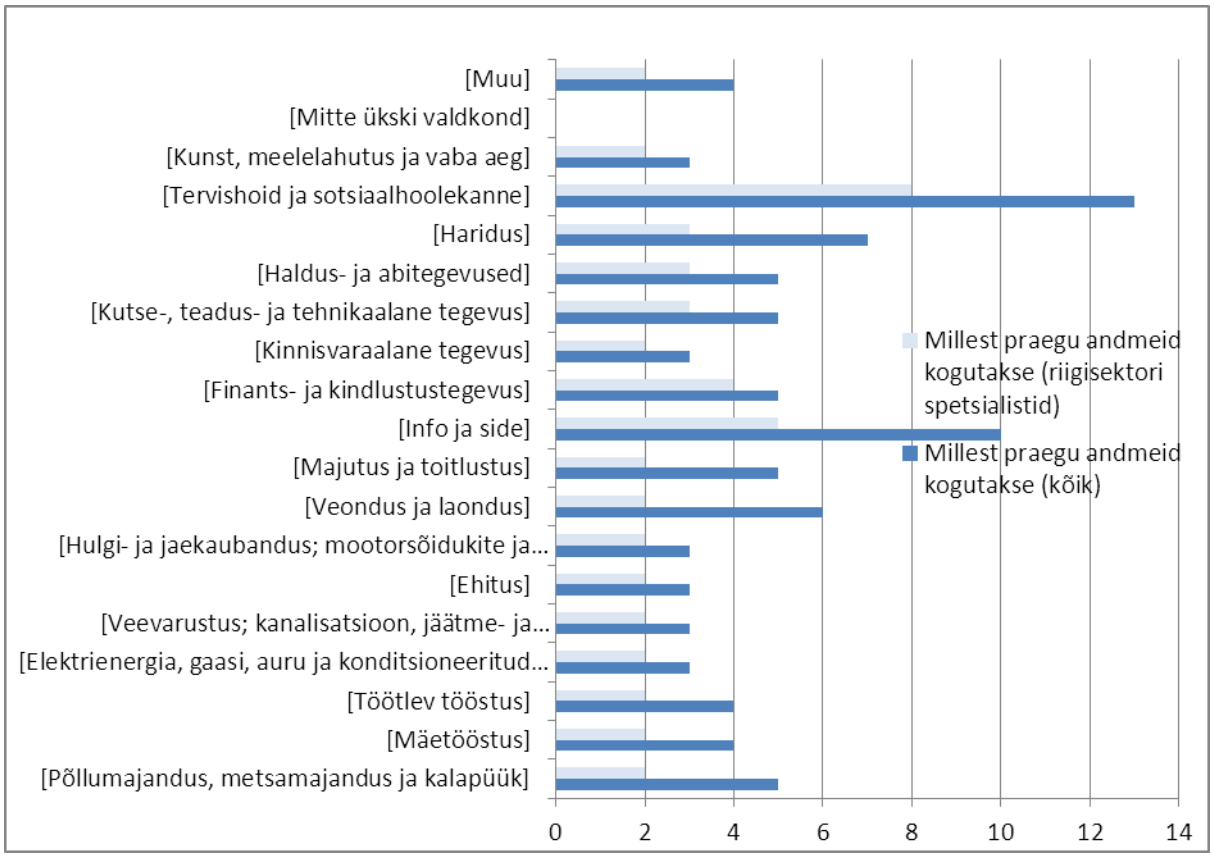
Märgiti, et andmeaidad võivad vähendada inimese koormust erinevatele küsitlustele vastamisel, samuti pakuvad teatud osas täpsemaid andmeid. Samuti avaldati arvamust, et ühiskond vajab ja soosib üha suurenevat andmeanalüüsi, samas pole kindel, kas seda tasub teha just aitade kaudu, vaid pigem vähemkohmakal viisil, sest andmeanalüüsi vajadused muutuvad üha detailsemaks ning neid vajadusi ette programmeerida (ja eriti järelprogrammeerida) on väga kulukas. Eitavalt vastanu kommenteeris oma vastust järgmiselt: "Andmeaitasid luuakse väga palju, kuid reaalselt on nende kasutus ikkagi väga madal. Andmete vastu puudub usaldus ning analüüsikiht on nende peal puudulik".

Kokkuvõttes võib järeldada, et ühiskonna mõju andmeaitade kasutuselevõtuks on vastanute arvates väga oluline. Samas võivad erinevad lähenemised olla erineva efektiivsusega: näiteks, kas olemasolevate andmeaitadega tööks on kõigepealt olemas ülesannete ja probleemide kompleks, mida loodav andmeait peab lahendama või on olukord vastupidine: kogutakse „igaks juhuks“ kokku mitmete andmekogude andmed ja alles seejärel hakatakse mõtlema, mida kogutud andmetega peale hakata.

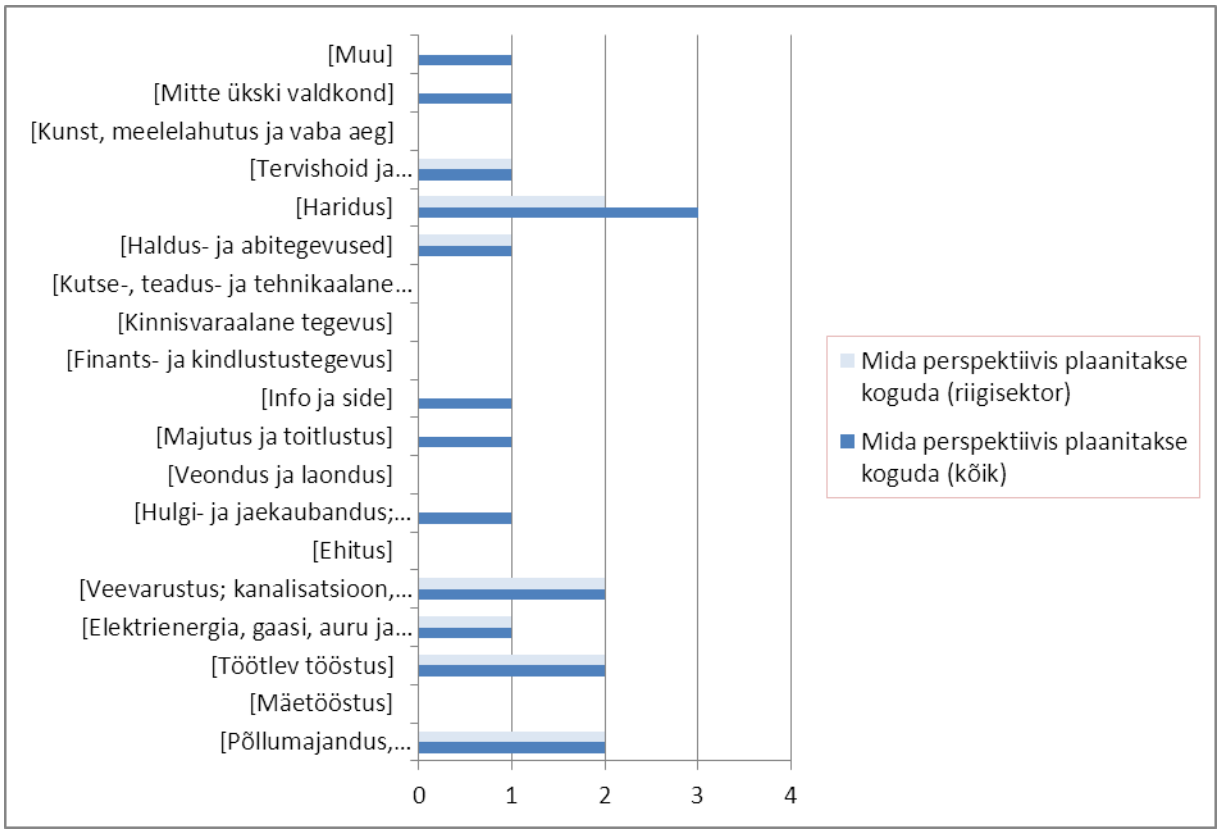
Käesolevas uurimuses küsiti anketeeritavatelt, et milliste valdkondade andmeid kogutakse andmeaitadesse praegu ja milliste omi planeeritakse koguda tulevikus (vt joonised 23, 24, 25). Kõigi küsitletute vastustest selgus, et praegu kogutakse peamiselt tervishoiu, info ja side, hariduse ning veonduse-laonduse valdkondade andmeid. Tulevikus planeeritakse peamiselt koguda lisaks hariduse, töötleva tööstuse ja põllumajanduse andmeid. Kui aga vaadelda eraldi riigisektori spetsialistide vastuseid, siis selgub, et riigisektori andmeaitadesse kogutakse eelkõige tervishoiu, info ja side, finants- ja kindlustustegevuse, hariduse ning halduse valdkondade andmeid. Valdkondi, millest andmeid üldse ei kogutud ei esinenud.



Joonis 23 Milliste valdkondade andmeid andmeaitadesse kogutakse praegu ja planeeritakse koguda tulevikus (kõik anketeeritud)



Joonis 24 Milliste valdkondade andmeid andmeaitadesse kogutakse praegu (riigisektor vs kogu valim)



Joonis 25 Milliste valdkondade andmeid andmeaitadesse planeeritakse koguda tulevikus (riigisektor vs kogu valim)

5 TEABEMATERJALIDE ANALÜÜSI TULEMUSED

5.1 TEADUS- JA TEHNOLOOGIATRENDIDE ANALÜÜS

Peamised trendid, mis moodustavad praeguse revolutsiooni andmeaitades ja andmekaevanduses, on spetsialistide arvates järgmised [McKinsey 2011]:

1. Läbilaskevõime tähtsuse oluline tõus.

Ühelt poolt suurenevad lähiajal oluliselt andmemahud, samal ajal soovib üha rohkem inimesi nii organisatsiooni seest kui ka väljastpoolt juurdepääsu nendele andmetele. See esitab kõrgendatud nõuded süsteemi läbilaskevõimele.

2. Reaalaja töötlus saab reaalsuseks.

On möödapääsmatu, et ettevõtted oma tegevuses hakkavad mõistma ja kasutama reaalajaandmeid.

3. Uute andmekeskuste (Data center) kasutamine on tugevalt lihtsustunud.

On tekkinud väga paindlikud meetodid süsteemide skaleeritavuse, mahtude ja reaktsiooniaja tõstmiseks.

4. Andmekaevandamine on lihtsustunud ja tema kasutamine laienenud.

On tekkinud üha laiem arusaamine andmekaevanduse võimalustest, samal ajal on andmekaevanduse meetodite kasutamine tunduvalt lihtsustunud.

5. „Big Data“ kasutamine hakkab filtreerima organisatsioone.

Juunis 2011 andis konsultatsioonifirma McKinsey and Company välja aruande [McKinsey 2011] „Big data: The next frontier for innovation, competition, and productivity“, milles ta ennustas, et järgnevatel aastatel ettevõtted upuvad andmevoogudesse. Samuti ennustas ta, et paljud, nagu näiteks, tervishoid, avalik sektor, tarneahelad ja tööstus, saavad suurt kasu nende andmete töötlusest. Seda liikumist tähistatakse terminiga „Big Data“.

6. Operatsioonilisi andmeid ja andmelao andmeid hakatakse koos töötlema.

Üha sagedamini tekib vajadus igapäevases töös kasutada värskeid koondandmeid andmelaost. Andmelao andmete kõrge kättesaadavus muutub üheks põhiliseks nõudeks.

7. Süsteemide läbilaskevõime suureneb järsult seoses uute mälude kasutuselevõtuga (Flash ja DRAM).

Käesolevas teadus ja tehnoloogiatrendide ülevaates vaatame lähemalt alljärgnevaid tehnoloogiatrende:

- "Big Data",
- NoSQL andmebaasid,
- virtuaalandmed.

Põhjendus selliseks valikuks on see, et ilmselt on "Big Data" kõige olulisem ja jõulisemalt areenile tulev käsitlus, mis mõjutab andmeaitade arengut. Kuna NoSQL liikumise põhiline motivatsioon on seotud just "Big Data" trendiga ja lisaks veel rakenduste efektiisusega, siis on selle valik samuti igati asjakohane. Virtuaalandmed on aga peaaegu paratamatu järeldus praegu toimuvale taristu tormisele arengule ja vajadusele otsustavalt tõsta andmeaitade läbilaskevõimet.

5.1.1 Big Data

„*Big Data*" on praegu analüütika spetsialistide põhitähelepanu all. „*Big Data*"-t võib defineerida, kui mahu, muutuvuse ja mitmekesisuse poolest suurt andmehulka, mida on võimatu efektiivselt töödelda ja hallata traditsiooniliste andmbaasisüsteemide abil.

Kolmeks „*Big Data*" põhiliseks omaduseks on seega:

- suur maht,
- muutuvus,
- mitmekesisus.

Tähtis on siin rõhutada, et raskuste tekitajaks ei ole siin vaid andmemaht , vaid ka informatsiooni allikad ja struktuur on oluliselt laienenud. Me saame üha rohkem infot erinevatest sensoritest, sotsiaalsete suhete informatsioon muutub üha kättesaadavamaks jne.

Info on väga erinevates formaatides, kaasaarvatud video, audio ja pildid, seega on info väga tihti mittestruktureeritud.

Andmeid tuleb töödelda vastavalt nende tekkimise ajalisele dimensioonile ja tihti reaalajas või reaalajalähedaselt.

5.1.2 NoSQL andmebaasid

NoSQL andmebaaside idee oli pakkuda lihtsat andmete haldamise mehhanismi, mis pakuks traditsioonilistest relatsioonilistest andmebaasidest paremat skaleeruvust ja käideldavust. Nagu juhtub tihti uute tehnoloogiate tulekuga, on praegu ka NoSQL andmebaasidega seoses suur segadus. Igaüks defineerib neid isemoodi, osadel on nendega seotud liialdatud lootused ja osa on kategooriliselt nende vastu. Nagu ikka elus on kõikidel osapooltel teatud määral õigus; kõik hea, mis selle lähenemisviisiga seotud, tuleb mingi hinnaga ja see hind ei pruugi igal juhul sobida.

Täpsustame NoSQL andmebaasidega seotud küsimusi. Esiteks tuleks aru saada NoSQL liikumise põhiolemusest, võrreldes teda just traditsioonilise andmebaasi käsitlusega. Traditsioonilises käsitluses on põhimureks rakenduste integratsioon ja terviklikus. Ühel andmebaasil saab realiseerida väga erinevaid rakendusi ja andmete terviklikus saavutatakse andmete normaliseerimisega. Normaliseeritud andmete korral on andmete uuendamine lokaliseeritud ja andmebaasis ei teki eriti lihtsalt vastuolusid. Mittenormaliseeritud andmete korral on tekivad vastuolud väga lihtsalt, kuna näiteks mingi isiku aadressi uuendamisel tuleb otsida üles kõik kohad, kus seda aadressi kasutatakse.

NoSQL liikumine tekkis just rakenduste loojate poolse initsiatiivina, kuna rakendustes kasutatavad struktuurid erinevad väga tihti andmebaasi tabelitest ja rakenduste loojad pidid pidevalt vaeva nägema vastavate teisendustega. Näiteks andmed ühe arve kohta paiknevad relatsioonilises andmebaasis mitmes tabelis ja kõiki neid peab eraldi käsitlema. Miks mitte hoida arvet kui ühte tervikut ja just sellisel kujul, kuidas teda rakenduses töödeldakse? Just see oligi esimene lähenemine NoSQL andmebaasidele. Loodi nn. võtme – väärtuse (*key-value*), tulba (*column*) ja dokumendi andmebaase, kus väärtuse ja dokumendi struktuur oli lähedane rakenduses kasutatavale.

Teiseks muudatuseks andmete käsitluses oli skaleeruvuse käsitlus. Traditsioonilist andmebaasi saab laiendada, suurendades kasutatavat serverit, kõvaketast ja mälu – niinimetatud skaleerimine üles.

NoSQL korral kasutatakse niinimetatud skaleerimist välja - kui andmete hulk suureneb, võetakse kasutusele uusi servereid (väga tihti virtuaalseid) või mälumassiive.

Suur muudatus on ka andmemudeli käsitluses. Traditsiooniline andmebaas nõuab ühtset skeemi kõikidele andmetele. Skeemi muutmine on sellisel juhul tõsine probleem. NoSQLis on skeemivaba käsitlus. Näiteks võtme – väärtuse andmebaasis võivad väärtuste struktuurid olla erinevad.

5.1.3 Andmete virtualiseerimine

Otsuste vastuvõtmiseks vajaliku andmete kättesaamise tagamine on alati olnud raske. Täna, kus meil on tegemist andmemahtude olulise suurenemisega, heterogeensete andmeallikate ning struktuuridega ja kiiresti muutuvate andmetega, on see eriti raske. Andmelao süsteemid on siiani olnud vastus sellele väljakutsele. Traditsiooniline lähenemine andmelaole, kus kord päevas või veelgi harvem tehakse ülekandeid operatiivbaasidest andmelattu, hakkab jalgu jääma kasvavatele nõuetele.

Andmete virtualiseerimine on kogum meetodeid ja tehnoloogiaid, mis võimaldavad organisatsioonidel tõsta andmete kättesaadavuse efektiivsust. Selle asemel, et tõsta andmeid ühest andmekogust teise, tehakse nad kättesaadavaks virtuaalsete kogumitena – luuakse nn. virtuaalne andmeladu. Muidugi on seda kerge öelda ja mitte nii kerge teha. Praegu on see saanud võimalikuks tänu uute tehnoloogiate esilekerkimisele.

5.2 RAHVUSVAHELISE PRAKTIKA ANALÜÜS

5.2.1 Rahvusvaheline taust

Andmaks käesolevas uuringus läbiviidud intervjuude ja ankeetide tulemustele (vt peatükk 4) rahvusvahelist tausta ja osadele küsimustele ka võrdlusvõimalust, kasutame mõningate andmeaitade alaste rahvusvaheliste uuringute tulemusi. Kõigepealt refereerime lühidalt Unisphere Research poolt Oracle sõltumatute kasutajate grupi (IOUG) liikmete hulgas 2011. a. septembris läbiviidud andmelaonduse alase küsitluse tulemusi [McKendrick 2011].

Uuringus osalesid 421 IOUG liiget, põhiliselt andmehaldajad ja spetsialistid. Enamus küsitletutest töötasid tarkvara- ja tehnoloogiafirmades, riigiasutustes, hariduses, finantsasutustes, tervishoius ja telekommunikatsioonifirmades.

Selle uuringu põhilised järeldused olid järgmised:

- 66% firmadest kasutab andmelaondust kui põhilist alust oma ärianalüüsi rakendusteks, kuid paljudel juhtudel jäävad need rakendused isoleerituks ja pole hästi seostatud äriprotsessidega. Enamus rakendusi on loodud organisatsiooni siseselt põhiliselt analüütikute ja tippjuhtide jaoks. Ainult 33% küsitletutest vastas, et nende firmades võimaldatakse laiemat juurdepääsu turundus- ja müügiinformatsioonile, st valdkonnale, kus andmelaondus on tõestanud oma kasulikkuse.

