

ANDMEHALDUSE JUHISED

Andmekirjelduse juhis

August 2023

Version 2.0

Dokumendi ajalugu

ver	muutuse sisu	autor	kuupäev
1.0	Juhise aluseks on 2019-20 Eesti andmehalduse metoodikaprojekti raames koostatud „Andmekirjelduse juhise“.	Raivo Ruusalepp, Kuldar Aas, Siim Aben; (Veiko Berendsen)	august 2020
1.1	Struktuuri ülevaatamine ja tagasiside sisse viimine	Veiko Berendsen	märts 2021
1.2	Toimetamine	Veiko Berendsen	oktoober 2021
1.3	Uus terminite ja sõnastike osa	Veiko Berendsen	05.04.2022
1.4	Sissejuhatava peatüki täiendused	Veiko Berendsen	nov-dets 2022
1.5	Sõnastike praktilise kirjeldus juhise osa lisamine	Veiko Berendsen	dets 2022
1.7	Ver 1.3 kuni 1.5 kommentaarid	Veiko Berendsen, Kuldar Aas	veebr-märts 2023
1.8	Andmete juhtrühmale saadetud versioon	Veiko Berendsen	13.03.2023
1.9	Märkuste sisseviimine ja toimetamine	Veiko Berendsen, Kuldar Aas, Aune Suitsmart	juuni- aug 2023
2.0	Uus lõppversioon	Veiko Berendsen	sept 2023

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

Sisukord

1	Sissejuhatus	5
1.1	Üldist	5
1.2	Andmekirjeldusest saadav kasu	6
1.3	Juhise koostamine	7
1.4	Andmekirjelduse juhise eesmärgid ja andmekirjelduse standard	8
1.5	Andmekirjelduse juhise sihtrühmad	9
1.6	Andmekirjelduse juhise ülesehitus	9
1.7	Andmehalduse ja andmekirjelduse õiguslik raamistik	10
1.8	Andmekirjelduse koostamise metoodilised alused ja alusstandardid	14
2	Andmekirjelduse põhimõtted	18
2.1	Andmekirjelduse ja metaandmete mõiste	18
2.1.1	Andmekirjeldus	18
2.1.2	Metaandmed	20
2.2	Andmed ja andmestikud	23
2.2.1	Andmed	23
2.2.2	Andmeelement	26
2.2.3	Andmestik	28
2.2.4	Andmetabel	31
2.2.5	Andmekataloog	34
2.3	Mõisted, terminid, sõnastikud	36
2.3.1	Termin ja mõiste	36
2.3.2	Andmekirjelduses kasutatavad sõnastikud	41
2.4	Tervikkirjeldus	43
3	Andmekirjelduse koostamise korraldamine	46
3.1	Üldist	46
3.2	Andmekirjelduse objektide piiritlemine	48
3.3	Andmekirjelduse sõnastike kaardistamine	51
3.4	Andmekirjelduse mudeli ja arhitektuuri kavandamine	53
3.5	Andmekirjeldusega seotud rollide määramine	54
3.6	Andmekirjelduse loomine ja täiendamine andmekirjelduse töövahendis	55
3.6.1	Üldist	55
3.6.2	Ärisõnastiku loomine ja andmekirjelduse töövahendiga sidumine	56
3.6.3	Andmekirjelduse koostamine	58
3.6.4	Andme- ja ärisõnastiku täiendamine	61
3.6.5	Andmekirjelduse kvaliteedikontroll	62
3.7	Andmekirjelduse edastamine	63
3.8	Andmekirjeldus seotuna andmehalduse ja organisatsiooni tegevustega	65
	Terminid ja mõisted	67
	Viited	69

