

Andmete haldus ja nende struktuur ning kasutamine EIS's

(Marek Popman, jaanuar 2024)

AS IS:

1. **ANDMED** on meil täna enamjaolt keskselt kogutud Vertica lahendused toimivasse andmebaasi. Andmebaasi tekivad andmed igapäevaselt kõikide asutuse osakondade tegevuse tulemusena (osaliselt ka neile veel endale teadmata) ja need tekkinud andmed on paljudel juhtudel ka struktureerimata (sõltub allikast ja kogumise viisist). Mingid andmed on täna veel ka erinevate kasutajate isiklikes kaustades või võrguketastel:
 - Probleemiks on täna meie asutuses osaliselt puudulik andmete **kogumise** loogika, kuna teenuste või toodete protsessid on kohati segased ja vastavad rollid ning vastutused osakonna siseselt jaotamata, näiteks puudub vastus küsimusele, kes vastutab osakonnas andmete eest.
 - Meie tänases andmelaos olevad andmed ei ole täna täies mahus **omavahel liidetavad**, kuna puudub kohati otsene omavaheline seotus (st puudub ühine tunnus nende liitmiseks ja keskne arhitektuuriloogika), mis on ilmnunud erinevate ülesannete täitmisel (näiteks teenuste ja eelarve kokku viimine, teenuste teenindusjuhtumid, mõjuhinnaang ja -analüüs jne)
 - Meie tänases asutuse andmemudelil puudub sisuline arusaam, milliseid andmed on vaja edaspidi **säilitada** ja milliseid mitte, kuna asutuse pikaajaline strateegia on selles lõigus liialt laiahaardeline ning ebapiisava detailsusastmega ja osakonnad tajuvad sisemiselt väga erineval tasemel seda sisulist kohustust
2. Mitmed meie asutuse osakonnad on **tegutsenud senini andmetega** vaid omal jõul ja teadlikumad osakonnad küsivad abi lahendamist vajavate ülesannete efektiivsemaks adresseerimiseks (näiteks mis andmeid selleks vaja on ja kust neid üldse leida on võimalik jne):
 - See olukord on tekitanud paralleelselt samade **tegevuste kordumist**, kus erinevad osakonnad tellivad andmeid ühest ning samast allikast ning ka kohati samaaegselt teadmata teise osakonna sarnaseid vajadusi. Sellest tulenevalt toimub sisuliste tegevuste dubleerimine, mis tekitab tõenäoliselt põhjendamata raha- ja ajakulu
 - Mitmed asutuse osakonnad on seni ise otsustanud, **kust ja kuidas andmeid hankida** ning neid tõlgendanud endale just parimal viisil kindlal ajahetkel. Sellisest andmekasutusest tekkinud uus teadmine on jäänud mõnedel juhtudel ka vaid osakonna tasemele
 - Täna on osakondades erineva tasemega arusaam **kuidas peaks käituma olukorras**, kus oleks vaja lahendada andmete keskne ülesanne osakonnas, millega võib kaasneda risk sattuda vastuollu andmekaitse seaduses (GDPR) sätestatud reeglitega
3. **AVAANDMETE** avaldamine on senini meie majas veel puudulik, kuna puudub selge arusaam ja metoodika, milliseid andmeid ja miks peab üldse avaldama.

Kõige eelneva tõttu on hetkel asutuses ebaefektiivne ressursside kasutus andmepõhiste tegevuste edendamiseks ning risk sattuda vastuollu ka andmekaitseseadusega, mis võib realiseeruda seetõttu ka suuremas kulutustes.

TO BE:

1. IGA meie maja osakond teab millise protsessi käigus millised andmed tekivad ning kes nendega osakonnas tegeleb (st on määratud osakonnas **andmehaldur*** ja **andmete omanik**** rollid):

- On määratud ja vajadusel koolitatud minimaalselt **1 töötaja igas andmeid kasutavas osakonnas** (hea lahend on 2 töötajat, et oleks ka tagatud asendatavus), kes omab osakonnas pakutavatest teenuste ning toodete protsessides detailselt väga head ülevaadet, kuna just seal tekivadki osakonna andmed
- Kõik meie majas olevad andmed on **ühtsel andmestandardist tuleneval struktuuril**, andmete dubleerimine on minimaalne või puudub üldse ning on tagatud loogika kuidas andmeid omavahel liita (st unikaalsed identifikaatortunnused lisatud)
- Eksisteerib kirjeldatud kokkulepe, millised andmed on **ärikriitilised** ja millised mitte, on selge arusaam milliseid andmeid on vaja **säilitada** pika perioodi vältel ja on määratud ka **nende andmete tundlikkuse tase**