- Andmeladude haldajad tegelevad üha keerulisema ja erilaadsema andmestikuga, mis sisaldab nii struktureeritud kui ka mittestruktureeritud andmeid ja mida genereeritakse üha suuremas mahus. Suurema osa küsitletute jaoks on andmeait kasutuses kui operatsiooniliste andmete hoidla, kuigi 42%. küsitletutest vastas, et nende andmeaita kasutatakse ka kui dokumentide hoidlat. 36 % küsitletutest ütlesid, et nad kasutavad andmeaitu info- ja tootmistehnoloogiaga seostatud andmete hoidmiseks ning sama palju küsitletutest salvestavad tekstifaile. Umbes pooled vastajatest kavatsevad oluliselt suurendada mittestruktureeritud andmete analüüsi lähema 5 aasta jooksul. Peaaegu 90% küsitletutest vastasid, et andmemahud on kasvanud ja nad eeldavad ka andmemahtude kasvamist tulevikus. 50% küsitletutest pole kindlad, kas nende andmelao lahendus on skaleeruv, rahuldamaks tuleviku nn ülisuurte andmekogumite töötluse nõudeid.

- Firmad jäävad oma andmelaonduse uuendamisplaanides konservatiivseks, kuid samas on näha trendi valmislahenduste (ka spetsiaalsete andmelao seadmete) kasutamise ja realiseerimise suunas. Et andmelao süsteem töötab tavaliselt eraldi tootmissüsteemist, siis umbes 40% küsitletutest soovivad näha rohkem nende süsteemide lähenemist ja andmeladude sulandumist olemasolevasse infrastruktuuri.

Selle uuringu läbiviijad on arvamusel, et andmeaitade kasutamine on muutuste lävel seoses ülisuurte andmemahtude tekkimisega ja otsustusprotsesside vajadusega reaalajalähedase või reaalajas informatsiooni järele. Seetõttu pakutakse selles uuringus nimetatud vajaduste rahuldamiseks välja järgmised põhitaktikad: õpi tundma uusi andmeaitade tehnoloogiaid, loo strateegia efektiivseks mittestruktureeritud andmete töötluseks (nt NoSQL andmebaasid), ava-andmeaitadepõhine analüütika paljudele kasutamiseks (näiteks läbi lihtsa kasutajaliidese), erguta otsustajaid iseseisvale tööle ärianalüüsi tarkvaraga, arenda välja strateegia andmeaitade integreerimiseks kogu organisatsiooni informatsioonilisse infrastruktuuri ja seo see äriprotsessidega.

Tehnilises mõttes räägitakse kaasajal rohkem andmeanalüüsi platvormist kui andmeaidast [Adrian 2010, Russom 2009]. Andmeaitade Instituut (The Data Warehousing Institute, TDWI) viis 2009. aasta mais läbi uuringu [Russom 2009] uue põlvkonna andmeaitade platvormidest. Andmeaida platvorm defineeritakse selles uuringus kui platvorm, mis koosneb ühest või mitmest serverist, operatsioonisüsteemist, andmebaasi haldussüsteemist, andmehoidlast ja võrgust. Andmeaida platvorm haldab andmeaita, mis koosneb metaandmetest, andmemudelist ja andmetest. Andmeait on loodud aruandluseks, infoanalüüsiks ja otsustusprotsessi tagamiseks. Andmeaita aga ei vaadelda selles kontekstis kui andmeaida platvormi enda osa.

Selle küsitluse eesmärk oli uurida just andmeaitade platvormide arengu uusi suundi ja aruande abil teavitada spetsialiste uute võimaluste tekkest ja nende kasutusaladest.

TDWI e-küsitlusele vastasid 417 andmeaitade spetsialisti, kelle seas ei olnud akadeemiliste organisatsioonide ega andmaitade tarkvaralahenduste müüjate esindajaid (nende vastused eemaldati). Seega koosnes valim 71% IT spetsialistidest, 23% konsultantidest ja 6% ärikasutajatest. Tegevusvaldkondadest olid põhiliselt esindatud finantsteenused, kindlustus, tervishoid, tarkvara ja telekommunikatsioon. Enamus vastajatest olid USA-st (53%) ja Euroopast (19%), esindades üsna ühtlaselt erineva suurusega organisatsioone ja firmasid.

Ka see uuring näitas andmemahtude kasvutendentsi: näiteks vastuseks küsimusele, milline on teie organisatsiooni andmemaht täna ja 3 aasta pärast, oli 10 TB andmeaitade omanike hulk kahekordistumas (17% -lt 34%-ni). Alla 500 GB suuruste andmeaitade omanike arv aga muutuks 21%-lt 5 %-ni.

Vaatleme meie uuringu kontekstiga seonduvat kahte huvitavat uue põlvkonna andmeaitade platvormide valikuvõimalust, mis esitati ka TDWI uuringus. Esimene neist on reaalaja andmeaidandus (Real Time Data Warehousing, RTDW). RTDW omab kõige suuremat prognoositud kasvutendentsi vaadeldud uute andmeaitade platvormide võimaluste hulgas. 92% küsitletutest vastasid, et nad kasutaksid seda võimalust lähema 3 aasta jooksul.

Reaalaja andmeaitade loomise võimalus on seotud teiste valikutega, nagu näiteks voogandmete laadimisega andmeaita, sest see on RTDW põhiline tingimus. Tavaliselt tähendab see värskete andmete hõivamist lähtesüsteemidest ja vajalike aruannete ja analüüside sagedast uuendamist andmeaidas. RTDW toetab ajatundlikku äripraktikat nagu näiteks operatsiooniline ärianalüüs, tellimisel tulemusjuhtimine jms. Seega peab selliseid praktikaid toetav andmeaida platvorm olema töös 24/7. Teine võimalus andmete ühendamiseks on see, kui lähteallikatest võetakse värskeid andmeid ainult siis, kui rakendus

või kasutaja seda nõuavad. Mõlemad suundumused näitavad võimaliku kasutuse kasvutrendi.

Teine tulevikku suunatud valikuvõimalus on teenustepõhise arhitektuuri ja veebiteenuste kasutamine. TDWI uuringu järgi vastasid 59% küsitletutest, et nad kasutaksid neid veebiteenuste tehnoloogiaid oma järgmises andmeaida lahenduses. Põhiliseks põhjuseks märgiti, et teenused muutuvad eelistatuks liidese tüübiks, sest neid on lihtne välja kutsuda ja kasutada erinevatel platvormidel võrreldes konkreetsete liidestega, mida pakuvad andme- ja rakenduste integratsiooni platvormid. Veebiteenused on kergesti taaskasutatavad ja vastavad andmekäitluse parimad tavad on juba välja kujunenud.

Vaadeldav uuring sisaldas ka nõuandeid, millest meile kasulikumad oleksid järgmised:

- Väldi andmeaida platvormi kokkupanemist ettevõttesiselt (kuigi 55% teevad seda praegu), vaid kasuta selleks pigem süsteemi integraatori või konsultandi abi.
- Kui andmemahud näitavad kiiret kasvutendentsi, siis oleks aeg valmistuda andmeaida lahenduse uuendamisele selleks, et tagada suurte andmemahtude töötlus.
- Jälgi järgmiste andmelaondusega seotud tehnoloogiate arengut: andmehaldus, andmekvaliteet ja andmete integratsioon.
- Eelda, et andmeanalüüs saab järgmise põlvkonna andmeaitade platvormide prioriteediks.

Gartneri järjekordse 2013. a. nn *Magic Qudrant* uuringu aruande järgi on tugevamateks platvormide pakkujateks Teradata, Oracle, IBM ja SAP [Gartner 2013]. Megatrendideks pakutakse loogiliste andmeaitade esilekerkimist, kuigi vastavate lahenduste pakkujaid on vähe. See mõiste viitab erinevate andmeaitade/andmeallikate loogilisele sidumisele üheks terviklikuks andmeaidaks, kusjuures iga üksik andmeait või andmeallikas jääb eraldiseisvaks andmehoidlaks. Teine suund on erinevate andmetüüpide (relatsiooniline/struktureeritud, poolstruktureeritud ja mittestruktureeritud) loogiline integreerimine. See on seotud ka ülisuurte andmemahtude integreerimisega; 10-14 andmeaida platvormide müüjat pakuvad selleks otstarbeks Hadoop-i liidest oma andmeaida tarkvarale.

5.2.2 Erinevate riikide praktika analüüs

Käesoleva uuringu raames analüüsime kolme välisriigi praktikaid seoses meie uuringu hüpoteeside ja küsimustikuga. Vaadeldavad riigid valisime ÜRO 2012. aasta e-riigi uuringu

edetabelis toodud e-riigi arenguindeksi järgi [UNPAN 2012]. Eesti (0.7987) on selles 20-ndal kohal, meie aga valisime riigid, mis kuuluvad oma kõrge indeksiga esimese 5 maa hulka selles tabelis. Need riigid on järgmised: Holland (0.9125) teisel kohal, Suurbritannia (0.8960) 3-ndal kohal ja Ameerika Ühendriigid (0.8687) 5-ndal kohal. Esimesel kohal oli Korea Vabariik (0.9283).

Iga vaadeldava riigi kohta toome eraldi välja andmete andmeaitadesse integreerimist puudutavate õiguslike regulatsioonide osa. Võimalusel uuritakse, kuidas ja millistes sektorites andmeaitu luuakse, analüüsime ka andmeaitade uusimaid trende ja suurte andmemahtude analüüsi mõju andmeaidandusele. Eraldi tuuakse välja iga maa parimad praktikad.

Nimetatud riikide praktika analüüs on tehtud kasutades erinevaid teabematerjale (põhiliselt saadaval veebis) ja e-intervjuusid, mis viidi läbi ühe esindajaga igalt nimetatud maalt. Intervjuude küsimused olid järgmised (tõlge Eesti keeles):

1. Kas teie riigis (eriti riigisektoris) on olemas eraldi seadusandlikud regulatsioonid andmeaitadesse andmete ühendamiseks?
2. Kas teie riigis loodavad andmeaidad pigem integreerivad erinevate lähteandmeallikate andmed ühte füüsilisse andmeaita või kasutakse virtuaalseid andmeaitu või andmeaitade koosvõime lahendusi jm tehnoloogiaid, mis ei nõua andmete laadimist ühte andmeaita?
3. Mis on uusim trend andmeaitade valdkonnas teie riigis? Kuidas seda mõjutab nn Big Data analüütika?
4. Milliseid häid soovitusi ja praktilisi kogemusi pakuksite Eesti avaliku sektori andmeaitade loomiseks?

Küsitlus esitati Hollandi, Inglismaa ja Ameerika Ühendriikide esindajatele (andmeaitade valdkonna asjatundjad ja pika praktikaga andmeaitade loojad), kes ka kõik vastasid.

5.2.2.1 Suurbritannia

Regulatsioonid. Andmekaitse seadus [DPA 1998] võeti Suurbritannia parlamendi poolt vastu 1998. a. See on põhiline seadus, mis reguleerib isikuandmete kaitset Suurbritannias ja on vastavuses EL andmekaitse direktiiviga aastast 1995 [EL andmekaitse direktiiv 1995], mis nõuab EL liikmesriikidelt inimestele õigust privaatsusele nende isikuandmete töötlemisel. 25.

jaanuaril 2012 avalikustas EL uue eelnõu, Euroopa Üldine Andmekaitse Seadus (European General Data Protection Regulation), mis hakkab asendama varasemat EL andmekaitse seadust. See seadus peaks asendama ka erinevates EL riikides kehtivad erinevad andmekaitse seadused [EL uus andmekaitse direktiiv 2012].

Meie intervjueeritav Suurbritanniast mainis ka, et põhiline andmeaitade valdkonna regulatsioon on andmekaitse seadus:

> „Otsestest regulatsioonidest andmete ühendamiseks andmeaitadesse pole kuulnud. Regulatsioonid on reeglina vastavale ärivaldkonnale kehtestatud Finantsjärelvalve (FSA), EU, Sarbanes-Oxley jne. poolt. Nendest regulatsioonidest tulenevad ka paljud nõuded aruandlusele, ajalooliste andmete säilitamisele, andmete töötlemisele jne."

Andmete integreerimine. Intervjueeritav ütles, et andmete integreerimiseks andmeaita on rõhuvaks trendiks endiselt ELT/ETL vahekiht algallikate ja andmelao vahel. Uurides Suurbritannia suuri andmeaitadel põhinevaid süsteeme, näiteks tervishoiu valdkonnas, võib samuti väita, et luuakse tsentraalseid andmeaitu, mis integreerivaid andmeid paljudest geograafiliselt hajutatud lähteallikatest. Tervishoiu süsteemis on põhiliseks teiseste kasutajate teenus (*The Secondary Uses Service*, SUS, http://www.ic.nhs.uk/sus). See on ühtne tervishoiu andmete repositoorium, mis tagab rahvusliku terviseteenuse (*National Health Service*) aruandluse ja andmeanalüüsi ning andmete visualiseerimise vajadused. SUS on andmeait, mis sisaldab patsientide andmeid, mis on anonüümitud või pseudonüümitud vastavalt kasutajate vajadustele. Rahvusliku terviseteenuse pakkujad saavad neid andmeid kasutada teisesteks juhtumiteks, st mitte otsese kliinilise ravi juhul. Seda repositooriumi saavad kasutada kõik tervishoiu teenuse pakkujad nii riigi kui erasektoris, kes saadavad oma andmeid sellesse repositooriumisse. Peale selle võivad kasutajateks olla ka vastavad riigi valitsusasutused või teised organisatsioonid. Ainult 3 indiviidi igast organisatsioonist saab loa SUS-i kasutada, juurdepääsuõigused on rollipõhised. Peale selle on loodud teisi tervishoiu andmeaitu: näiteks vaba juurdepääsuga haiglate episoodide statistika (*Hospital Episodes Statistics*, HES, www.hesonline.nhs.uk), mis sisaldab andmeid haiglatesse vastuvõtu kohta üle kõigi Inglismaa haiglate. Suurbritannia tervishoiu süsteemis on veel teisigi andmeaitu; näiteks ESR (www.electronicstaffrecord.nhs.uk), mille abil tehakse tervishoiu tööjõu statistikat ja analüüsi. See andmebaas koondab kogu Suurbritannia tervishoiutöötajate arvu, oskuste, palkade jms andmeid nii rahvuslikul, regionaalsel kui organisatsioonilisel tasandil.

Andmekaitse seaduse kohaselt saadetakse igale tervishoiutöötajale kiri, mis teatab, mis tüüpi andmeid tema kohta kogutakse ja kuidas neid kasutatakse.

Andmeaidanduse trendid. Meie intervjueeritav selgitas, et andmeaitade ärikriitilisuse tase Suurbritannias on viimasel ajal selgelt muutunud. Nõutavad reaalajas toimuvad laadimised ja rakenduste otsepäringud internetis on kindlasti andnud andmeaida mõistele teise varjundi kui see oli varem. Samuti on esile kerkinud vajadus nii poolstruktureeritud kui struktureerimata andmete töötluseks ja nende uute lahenduste integratsiooniks olemasolevate andmeaitadega.

Suurbritannia valitsuse ja rahvusliku auditi mõnede 2011. aasta aruannete ja uurimuste kohaselt on Suurbritannia IT infrastruktuuri trendideks virtualiseerimine, agiilne süsteemiarendus ja ärianalüüs [IT audit 2011]. Neil on ilmselt mõju ka andmeaitade loomisele. Serveri ja/või lauaarvuti virtualiseerimist kasutavad juba praegu paljud valitsusasutused: näiteks Londoni transport, Töö ja pensioni ministeerium, Keskkonnaministeerium, Rahvuspoliitika täiustamise ministeerium jm.

PASC aruanne soovitas valitsusel kasutada agiilseid meetodeid infosüsteemide loomiseks ja arenduseks. Soovitati kasutada ka pigem väiksemaid ja keskmise suurusega teenusepakkujaid, kes on paindlikumad kui suured müüjad. See peaks võimaldama ka kiiremini kasutusele võtta uusimaid tehnoloogiaid. Soovitati ka lihtsustada riigihanke protsessi ja vähendada lepingute mahtu ning õppida erasektorilt [McKenna 2011].

Rahvuslik riigikontroll (*National Audit Office*, NAO) soovitas oma "*Landscape Review*" dokumendis muuhulgas ka laiemat ja paremat ärianalüüsi kasutamist IKT kulude ja jõudluse hindamiseks valitsuses [McKenna 2011]. Ka mainiti, et valitsus on väga vähe süstemaatiliselt kasutanud ärianalüüsi võimalusi ja vahendeid selleks, et toetada oma kõrgetasemelisi otsuseid. Riigikontroll uurib ärianalüüsi kasutavust eelkõige otsustuste parendamiseks geo-infosüsteemide valdkonnas, Keskkonnaministeeriumi, Toidu- ja maaelu ministeeriumi ja Kaitseministeeriumi haldusalas.

Andmeaidanduse head tavad. Meie intervjueeritav Suurbritanniast arvas, et andmeaitade loomisel võiks lähtuda paljudest lihtsatest reeglitest, mis on oma paikapidavust erasektoris hästi tõestanud ja andis meile järgmised näpunäited:

- Tellijal olgu selge arusaam, kellele ja miks andmeladu luuakse, millistes otsustusprotsessides soovitakse lahendust kasutada.
- Olgu selge, kes on projekti tellija ja kes on kogu projekti omanik.

- Peaks selgelt sätestama tellija suhte andmete algallikate omanike suhtes.
- Peaks looma kõigile osapooltele üheselt mõistetava sõnastiku.

Need esmased reeglid määravad väga tihti kogu juurutuse kulu ja edukuse.

Avaandmed. Euroopa Liidu direktiivi avaliku sektori informatsiooni taaskasutuse alal [PSI-direktiiv 2003] rakendati Suurbritannias 2005. aastal [UK PSI 2005] ja vastav kesk- ja omavalitsuste avaandmete portaal (data.gov.uk) pakub praegu vabalt kasutamiseks 9234 andmestikku. Suurbritannias on ka hästi arenenud lingitud andmete loomine ja neile juurdepääsu tagamine (vt http://data.gov.uk/linked-data). Peale selle avas Suurbritannia kaardistamise amet Ordance Survey osa oma andmeid 2010. aastal luues vastava portaali (www.ordnancesurvey.co.uk/oswebsite/products/os-opendata.html).

Listpoint (www.listpoint.co.uk) viis läbi avaandmete uuringu 2012. a. (esitlus on saadaval nende kodulehel, vt ka [Listpoint 2012]). Sellele küsitlusele vastasid 1017 keskvalitsuse ja kohalike valitsuste, tervishoiu-, riigikaitse- ja politseiasutuste töötajat. Uuringu põhitulemus oli, et riigiametnikel puudub piisav teadlikkus avaandmetest ja riigi avaandmetega seotud plaanidest. 78% vastajatest ei tea riigi avaandmete plaanist ja selle kasulikkusest. Nad (66%) ei saa aru, mis on nende individuaalne roll riigi avaandmete vallas. Eksperdid (50%) tunnetavad, et erasektori juurdepääs andmestandarditele ja riigi andmekogumitele on vajalik uute töökohtade ja paremate teenuste pakkumiseks. Samas suur osa riigiametnikke ei näe, et andmete avamine erasektorile oleks prioriteet.