Joonised

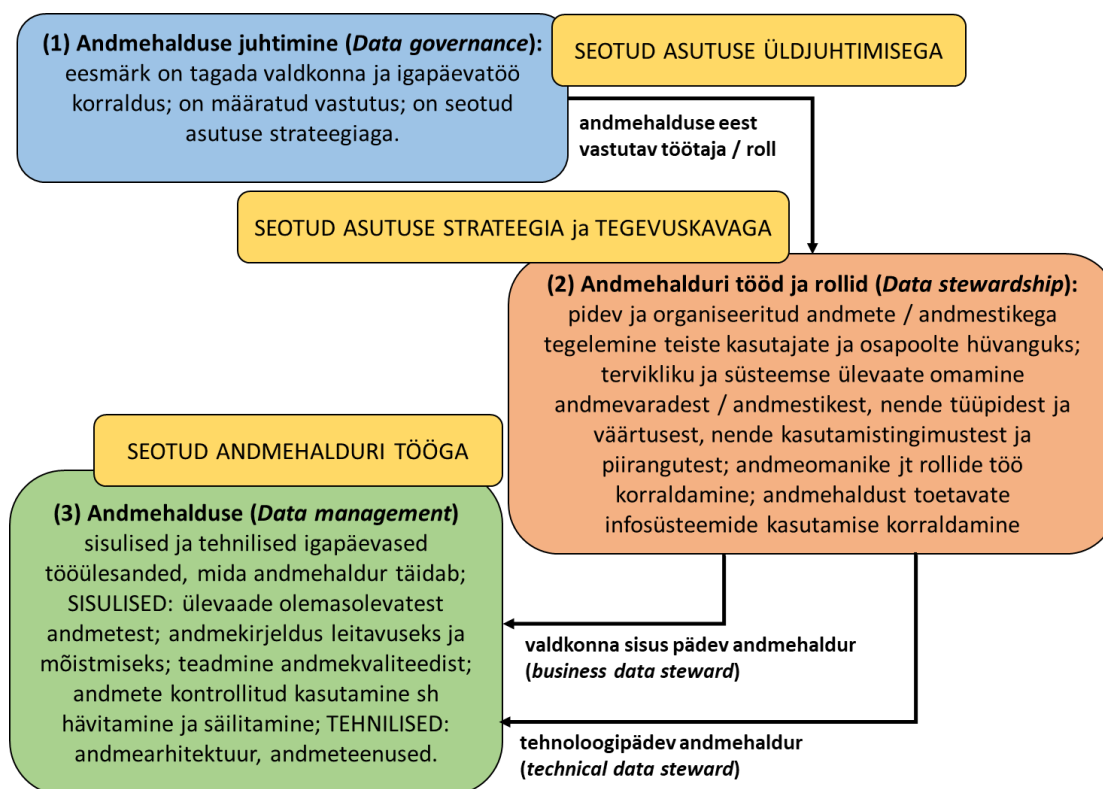
Joonis 1. Kolm vaadet andmehaldusele	5
Joonis 2. Juhise praktilise kasutamise soovitus.....	10
Joonis 3. Teabe/andmete inim- ja masinloetavus	12
Joonis 4. Valdkonnad, asutused ja juhised.....	13
Joonis 5. Andmekirjelduse koht andmehalduses	16
Joonis 6. Andmekirjelduse seos andmetega.....	19
Joonis 7. Metaandmete rühmad	21
Joonis 8. Andmete ja informatsiooni käsitus EVS-ISO/IEC 2382-1:1998 (01.01.02) põhjal.....	24
Joonis 9. Andmete ja informatsiooni seoste käsitus (L. Floridi).....	26
Joonis 10. Andmeelement ISO/IEC 11179 järgi	27
Joonis 11. Andmestik seotuna andmekataloogi ja andmeelemendiga	29
Joonis 12. Tabeli liigid (lihtsustatud tüpoloogia)	32
Joonis 13. Lihttabel ehk üksikandmed ridades ja veergudes.....	32
Joonis 14. Vahetabeli näide (tunnuste teisendus ja rühmitamine)	33
Joonis 15. Kahedimensionaalne tabel	33
Joonis 16. Tabeli esitamine lõikena	33
Joonis 17. Andmekuup, selle dimensioonid ja üksik väärtus.....	34
Joonis 18. Mitmedimensionaalne tabel ehk tabelina vormistatud andmekuup	34
Joonis 19. Tegelikuse ja keele seos mõiste ja terminiga (ISO 1087)	40
Joonis 20. Andmekirjelduse ja ärimõistete sõnastike üldine mudel	42
Joonis 21. Tervikliku andmekirjelduse komponendid.....	44
Joonis 22. Sõnastikevahelise infovahetuse mudel andmekirjelduses	45
Joonis 23. Andmehalduse ja andmekirjelduste seos organisatsiooni muude tegevusega	46
Joonis 24. Andmekirjelduse haldamise tegevused.....	47
Joonis 25. Andmekirjelduse objektide piiritlemise tegevused	49
Joonis 26. Andmekirjelduse koostamine ja täiendamine (protsess)	56