2. Iga meie maja osakond teab lahenduskäiku (minimaalselt on vähemalt olemas teadmine, kus see on kirjas) **kuidas andmete põhists ülesannet sisult lahendada** ja on olemas vastused kõigile olulistele küsimustele- näiteks mida saab ja võib teha iga töötaja ise ning kelle poole pöörduda vajadusel abi saamiseks:

- Andmed on lihtsasti **sirvitavad ühest kokkulepitud kohast** ja need omavad selgesti loetavaid kirjeldusi nendega toimetamiseks (kirjelduse detailsusaste kõrge)
- On olemas asutuseülene **andmemudeli lahendus**, mis tagab pikas vaates kõigi osakondade huvid ja mis omakorda tagab andmete riskasutuse võimaluse nii majasiseselt kui ka majaväliselt ning uute andmete tellimisel on üheselt arusaadav majaülene protsess, mis aitab tagada erinevate riskide maandused (nt vastavus andmekaitseseadusele, kas on tegemist avaandmetega või kas andmed sobivad riskasutusse jne)

3. AVAANDMED on selgesti andmestandardi abil defineeritud igas osakonnas ja need on kokkulepitud viisil ning kohas ka õigeaegselt avaldatud.

** Äriprotsesse esindav roll andmehalduse alal:*

- *andmete sisu, konteksti ja metaandmete eest vastutaja;*
- *kohustused sõltuvad kontekstist ja võivad osaliselt kattuda andmekäitleja omadega*

Andmekäitleja:

- *andmeaida arendaja*
- *andmebaasiülem*
- *andmemodelleerija*

**** osakonnas andmete eest vastutav isik, kellel on [andmete](#) kohta vastutus ja õigused**

MKM pakutav andmestandardil baseeruv tööriist - RIHAKE

Lihtsustades on **RIHAKE** asutusse paigaldatav andmehalduse rakendus, mis aitab asutustel oma andmestikke korrastada.

[RIHAKE võimaldab andmekirjelduse standardi kohaselt:](#)

- Kirjeldada asutuse andmestikke
- Kirjeldada andmestikes kasutatavaid klassifikaatoreid ja loendeid
- Koostada andme- ja ärisõnastikke
- Edastada andmekirjeldusi teistesse süsteemidesse

RIHAKE aitab asutustel luua ühtset struktuuri ja formaati oma andmete jaoks, mis omakorda aitab parandada andmete kvaliteeti, juurdepääsetavust ja kasutatavust.

Kõik vajalikud juhised riigiasutuse andmestandardi loomiseks ja edasiseks kasutamiseks asuvad siin - [Juhised | Kratid](#)

RIHAKE elluviimisele aitavad kaasa Statistikaamet ja Majandus- ja Kommunikatsiooniministeerium. Mõlema asutuse poolt on tagatud tugi RIHAKEse juurutamisele ja edaspidi on tagatud vajaduspõhine suhtlus, nõustamine ning muu tugi.

Kuidas edasi ehk 3 võimalikud arengusuunda

1. Kui meie asutus jätkab tegevust **AS IS** olukorra kirjelduse alusel ja **ei võta kasutusele ühtki standardiseeritud** andmemudelil baseeruvat tööriista, siis on risk järgmiste riskide realiseerumiseks:

- **Andmete haldamise keerukus:** andmete kogumine ja säilitamine muutub aina keerulisemaks kuna puudub ühtne struktuur ja formaat, mis omakorda põhjustab andmete kvaliteedi langust ning veelgi keerukamat edasist andmete analüüsi erinevate ülesannete lahendamiseks
- **Ressursside ebaefektiivne kasutamine:** kuna erinevad asutuse osakonnad tegelevad paralleelselt kohati samade andmepõhiste ülesannetega, võib see põhjustada jätkuvalt ressursside ebaefektiivset kasutamist ja suurendada asutusele täiendavat kulu veelgi
- **Andmekaitseaduse nõuete rikkumise risk:** kui asutus ei taga andmekaitse põhimõtteid järgivat süsteemi andmete kogumisel, töötlemisel ja säilitamisel, siis võib see kaasa tuua andmekaitseaduse nõuete rikkumise ja sellest tulenevad trahvid
- **Avaandmete avaldamise probleemid:** kui asutusel puudub selge arusaam ja metoodika, mida ja miks peab andmete poolelt avaldama, siis võib see takistada avaandmete tekkimist ja kasutamist ka edaspidi
- **Andmete integreerimise probleemid:** kui andmed ei ole omavahel kergesti liidetavad, võib see takistada andmete tõhusat kasutamist analüüsi- või automatiseerimisülesannete lahendamisel, mis väljendub täiendavas kulus ebaefektiivse toimimise tõttu.