5.2.2.2 Holland

Regulatsioonid. Hollandis reguleerib andmete integreerimist andmeaitadesse ainult Hollandi Andmekaitse Seadus (Wet bescherming persoonsgegevens (Wbp) inglise keeles Dutch Data Protection Act). See akt võeti vastu 2000. aastal ja ta jõustus 1. Septembril 2001. a. Nimetatud seadus reguleerib igasugust personaalsete andmete kasutust alates andmete kogumisest kuni andmete hävitamiseni [Hof 2007].

Meie intervjueeritav ütles, et see akt on praktikas “oluliseks raskuste allikaks” andmeaitade loomisel. Praegu luuakse Hollandis keskset andmeaita, mis sisaldaks iga Hollandlase meditsiinilist informatsiooni ning seoses sellega on käimas väga ägedad andmete privaatsust puudutavad diskussioonid.

Andmete integreerimine. Andmete integreerimise kohta andmeaitadesse vastas meie intervjueeritav, et põhiliselt kasutatakse virtuaalseid andmeaitu, kus iga läheteallika omanik

vastutab oma andmekvaliteedi eest, võimaldades andmetele juurdepääsu nendele ametkondadele, kellel on õigus vastavaid andmeid vaadata. Intervjueeritav põhjendas virtuaalsete andmeaitade kasutamist sellega, et vajalike andmete omanikeks on tihti erinevad organisatsioonid ja tunduvalt lihtsam on saada juurdepääsuõigus andmetele kui luba kopeerida andmeid ühte suurde tsentraalsesse andmeaita. Ta ütles, et mõnikord nad proovivad andmeid erinevates lähteallikatest kopeerida (laadida) ühte andmeaita, tuues näiteks praegu Hollandis käimasoleva patsientide andmeaida projekti "*Electronic Patient Dossier*". Kahjuks olevat see projekt praegu viinud poliitiliste probleemideni ja saanud halva meediatähelepanu objektiks. Need asjaolud aga takistavad projekti lõpuleviimist.

Andmeaidanduse trendid. Intervjueeritava arvates pole Hollandis palju firmasid, kellel oleks vaja töödelda väga suuri andmemahtusid. Siiski on märgata olulist huvi selliste ülisuurte andmemahtude töötlemise tehnoloogiliste lahenduste vastu nagu Hadoop, SAP HANA ja ka Datavault.

Andmeaidanduse head tavad. Meie intervjueeritav soovitas vaadata ja uurida andmeaitade haldussüsteemi BIREADY andmeaitade loomiseks.

Avaandmed. Erinevalt Suurbritanniast pole Hollandis PSI direktiivi rakendamiseks ühtset poliitilist huvi. Erinevad avaliku sektori asutused võtavad ette erinevaid samme avaandmete osas, kuid kogupilt on fragmentaarne. Siseministeerium on pandud tegelema andmete avamisega taaskasutuse eesmärgil, töötades nii regulatsioonidega kui stimuleerides altpoolt tulevat initsiatiivi [Zijlstra 2010]. Avaandmete portaal (data.overheid.nl) avati Hollandis 2011. aasta septembris ja praegu on seal 5193 andmestikku avatud taaskasutuseks.

5.2.2.3 Ameerika Ühendriigid

Regulatsioonid. Ameerika Ühendriikides pole sellist keskset andmekaitse seadust nagu EL-s. Ka meie intervjueeritav vastas, et ta ei tea ühtegi spetsiaalset andmeaitade loomist ja kasutamist puudutavad seadust. Andmekaitse on Ameerika Ühendriikides pigem sektoripõhine ja isereguleeruv kui riigi poolt keskselt reguleeritud. Näiteks on Ameerika Ühendriikides seetõttu palju erinevaid isikuandmete privaatsust puudutavaid seadusi, mis reguleerivad privaatsust teatud valdkonnas (näiteks, videoprivaatsuse kaitse akt, kaabeltelevisiooni kaitse akt, krediidiasutuste aruandluse akt jms) [Privaatsuse seadused]. Privaatsuse Akt 1974. aastast (The Privacy Act) kitsendab küll personaalse informatsiooni kogumist ja levitamist riigiasutuste poolt, kuid ei kitsenda seda erasektoris. Samas Föderaalne Kaubanduskomisjon loodab, et tööstuse isereguleeruvus on kõige efektiivsem

moodus tagada õiglane infomatsiooni töötlemise tava. Iseregulatsioon baseerub Õiglase informatsiooni tava koodeksil (*The Code of Fair Information Practices*, FIP) aastast 1972 [FIP 1972].

Seetõttu on Ameerika Ühendriikides tekkinud situatsioon, kus riigiasutused ostavad andmeaitu omavatelt firmadelt (näiteks ChoicePoint) oma kodanike kohta käivat personaalset informatsiooni. CoicePoint omas 2002. aastal kokkuleppeid paljude föderaalametitega isikuandmete müümiseks [EPIC 2003]. Et avalike andmete avatuse aste sõltub riigi ja föderaalseadustest, siis võivad avalikuks osutuda väga erinevad isikut puudutavad andmed, mida koguvad oma andmeaitadesse erafirmad selleks, et neid taaskasutada. Kui andmed on juba avalikud, siis võib neid vabalt kasutada, näiteks ühendada andmeaita, agregeerida ja ka müüa. Selline äri on Ameerika Ühendriikides seaduslik ja viimasel ajal ka levinud. Samas on paljud arvamusel, et isikuandmete kaitset peaks tõhustama. Uus EL isikuandmete kaitse direktiiv hakkab mõjutama ka Ameerika Ühendriikide firmasid, kuna selle regulatsioonid laienevad ka kõigile välisfirmadele, kes töötlevad EL elanike andmeid.

Andmete integreerimine. Ameerika Ühendriikides on paljudes sektorites oma kesksed andmeaidad, mis pakuvad teisestele ja ka erinevate andmebaaside andmetele avalikku tasuta või tasulist juurdepääsu. Näiteks võib tuua Ameerika Ühendriikide keskkonnakaitse agentuuri keskkonna andmelao portaali (http://www.epa.gov/enviro/html/topics.html), mis koondab ühte kohta juurdepääsu erinevatele keskkonnaandmetele. Suurimaks andmeaidaks on 1996. aastal loodud tulumaksu teenistuse (*Internal Revenue Service*, IRS) andmeait (CDW), mis kogub tuludeklaratsioonide andmeid, mille maht on juba üle 1 Petabaidi [CDW 2012]. IRS CDW on loodud kasutades SAP Sybase IQ platvormi veerupõhist tarkvaralahendust ja spetsiaalseid ainult lugemiseks mõeldud mäluseadmeid (näiteks SATA) [Butler 2012]. On olemas näiteks ka tervishoiu ressursside ja teenuste administreerimise andmeait (*Health Resources and Services Administration Data Warehouse*, datawarehouse.hrsa.gov/), päikesekiirguse ja selle ennustamise andmeait (www.solardatawarehouse.com/Data.aspx) ja palju teisi mingi sektori andmete põhiseid andmeaitu. Meie intervjueeritav selgitas, et sõltuvalt valdkonnast ja andmeanalüüsi tüübist kasutatakse mõlemat lähenemist, nii virtuaalseid andmeaitu kui ka ühte kesksesse füüsilisse andmeaita andmete koondamist. Ta mainis, et andmete füüsiline integreerimine parendab analüüsi efektiivsust, kuid eeldab kõrgjõudlusega arvutusvõimsust ja spetsiaalseid mäluseadmeid. Virtualiseerimine on paindlik, kuid efektiivsust viib alla andmeanalüüsi ajal toimuv päringute haldus.

Andmeaidanduse trendid. Ameerika Ühendriigid esitasid väga jõuliselt oma suundumuse ülisuurte andmemahtude töötluse tehnoloogiate uurimisele, arendamisele ja kasutamisele nn *Big Data Initiative* raames [Whitehouse 2012]. Andmepõhine otsuste vastuvõtmine mängis väga suurt rolli ka president B. Obama tagasivalimisel presidendiks. Eduka valimiskampaania planeerimisel ja läbiviimisel kasutati ülisuurtel andmehulkadel (sisaldasid ka sotsiaalmeediast kogutud andmeid) põhinevate andmeanalüüside ja prognooside tulemusi [Scherer 2012].

Andmeaidanduse head tavad. Kahjuks ei osanud meie intervjueeritav Eestile midagi soovitada ja seega koondasime heade tavade alla Ameerika Ühendriikide tulumaksu andmeaida tegijate soovitused [Butler 2012], mis on järgmised:

- Loo andmeaida meeskond, mis koosneb erinevate täiendavate oskustega inimestest, näiteks IT ja andmeanalüüsi oskustega inimestest (statistikud, ökonomistid, teadlased, masinõppe eksperdid).
- Hoia fookus andmekvaliteedil isegi kui andmemahud kasvavad, sellega tagad kasutajate usalduse.
- Realiseeri õige keerukusega valitsemistava muutuste haldamiseks nii, et eeskirjad ja protseduurid ei lämmataks muutust ennast.
- Ära investeeri üleshaibitud tehnoloogiatesse, kui need ei vasta ärivajadustele.

Avaandmed. Ameerika Ühendriikides loodi föderaalvalitsuse avaandmete portaal (data.gov) 2009. aastal ja praegu on seal 373029 andmestikku, mis on avalikuks kasutamiseks. Eraldi portaal on loodud valitsuse kulutuste avalikustamiseks (vt recovery.gov, USAspending.gov), lisaks on veel teisigi andmete avamise initsiatiive nii energeetika, hariduse, avaliku julgeoleku kui tervishoiu (vt healthdata.gov) valdkondades [Obama 2011].

5.2.3 Eesti rahvusvaheliste uuringute ja praktika taustal

Tabel 3 võrdleb eespool toodud rahvusvaheliste uuringute, välismaise praktika analüüsi ja meie uuringu tulemusi põhilistes küsimustes ja trendides. Samas esitame ka soovitused Eestile.

Tabel 3 Rahvusvaheliste uuringute ja Eesti uuringu tulemuste võrdlus

Rahvusvaheline andmeaitade uuring või praktika	**Eesti andmeaitade uuring**	**Soovitused Eestile**
Paljude juhtudel jäävad ärianalüüsi rakendused isoleerituks ja pole hästi seostatud äriprotsessidega [McKendrick 2011].	Andmeaidad pole loodud toetama äriprotsesse ja andmeanalüüs pole sinna integreeritud [intervjuu erasektori spetsialistid]. Üle 90% anketeeritutest arvab, et andmeaidad võivad aidata kaasa otsuste vastuvõtmise kvaliteedi parendamisele [ptk 4.3.2].	Ärianalüüsi rakendused peaks juba projekti algatamisel planeerima toetamaks äri- ja otsustusprotsesse.
Paljud rakendused on loodud organisatsiooni siseselt põhiliselt analüütikute ja tippjuhtide jaoks [McKendrick 2011].	Sageli on andmelao kasutajatering väga kitsas. Meie arvates, peab andmelaol olema taga ikkagi kogu kasutajate fännklubi, vastutajad ja protsessid [intervjuu erasektori spetsialistid].	Tuleks oluliselt laiendada ärianalüüsi rakenduste kasutajate gruppe nii organisatsiooni sees kui väljaspool. Näiteks erineva tasemega juhid, analüütikud, statistikud, teadlased jt huvigrupid võiksid andmeloa analüüsi kasutada vastava kasutusõigusega või vabalt (sõltub andmetest).
Umbes pooled vastajatest kavatsevad oluliselt suurendada mittestruktureeritud andmete analüüsi lähema 5 aasta jooksul [McKendrick 2011].	Nähakse relatsiooniliste andmebaaside ning struktureeritud pärandandmebaaside andmete osatähtsuse vähenemist (46% ja 27%) ja XML andmete ning veebi logide osatähtsuse suurenemist (vastavalt 46% ja 22%). Suureneb ka vido/audio andmete kasutamine [ptk 4.5.1].	Mitte-struktureeritud andmete töötlemiseks võiks kasutada vajadusel NoSQL liidestamist olemasoleva andmeaida platvormiga.
40% küsitletustest soovivad näha rohkem andmeladude sulandumist olemasolevasse infrastruktuuri [McKendrick 2011].	Andmeaida platvormi integreeritavusega olemasolevasse IT keskkonda oli rahul 57% vastajatest [ptk 4.5.2].	Projekti ja tellimuse planeerimisel oleks vaja pöörata tähelepanu ka andmeaida platvormi integreeritavusele olemasolevasse infrastruktuuri.
Kasvav vajadus reaalaja või reaalajalähedaste andmeaitade järele	Eesti Statistikaamet on ühinemas EL SensusHub projektiga [SensusHub 2013], mille tulemusena tuleb meie statistikaametil	Vajadus kiiresti ja andmetepõhiselt otsustada tingib ka vajaduse reaalaja andmeaitade

[Russom 2009].	esitada rahvastikustatistika andmeid EL nõudmisel (on-demand) ja vajadusel vähemalt 10 min värskusega (nn pull technology). See on reaalajalähedane nõue andmete järele [intervjuu riigisektori spetsialistiga]. Anketeeritud märkisid, et andmeaitade kasutamine võimaldaks liikuda rohkem reaalajas info põhjal tehtavate ja terviklikumate otsusteni [ptk 4.3.2].	järele. Taoline vajadus on Eesti asutustel juba tekkinud ja sellest sõltuvalt tuleb üle minna uuele mõtteviisile andmeanalüüsi läbiviimisel ja vastavatele nn *Big Data* tehnoloogiatele.
Veebiteenuste üha suurem kasutamine erinevate andmete ja rakendusplatvormide liidestamisel [Russom 2009].	X-tee teenuseid kasutatakse andmete integreerimiseks juba praegu, küll mitte veel andmeaitade endi liidestamiseks.	Veebiteenuste kasutamine andmeaitade ja teiste andmeallikate liidestamisel ning andmete integreerimisel.
Andmemahtude prognoositav kasv [Russom 2009, McKendrick 2011, Ganter 2013].	Eestis pole eriti asutusi, kus oleks üle paari TB andmeid. Andmemahud pole tähtsad, pigem on tähtis, et päringud jookseksid talutava kiirusega. Üliioluline on kompressioonitehnika. Vastupidi üldlevinud ettekujutusele, et andmeladu, mis koondab mitu operatiivandmebaasi muutub mahult väga suureks, võimaldab just kompressiooni tehnikate kasutamine luua andmelao, mis on väiksem kui tema lähteallikate summa [intervjuu erasektori spetsialistiga] Ligi neli viiendikku anketeerituist vastas, et andmeaida maht jääb alla 10 TB piiri [ptk 4.3.3]. Alla 15% küsitletuist oli kursis nn *Big Data* analüüsi võimalustega.	Vajadusel andmeaida platvorm välja vahetada või kasutada Hadoop jt liideseid.
Eelda, et andmeanalüüs saab järgmise põlvkonna andmeaitade platvormide prioriteediks [Russom	Keerukamaid andmeanalüüse nagu näiteks andmekaevandus (30%), ennustav analüüs (60%), on-line analüüs (70%) teevad juba praegu suhteliselt paljud erasektori firmad. Riigisektoris vastavalt (6, 41 ja 41 %) teevad	Tõsta teadlikkust andmeanalüüsi võimalustest ja uutest trendidest riigisektoris.

2009].	neid praegu ja planeerivad tulevikus teha vastavalt 59, 29 ja 29% vastanutest (riigisektori spetsialistidest) [ptk 4.3.3].	
Andmeaitade loogiline sidumine on megatrend [Gartner 2013].	Intervjuude ja ankeetide põhjal selgus, et Eestis on valdav andmeait, mis füüsiliselt integreerib andmed lähteallikatest ühte hoidlasse analüüsi tarbeks.	Võiks laiendada praegust Eestis põhiliselt levinud klassikalist andmeaida käsitlust analüütilise platvormi ja virtuaalsete jt andmeaitade definitsioonide suunas. Soovitame ka andmeaitade/andmestike linkimist kasutades lingitud andmete standardeid (RDF, SPARQL jt).
Suurbritannia valitsus on väga vähe süstemaatiliselt kasutanud ärianalüüsi võimalusi ja vahendeid selleks, et toetada oma kõrgetasemelisi otsuseid [McKenna 2011].	Põhiprobleem, miks andmeaitade kasutamine juhtimisel on nõrk seisneb selles, et andmeait pole loodud juhtimisprotsessi toetavana [intervjuu erasektori spetsialistiga].	Soovitame laiemat ja paremat ärianalüüsi kasutamist kõigis valdkondades, kaasa arvatud riigi juhtimine (selleks sobib näiteks andmetepõhine otsuste vastuvõtmine)
Tuleks luua selged reeglid andmeaida loomise projekti tellimiseks ja läbiviimiseks [intervjueeritav Inglismaalt].	Riigisektoris on vähene teadlikkus andmeaitade valdkonna kohta. Ei ole reeglina selget ettekujutust sellest, et mis on andmeladu, kuidas seda ehitatakse, kuidas projekti läbi viia, kuidas vastavaid hankeid korraldada. Andmelao lahenduste hanked peaksid olema mõnevõrra spetsiifilised. Hangetel kohtab pealisehitusena tihti nõudeid nt objektorienteerituse vmt enda arendatud tarkvaralahenduste kohta. Tundub, et hanke tegijad ei tunne andmeaitade tehnoloogiat. Sageli tehakse ühe hankega analüüs ja realisatsioon. Analüüsi käigus aga alles selgub tegelik skoop ja tööde maht [intervjuu erasektori spetsialistiga]. Kaks kolmandikku anketeerituist vastas, et avaliku sektori asutused võiksid andmeaitade loomisel ja halduses õppida	Tõsta teadlikkust andmeanalüüsi ja seda võimaldavate tehnoloogiate osas. Lihtsustada riigihanke protsessi, tehes selle paindlikumaks ja kohandades andmeaida loomise tsüklitega. Erasektoris on andmeaitade tellimused palju dünaamilisemad, seega võiks riik õppida andmeaida projektide tellimist erasektorilt. Peaks selgelt sätestama tellija suhte andmete algallikate omanike suhtes. Tellida väiksematelt ja keskmise suurusega teenusepakkujailt, sest see peaks võimaldama kiiremini kasutusele võtta uusimaid

	Eesti erasektori praktikatest [ptk 4.3.3].	tehnoloogiaid.
Peaks looma kõigile andmelao projekti osapooltele üheselt mõistetava sõnastiku [intervjueeritav Inglismaalt].	Sellega (st nn *Big Data* analüütikaga) seonduvad ka andmete sisu kvaliteet ja andmete tähenduse kvaliteet ehk andmete semantika, mis vajavad korrastamist ja suurt tähelepanu [intervjuu erasektori spetsialistid]. Umbes 20% anketeeritutest on olnud probleeme oma valdkonna andmeaitade semantilise koosvõimega teiste andmekogudega (ka andmeaitadega). Märgitakse, et kui andmeaida koosseis ja semantika on ette antud, siis on suur töö leida semantiline kooskõla andmekogude andmetga, mis on andmeaidale andmete andjateks [ptk 4.3.3].	Vastavate inim- ja masinloetavate sõnastike (ontoloogiate) loomine.
Hoia fookus andmekvaliteedil isegi kui andmemahud kasvavad, sellega tagad kasutajate usalduse [Butler 2012].	Teades, milline on andmete kvaliteet, saab anda usaldatavuse protsendi aruannete jaoks. Andmeaida põhjal tehtud aruandluse andmekvaliteet on alati parem kui operatiivsüsteemide põhjal tehtud aruandluse andmekvaliteet. Andmeaida vastutavad ja volitatud töötajad saavad takistada valede andmete andmeaita sattumist [intervjuu erasektori spetsialist]. Suhteliselt väiksem oli rahulolu andmete sidususe, kättesaadavuse ja selgusega. Andmeaitade andmekvaliteedi probleemide põhilised allikad on kvaliteediprobleemid andmeaitade aluseks olevates andmekogudes [ptk 4.2.2].	Eriliselt tähtsustada andmekavaliteedi rolli andmeaitade andmeanalüüsi usaldusväärsete tulemuste saamiseks. Töötada välja meetmed/reeglid andmekvaliteedi tagamiseks andmekogudes ja muutuste haldamiseks.
Avaandmestike publitseerimine: Suurbritannia 9234, Holland 5193 ja Ameerika Ühedriigid 373029	Publitseeritud on 6 avaandmestikku, nendest osa näidised (vt opendata.riik.ee). Alates 1. aprillist 2013 jõustub Eestis EL PSI-direktiiv.	Alustada avaandmestike publitseerimist kõigis riigiasutustes. Töötada välja meetmed huvi tekitamiseks andmete avalikustamise vastu. Luua veel pilootrakendusi

andmestikku		avaandmete baasil ja publitseerida parimad praktikad. Alustada lingitud avaandmete loomist.
Suurbritannia eksperdid (50%) tunnetavad, et erasektori juurdepääs andmestandarditele ja riigi andmekogumitele on vajalik uute töökohtade ja paremate teenuste pakkumiseks. Samas suur osa Suurbritannia riigiametnikke ei näe, et andmete avamine erasektorile oleks prioriteet [Listpoint 2012**].**	Kaks kolmandikku küsitletutest vastasid, et nende valdkonna andmed võiksid olla kas täielikult või osaliselt avaandmed [ptk 4.3.3].	Tõsta riigiametnike teadlikkust avaandmetest, riigi plaanidest selles valdkonnas, nende rollist avaandmete tagamisel/kasutamisel ja avaandmete vajalikkusest.

5.2.4 Järeldused.

Välismaa ja Eesti praktika on paljuski sarnased. Näiteks on sarnasused järgmistes valdkondades: ärianalüüsi isoleeritus äri ja juhtimisprotsessidest, rakenduste kitsas kasutajaskond organisatsioonis ja väljaspool, mittestruktureeritud andmete üha suurenev kasutamine, suurenev nõudmine reaalajalahenduste järele, andmekvaliteedi tähtsustamine, andmete üheselt mõistetava sõnastiku vajadus, andmeaitade projekti tellimine ja läbiviimine.

Erinevused aga on järgmistes valdkondades: andmeaida ja analüütilise platvormi käsitlus, suurte või ülisuurte andmemahtude analüüsimine, erinevate (eriti keerukamate) andmeanalüüsi meetodite kasutamine, loogiliste andmeaitade trendid, avaandmete publitseerimine.

Huvitav on see, et vaid kaks anketeeritut arvas otseselt, et teiste riikide lahendusi ja regulatsioone peaks kasutama Eesti riigisektori andmeaitade loomisel. Samas enamus (74%) vastas sellele küsimusele, et „Ei oska öelda“, mis näitab rahvusvahelisse praktika mitte tundmist. Küsimusele „Kas andmeaitade tehnoloogia areng teistes riikides sunnib meie asutustes ka sellega tegelema?“ vastasid 41% riigi- ja erasektori andmeaitade spetsialistidest

eitavalt, kusjuures riigisektori spetsialistidest vastas eitavalt 53 %. Nagu aga nägime uuringu tulemustest on Eestil teiste riikidega palju sarnasusi andmelaonduse valdkonnas seda nii regulatsioonide kui lahenduste ning tulevikuvisioonide osas.

6 PAKUTAVAD LAHENDUSED JA ETTEPANEKUD

Järgnevas pakutakse ettepanekud selle kohta, mida peaks Eestis andmeaitade seadusandluse, kasutamise ja organisatsiooni ning tehnoloogia alal edasi tegema. Arvestatakse püstitatud uurimisküsimusi ja intervjuudes ja anketeerimisel saadud tulemusi ning arendatakse edasi eelmistes jaotistes läbi viidud analüüsi.

Lähtutakse kehtivast ja perspektiivsest Eesti ja Euroopa Liidu andmekogusid puudutavast seadusandlusest, vajadusel ka väljaspool EL olevate riikide vastavast seadusandlusest. Arvestatakse andmeaitade loomisega seonduvaid olulisi eesmärke, mis on toodud „Eesti infoühiskonna arengukava 2013" jaotises 4.3.1:

- paberivaba asjaajamise ja haldustoimingute ning menetluste automatiseerimiseks vajalikud muutused õiguslikus keskkonnas ning organisatsioonide juhtimises on analüüsitud ning rakendatud;
- viiakse läbi uuringuid, mis käsitlevad infoühiskonna arengu ning arendamisega seonduvaid aspekte nii majanduse, ühiskonna kui indiviidi tasandil;
- kasutatakse riigi infosüsteemis kogutud andmeid otsuste kvaliteeti tõstmiseks.

Võetakse arvesse riigi infosüsteemide koosvõime raamistikke, so veebide, infoturbe, tarkvara ja Eesti IT koosvõime raamistikke, milledest viimane näeb ette semantikatehnoloogiate arendamise ja kasutuselevõtu lähimate aastate jooksul, olles seotud andmeaitade kui andmekogumite koosvõime probleemidega. Arvestatakse teiste riikide kogemust andmeaitade loomise, halduse ja kaitse alal.

6.1 ANDMEAITADE SEADUSANDLUS JA STANDARDID

Jaotises esitatakse regulatsioonide täiendamise üldine SWOT analüüs, vaadeldakse konkreetseid probleeme andmeaitade regulatsioonide osas (andmete dubleerimine, isikuandmete kaitse, andmeaitade loomine ja sätestamine RIHAs) ning pakutakse võimalikke andmeaitade valdkonna üle võetavaid standardeid.

Soovitame tehtud SWOT analüüsi rakendada eriti nende andmeaitade probleemidest lähtuvate uute regulatsioonide osas, mis võivad lisanduda käesoleva uuringu soovitustele.

Käesolevasse jaotisesse on koondatud kõik peaasjalikult regulatsioonide kohta käivad soovitused (kuigi tänu temaatilisele grupeerimisele tuuakse kohati sisse ka organisatoorseid ja tehnoloogilisi aspekte).

6.1.1 Regulatsioonide täiendamise vajadused: üldine analüüs

Selles ja järgnevates jaotistes vastatakse küsimusele, milliseid regulatsioone oleks andmeaitade loomise, ülalpidamise ja omavahelise suhtlemisega seoses vaja muuta, täiendada või lisada ning mida peaks eelnevast lähtudes Eestis andmeaitade seadusandluse alal edasi tegema.

Intervjuude ja ankeetide analüüs näitas, et küsimustele uute regulatsioonide vajalikkuse kohta vastas jaatavalt üldjuhul vähem ning eitavalt - rohkem vastajaid. Muutmise peamised valdkonnad on seotud andmeaitade määratlemisega, andmete privaatsusega andmeaitades ning andmeaitade loomise ja haldamisega. Riigi esindajad leiavad intervjuudes, et seadusandlike regulatsioonide muudatused on vajalikud, erasektori esindajad seevastu ei näe vajadust muudatusteks.

Kokkuvõttes, intervjuude ja ankeetküsitluste hinnangud regulatsioonide muutmise vajadusele ei ole kaugeltki ühesed.

Nagu märgitud jaotises 4.1, jagunesid muudatuste ettepanekud intervjuudes ja ankeetides järgnevatesse põhigruppidesse.

- Andmeaidad tuleks määratleda regulatsioonide tasemel. Näiteks, luua normid andmeaitade loomiseks ja halduseks, tuua eraldi välja andmelaondus kui eraldi eesmärkidega ja ülesehitusega andmekogundus või tuua andmeaidad seaduses välja kui teatud eri klass andmekogusid, millel on rida spetsiifilisi omadusi.
- Tuleks sätestada andmeaidas paiknevate andmete privaatsuse nõuded. Näiteks, reguleerides selgemalt andmeaidas paiknevate andmete privaatsust, täpsustades riikliku statistika seaduse § 34 lõiget 3, või luues tervise infoga seotud andmete osas kindlustunde selles, kuivõrd eri andmeaitade andmeid võib siduda.
- Tuleks sätestada andmeaitade registreerimine ja haldamine, sh nende registreerimine RIHAs. Näiteks, defineerides RIHA määruses andmeaida kui ühe andmekogu liigi ja vastavalt sellele kehtestades RIHA eeskirjad andmeaitadele.

Muudatused võivad esmajoones olla vajalikud järgmistes seadusandlikes aktides.

- Avaliku teabe seadus (sh andmekogu määratlus § 43^1 ning keeld kasutada ühtede ja samade andmete kogumiseks eraldi andmekogusid § 43^3).
- Isikuandmete kaitse seadus (sh privaatsuse nõuded).
- Vabariigi Valitsuse määrus "Riigi infosüsteemi haldussüsteem" (sh andmeaitade registreerimine RIHAs).
- Riikliku statistika seadus (sh § 34 lõige 3).
- Muud seadusandlikud aktid vastavalt vajadusele.

Analüüsides regulatsioonide täiendamise vajadusi, tuleks arvestada, et ei saa luua kitsalt andmeaitadest lähtuvaid regulatsioone, vaid lähtuda tuleks üldistest eesmärkidest ja põhimõtetest, sealhulgas nii soovist arendada riigi pakutavaid teenuseid ja teha paremaid otsustusi kui ka vajadusest kaitsta inimeste privaatsust.

Lisanduvad regulatsioonid võivad tuua selgust andmeaitade kasutamise osas. Samas on võimalik, et regulatsioonide kavandamise ja kehtestamise käigus tekitatakse hoopis täiendavaid kitsendusi ja bürokraatiat.

Tehnoloogilise arengu tingimustes on andmeaitade määratlus pidevas muutumises. Kui sätestada andmeaidad seadusandluses kui teatud eri klass andmekogusid, millel on rida spetsiifilisi omadusi (*"taking the data to the analysis"*), siis selline definitsioon võib hoopis takistada tehnoloogilist arengut koosvõime, andmete virtualiseerimise, pilvetehnoloogia, mobiilsete tehnoloogiate ja suurte toorete andmete baaside suunas, mida vastavalt vajadusele kasutatakse ja järk-järgult kustutatakse (*"taking the analysis to the data "*).

Tuleks arvestada ka võimalikke eelseisvaid muudatusi seadusandluses, eelkõige käimasolevat Euroopa Komisjoni isikuandmete kaitse reformi (isikuandmete kaitse üldmäärust ja direktiivi) ning selle arutelu, sealhulgas Eesti seisukohti isikuandmete kaitset puudutavate Euroopa Komisjoni algatuste suhtes.

Kui Euroopa Komisjoni isikuandmete kaitse reform annab tulemusi, võib see oluliselt muuta andmeaitadega seotud regulatsioone. Kuna vastava arutelu tulemused ei ole teada, võiks sellega seotud regulatsioonide tekitamisega olla pigem ettevaatlik.

Otstarbekas on analüüsida võimalikke lisanduvaid kulutusi seoses uute regulatsioonidega ning kaaluda OIOO (üks sisse, üks välja) tüüpi metoodika rakendamist, kus uute regulatsioonide puhul lisanduvad kulutused kompenseeritakse olemasolevate regulatsioonide eemaldamisest tekkiva kokkuhoiuga.

Arvestades küsitletute erinevaid arvamusi, käimasolevat kohati vastuolulist diskussiooni ning erinevaid poolt- ja vastuargumente, esitame ülaltoodud andmeaitade regulatsioonide täiendamise ettepaneku SWOT (tugevused, nõrkused, võimalused, ohud) analüüsi abil.

Andmeaitade regulatsioonide täiendamise eesmärgid on järgmised.

- Arendada riigi pakutavaid teenuseid.
- Teha paremaid otsustusi.
- Jätkuvalt kaitsta inimeste privaatsust.
- Luua selgus selles, millisel määral on võimalik ja otstarbekas luua ühtede ja samade andmete kogumiseks eraldi andmekogusid.
- Luua selgus isikuandmete kasutamise osas andmeaitades.

Järgnevas tabelis on toodud andmeaitade regulatsioonide täiendamise võimalikud tugevused, nõrkused, võimalused ja ohud seoses püstitatud eesmärkidega.

	Kasulikud eesmärkide suhtes	Kahjulikud eesmärkide suhtes
Sisemised	Tugevused Andmeaitade lihtsam loomine. Selgus isikuandmete kasutamise osas. Parem ülevaade olemasolevatest andmeaitadest.	Nõrkused Regulatsioon osutub lõpptulemusena kitsendavaks. Tekivad lisatöö, -kulutused ja -bürokraatia andmeaitade loomisel ja haldamisel.
Välised	Võimalused Paremad teenused ja otsustused. OIOO metoodika kasutamine kulutuste kompenseerimiseks.	Ohud Oht privaatsusele, isikuandmete leke suurtest andmeaitadest. Euroopa Komisjoni isikuandmete kaitse reform tühistab regulatsioonide muudatused. Uued regulatsioonid osutuvad kitsendavaks uute tehnoloogiate suhtes.

Järgnevates alajaotistes toodud regulatsioonide täiendamise ettepanekute puhul on sellist analüüsi arvestatud, kuid iga konkreetse seadusemuudatuse puhul on soovitatav see analüüs lühidalt üle teha. Kindlasti tuleks SWOT analüüs läbi viia võimalike andmeaitade temaatikast lähtuvate seadusemuudatuste puhul, mis lisanduvad käesoleva uuringu ettepanekutele.

Ettepanek. Arvestada iga konkreetse regulatsiooni muutmisel ja tekitamisel andmeaitade jaoks nii sellise tegevuse tugevusi ja võimalusi kui ka nõrkusi ja ohte, muuhulgas tehnoloogilisi arenguid, eelseisvaid muudatusi seadusandluses, võimalikke lisakulutusi ja privaatsuse probleeme. Kindlasti tuleks SWOT analüüs läbi viia võimalike andmeaitade temaatikast lähtuvate seadusemuudatuste puhul, mis lisanduvad käesoleva uuringu ettepanekutele.

6.1.2 Andmete dubleerimine andmeaitades ja avaliku teabe seadus

Vaatame eraldi andmete dubleerimise probleemi andmeaitades. Vastavalt avaliku teabe seaduse § 43^3 lõikele 2 on keelatud asutada ühtede ja samade andmete kogumiseks eraldi andmekogusid. See säte võib põhjustada seda, et loodav andmeait on mõnikord ajutise iseloomuga ja kuulub likvideerimisele pärast aruannete ja andmetöötlusülesannete valmimist.

Analüüsime andmete dubleerimise probleemi andmeaitades avaliku teabe seaduse kontekstis. Avaliku teabe seaduse § 43^1 lõige 2 ütleb: "Andmekogus töödeldavate korrastatud andmete kogum võib koosneda ka üksnes teistes andmekogudes sisalduvatest unikaalsetest andmetest". See säte näitab, et tuleks teha vahet andmete töötlemisel ja nende kogumisel. Kui andmeid vaid töödeldakse (nagu andmeaitades, kus andmed moodustatakse teiste andmekogude baasil ning neid spetsiaalselt ei koguta), siis on dubleerimine lubatud.

Sama mõtet arendab edasi avaliku teabe seaduse § 43^6 lõige 2: "Andmete töötlemisel, mida kogub põhiandmetena teine riigi infosüsteemi kuuluv andmekogu, tuleb aluseks võtta vastava teise andmekogu põhiandmed". See säte kinnitab uuesti, et on lubatud andmete töötlemine teiste andmekogude andmete põhjal, sealhulgas andmeaitades.

Avaliku teabe seaduse § 43^3 lõikes 2 käsitletakse andmete kogumist. Avaliku teabe seadus ei määratle andmete kogumist otseselt, kuid seaduse mõttest, ülaltoodud sätetest ning andmeaida määratlusest (andmeait kui teisene andmekogu) võib järeldada, et andmeaitades ei toimu andmete (esmast) kogumist ja seega käesolevat lõiget ei tuleks andmeaitade puhul rakendada.

Nagu ikka seadusandluse "hallide alade" puhul annab lõpliku lahenduse kohtupraktika, kui peaks tekkima vajadus selle rakendamiseks. Seni pole teada ühtegi andmete dubleerimisega

seotud kohtukaasust. Segaduste vältimiseks on otstarbekas seadust täpsustada, et vältida vääritimõistmisi.

Kokkuvõttes, käesoleva uuringu hinnangul ei tohiks avaliku teabe seaduse § 43^3 lõikes 2 sätestatu takistada andmeaitade loomist ja pikaajalist haldamist. Muuhulgas näitab käesolev analüüs, et avaliku teabe seadus (sh § 43^1 lõige 2) hõlmab ka andmeaidad kui ühe andmekogude liigi. Kui dubleerimisega seotud regulatsioone kaaluda, tuleks seda teha andmekogude jaoks laiemalt, mitte vaid andmeaitade jaoks.

Ettepanek. Määratleda avaliku teabe seaduse tekstis ilmutatult andmete kogumise mõiste, tekitamaks ühese arusaamise sellest, et avaliku teabe seaduse § 43^3 lõige 2 ei puuduta andmete ülekannet teistest andmekogudest ning ei takista seega andmeaitade loomist ja kasutamist.

6.1.3 Andmeturbe probleemid andmeaitades ning isikuandmete kaitse seadus

Lisaks dubleeriva andmekogumise probleemile võib andmeaitade ajutine iseloom olla tingitud andmeturbe probleemidest, mida ei osata lahendada. Paljudel juhtudel on tegemist ka isikuandmete töötlusega, mis on eriti tundlik seoses asjaoluga, et kõikvõimalikele andmeturvamise meetoditele vaatamata säilib ikkagi oht, et isikuandmeid on võimalik kaudsete tunnuste baasil ära tunda.

Analüüsime lühidalt vajadust muuta kehtivat seadusandlust, lähtudes andmeturbe ja isikuandmete kaitse probleemidest andmeaitades. Lähtume seejuures üldistest eesmärkidest ja põhimõtetest, sealhulgas nii soovist arendada riigi pakutavaid teenuseid ja teha paremaid otsustusi kui ka vajadusest kaitsta inimeste privaatsust.

Andmeturbe probleeme käsitletakse eelkõige järgmistes seadustes ja määrustes.

- Avaliku teabe seadus.
- Isikuandmete kaitse seadus.
- Vabariigi Valitsuse määrus "Infosüsteemide turvameetmete süsteem".

Eelmises jaotises toodud analüüs näitas, et avaliku teabe seadus hõlmab ka teiseseid andmekogusid, sealhulgas andmeaitasid. Küsimus pole siis niivõrd selles, kas oleks vaja eraldi regulatsioone andmeaitade jaoks, kuivõrd selles, kas on vaja eraldi regulatsioone teiseste andmekogude jaoks. Vastavalt avaliku teabe seaduse § 43^6 lõikele 2 ning kehtivale praktikale

peavad väga paljud andmekogud kasutama teistes andmekogudes kogutavaid põhiandmeid. Konkreetse andmekogu puhul võib kogutavate põhiandmete osakaal töödeldavatest andmetest olla väike või puududa. Teiseses andmekogus põhiandmeid ei koguta - kõik töödeldavad andmed saadakse välistest andmekogudest.