2. Kui meie asutus võtab kasutusele mõne teise **alternatiivse andmestandardil baseeruva tööriista*****, siis võivad tekkida järgmised olukorrad:

- **Andmestandardi sobivus:** uue sisult meile vööra andmestandardil baseeruva tööriista kasutuselevõtt sõltub suuresti sellest, kui hästi see vastab meie asutuse vajadustele ehk see vajab rohkem eeltööd ning kui see ei ole nii paindlik ega kohandatav kui RIHAKE, võib see põhjustada probleeme andmete struktureerimisel ning haldamisel ka edaspidi
- **Üleminekuperiood:** uuele standardile üleminek nõuab tõenäoliselt rohkem aega ja ressursse, kuna puudub sisuline ja tasuta nõustamise võimalus ning see võib hõlmata olemasolevate andmete migreerimist, personali koolitamist ja süsteemi uuendamist laiemalt ning veelgi kulukamalt
- **Ühilduvus teiste süsteemidega:** kui uus standard ei ole ühilduv teiste riigiasutuses kasutatavate süsteemidega, siis võib see põhjustada integreerimisprobleeme ning see võib mõjutada andmete liikumist süsteemide vahel ja vähendada andmetöötluste tõhusust ka edaspidi
- **Andmekaitse ja privaatsus:** alternatiivse standardi kasutuselevõtt peab tagama, et järgitakse kõiki andmekaitse ja privaatsuse nõudeid ning juhul, kui see ei paku piisavaid turvameetmeid, võib see suurendada andmepõhiste riskide, nagu andmelekke või volitamata juurdepääsu, tõenäosust

3. ***Kui asutus võtab kasutusele RIHAKEse, võib see tuua kaasa mitmeid positiivseid muutusi terves asutuses:***

- **Andmete kvaliteedi parandamine:** RIHAKE aitab luua ühtset struktuuri ja formaati andmetele, mis parandab andmete kvaliteeti ja sellest saadavat väljundit
- **Andmete juurdepääsetavuse ja kasutatavuse suurendamine:** kuna andmed on korraldatud ja struktureeritud, on neid lihtsam kasutada ja analüüsida, mis omakorda suurendab rahulolevate andmekasutajate hulka
- **Ressursside efektiivsem kasutamine:** kuna andmete kogumine ja säilitamine on korraldatud, saab asutus oma ressursse tõhusamalt kasutada ehk on eeldused paremaks ressursside kasutuseks
- **GDPR-i nõuete täitmine:** RIHAKE aitab asutusel järgida andmekaitse põhimõtteid, vähendades seeläbi andmekaitseeaduse nõuete rikkumise riski
- **Avaandmete tõhusam kasutamine:** RIHAKE aitab asutusel mõista, milliseid andmeid tuleks avaldada ning tagab ka avaandmete tekkimise eeldused, mis omakorda võimaldab avaandmete kasutamist ka laiemalt riigis
- **Andmete integreerimise hõlbustamine:** Kuna andmed on omavahel liidetavad, on neid lihtsam integreerida teiste süsteemidega, mis omakorda suurendab loodava info lisandväärtust erinevate ülesannete täitmisel

Eelneva kokkuvõtteks, kui meie asutus ei võta edaspidi kasutusele andmete struktureerimiseks mõnda andmestandardil baseeruvat tööriista, siis suure tõenäosusega ei suuda meie asutus ei lühikeses ega ka pikas vaates täita / kasutada kõiki andmepõhiseid ootusi või võimalusi ning väga keeruliseks võib osutuda ka on edaspidi kaasata võimaliku AI (tehisintellekt) ootusi või võimalusi ning suure tõenäosusega jätkuvad **AS IS** osas kirjeldatud olukorrad.