Kas põhiandmete mittekogumine tekitab vajaduse eraldi regulatsiooni järele? Antud uuringu eesmärkide kohaselt peaks uued regulatsioonid lihtsustama andmeaitade asutamist ja pidamist, mitte tegema neid keerukamaks. Selles kontekstis tuleks küsida, kas põhiandmete mittekogumine võimaldab mingeid seadusandlusest tulenevaid nõudeid lihtsustada või vähendada.

Avaliku teabe seadus märgib põhiandmeid § 43^6 raames ning rakendussätetes 7. peatükis. Isikuandmete kaitse seadus ei käsitle põhiandmeid. Riikliku statistika seaduses puudutab põhiandmeid § 50 lõige 2 "Statistikaamet hindab 2011. aastal toimuva rahva ja eluruumide loenduse tulemuste põhjal andmekogude põhiandmete kvaliteeti ning teeb vajaduse korral andmekogu vastutavale töötlejale ettepanekuid andmete kvaliteedi parandamiseks". Vabariigi Valitsuse määruses "Riigi infosüsteemi haldussüsteem" käsitletakse põhiandmeid järgmistes paragrahvides: § 5 "Riigi infosüsteemi haldamise põhimõtted", § 7 "Andmekogu dokumentatsiooni kooskõlastamine", § 10 "Andmekogu registreerimine ning andmekogus kogutavate andmete koosseisu muutmise registreerimine RIHA-s" ning § 18 "RIHA andmekogude alamregister". Vastavalt ülalmainitud õigusaktidele lihtsustab põhiandmete mittekogumine põhiandmetega seotud tegevusi (näiteks, puudub vajadus põhiandmete kindlaksmääramiseks vastavalt avaliku teabe seaduse § 43^6 lõikele 3), kuid ükski ülalmainitud paragrahv ei võimalda muid lihtsustusi seoses põhiandmete puudumisega. See on ka arusaadav, sest näiteks isikuandmete kaitse vajadused ei sõltu sellest, kas andmeid kogutakse antud andmekogus põhiandmetena või saadakse välistest andmekogudest.

Teine oluline lihtsustus andmeaitade pidamisel oleks võimalik andmeaitade puhul, mis sisaldavad vaid agregeeritud andmeid ning ei võimalda tuvastada konkreetset isikut. Isikuandmete kaitse seadus § 4 ütleb: "Isikuandmed on mis tahes andmed tuvastatud või tuvastatava füüsilise isiku kohta, sõltumata sellest, millisel kujul või millises vormis need andmed on". Seega kui andmeaida andmed on agregeeritud ja ei võimalda isikut tuvastada, pole enamikku isikuandmete kaitse seaduse sätteid vaja rakendada ning eraldi andmeaitade kohta käivat regulatsiooni ei ole vaja.

Kokkuvõttes, käesoleva uuringu hinnangul ei tohiks põhiandmete mittekogumine andmeaitades tekitada vajadust uue regulatsiooni järele. Kui sellist regulatsiooni kaaluda,

tuleks seda teha andmekogude jaoks laiemalt, mitte vaid andmeaitade jaoks. Tuleks vältida muudatusi regulatsioonides, mis tekitavad uusi kitsendusi.

Ettepanek. Vältida võimalust mööda lisanduvaid kitsendusi isikuandmete kasutamises, näiteks seoses käimasoleva Euroopa Komisjoni isikuandmete kaitse reformiga (isikuandmete kaitse üldmäärus ja direktiiv).

6.1.4 Andmeaida asutamine, registreerimine ja pidamine

Märgime kõigepealt, et andmeaida asutamine, RIHAs registreerimine ja pidamine on reguleeritud ka olemasoleva seadustiku raames. Tõepoolest, avaliku teabe seaduse § 43^1 põhjal on andmeait andmekogu, mida tuleb näiteks seaduses ettenähtud korras registreerida.

Nagu ka eelmises jaotises, tekib küsimus, kas põhiandmete mittekogumine andmeaitades õigustab andmeaitade asutamise, RIHAs registreerimise ja pidamise seadusandlike üldiste eriregulatsioonide kehtestamist. Analoogiliselt eelmises jaotises esitatuga võib sellele küsimusele vastata eitavalt.

Samas tuleb konkreetne andmeait asutada vastavalt avaliku teabe seaduse § 43^3. Muuhulgas, andmekogu asutatakse seadusega või selle alusel antud õigusaktiga (§ 43^3 lõige 1). Samas viitab § 43^3 lõikes 1 toodud õigusakti link Riigi Teatajas Vabariigi Valitsuse määrusele "Konsulaarametniku ametitoimingute andmekogu asutamise ja pidamise kord", mis antud kontekstis ei ole asjakohane. Seadusest ei tulene, millise taseme õigusaktiga tuleks andmekogu asutada.

Andmeaitade puhul tuleks lihtsustava tegurina arvesse võimalikult madala taseme õigusaktid. Vastavalt vabariigi valitsuse seaduse § 50 on näiteks ministri käskkiri õigusakt, samas ameti või inspektsiooni peadirektori käskkirja kohta seda ei väideta (§ 74). Haldusmenetluse seadus ei täpsusta, mis on õigusakt. Riigi Teataja seadusest võiks välja lugeda, et kõik õigusaktid avaldatakse Riigi Teatajas, samas täit kindlust selle kohta ei ole. Siit ettepanek täpsustada avaliku teabe seaduses, millised õigusaktid on andmekogu (sh andmeaida) asutamiseks lubatavad. Muuhulgas tuleks korrigeerida Riigi Teatajas avaliku teabe seaduse § 43^3 lõikes 1 toodud õigusakti linki. Õigusakti tase võiks olla minimaalselt selline, mis hõlmab andmeaita kuuluvate andmeallikate valdkondi.

Intervjuudes ja ankeetides jagunesid vastused andmeaitade registreerimise vajalikkuse küsimusele RIHAs ligikaudu võrdselt (12 registreerimise poolt, 13 vastu, 9 ei osanud öelda või

ei vastanud). See on ka arusaadav, sest nagu märgitud jaotises 4, on RIHAs registreerimisel nii positiivseid kui ka negatiivseid külgi.

Samas on praegu raskendatud avaliku teabe seaduse § 43^6 lõike 2 täitmine, sest RIHAst ei ole hetkel lihtsalt välja loetav, millised andmekogud milliseid põhiandmeid koguvad ja kuidas tuleks vastava andmekogu poole pöörduda. Sellel on mitmeid põhjuseid: semantilise koosvõime raamistiku ebapiisav rakendamine (näiteks, puuduvad valdkondade andmete kirjeldused, ja andmesõnastikud ja ontoloogiad), RIHAs olevate andmete ebatäpsus (näiteks, paljude andmekogude puhul puudub andmete koosseis) ning RIHA päringute ebapiisavus (pole võimalik teha päringuid põhiandmete kohta). Seega võiks soovitada edasist tööd kõigis nimetatud valdkondades.

Kokkuvõttes, käesoleva uuringu hinnangul ei tohiks põhiandmete mittekogumine tekitada vajadust uute üldiste andmeaida asutamise, registreerimise ja pidamise regulatsioonide järele. Kui selliseid regulatsioone kaaluda, tuleks seda teha andmekogude jaoks laiemalt, mitte vaid andmeaitade jaoks. Samas tuleks olemasolevad ja loodavad andmeaidad viia kooskõlla õigusaktidega. Otstarbekas on teha järgmist.

Ettepanek. Täpsustada avaliku teabe seaduses, millised õigusaktid on andmekogu (sh andmeaida) asutamiseks lubatavad (muuhulgas, korrigeerida Riigi Teatajas avaliku teabe seaduse § 43^3 lõikes 1 toodud õigusakti linki). Andmeaitade puhul tuleks lihtsustava tegurina kasuks võimalikult madala taseme õigusakti kasutamine. Seega võiks õigusakti tase olla minimaalselt selline, mis hõlmab andmeaita kuuluvate andmeallikate valdkondi.

Ettepanek. Peale eelmise ettepaneku elluviimist viia olemasolevad andmeaidad kooskõlla õigusaktidega, asutades need vastavalt avaliku teabe seaduse § 43^3.

Ettepanek. Võimaldamaks saada andmeaitade jaoks infot riigi infosüsteemi põhiandmete kohta, luua RIHAs põhiandmete kohta käivad päringud, täpsustada RIHAs olevaid andmeid ning rakendada järjekindlamalt semantilise koosvõime raamistikku. Soovitame inim- ja masinloetavate sõnastike (ontoloogiate) loomist andmeaita integreeritavate andmete tähendusest arusaamiseks.

6.1.5 Andmete kvaliteet ja andmeaitade omavaheline suhtlus

Andmekogu andmete kvaliteedi eest vastutavad andmekogu vastutavad ja volitatud töötlejad. Kuna andmeait liidab kokku erinevate andmekogude andmed, tekivad küsimused, kuidas jaguneb vastutus andmeaidas olevate andmete kvaliteedi osas, kuidas andmeaidad

omavahel peaksid suhtlema ning kas selles valdkonnas vaja lisada seadusandlusse uusi regulatsioone.

Intervjuude ja ankeetide järgi otsustades hinnatakse rahulolu andmete kvaliteediga andmeaitades paremaks, kui seda enne uuringut oleks võinud oletada (vt jaotis 4.2.2 ülal). Suhteliselt väiksem oli seejuures rahulolu andmete sidususe, kättesaadavuse ja selgusega. Anketeeritute arvates peaksid kvaliteedi eest vastutama vastavate andmekogude vastutavad ja volitatud töötajad. Küsitlus ei näidanud selget uute regulatsioonide vajadust andmeid ja nende kvaliteeti puudutava vastutuse sätestamise osas andmeaida tasandil.

Arvestades vajadust andmeanalüüsi usaldusväärsete tulemuste saamiseks, tuleks samas andmeaitade andmekvaliteedi rolli eriliselt tähtsustada nii seadusandlikul kui ka iga konkreetse projekti tasemel.

Olemasolevad õigusaktid vastutust andmete kvaliteedi eest ei sätesta. Avaliku teabe seaduse § 43^4 seab vastutavale ja volitatud töötlejale ülesanded, kuid need ei sisalda vastutust andmete kvaliteedi eest. Vabariigi Valitsuse määrus "Riigi infosüsteemi haldussüsteem" ei kehtesta nõudeid andmete kvaliteedile, kuigi neid on põhimõtteliselt võimalik RIHAsse sisestada - vastavalt määruse § 18 lõike 2 punktile 5 kuuluvad RIHA andmekogude alamregistrisse kantavate andmekogu üldandmete koosseisu andmekogu tehnilise kirjelduse dokumendid, kus sisalduvad andmekogu arhitektuuri, talitlusprotsessi, koosvõime nõuetele vastavuse, haldamise reeglite kirjeldused ja muud olulised andmekogu kohta käivad tehnilised kirjeldused ning andmekogu põhimäärus või selle kavand. Isikuandmete kaitse seaduse § 6 toob ära isikuandmete kvaliteedi põhimõtte, kuid ei sätesta vastutust muude andmete kvaliteedi eest. Riikliku statistika seadus sisaldab mitmeid andmete kvaliteediga seotud sätteid, mis käsitlevad eelkõige vastavust riikliku statistika kvaliteedikriteeriumidele.

On otstarbekas ilmutatult sätestada õigusaktides kooskõlastatud ja kontrollitavad nõuded andmete kvaliteedile.

Andmeaitade omavahelist suhtlust käsitleti ülal, sealhulgas dubleerimist käsitlevas jaotises. Vabariigi Valitsuse määruse "Riigi infosüsteemi haldussüsteem" § 5 lõike 1 punkt 6 sisaldab määratluse: "andmevahetuse teenusekeskuse põhimõte – andmevahetus (ristkasutus) erinevate andmekogude ja töötlejate vahel realiseeritakse andmeteenuste põhjal". Isikuandmete kaitse seaduse § 5 märgitakse: "Isikuandmete töötlemine on iga isikuandmetega tehtav toiming, sealhulgas isikuandmete kogumine, salvestamine, korrastamine, säilitamine, muutmine ja avalikustamine, juurdepääsu võimaldamine isikuandmetele, päringute teostamine ja väljavõtete tegemine, isikuandmete kasutamine,

edastamine, ristkasutamine, ühendamine, sulgemine, kustutamine või hävitamine, või mitu eelnimetatud toimingut, sõltumata toimingute teostamise viisist ja kasutatavatest vahenditest". See määratlus puudutab isikuandmeid. Üldist andmete töötlemise määratlust õigusaktides ei ole, kuid see oleks otstarbekas anda.

Küsimusele "Kas andmeaitade omavaheliseks suhtlemiseks on seadusandlusse vaja lisada uusi regulatsioone?" vastasid nendest anketeeritutest, kes selle kohta hinnangu andsid (vastasid kas "Jah" või "Ei") kokku ligi kolmandik jaatavalt, üle kahe kolmandiku eitavalt (jaotis 4.1.2).

Arvestades anketeeritute hinnanguid, dubleerimise kohta eelpool tehtud ettepanekuid, andmete töötlemise toimingute laia valikut (millest andmekogude vaheline suhtlemine on vaid üks komponent) ning seda, et põhiandmete mittekogumine andmeaitades ei tohiks tekitada vajadust uue regulatsiooni järele, ei soovita käesolev uuring luua uusi regulatsioone spetsiifiliselt andmeaitade omavahelise suhtlemise jaoks. Kui selliseid regulatsioone kaaluda, tuleks neid teha andmekogude ja andmetega tehtavate toimingute jaoks laiemalt, mitte vaid andmeaitade ja andmeaitade omavahelise suhtlemise jaoks.

Lähtudes ülaltoodud analüüsist tehakse käesolevas uuringus järgmised ettepanekud.

Ettepanek. Arvestades vajadust andmeanalüüsi usaldusväärsete tulemuste saamiseks, tuleks andmeaitade andmekvaliteedi rolli eriliselt tähtsustada nii seadusandlikul kui ka iga konkreetse projekti tasemel.

Ettepanek. Sätestada avaliku teabe seaduses ilmutatult vastutava ja volitatud töötleja vastutus andmekogu andmete kvaliteedi eest ja nõue kehtestada ning kooskõlastada kriteeriumid, mille alusel hinnatakse andmete kvaliteeti. Töötada välja meetmed/reeglid andmekogude andmekooseisude muutuste haldamiseks ja nendest teavitamiseks.

Ettepanek. Lisada Vabariigi Valitsuse määrusse "Riigi infosüsteemi haldussüsteem" RIHA andmekogude alamregistrisse kantavate andmete koosseisu kriteeriumid, mille alusel hinnatakse andmete kvaliteeti.

Ettepanek. Määratleda avaliku teabe seaduses või muudes õigusaktides andmete töötlus.

6.1.6 Statistilise üksuse tuvastamine ja riikliku statistika seadus

Riikliku statistika seaduse §34 lõige 3 ütleb: "Statistiline üksus käesoleva seaduse tähenduses on andmete alusel kaudselt tuvastatav, kui otsest tuvastamist võimaldavate tunnuste

puudumisel on võimalik statistilist üksust tuvastada muude andmete alusel. Et otsustada, kas statistiline üksus on tuvastatav, võetakse arvesse kõik võimalused, mida kolmas isik võib eeldatavasti kasutada nimetatud statistilise üksuse tuvastamiseks".

Toodud viide kõikidele võimalustele, mida kolmas isik võib eeldatavasti kasutada, võib töötleja või riikliku statistika tegija ette seada raskesti lahendatava ülesande. Võimalusi võib olla väga palju, eriti seoses andmekaevandamise ja sotsiaalse meedia andmete analüüsiga. Otstarbekas on võimaluste ringi kitsendada.

Ettepanek. Piiritleda riikliku statistika seaduse §34 lõikes 3 kasutatavad võimalused, näiteks kas kasutatavate võimaluste otstarbekuse kriteeriumi abil või loetledes tuvastamise võimaluste klassid.

6.1.7 Standardid

Enamus ankeetidele vastanud spetsialistidest leidis, et Eesti peaks kasutama rahvusvahelisi standardeid avaliku sektori andmeaitade osas. Põhjendusena toodi seda, et rahvusvaheliste kokkulepete järgimine tuleb üldiselt kasuks, et saab üle võtta häid praktikaid, et saab vältida liigset tööd ("jalgratta leiutamist") ning et standardite kasutamine loob aluse andmete ristkasutuseks.

Andmeladudega seonduvalt on muuhulgas kasutatavad kaks tervishoiu valdkonna standardit: ISO/TR 22221:2006 (Health informatics - Good principles and practices for a clinical data warehouse) ja ISO/TS 29585:2010 (Health informatics -- Deployment of a clinical data warehouse).

Standard ISO/TR 22221:2006 keskendub kliiniliste andmeaitade ja seotud teenuste loomisele, mis säilitavad kliinilisi andmeid või võimaldavad neile juurdepääsu teisese kasutamise eesmärgil. Standardi eesmärk on määratleda kliinilise andmeaida loomise, kasutamise, hoolduse ja kaitse põhimõtted ja praktikad.

Standard ISO/TS 29585:2010 käsitleb andmeaitade kavandamist ja loomist, andmete agregeerimist ja modelleerimist ning andmeaitade arhitektuuri ja tehnoloogiat. Kavandamise ja loomise osa keskendub andmeaitade eduka rakendamise nõuetele ja protseduuridele. Andmete agregeerimise ja modelleerimise osa käsitleb andmete valiku ja agregeerimise meetoditele, võimaldamaks edukat otsustamist. Arhitektuuri ja tehnoloogiate osa kirjeldab kliiniliste andmeaitade arhitektuure, andmekaevandamise meetodeid ja visualisatsiooni.

Tervishoiu valdkond on üks olulisemaid andmeaitade rakendusalasid. Kirjeldatud standardid võivad olla kasulikud ka teiste valdkondade andmeaitade arendamiseks. Arvestades samuti ankeetide vastajate toetust rahvusvaheliste standardite kasutamisele avaliku sektori andmeaitade osas, teeme ettepaneku kaaluda rahvusvahelise standardite kasutamist andmeaitade loomisel, vajadusel ka tõlkimist eesti keelde ning ülevõttu eesti standardiks. Selliste standardite näited on ISO/TR 22221:2006 (Health informatics - Good principles and practices for a clinical data warehouse) ja ISO/TS 29585:2010 (Health informatics -- Deployment of a clinical data warehouse), samuti rahvusvahelised koosvõime standardid.

Ettepanek. Kaaluda konkreetsete andmeaitade loomisel rahvusvaheliste, sh valdkondlike standardite kasutamist. Võimalusel tõlkida eesti keelde ning võtta üle eesti standardiks olemasolevad tervishoiu valdkonna standardid ISO/TS 29585:2010 ja ISO/TS 29585:2010.

6.2 ANDMELAONDUSE ARENDAMISE ORGANISATSIOONILISED ASPEKTID

Lisaks andmeaitade regulatsioonide kohta käivatele soovitustele (vt jaotis 6.1) ja avaliku sektori andmete andmeaitadesse kogumise, töötlemise ja väljastamise metoodilistele soovitustele (vt jaotis 6.3) käsitleme ka andmelaonduse organisatsioonilisi aspekte.