Kui aga asutus otsustab kasutusele võtta mõne teise andmestandardil toimiva tööriista kui **RIHAKE**, siis tuleb hoolikalt kaaluda kõiki neid tegureid, mis tagavad samuti sujuva ülemineku (sh ajaliselt) ja andmete tõhusa haldamise pikas vaates ning ühildumise teiste andmekogudega (sh riiklikul tasemel). Kui omal jõul hakata tundma õppima ja katsetama meile sisult tundmatut andmestandardit (sh puudub sellel sarnane nõustamise tugi nagu RIHAKEse puhul), siis see võib suure tõenäosusega pikendada ajaliselt üleminekuperioodi

ja ei pruugi täielikult tagada lõplikult sobivust riiklikult soovitud arengutega (nt väljundiks ühine riiklik andmekogude võrgustik, andmete teabevärs jne).

Lisaks suurenevad suure tõenäosusega ka otseselt selle üleminekuperioodil tehtavad kulud (nt asjaga seotud töötajate palgakulu või sisse ostetavad teenused nõustamiseks).

Kokkuvõtteks võib väita, et RIHAKese kasutuselevõtt aitaks asutusel parandada andmete haldamist, suurendada ressursside efektiivsust, vähendada andmekaitseaduse rikkumise riski, parandada avaandmete avaldamist ja hõlbustada andmete integreerimist ning luua aluse tehisintellekti lahenduste asutusesiseks kasutusele võtuks.

***** Näited RIHAKe sarnastest andmestandardile vastavatest tööriistadest:**

Collibra - See on andmehalduse tarkvaraplatvorm, mis aitab asutustel hallata ja valitseda oma andmevarasid. See pakub laia valikut võimalusi, sealhulgas andmekataloogimist, andmekvaliteedi haldamist, andmeliini jälgimist ja andmeprivaatsust ning vastavust.

ACCURITY - on kõikehõlmav andmete intelligentsuse platvorm, mis pakub ettevõtte tasemel mõistmist ja täielikku usaldust teie andmete vastu

OvalEdge - OvalEdge on andmekataloog, mis loob organisatsiooni kõikide andmeallikate põhjaliku koostise paremaks andmetele juurdepääsuks ja nende analüüsimiseks

RIHAKese kasutuselevõtu ajaline tegevuskava (esialgne prognoos)

I ETAPP

veebruar - mai 2024:

- Selles etapis peab tekkima tunnetus, kuidas toimub andmete RIHAKese andmestandardiga vastavusse viimine (tõenäoliselt alles siin ilmnevad võimalikud kitsaskohad ja ka lahendused) ning tekib kogu protsessist hea ülevaade, millise alusel saab juba täpsemalt prognoosida vajaliku sisemist ressurssi igas osakonnas eraldi edaspidiseks standardi juurutamiseks.
- Kuna kohe kõiki andmeid RIHAKese standardile vastavaks viia ei ole tehniliselt võimalik, siis alustame samm-sammult ehk valime esmalt andmed, mis annavad meile esmase tunnetuse kõigis võimalikes oodatavates väljundites (sh avaandmete tekkimine ja nende avaldamine).
- Selles etapis aitavad meid Statistikaameti 3 andmespetsialisti (koostöö on nende poolt tasuta, sõlmime kõigi nendega NDA lepingu), kes siis koondavad esmalt meie poolt valitud andmestiku vastu RIHAKese andmestandardi loogikat, peale mida teeme ühistel koosolekutel tehtust ülevaated ja vajalikud täiendused ning uued kokkulepped edasiseks
- Selle etapi jooksul peaks tekkima nn esmane lihasmälu, kuidas järgmiste osakondade andmestikega toimetada ja selle arusaama ning praktika tekkimisel saame käivitada kõigis osakondades sarnase tegevuse, kui selleks ajaks on loodud ka muud eeldused (st minimaalselt määratud osakonnas andmete omanik või andmehaldur võtmerollid)

- Selle etapi lõpuks on meie asutuses olemas hea tasemel arusaam, kuidas kõigis osakondade andmed vastu RIHAKese standardiseerida ja kui palju selleks vajame lisaressurssi töötajate näol ning palju see ka otseselt lisa finantskulus võib väljenduda

II ETAPP

juuni – detsember 2024:

- Sellel etapis jätkub koostöö Statistikaameti andmespetsialistidega, kuni on kõigi meie asutuse osakondade andmed viidud vastavusse RIHAKese andmestandardi loogikaga, st igas osakonnas tekkinud arusaam, kuidas seda olemasolevate ja ka uute lisanduvate andmetega sisuliselt ning välise abita juba ise korraldada
- Kõigis meie asutuse osakondades, **kus tekivad andmed või kasutatakse andmeid**, on tekkinud selge arusaam ja tunnetus just neile vajalikest andmetest, nende andmete haldusest ja vajadusest edaspidi ning on selge teadlikkus ning ka rollijaotus nendega toimetamiseks
- Selle etapi lõpuks on RIHAKe andmestandard suures osas (min 70% meie asutuse andmestikest) integreeritud meie asutuse andmetega ja on tekkinud ligilähedane olukord, mis on kirjeldatud **TO BE** ülevaates
- Kogu etapi jooksul toimub jooksvalt sisuline hinnang (nt iga kalendrikuu lõpus) juba tehtule ja vajadusel korrigeeritakse andmestandardi juurutamisel igapäevast praktikat vastavat osakondade äripoolsele loogikale

III ETAPP

2025 I poolaasta:

- Vajadusel jätkub koostöö Statistikaameti andmespetsialistidega, kuid tõenäoliselt vähem intensiivsemas vormis (st vajaduspõhine nõustamine, peamiselt keerukamad olukorrad ja lahendid), kuna eelduste kohaselt suudame sellel ajajärgul oma maja ressursidega lahendada enamus andmestandardiga seotud väljakutsed
- Etapi lõpuks oleme eelnevalt märgitud 14 osakonnas saavutanud RIHAKese andmestandardi juurutamise ja kasutamise tasemel, mis väljendub väga ligilähedases **TO BE** olukorra kirjelduses
- Siinkohal võetakse kokku kõik senini tehtud ja analüüsitakse esialgset vastu hetke olukorda sellel ajahetkel toimuvaga andmetes (teeb seda STAR tiim). Vajadusel lepitakse kokku jätkutegevused.

Vajalikud lisa ressursid RIHAKE juurutamiseks (esialgne prognoos, täpsustub peale I etapi)

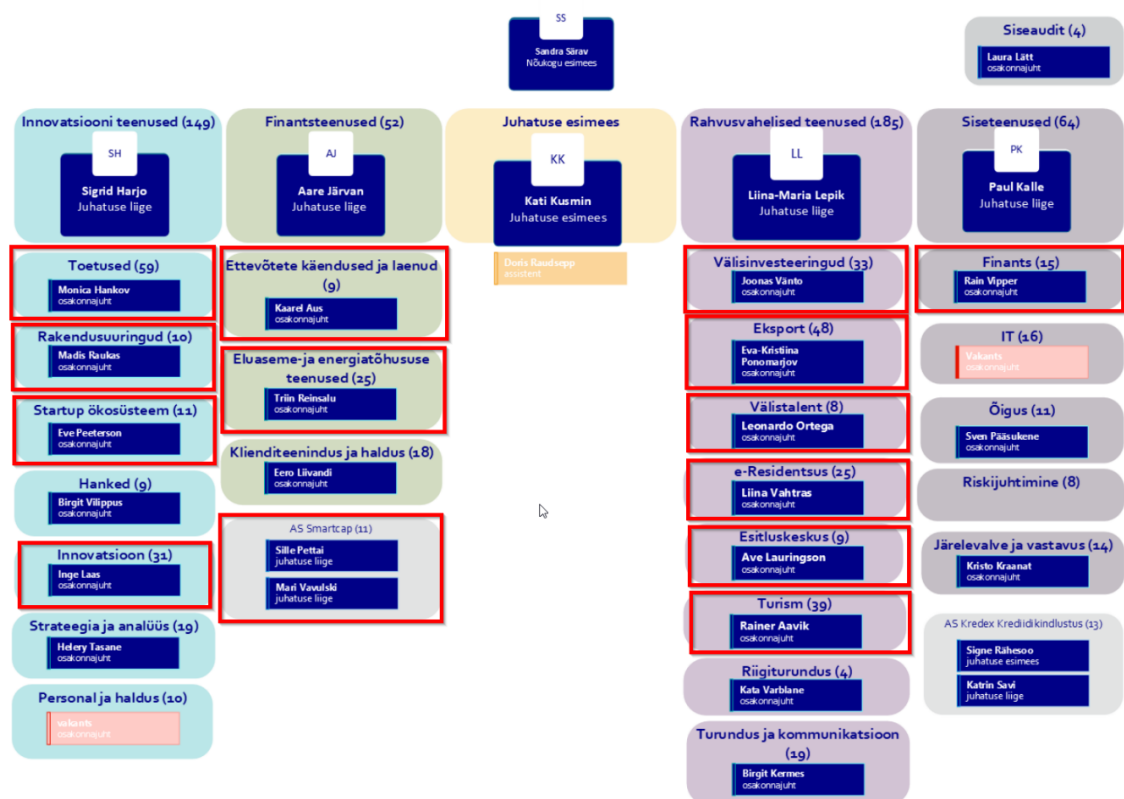
1. Kuna igas andmeid kasutavas osakonnas tuleb määrata minimaalselt **1 töötaja** (hea lahend on siiski 2 töötajat, kuna vaja on ka asendajat) tegelema osakonnas nende andmetega, siis selle töötaja(te) tööajast tekib kuni 10% ulatuses lisakohutusi mingitel ajahetketel (*hinnangu aluseks senine RIHAKEse kasutajate tagasiside*), kuid seda mitte igas kalendrikuus (see suurusjärk saab sõltuma osakonna andmete mahust ja keerukusest ning andmeallikate arvust).