Ettepanekuid selle kohta, kes käesoleva uuringu soovitusi ellu viib ja kuidas seda tehakse, on osaliselt esitatud ettepanekute juures. Üldine põhimõte on, et ettepanekud viib ellu vastava tegevuse teostaja, jälgides seadusi, standardeid, häid praktikaid jne. Käesoleva uuringu tulemusena pakume täiendavalt välja järgmised andmelaonduse organisatoorse korralduse viisid, mis on koondatud lähedaste tegevuste gruppidesse.

Seadusandluse korrastamine vastavalt käesolevas uuringus toodud ettepanekutele. Selle tegevuse peaksid läbi viima Majandus ja Kommunikatsiooniministeerium, AKI, Riigikantselei, Justiitsministeerium, vastavalt vajadusele muud osapooled.

RIHA kontseptuaalne korrastamine (mõistete infosüsteem, andmekogu jms kooskõlastamine) ja päringute täiustamine. Nende tegevuste läbiviijaks võiks olla RIA, kes korraldab vastava riigihanke.

Erineva tasemega spetsialistide ja otsustajate teadlikkuse tõstmine ärianalüüsi võimalustest andmeaitade tehnoloogia abil. Selle eesmärgi täitmiseks on soovitatav korraldada koolituskursusi ja seminare (sealhulgas nt programmi „Infoühiskonna teadlikkuse tõstmine“

raames), töötada välja ja teha vabalt kättesaadavaks andmeaitade teemaline juhendmaterjal jne. Selliseid tegevusi võiks korraldada RIA eestvedamisel vastava hanke abil. Koolitused peaksid hõlmama erinevaid sihtgruppe (andmeaida kasutajad, nt juhid; andmeaida tellijad; valdkonna asjatundjad) ja katma vähemalt järgmisi teemasid.

- Kaasaegse andmeanalüüsi võimalused äri- ja otsustusprotsesside toetamisel, andmeaida tehnoloogiad ja nende uued suundumused (nt Big Data, virtuaalsed ja reaalaja andmeaidad jt).
- Andmeaitade ja teiste andmestike koosvõimetehnoloogiad (linkimine, veebiteenused jm).
- Andmeaitade riigihangete ettevalmistamine.
- Andmeaitade loomise metodoloogia, head tavad, erasektori ja rahvusvaheline praktika.

Meetmete väljatöötamine huvi tekitamiseks andmestike avalikustamise vastu kõigis riigiasutustes (nii esmased kui teisesed andmestikud ja ka analüüside tulemused). Avaandmestikud soodustavad suurel määral andmeaitade loomist. Andmeaitu on lihtsam luua hästi kättesaadavate avaandmete baasil. Seepärast soovitame alustada avaandmestike (eelkõige andmekogude, seejärel ka andmeaitade) publitseerimist kõigis riigiasutustes. Selle soodustamiseks võiksid RISO ja RIA korraldada infopäevi ja seminare. Sihtgruppideks võiksid olla eelkõige andmekogude ja andmeaitade omanikud. Seminarid peaksid tõstma riigiametnike teadlikkust riigi plaanidest avaandmete valdkonnas ja avaandmete kasulikkusest (sh seoses andmeaidandusega). Riigiasutused (andmete omanikud) peaksid kindlaks määrama erinevate ametnike rollid nende poolt hallatavate andmete avalikustamisel ja ka avaandmete kasutamisel. Andmete avalikustamisele peaks asutustes seadma kõrge prioriteedi. Asutustes tuleks luua pilootrakendusi avaandmete baasil, publitseerida parimad praktikad ning alustada lingitud avaandmete loomist.

Riigihangete läbiviimine. Lähtudes rahvusvahelisest kogemusest tuleks riigihanke läbiviimisel arvestada pakkumuskutse tehnilises kirjelduses andmeaida loomise tsüklitega. Seda peaksid tegema eelkõige andmeaida projekti osapooled, sätestades selgelt andmeaida projekti tellija suhte andmete algallikate omanikega.

Ärianalüüsi kättesaadavaks tegemine. Et teha ärianalüüs laiemalt kättesaadavaks, tuleks oluliselt laiendada ärianalüüsi rakenduste kasutajate gruppe nii organisatsiooni sees kui väljaspool. Näiteks erineva tasemega juhid, analüütikud, statistikud, teadlased jt huvigrupid

võiksid andmelao analüüsi kasutada vastava kasutusõigusega või vabalt (sõltub andmetest). Soovitame planeerida ärianalüüsi rakendused juba projekti algatamisel, toetamaks erineva tasemega äri- ja otsustusprotsesse ja seega rahuldamaks erinevate kasutajagruppide vajadusi.

Andmestike linkimine. Rahvusvahelist kogemust arvestades soovitame andmeaitadega tegelevatele osapooltele rakendada andmeaitade/andmestike linkimist, kasutades lingitud andmete standardeid (RDF, SPARQL jt) ja veebiteenuste kasutamist andmeaitade ja teiste andmeallikate liidestamisel ning andmete integreerimisel.

Järgnevad uuringud. Andmeaitade, *Big Data*, avaandmete, ärianalüüsi, semantika ja muud uuringus käsitletud teemad arenevad kiiresti edasi. Soovitame IT uute suundade kasutamisele pühendatud uuringuid tulevikus perioodiliselt läbi viia.

Ettepanek. Soovitame tõsta spetsialistide teadlikkust andmelaonduse valdkonnas ja selle tehnoloogia perspektiivide osas, korraldades koolituskursusi ja avaandmete teemalisi seminare, töötades välja ja tehes vabalt kättesaadavaks andmeaitade teemalisi juhendmaterjale jne. Ühtlasi võiks laiendada praegust Eestis põhiliselt levinud klassikalist andmeaida käsitlust analüütilise platvormi ja loogiliste/virtuaalsete jt andmeaitade käsitlustega.

Ettepanek. Et andmeaitu on lihtsam luua hästi kättesaadavate avaandmete baasil, siis soovitame alustada avaandmestike (eelkõige andmekogude, seejärel ka andmeaitade) publitseerimist kõigis riigiasutustes. Töötada välja meetmed huvi tekitamiseks andmete avalikustamise vastu. Luua pilootrakendusi avaandmete baasil ja publitseerida parimad praktikad. Alustada lingitud avaandmete loomist.

Ettepanek. Lihtsustada riigihanke protsessi, arvestades pakkumuskutse tehnilises kirjelduses andmeaida loomise tsüklitega. Andmeaida projekti käivitamise käigus peaksid osapooled selgelt sätestama andmeaida projekti tellija suhte andmete algallikate omanikega.

Ettepanek. Tuleks oluliselt laiendada ärianalüüsi rakenduste kasutajate gruppe nii organisatsiooni sees kui väljaspool. Näiteks erineva tasemega juhid, analüütikud, statistikud, teadlased jt huvigrupid võiksid andmelao analüüsi kasutada vastava kasutusõigusega või vabalt (sõltub andmetest). Soovitame ärianalüüsi rakendused juba projekti algatamisel planeerida toetamaks erineva tasemega äri- ja otsustusprotsesse ja seega rahuldamaks erinevate kasutajagruppide vajadusi.

Ettepanek. Vajadus kiiresti ja andmetepõhiselt otsustada tingib ka vajaduse reaalaja andmeaitade järele. Taoline vajadus on Eesti asutustel juba tekkinud ja sellest sõltuvalt tuleb vajadusel üle minna uuele mõtteviisile andmeanalüüsi läbiviimisel ja vastavatele nn Big Data

tehnoloogiatele.

Ettepanek. Soovitame andmeaitade/andmestike linkimist kasutades lingitud andmete standardeid (RDF, SPARQL jt) ja veebiteenuste kasutamist andmeaitade ja teiste andmeallikate liidestamisel ning andmete integreerimisel.

Ettepanek. Soovitame õppida erasektori ja välismaa praktikatest ning headest tavadest andmelaonduse, ülisuurte andmemahtude töötlemise ja analüüsi, avaandmete ja nende linkimise valdkondades.

6.3 ANDMEAITADE KASUTAMINE JA TEHNOLOOGIA

Käesolevasse jaotisesse on koondatud peaasjalikult andmeaitade organisatoorsete ja tehnoloogilise aspektide kohta käivad soovitused (kuigi tänu temaatilisele grupeerimisele tuuakse sisse ka regulatsioonide küsimusi).

Kuna andmeaitade kasutamise ja tehnoloogia kohta käivaid materjale on väga palju, keskendutakse järgnevas eelkõige intervjuudes ja anketeerimisel saadud Eesti kogemusele, süstemaatiseerides ja täiendades seda vastavalt maailma parimatele praktikatele.

6.3.1 Metoodilised soovitused andmete kogumiseks, töötlemiseks ja väljastamiseks

Metoodilisi soovitusi andmete kogumiseks, töötlemiseks ja väljastamiseks avaliku sektori andmeaitadest võib väga erinevalt liigendada, näiteks organisatsiooni, süsteemi etappide või tehnoloogia põhiselt. Käesolevas uuringus kasutatakse soovituste liigendamiseks standardi "EVS-ISO/IEC 12207:2009. Süsteemi- ja tarkvaratehnika. Tarkvara elutsükli protsessid (ISO/IEC 12207:2008)" protsesse. See standard ei nõua mingi konkreetse elutsükli mudeli kasutamist, küll aga tuleks mingi sobiv elutsükli mudel iga konkreetse projekti jaoks määratleda.

Käesolevas on aluseks võetud üldine andmeaida elutsükkel, mis koosneb järgmistest etappidest: algatamine, väljatöötamine, kasutamine ja hooldus. Seejuures võib kasutada mitmesuguseid elutsükli mudeli tüüpe, näiteks inkrementarendus, evolutsioonarendus, spiraalmudel jt.

Andmeaida algatamisel on ankeetides ja intervjuudes esitatud küsimusele "Kas Teil oleks häid soovitusi neile, kes alustavad oma valdkonnas andmeaida loomist?" antud vastuste põhjal soovitatav teha järgmist.

- Määratleda andmeaida loomise vajadus, sõnastades selle rolli, tähtsuse ja kasu tulevaste kasutajate jaoks ning nende kasutajate kaudu laiemalt loodavad hüved. See peab olema konkreetne, kasutaja ja huvitatud osapoolte poolt heaks kiidetud.
- Hinnata tasuvust (nt ROI või Cost/Benefit meetodil).
- Soovitatavad etapid: (1) läbida koolitus - andmelao loomise eesmärk ja metoodika; (2) õnneliku kliendi külastus, võimalusel ka vaadata lahendust, uurida edulugu; (3) fantaseerimine – piiranguteta soovide avaldamine, ajurünnakud jmt; (4) eesmärgi püstitus; (5) ülesandepüstitus. Need etapid võiks läbida soovitavalt koostöös potentsiaalse realiseerija või konsultandiga. See võiks toimuda enne hanget.
- Alustada koolitusest, et inimesed saaksid aru ja räägiksid ühest asjast.
- Määratleda vastutavad, aruandekohustuslikud, konsulteeritavad ja informeeritavad osapooled (nt vastutusmaatriks, *RACI chart*).
- Selgelt lahus hoida parenduslikud (kiirem, täpsem, kvaliteetsem, odavam analüüs) ja uuenduslikud (uued teadmised ja võimalused) aspektid.
- Hästi läbi mõelda, mida soovitakse andmeaidast kätte saada; samuti andmekoosseis, aruanded; kas on delikaatseid isikuandmeid; käideldavus, terviklus, konfidentsiaalsus; andmete puhastamine.
- Võimalusel on soovitatav jälgida kindlat andmete standardit.
- Tuleks eeldada, et nende andmekogude andmetega, mida andmeait kasutab, tuleb lahendada kvaliteediprobleeme. Selleks tuleks planeerida aega ja ressursse.
- Selgitada välja, kas andmaitade andmeid saab välistelt osapooltelt ja kui ei saa, siis kas probleem on tehniline (mida saab parandada) või on küsimus pigem usalduses.
- Uue andmekogu loomisel võiks juba mõelda andmeaida peale, pärast on selle liitmine keeruline või meelevaldne.
- Jälgida seadusandluses ettenähtud nõudeid andmeaida asutamisele, registreerimisele, turbele jne, võimalusel kasutada vastavaid standardeid (vt ka eelmine alajaotis).

<u>Andmeaitade väljatöötamisel, kasutamisel ja hooldusel</u> tuleks ankeetide, intervjuude ning muude materjalide põhjal soovitada järgmist.

- Jälgida mingit praktikas järele proovitud andmeaitade arendamise metoodikat.
- Alustada etapiviisiliselt (väikese mahuga).

- On vaja „tõlki“ IT ja äriinimeste vahel. Kasulikud on andmeaitade tehnoloogiate pakkujate ülevaated. Vajalik vaadata nii andmeaida tarkvara platvorme kui ka seda, kas meil pakutakse tuge ja milline on selle toe kvaliteet.

- Andmete laadimise käigus ei tohi teha andmete teisendusi. Andmed tuleks üheselt kanda operatiivsüsteemist andmeaita. Kõik puuduvate näitajate arvutused tuleb teha peale andmete ülekandmist andmeaidas.

- Kuna probleemid võivad olla andmete hõive ja kvaliteediga, kaaluda automaatset andmehõivet ja andmepuhastust.

Võimalik on kasutada muuhulgas järgmisi metoodikaid.

- Eestis on edukalt kasutatud Kimballi metoodikat, mille puhul andmeaida andmed koosnevad põhifaktide tabeli(te)st ning nendega seotud kontekstiandmeid sisaldavatest "dimensioonidest" (vt nt Ralph Kimball, Margy Ross. *The Data Warehouse Toolkit* , Second Edition, John Wiley and Sons, Inc, 2002).

- Maailmapraktikas on palju kasutatav ka normaliseeritud lähenemine, mille puhul andmed andmeaidas normaliseeritakse nagu tavapärases andmebaasis (vt nt William Inmon. *Building the Data Warehouse*, John Wiley and Sons, 2005).

- Kui on vaja kasutada ka struktureerimata andmeid (nt e-posti), siis võib kasutada spetsiaalseid tehnikaid selliste andmete kogumiseks ja kasutamiseks (vt nt William Inmon, Krish Krishnan. *Building the Unstructured Data Warehouse*, John Wiley and Sons, 2011).

- Eelmiste kõrval võib kasutada ka tootjate poolt pakutavaid metoodikaid. Eriti kehtib see siis, kui suund on pigem andmete virtualiseerimisele, suurte töötlemata andmekogumite kasutamisele ning andmete analüüsile, sest vastavad metoodikad on alles kujunemisel.

Ülaltoodud soovitused kehtivad ka siis, kui on vaja analüüsida vajadust andmeaitade andmehõive laiendamiseks ja uute andmeaitade loomiseks. Tõepoolest, andmeaida loomine ja pidamine on kulukas ja töömahukas ning õigustatud siis, kui tulud ületavad kulutusi. Selle poolest ei erine andmeait teistest andmekogudest. Seega saab andmeaida loomise otsuse teha asutus ise, lähtudes ülaltoodud soovitustest ning oma võimalustes ja vajadustest. Ankeetide ja intervjuude põhjal võib järeldada, et enamasti on ressursid andmeaitade loomiseks kas olemas või neid saab vajadusel hankida (vt jaotis 4.4), samas enamikus organisatsioonides jääb andmeaitade arv ka lähitulevikus põhiliselt vahemikku 1-3, üksikjuhtudel rohkem.

Eeltoodust võib järeldada, et andemaitade loomist spetsiaalselt laiendada ei ole vaja, küll aga on mõtet andmeaitade võimalusi ja tehnoloogiat laiemalt tutvustada, et asutused oleksid sellest teadlikud ning oskaksid vajadusel andmaida projekti algatada.

Ettepanek. Andmete kogumisel, töötlemisel ja väljastamisel avaliku sektori andmeaitadest on otstarbekas määratleda mingi sobiv elutsükli mudel. Näiteks võib aluseks võtta üldise andmeaida elutsükli, mis koosneb järgmistest etappidest: algatamine, väljatöötamine, kasutamine ja hooldus. Seejuures võib kasutada mitmesuguseid elutsükli mudeli tüüpe, näiteks inkrementarendus, evolutsioonarendus, spiraalmudel jt.

Ettepanek. Andmeaida algatamisel tuleks andmeaida kasutajate ja teiste osapoolte koostöös määratleda ja heaks kiita andmeaida loomise vajadus, sõnastades selle rolli, tähtsuse ja kasu tulevaste kasutajate jaoks ning nende kasutajate kaudu laiemalt loodavad hüved; hinnata tasuvust; läbida koolitus, mille esmane teema on andmelao loomise eesmärk ja metoodika; tutvuda edukate lahendustega; koguda soove, korraldada, ajurünnakud jmt; püstitada eesmärgid; püstitada ülesanne; määratleda vastutavad, aruandekohustuslikud, konsulteeritavad ja informeeritavad osapooled; hoida lahus parenduslikud (kiirem, täpsem, kvaliteetsem, odavam analüüs) ja uuenduslikud (uued teadmised ja võimalused) aspektid; läbi mõelda andmekoosseis ja aruanded; analüüsida, kas andmeaidas on delikaatseid isikuandmeid; spetsifitseerida käideldavuse, tervikluse, konfidentsiaalsuse nõuded; võimalusel jälgida kindlat andmete standardit; eeldada, et nende andmekogude andmetega, mida andmeait kasutab, tuleb lahendada kvaliteediprobleeme ning planeerida selleks aega ja ressursse; selgitada välja, kas andmaitade andmeid saab välistelt osapooltelt ja kui ei saa, siis kas probleem on tehniline (mida saab parandada) või on küsimus pigem usalduses; võimalusel mõelda andmeaida peale juba uue andmekogu loomisel; jälgida seadusandluses ettenähtud nõudeid andmeaida asutamisele, registreerimisele, turbele jne; võimalusel kasutada vastavaid standardeid.

Ettepanek. Andmeaitade väljatöötamisel, kasutamisel ja hooldusel tuleks jälgida mingit praktikas järele proovitud andmeaitade arendamise metoodikat; alustada etapiviisiliselt, väikese mahuga; soodustada kommunikatsiooni IT poole ja äriinimeste vahel, tutvuda andmeaitade tehnoloogiate pakkujate ülevaadetega; vaadata nii andmeaida tarkvara platvorme kui ka seda, kas meil pakutakse tuge ja milline on selle toe kvaliteet; soovitatavalt mitte teha andmete teisendusi nende andmeaita laadimise käigus - kõik puuduvate näitajate arvutused tehakse peale andmete ülekandmist andmeaidas; kuna probleemid võivad olla andmete hõive ja kvaliteediga, kaaluda automaatset andmehõivet ja andmepuhastust.

Ettepanek. Andmeaitade puhul tuleks valida sobiv metoodika, sh nt Kimballi metoodika, mille puhul andmeaida andmed koosnevad põhifaktide tabeli(te)st ning nendega seotud kontekstiandmeid sisaldavatest "dimensioonidest"; normaliseeritud lähenemine, mille puhul andmed andmeaidas normaliseeritakse nagu tavapärases andmebaasis; spetsiaalsed tehnikad struktureerimata andmete kogumiseks ja kasutamiseks; tootjate poolt pakutavaid metoodikad, eriti kui suund on pigem andmete virtualiseerimisele, suurte töötlemata andmekogumite kasutamisele ning andmete analüüsile.

Ettepanek. Andmeaitade andmehõivet ja uute andemaitade loomist spetsiaalselt laiendada ei ole vaja, küll aga on mõtet andmeaitade võimalusi ja tehnoloogiat laiemalt tutvustada, et asutused oleksid sellest teadlikud ning oskaksid vajadusel andmaida projekti algatada.

6.3.2 Isikuandmete kaitse ja andmete kodeerimine

Andmeaitade juurutamisel riigisektori infosüsteemides on oluline, millised andmed on vaja andmeaita sisestamiseks kodeerida (eelkõige isikuandmete kaitsest lähtuvalt) ja millised mitte.

Avaliku teabe seadus, Vabariigi Valitsuse määrus "Riigi infosüsteemi haldussüsteem" ja riikliku statistika seadus ei sätesta kodeerimise vajadust.

Isikuandmete kaitse seadus iseenesest ei keela isikuandmete töötlemist teisestes andmekogudes, sealhulgas andmeaitades. Vastavalt isikuandmete kaitse seaduse § 5 on isikuandmete töötlemine iga isikuandmetega tehtav toiming, sealhulgas isikuandmete kogumine, salvestamine, korrastamine, säilitamine, muutmine ja avalikustamine, juurdepääsu võimaldamine isikuandmetele, päringute teostamine ja väljavõtete tegemine, isikuandmete kasutamine, edastamine, ristkasutamine, ühendamine, sulgemine, kustutamine või hävitamine.

Isikuandmete töötlemisel tuleb jälgida vastavaid põhimõtteid (isikuandmete kaitse seaduse § 6, sealhulgas turvalisuse põhimõte) ja isikuandmete töötlemise lubatavust (sh isikuandmete kaitse seaduse § 10 kuni 14 sätestatu, nt andmesubjekti nõusolek või töötlemine avaliku ülesande täitmise käigus). Kui isikuandmete töötlemise nõuded on rahuldatud, siis pole vastavaid isikuandmeid vaja andmeaita sisestamiseks kodeerida.

Kodeerimise nõude sätestab isikuandmete kaitse seaduse § 16 lõige 1 isikuandmete töötlemisel teadusuuringu või riikliku statistika vajadusteks: "Andmesubjekti nõusolekuta

võib teadusuuringu või riikliku statistika vajadusteks töödelda andmesubjekti kohta käivaid andmeid üksnes kodeeritud kujul. Enne isikuandmete üleandmist teadusuuringu või riikliku statistika vajadustel töötlemiseks asendatakse isiku tuvastamist võimaldavad andmed koodiga. Tagasikodeerimine ja selle võimalus on lubatud ainult täiendavate teadusuuringute või riikliku statistika vajadusteks. Isikuandmete töötleja määrab nimeliselt isiku, kellel on ligipääs tagasikodeerimist võimaldavatele andmetele".

Sama paragrahvi lõige 2 räägib eelmisele lõikele vastu, tuues sisse ka olukorra, kus teadusuuringu või riikliku statistika vajadusteks on lubatud andmesubjekti nõusolekuta tema kohta käivate andmete töötlemine andmesubjekti tuvastamist võimaldaval kujul.

Sama paragrahvi lõige 4 sätestab: "Kogutud isikuandmeid on lubatud töödelda teadusuuringu või riikliku statistika vajadusteks, olenemata sellest, millisel eesmärgil neid isikuandmeid algselt koguti", täpsustamata, kas jutt on kodeeritud või kodeerimata andmetest, töötlemisest andmesubjekti nõusolekul või ilma nõusolekuta.

Need sõnastused on otstarbekas täpsustada.

Ettepanek. Kasutada andmeaitades kodeerimist isikuandmete töötlemisel andmesubjekti nõusolekuta teadusuuringu või riikliku statistika vajadusteks. Muudel juhtudel jälgida isikuandmete töötlemisel vastavaid põhimõtteid (isikuandmete kaitse seaduse § 6, sealhulgas turvalisuse põhimõte) ja isikuandmete töötlemise lubatavust (sh isikuandmete kaitse seaduse § 10 kuni 14, nt andmesubjekti nõusolek või töötlemine avaliku ülesande täitmise käigus).

Ettepanek. Kui andmeaitade jaoks on tihti vaja kasutada kodeerimist ka muul juhul kui isikuandmete töötlemisel andmesubjekti nõusolekuta teadusuuringu või riikliku statistika vajadusteks, kaaluda vastava täienduse tegemist isikuandmete kaitse seadusesse.

Ettepanek. Täpsustada ja muuta mittevasturääkivaks isikuandmete kaitse seaduse § 16, sealhulgas § 16 lõigete 1, 2 ja 3 sõnastused.

6.3.3 Andmeaitade tehnoloogia

Andmeaitade arendamiseks on vaja teadmist andmeaitade tehnoloogiast, sealhulgas sellest, milliseid infotehnoloogilisi vahendeid ja keskkondi peaks kasutama. Märgime kõigepealt, et spetsiaalsed andmeaitade tehnoloogiad ei pruugi alati olla vajalikud. Küsimusele "Kas erinevate andmekogude andmete ühendamiseks ühte andmeaita on vaja eraldi

tehnoloogilisi lahendusi?" vastati kolmandikul juhtudel eitavalt (jaotis 4.4). Kui on võimalik läbi ajada olemasolevate vahenditega, näiteks kui kasutatavad andmekogud on samal platvormil, siis on see efektiivne lahendus.

Siiski vastasid ligi pooled anketeerituist samale küsimusele jaatavalt (on vaja või mõnikord on vaja eraldi tehnoloogilisi lahendusi). Põhjenduseks märgitakse, et lähteandmebaasid on erinevate andmebaasimootorite peale ehitatud ning on vaja andmete erinevaid struktuure ühendavaid/tõlgendavaid liideseid. Lahendustena mainitakse kolmel korral X-tee lahendusi, kahel korral - ETL (*Extract, Transform and Load*) tööriistu / raamistikke. Paljudel juhtudel seega tuleks kaaluda andmeaitade loomiseks eraldi tehnoloogilisi lahendusi.

Tootjatest märgiti intervjuudes Sybase, Oracle, SAP ja Microsofti tarkvara, samuti mitmesuguseid kombinatsioone nendest ja muudest tarkvaraplatvormidest. Andmeaida halduseks kasutatava tarkvara platvormi omadustega ollakse valdavalt kas täiesti või osaliselt rahul. Arvestades seda tulemust ning seda, et tarkvara valik sõltub paljudest teguritest, sealhulgas ka lähteandmebaaside platvormidest, ei näe käesoleva uuringu autorid põhjust pakkuda mingi konkreetse andmeaitade tarkvara tootja eelistust.

Perspektiivselt hindasid anketeeritud kõige olulisemateks andmeaitade tarkvara omadusteks päringute jõudlust, häid administreerimisvahendeid, tõrketaluvust, integreeruvust olemasolevasse IT keskkonda ja ärianalüüsi süsteemide toetust. Kõige vähemtähtsamaks hinnati andmete kompressiooni, pilvearvutuse toetust ja suurte andmemahtude toetust. Arvestades tendentsi andmebaasidest võetud andmete osatähtsuse vähenemisele ning erinevatest andmeallikatest pärinevate osaliselt mittestruktureeritud andmete suuremale kasutusele, tuleks andmeaida tarkvara valikul siiski kaaluda ka neid omadusi.

Kuna andmete virtualiseerimine on andmeaitades üha rohkem kasutuses, tekib küsimus, millal peaks seda eelistama andmete koondamisele ühte kesksesse andmeaita. Jaotistes 5.1.3, 5.2.1 ja 5.2.2.3 (USA kogemus) esitatud kirjelduste põhjal võib pakkuda järgmised põhimõtted valikuks keskse või virtuaalse andmeaida tehnoloogiate vahel.

- Sõltuvalt valdkonnast ja andmeanalüüsi tüübist kasutatakse mõlemat lähenemist, nii virtuaalseid andmeaitu kui ka ühte kesksesse füüsilisse andmeaita andmete koondamist. Otsus tuleb teha iga andmeaida puhul eraldi ning see sõltub konkreetsest ülesandest, analüüsi vajadustest, lähteandmete kättesaadavusest jne.
- Andmete füüsiline integreerimine parendab analüüsi efektiivsust, kuid eeldab kõrgjõudlusega arvutusvõimsust ja spetsiaalseid mäluseadmeid.

- Virtualiseerimine on paindlik, kuid efektiivsust viib alla andmeanalüüsi ajal toimuv päringute haldus.
- Virtualiseerimist tasub kaaluda, kui andmeaidas on kiiresti muutuvad andmed, mida on vaja operatiivselt kajastada, kui on tegemist reaalaja andmeaitadega, kui on vaja kasutada voogandmeid jne.

Ettepanek. Lihtsamatel juhtudel kaaluda andmeaitade loomisel võimalust läbi ajada olemasolevate tehnoloogiliste vahenditega, näiteks kui kasutatavad andmekogud on samal platvormil.

Ettepanek. Kui lähteandmebaasid on erinevate andmebaasimootorite peale ehitatud või liidestamata, tuleks kaaluda andmete erinevaid struktuure ühendavaid/tõlgendavaid liideseid. Seejuures võib kasutada X-tee lahendusi, spetsiaalset andmeaitade tarkvara, ETL (*Extract, Transform and Load*) tööriistu jm.

Ettepanek. Kuna andmeaida halduseks kasutatava tarkvara platvormi omadustega ollakse valdavalt kas täiesti või osaliselt rahul, siis käesolevas uuringus ei pakuta konkreetse andmeaitade tarkvara tootja eelistust. Uute andmeaitade tarkvara valikul tuleks eelkõige arvestada omadusi, mida küsitlusel hinnati kõige olulisemaks: päringute jõudlust, häid administreerimisvahendeid, tõrketaluvust, integreeruvust olemasolevasse IT keskkonda ja ärianalüüsi süsteemide toetust. Arvestades tendentsi andmebaasidest võetud andmete osatähtsuse vähenemisele ning erinevatest andmeallikatest pärinevate osaliselt mittestruktureeritud andmete suuremale kasutusele, tuleks andmeaida tarkvara valikul kaaluda ka hetkel vähemtähtsamaks hinnatud andmete kompressiooni, pilvearvutuse toetust ja suurte andmemahtude toetust.

Ettepanek. Teha valik keskse või virtuaalse andmeaida tehnoloogiate vahel iga andmeaida puhul eraldi, sõltuvalt konkreetsest ülesandest, analüüsi vajadustest, lähteandmete kättesaadavusest jne.

7 KOKKUVÕTE

Uuringu põhihüpoteesiks oli, et riigi infosüsteemi andmeaitade valdkonnas on vajalikud nii poliitilised, seadusandlikud, organisatoorsed kui tehnoloogilised muutused.

Uuringu tulemused näitasid, et põhihüpotees pidas paika poliitiliste, seadusandlike ja organisatoorsete muutuste vajalikkuse osas. Tehnoloogiliste muutuste osas on pigem vaja jälgida tehnoloogia arengutrende äriprotsesside ja otsustusprotsesside muutuvate vajaduste rahuldamiseks.

Seoses muutuste vajadusega pakume välja terve rea lahendusi ja ettepanekuid nii poliitika, seadusandluse kui andmeladude organiseerimise valdkonnas. Ettepanekute juures on toodud ka soovitused selle kohta, kes vastavaid ettepanekuid peaks ellu viima. Üldine põhimõte on, et ettepanekud viib ellu vastava tegevuse teostaja, jälgides seadusi, standardeid, häid praktikaid jne. Seadusandlike regulatsioonidega seotud soovitused peaksid läbi viima Majandus- ja Kommunikatsiooniministeerium, AKI, Riigikantselei, Justiitsministeerium, vastavalt vajadusele muud osapooled.

Allpool esitatud soovituste ja ettepanekute koostamisel lähtuti anketeerimise analüüsi tulemustest, intervjuudel räägitust, teabematerjalide analüüsist, välismaisest kogemusest (ka välisintervjuude tulemustest) ja isiklikest ekspertteadmistest ning kogemustest.

Uuringus pakutavad lahendused ja ettepanekud on järgnevad.

- Arvestada iga konkreetse regulatsiooni muutmisel ja tekitamisel andmeaitade jaoks nii sellise tegevuse tugevusi ja võimalusi kui ka nõrkusi ja ohte, muuhulgas tehnoloogilisi arenguid, eelseisvaid muudatusi seadusandluses, võimalikke lisakulutusi ja privaatsuse probleeme. Kindlasti tuleks SWOT analüüs läbi viia võimalike andmeaitade temaatikast lähtuvate seadusemuudatuste puhul, mis lisanduvad käesoleva uuringu ettepanekutele.
- Määratleda avaliku teabe seaduse tekstis ilmutatult andmete kogumise mõiste, tekitamaks ühese arusaamise sellest, et avaliku teabe seaduse § 43^3 lõige 2 ei puuduta andmete ülekannet teistest andmekogudest ning ei takista seega andmeaitade loomist ja kasutamist.

- Vältida võimalust mööda lisanduvaid kitsendusi isikuandmete kasutamises, näiteks seoses käimasoleva Euroopa Komisjoni isikuandmete kaitse reformiga (isikuandmete kaitse üldmäärus ja direktiiv).

- Täpsustada avaliku teabe seaduses, millised õigusaktid on andmekogu (sh andmeaida) asutamiseks lubatavad (muuhulgas, korrigeerida Riigi Teatajas avaliku teabe seaduse § 43^3 lõikes 1 toodud õigusakti linki). Andmeaitade puhul tuleks lihtsustava tegurina kasuks võimalikult madala taseme õigusakti kasutamine. Seega võiks õigusakti tase olla minimaalselt selline, mis hõlmab andmeaita kuuluvate andmeallikate valdkondi.

- Peale eelmise ettepaneku elluviimist viia olemasolevad andmeaidad kooskõlla õigusaktidega, asutades need vastavalt vastavalt avaliku teabe seaduse § 43^3.

- Võimaldamaks saada andmeaitade jaoks infot riigi infosüsteemi põhiandmete kohta, luua RIHAs põhiandmete kohta käivad päringud, täpsustada RIHAs olevaid andmeid ning rakendada järjekindlamalt semantilise koosvõime raamistikku. Soovitame inim- ja masinloetavate sõnastike (ontoloogiate) loomist andmeaita integreeritavate andmete tähendusest arusaamiseks. Neid tegevusi peaks läbi viima RIA.

- Arvestades vajadust andmeanalüüsi usaldusväärsete tulemuste saamiseks, tuleks andmeaitade andmekvaliteedi rolli eriliselt tähtsustada nii seadusandlikul kui ka iga konkreetse projekti tasemel.

- Sätestada avaliku teabe seaduses ilmutatult vastutava ja volitatud töötleja vastutus andmekogu andmete kvaliteedi eest ja nõue kehtestada ning kooskõlastada kriteeriumid, mille alusel hinnatakse andmete kvaliteeti. Töötada välja meetmed/reeglid andmekogude andmekooseisude muutuste haldamiseks ja nendest teavitamiseks.

- Lisada Vabariigi Valitsuse määrusse "Riigi infosüsteemi haldussüsteem" RIHA andmekogude alamregistrisse kantavate andmete koosseisu kriteeriumid, mille alusel hinnatakse andmete kvaliteeti.

- Määratleda avaliku teabe seaduses või muudes õigusaktides andmete töötlus.

- Piiritleda riikliku statistika seaduse §34 lõikes 3 kasutatavad võimalused, näiteks kas kasutatavate võimaluste otstarbekuse kriteeriumi abil või loetledes tuvastamise võimaluste klassid.

- Kaaluda konkreetsete andmeaitade loomisel rahvusvaheliste, sh valdkondlike standardite kasutamist. Võimalusel tõlkida eesti keelde ning võtta üle eesti

standardiks olemasolevad tervishoiu valdkonna standardid ISO/TS 29585:2010 ja ISO/TS 29585:2010.

- Soovitame tõsta spetsialistide teadlikkust andmelaonduse valdkonnas ja selle tehnoloogia perspektiivide osas, korraldades koolituskursusi ja temaatilisi seminare, töötades välja ja tehes vabalt kättesaadavaks andmeaitade teemalisi juhendmaterjale jne. Koolituste raames võiks ühtlasi laiendada praegust Eestis põhiliselt levinud klassikalist andmeaida käsitlust analüütilise platvormi ja loogiliste/virtuaalsete jt andmeaitade käsitlustega. Koolituste läbiviimiseks võiks RIA korraldada riigihanke.

- Et andmeaitu on lihtsam luua hästi kättesaadavate avaandmete baasil, siis soovitame alustada avaandmestike (eelkõige andmekogude, seejärel ka andmeaitade) publitseerimist kõigis riigiasutustes. Töötada välja meetmed huvi tekitamiseks andmete avalikustamise vastu. Luua pilootrakendusi avaandmete baasil ja publitseerida parimad praktikad. Alustada lingitud avaandmete loomist. Selle soodustamiseks võiksid RISO ja RIA korraldada infopäevi ja seminare. Seminarid peaksid tõstma riigiametnike teadlikkust riigi plaanidest avaandmete valdkonnas ja avaandmete kasulikkusest (sh seoses andmeaidandusega). Riigiasutused (andmete omanikud) peaksid kindlaks määrama erinevate ametnike rollid nende poolt hallatavate andmete avalikustamisel ja ka avaandmete kasutamisel. Andmete avalikustamisele peaks asutustes seadma kõrge prioriteedi. Asutustes tuleks luua pilootrakendusi avaandmete baasil, publitseerida parimad praktikad ning alustada lingitud avaandmete loomist.

- Lähtudes rahvusvahelisest kogemusest tuleks lihtsustada riigihanke protsessi, arvestades pakkumuskutse tehnilises kirjelduses andmeaida loomise tsüklitega. Seda peaksid tegema eelkõige andmeaida projekti osapooled, sätestades selgelt andmeaida projekti tellija suhte andmete algallikate omanikega.

- Tuleks oluliselt laiendada ärianalüüsi rakenduste kasutajate gruppe nii organisatsiooni sees kui väljaspool. Näiteks erineva tasemega juhid, analüütikud, statistikud, teadlased jt huvigrupid võiksid andmelao analüüsi kasutada vastava kasutusõigusega või vabalt (sõltub andmetest). Soovitame ärianalüüsi rakendused juba projekti algatamisel planeerida toetamaks erineva tasemega äri- ja otsustusprotsesse ja seega rahuldamaks erinevate kasutajagruppide vajadusi.

- Vajadus kiiresti ja andmetepõhiselt otsustada tingib ka vajaduse reaalaja andmeaitade järele. Taoline vajadus on Eesti asutustel juba tekkinud ja sellest sõltuvalt tuleb vajadusel üle minna uuele mõtteviisile andmeanalüüsi läbiviimisel ja

vastavatele nn Big Data tehnoloogiatele. Soovitame selle teema lülitada koolituste programmi.

- Soovitame andmeaitade/andmestike linkimist kasutades lingitud andmete standardeid (RDF, SPARQL jt) ja veebiteenuste kasutamist andmeaitade ja teiste andmeallikate liidestamisel ning andmete integreerimisel. See teema peaks olema üks koolitusprogrammi osa.