Rahaliselt võiks siin arvestada kuni 10% asutuse keskmisest töötaja kalendrikuu palga arvestusest (nt **www.teatmik.ee** andmetel on see hetkel ca 2538 eurot) lisapreemiat ühes majandusaastas ehk:

2 töötaja puhul x 250 = 500 lisa euroga osakonna kohta kord majandusaastas

1 töötaja puhul 250 lisa euroga kord majandusaastas

Meie asutuses on hetkel kehtiva struktuuri alusel ja esialgsel hinnangul andmete haldurite vajadus vähemalt 14 osakonnas (joonis 1), millise valiku aluseks on igapäevaselt vajalikud nn äriandmed.



Joonis 1. EIS kehtiv struktuur 2024

Tõenäoliselt on vajadus 2 töötaja kaasamiseks vaid suurema arvuga osakondades, nagu **turism, toetused, eksport ning eluaseme- ja energiatõhususe** osakond ehk kokku **4 osakonnas**. Ülejäänutes **10 osakonnas**, mis on vähesemate arvuga töötajate ja ka pakutavate teenuste mõistes vähem intensiivsema andmevoo haldusega, peaks esialgu piisama vaid 1 töötaja kaasamisest.

Kokku 1 aastas püsikulude kasv kogu asutusele:

4 osakonda (2 töötajat) x 500 eurot majandusaastas = **2 000 eurot**

10 osakonda (1 töötaja) x 250 eurot majandusaastas = **2 500 eurot**

Esialgse hinnagu alusel suureneks kokku ca 4 500 eurot otsekulud töötajate palgakasvu näol ühes majandusaastas (see täpsustuks siiski veel peale I etapi läbimist).

Tõenäoliselt on mõistlik siinkohal lisatasu määramisega igale vastutavale töötajale kord majandusaastas (sisult jäägu see otsustamiseks osakonna põhiselt, kuna töökoormus saab olema erinev osakonniti just andmete mahu ja allikate keerukuse erinevuse tõttu).

2. Uue lisanduva rolliga töötajaid on vaja ka mingi tasemeni koolitada, kuid selle lõigu osas saame sisemiselt arvestada olemasolevate ressurssidega, st koolituse teevad selles lõigus tõenäoliselt STAR tiimi andmetega igapäevaselt kokkupuutuvad töötajad (täpsustub töö käigus, kes konkreetselt seda teeb).
3. Lõpliku ja detailne töö sisukirjeldus andmete omaniku ning andmehalduri rollis selgub peale I etapi RIHAKE juurutamist.
4. Töökorralduslikult on andmehaldus ja IT-haldus üksteisest eraldatud. Andmehaldus keskendub andmete elutsüklile, mis hõlmab printsiipe (standardite, andmekirjelduste ja andmearhitektuuri ning mudelite määratlemist), otsustusfunktsioone (probleemide haldust ja teavitamist) ja töökorralduslikke funktsioone (sisu- ja administratiivseid funktsioone). IT poole eest vastutajad või IT asutuste juhid aga vastutavad andmete haldamist toetavate IT rakenduste, andmebaaside, riistvara ning üldise tehnilise arhitektuuri eest.
5. Meie asutuse **IT pooled püsi- ja muutuvkulud** tõenäoliselt ei suurene, kuna RIHAKEse rakenduse edasise halduse ja arenduse eest vastutab selle omanik MKM.

„Standardized data allows a shared data language across every team and tool. So it’s like walking into the United Nations and everyone speaking American English.”