- Soovitame õppida erasektori ja välismaa praktikatest ning headest tavadest andmelaonduse, ülisuurte andmemahtude töötlemise ja analüüsi, avaandmete ja nende linkimise valdkondades. Soovitame selle teema lülitada koolituste programmi.

- Andmete kogumisel, töötlemisel ja väljastamisel avaliku sektori andmeaitadest on otstarbekas määratleda mingi sobiv elutsükli mudel. Näiteks võib aluseks võtta üldise andmeaida elutsükli, mis koosneb järgmistest etappidest: algatamine, väljatöötamine, kasutamine ja hooldus. Seejuures võib kasutada mitmesuguseid elutsükli mudeli tüüpe, näiteks inkrementarendus, evolutsioonarendus, spiraalmudel jt. Soovitame selle teema lülitada koolituste programmi.

- Andmeaida algatamisel tuleks andmeaida kasutajate ja teiste osapoolte koostöös määratleda ja heaks kiita andmeaida loomise vajadus, sõnastades selle rolli, tähtsuse ja kasu tulevaste kasutajate jaoks ning nende kasutajate kaudu laiemalt loodavad hüved; hinnata tasuvust; läbida koolitus, mille esmane teema on andmelao loomise eesmärk ja metoodika; tutvuda edukate lahendustega; koguda soove, korraldada, ajurünnakud jmt; püstitada eesmärgid; püstitada ülesanne; määratleda vastutavad, aruandekohustuslikud, konsulteeritavad ja informeeritavad osapooled; hoida lahus parenduslikud (kiirem, täpsem, kvaliteetsem, odavam analüüs) ja uuenduslikud (uued teadmised ja võimalused) aspektid; läbi mõelda andmekoosseis ja aruanded; analüüsida, kas andmeaidas on delikaatseid isikuandmeid; spetsifitseerida käideldavuse, tervikluse, konfidentsiaalsuse nõuded; võimalusel jälgida kindlat andmete standardit; eeldada, et nende andmekogude andmetega, mida andmeait kasutab, tuleb lahendada kvaliteediprobleeme ning planeerida selleks aega ja ressursse; selgitada välja, kas andmaitade andmeid saab välistelt osapooltelt ja kui ei saa, siis kas probleem on tehniline (mida saab parandada) või on küsimus pigem usalduses; võimalusel mõelda andmeaida peale juba uue andmekogu loomisel; jälgida seadusandluses ettenähtud nõudeid andmeaida asutamisele, registreerimisele, turbele jne; võimalusel kasutada vastavaid standardeid.

- Andmeaitade väljatöötamisel, kasutamisel ja hooldusel tuleks jälgida mingit praktikas järele proovitud andmeaitade arendamise metoodikat; alustada etapiviisiliselt, väikese mahuga; soodustada kommunikatsiooni IT poole ja äriinimeste vahel, tutvuda andmeaitade tehnoloogiate pakkujate ülevaadetega; vaadata nii andmeaida tarkvara

platvorme kui ka seda, kas meil pakutakse tuge ja milline on selle toe kvaliteet; soovitatavalt mitte teha andmete teisendusi nende andmeaita laadimise käigus - kõik puuduvate näitajate arvutused tehakse peale andmete ülekandmist andmeaidas; kuna probleemid võivad olla andmete hõive ja kvaliteediga, kaaluda automaatset andmehõivet ja andmepuhastust.

- Andmeaitade puhul tuleks valida sobiv metoodika, sh nt Kimballi metoodika, mille puhul andmeaida andmed koosnevad põhifaktide tabeli(te)st ning nendega seotud kontekstiandmeid sisaldavatest "dimensioonidest"; normaliseeritud lähenemine, mille puhul andmed andmeaidas normaliseeritakse nagu tavapärases andmebaasis; spetsiaalsed tehnikad struktureerimata andmete kogumiseks ja kasutamiseks; tootjate poolt pakutavaid metoodikad, eriti kui suund on pigem andmete virtualiseerimisele, suurte töötlemata andmekogumite kasutamisele ning andmete analüüsile. Soovitame selle teema lülitada koolituste programmi.

- Andmeaitade andmehõivet ja uute andmemaitade loomist spetsiaalselt laiendada ei ole vaja, küll aga on mõtet andmeaitade võimalusi ja tehnoloogiat laiemalt tutvustada, et asutused oleksid sellest teadlikud ning oskaksid vajadusel andmaida projekti algatada. Soovitame selle teema lülitada koolituste programmi.

- Kasutada andmeaitades kodeerimist isikuandmete töötlemisel andmesubjekti nõusolekuta teadusuuringu või riikliku statistika vajadusteks. Muudel juhtudel jälgida isikuandmete töötlemisel vastavaid põhimõtteid (isikuandmete kaitse seaduse § 6, sealhulgas turvalisuse põhimõte) ja isikuandmete töötlemise lubatavust (sh isikuandmete kaitse seaduse § 10 kuni 14, nt andmesubjekti nõusolek või töötlemine avaliku ülesande täitmise käigus).

- Kui andmeaitade jaoks on tihti vaja kasutada kodeerimist ka muul juhul kui isikuandmete töötlemisel andmesubjekti nõusolekuta teadusuuringu või riikliku statistika vajadusteks, kaaluda vastava täienduse tegemist isikuandmete kaitse seadusesse.

- Täpsustada ja muuta mittevasturääkivaks isikuandmete kaitse seaduse § 16, sealhulgas § 16 lõigete 1, 2 ja 3 sõnastused.

- Lihtsamatel juhtudel kaaluda andmeaitade loomisel võimalust läbi ajada olemasolevate tehnoloogiliste vahenditega, näiteks kui kasutatavad andmekogud on samal platvormil.

- Kui lähteandmebaasid on erinevate andmebaasimootorite peale ehitatud või liidestamata, tuleks kaaluda andmete erinevaid struktuure ühendavaid/tõlgendavaid liideseid. Seejuures võib kasutada X-tee lahendusi, spetsiaalset andmeaitade tarkvara, ETL (*Extract, Transform and Load*) tööriistu jm.

- Kuna andmeaida halduseks kasutatava tarkvara platvormi omadustega ollakse valdavalt kas täiesti või osaliselt rahul, siis käesolevas uuringus ei pakuta konkreetse andmeaitade tarkvara tootja eelistust. Uute andmeaitade tarkvara valikul tuleks eelkõige arvestada omadusi, mida küsitlusel hinnati kõige olulisemaks: päringute jõudlust, häid administreerimisvahendeid, tõrketaluvust, integreeruvust olemasolevasse IT keskkonda ja ärianalüüsi süsteemide toetust. Arvestades tendentsi andmebaasidest võetud andmete osatähtsuse vähenemisele ning erinevatest andmeallikatest pärinevate osaliselt mittestruktureeritud andmete suuremale kasutusele, tuleks andmeaida tarkvara valikul kaaluda ka hetkel vähemtähtsamaks hinnatud andmete kompressiooni, pilvearvutuse toetust ja suurte andmemahtude toetust.

- Teha valik keskse või virtuaalse andmeaida tehnoloogiate vahel iga andmeaida puhul eraldi, sõltuvalt konkreetsest ülesandest, analüüsi vajadustest, lähteandmete kättesaadavusest jne. Soovitame seda teemat käsitleda ka koolitustel.

8 LÜHENDID JA MATERJALIDE LOETELU

8.1 LÜHENDID JA SÕNASELETUSED

1. **Andmeait, andmeladu (kitsam määratlus)** - kindlale valdkonnale (või probleemile) orienteeritud, teisene, integreeritud, ajast sõltuv, püsiv andmekogum, mille eesmärgiks on toetada otsuste tegemist.
2. **Andmeait, andmeladu (laiem määratlus)** - andmete kasutamise meetodite, tehnoloogiate ja praktikate kompleks, mille eesmärk on teha paremaid otsustusi ning pakkuda paremaid teenuseid, säilitades andmesubjektide privaatsuse ning luues võimalused andmete analüüsiks. Sellist andmeaita võib realiseerida väga erinevalt, sealhulgas andmete koondamise, koosvõime, andmete virtualiseerimise, pilvetehnoloogiate, mobiilsete tehnoloogiate, suurte töötlemata andmekogumite kasutamise ja muude vahenditega.
3. **ETL** - *Extract, Transform and Load*.
4. **IT** - infotehnoloogia.
5. **OIOO** - üks sisse, üks välja (*One-In, One-Out*) metoodika, mille puhul uute regulatsioonide puhul lisanduvad kulutused kompenseeritakse olemasolevate regulatsioonide eemaldamisest tekkiva kokkuhoiuga.
6. **RACI chart** - vastutusmaatriks, mis määratleb vastutavad, aruandekohustuslikud, konsulteeritavad ja informeeritavad osapooled *(responsible, accountable, consulted, informed - RACI)*.
7. **RIHA** - riigi infosüsteemi haldussüsteem.
8. **SWOT** - tugevused (*strengths*), nõrkused (*weaknesses*), võimalused (*opportunities*), ohud (*threats*).

8.2 KASUTATUD MATERJALID

1. Seadused, määrused
   1.1. Avaliku teabe seadus
   1.2. Isikuandmete kaitse seadus

1.3. Riikliku statistika seadus

1.4. Vabariigi Valitsuse seadus

1.5. Haldusmenetluse seadus

1.6. Riigi Teataja seadus

1.7. Riigi infosüsteemi haldussüsteem. Vabariigi Valitsuse määrus. Vastu võetud 28.02.2008 nr 58.

1.8. Infosüsteemide turvameetmete süsteem. Vabariigi Valitsuse määrus. Vastu võetud 20.12.2007 nr 252.

2. Regulatsioonide ettepanekud ja nende arutelu

2.1. Euroopa Parlamendi ja nõukogu määrus üksikisikute kaitse kohta isikuandmete töötlemisel ja selliste andmete vaba liikumise kohta (tekstis "isikuandmete kaitse üldmäärus"), [COM (2012) 11].

2.2. Euroopa Parlamendi ja nõukogu direktiiv üksikisikute kaitse kohta seoses pädevates asutustes isikuandmete töötlemisega kuritegude tõkestamise, uurimise, avastamise ja nende eest vastutusele võtmise või kriminaalkaristuste täitmisele pööramise eesmärgil ning selliste andmete vaba liikumise kohta (tekstis "isikuandmete kaitse direktiiv"), [COM (2012) 10].

2.3. Eesti seisukohad isikuandmete kaitset puudutavate Euroopa Komisjoni algatuste suhtes. Riigikantselei, 29.03.2012 nr 2-5/12-00274-4.

2.4. One-In, One-Out (OIOO) Methodology. Department for Business, Innovation and Skills, London, July 2011. [WWW] https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/31616/11-671-one-in-one-out-methodology.pdf (7.03.2013).

2.5. Proposal for an EU Data Protection Regulation - government impact assessment. Ministry of Justice (United Kingdom), 2012. [WWW] https://consult.justice.gov.uk/digital-communications/data-protection-proposals-cfe (7.03.2013).

3. Ametlikud standardid

3.1. ISO/TS 29585:2010. Health informatics -- Deployment of a clinical data warehouse. [WWW] http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=45582 (8.03.2013).

3.2. ISO/TR 22221:2006. Health informatics - Good principles and practices for a clinical data warehouse. [WWW] http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=40783 (8.03.2013).

3.3. EVS-ISO/IEC 12207:2009. Süsteemi- ja tarkvaratehnika. Tarkvara elutsükli protsessid (ISO/IEC 12207:2008).

4. Muud materjalid

4.1. Ralph Kimball, Margy Ross. *The Data Warehouse Toolkit* , John Wiley and Sons, Inc, 2002.

4.2. William Inmon. *Building the Data Warehouse*, John Wiley and Sons, 2005

4.3. William Inmon, Krish Krishnan. *Building the Unstructured Data Warehouse*, John Wiley and Sons, 2011.

4.4. Mark A. Beyer, Donald Feinberg, Merv Adrian, Roxane Edjlali , Magic Quadrant for Data Warehouse Database Management Systems, Gartner, **6 February 2012** ID:G00219281.

4.5. Handbook on Data Quality Assessment Methods and Tools. Manfred Ehling and Thomas Körner (eds). European Commission, Wiesbaden, 2007.

4.6. Mark Scott. The Shortcut Guide to Large Scale Data Warehousing and Advanced Analytics. [WWW] http://nexus.realtimepublishers.com/sgldw.php (28.2.2013).

4.7. Volume, Velocity, Variety, Value: Delivering the Elusive Fourth V out of Big Data. TIBCO. [WWW] http://resources.idgenterprise.com/original/AST-0077653_TIBCO_Ebook_Final__Big_Data_print_versionv4.pdf (2.03.2013).

4.8. PRIA andmeaida analüüs. Lõpparuanne. Logica, 31.10.2008. http://enos.itcollege.ee/~aarak/**PRIA**%20**andmeaida**%20**analüüsiaruanne**.docx *(26.03.2013)*

Viidatud kirjanduse loetelu

Adrian 2010. M. Adrian and C. White, Analytic Platforms: Beyond the Traditional Data Warehouse, BeyeNETWORK Custom Research Report Prepared for Vertica, 2010 TechTarget, BI Research, IT Market Strategy.

Basandra 2013, Basandra, Suresh (2013-02-17). Database, Data Warehouse and Business Intelligence Questions and Answers (Kindle Locations 7371-7385). Basandra Publications. Kindle Edition.

Butler 2012. J. Butler, Big Data and Analytics at the IRS, https://www-950.ibm.com/events/wwe/grp/grp004.nsf/vLookupPDFs/Jeff%20Butler's%20Presentation/$file/Jeff%20Butler's%20Presentation.pdf (01.03.2013)

CDW 2012. Compliance Data Warehouse (CDW) -Internal Revenue Service, http://www.techamericafoundation.org/content/wp-content/uploads/2012/10/Final-Big-Data-Case-Study-IRS-Compliance-Data-Warehouse.pdf (01.03.2013)

CensusHub 2013. https://webgate.ec.europa.eu/fpfis/mwikis/sdmx/index.php/Census_Hub (01.03.2013)

Davis 2011, Davis, Judith R.; Eve, Robert (2011-09-19). Data Virtualization: Going Beyond Traditional Data Integration to Achieve Business Agility (Kindle Locations 336-352). Nine Five Zero. Kindle Edition

DPA 1998. Data Protection Act of UK, https://www.gov.uk/data-protection/the-data-protection-act (21.03.2013)

EL andmekaitse direktiiv 1995. Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data, http://eur-lex.europa.eu/smartapi/cgi/sga_doc?smartapi!celexapi!prod!CELEXnumdoc&lg=EN&numdoc=31995L0046&model=guichett (22.03.2013)

EL uus andmekaitse direktiiv 2012. European new data protection law. http://www.mlawgroup.de/news/publications/detail.php?we_objectID=227 (20.03.2013)

EPIC 2003. Electronic Privacy Information Centre, 2003. http://epic.org/privacy/profiling/datamining3.25.03.html (10.03.2013)

Evans 2013, Bob Evans. Oracle. Data Warehouse 2.0: The 10 Top Trends Driving the Revolution http://www.forbes.com/sites/oracle/2013/01/14/data-warehouse-2-0-the-10-top-trends-driving-the-revolution/ retrieved 21 march 2013

FIP 1972. The Code of Fair Information Practices, http://epic.org/privacy/consumer/code_fair_info.html

Brust 2013, Gartner releases 2013 data warehouse Magic Quadrant, http://www.zdnet.com/gartner-releases-2013-data-warehouse-magic-quadrant-7000010796/ (21.03.2013)

Hof. 2007. S. van der Hof, The Status of eGovernment in the Netherlands, Electronic Journal of Comparative Law, vol. 11.1 (May 2007), http://www.ejcl.org, http://www.ejcl.org/111/art111-13.pdf (21.03.2013)

IT audit 2011. Government IT spending report could spark virtualization option, http://www.computerweekly.com/news/2240039353/Government-IT-spending-report-could-spark-virtualisation-adoption (18.03.2013)

Listpoint 2012. Listpoint open data survey, http://www.publictechnology.net/news/open-datas-potential-remains-closed-public-sector-staff/37572 (08.03.2013)

McKendrick 2011. J. McKendrick, A New Dimension To Data Warehousing:2011 IOUG Data Warehousing Survey, Unisphere Research, a Division of Information Today, Inc. September 2011, http://www.oracle.com/us/solutions/datawarehousing/2011-ioug-data-warehousing-survey-522175.pdf (13.03.2013)

McKenna 2011. B. McKenna, National Audit Office advocates strategic role for BI in government IT, Computer Weekly, Monday 28 February 2011, http://www.computerweekly.com/news/2240032829/National-Audit-Office-advocates-strategic-role-for-BI-in-government-IT (21.03.2013)

McKinsey 2011, McKinsey and Company „Big data: The next frontier for innovation, competition, and productivity“ june 2011. http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation retrieved 21 march 2013

Obama 2011. B. Obama, Technology, http://www.whitehouse.gov/issues/technology (18.03.2013)

Privaatsuse seadused. United States Privacy Laws, http://www.informationshield.com/usprivacylaws.html (20.03.2013)

PSI-directive 2003. DIRECTIVE 2003/98/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 17 November 2003 on the re-use of public sector information, http://ec.europa.eu/information_society/policy/psi/docs/pdfs/directive/psi_directive_en.pdf (16.03.2013)

Russom 2009 P. Russom, Next generation data Warehouse Platforms, TDWI Best Practises report, 2009

Scherer 2012. M. Scherer, How Obama's data crunchers helped him win, Time, http://edition.cnn.com/2012/11/07/tech/web/obama-campaign-tech-team (21.03.2013)

Zijlstra 2010. T. Zijlstra, Topic Report no. 17 State of Play: PSI in the Netherlands, http://epsiplatform.eu/content/topic-report-no-17-state-play-psi-netherlands (21.03.2013)

UK PSI 2005, UK PSI, http://www.legislation.gov.uk/uksi/2005/1515/contents/made

UNPAN 2012. United Nations Public Administration Network?s Electronic Government Development Index for 2012, http://unpan3.un.org/egovkb/print/printpage.asp?ref=http://unpan3.un.org/egovkb/global_reports/12report.htm (21.03.2013)

Whitehouse 2012, Big Data Press Release, http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release.pdf (19.03.2013)

9 LISAD

9.1 INTERVJUUDE ANALÜÜS

Eraldi dokument „Intervjuude_analüüs_2012-12-19.pdf“

Lisaks intervjuude tulemused 8-s failis: „Intervjuu_*_anon.pdf“

9.2 ANKETEERIMISE TULEMUSED

Eraldi dokumendid ankeetide töötlusest sihtrühmade lõikes:

„Ankeet_koond_kasutaja_2013-03-05.doc“

„Ankeet_koond_spetsialist_2013-03-05.doc“

„Ankeet_koond_spetsialist_erasektor_2013-03-22.doc“

„Ankeet_koond_spetsialist_riigisektor_2013-03-22.doc“

9.3 ANKEETIDE KÜSIMUSTIK SIHTRÜHMADE LÕIKES

Eraldi dokumendid

„Ankeet_küsimused_kasutaja.doc“ ja „Ankeet_küsimused_spetsialist.doc“

9.4 VAHESEMINARI ETTEKANNE 12.12.2012

Eraldi failis „Vaheseminar.ppt“

9.5 LÕPPSEMINARI ETTEKANNE 11.04.2013

Eraldi failis „Lõppseminar.ppt“