1 Sissejuhatus

1.1 Üldist

Statistikaamet koostööpartneritega koostab ja annab andmehaldusalaseid soovituslikke juhiseid. Juhised moodustavad metoodilise ja praktilise raamistiku andmehalduse korraldamiseks asutustes ja üleriigiliselt. Käesolev andmekirjelduse juhise on selle andmehalduse raamistiku ehk andmehaldust korraldava süsteemi osa.

Andmehaldust saab vaadelda kolmes vaates: juhtimisvaates (*Data Governance*), andmehalduri tööde ja rollide vaates (*Data Stewardship*) ja andmehalduse igapäevaste tööülesannete vaates (*Data Management*). Joonisel 1 on esitatud need kolm vaadet andmehaldusele. Iga andmehaldusega seotud konkreetne suurem tööloik on seotud nii juhtimisvaate ja laiema strateegilise eesmärgiga kui ka konkreetsema vastutusega terviku ja iga osa eest eraldi.



Joonis 1. Kolm vaadet andmehaldusele

Andmekirjelduse koostamine ja haldamine on üks olulisemaid ja töömahukamaid andmehalduse tegevusi. Andmekirjelduste olemasolu teeb andmed leitavaks ja mõistevaks ning toetab juhtimisotsuseid. Käesolevas juhises on täpselt näidatud, millistele andmete osadele kirjeldust koostada. Arvestada tuleb samuti sellega, et andmekirjeldused on seotud teiste andmehalduse tegevustega, nagu ka sellega, et tehniline ehk IT-vaade andmetele ja andmehaldusele pole täpselt sama, mis n-ö äri- ja äri-vaade. Üheks olulisemaks küsimuseks üldse andmehalduse korraldamisel ongi see, kuidas n-ö juhtimislik vaade oskab ühitada valdkondliku andmetest arusaamise pädevuse ning tehnoloogilise pädevuse. Pädevus ehk sisulised ja

tehnilised oskused on andmehalduse korraldamisel peamine, kuid praktika on näidanud ka seda, et volituste andmiseta ja vastutuse panemiseta on edenemine kesine.

Andmeid hallatakse metaandmete abil. Andmekirjelduse koostamine tähendabki andmetele neid kirjeldavate metaandmete lisamist ja nende metaandmete abil andmete haldamist. Andmekirjeldus koostatakse ehk metaandmed lisatakse erineva ulatusega andmeüksuste tüüpidele. Väikseim selles juhises käsitletav kirjeldatav üksus on andmeelement. Teine keskne kirjeldatav üksus on andmestik. Andmestik ja andmeelement on omavahel hierarhiliselt seotud, kuid on veel mitu andmestruktuuri, mille kirjeldamine on praktiliselt vajalik. Need on näiteks andmekirje, massiiv, andmebaas, andmeteenus ja andmekogu. Kõiki hallatavaid andmeobjekte võib nimetada varadeks.¹

1.2 Andmekirjeldusest saadav kasu

Ikka küsitakse juhtimisvaates, mis kasu me sellest saame, kui meil on koostatud andmekirjeldus? Eriti pelutav on see, kui selgub, et andmekirjelduse tegemine nõuab pädevust ja aega ega ole kergesti automatiseeritav. Lühidalt: mis on andmekirjelduse väärtuspakkumine?

Andmekirjelduse olemasolu võimaldab andmete leidmist, kasutamist ja nende mõistmist nii inimesele kui ka masinale (arvutile). Inimesed saavad aru sõnadest, nende tähendusest ehk mõistetest ning oskavad kasutada oskussõnu ehk termineid. Oskussõnavara ehk terminoloogia paremaks kasutamiseks koostatakse erinevaid oskussõnastikke. Arvutitele muudetakse andmed mõistetavaks semantilise veebi tehnoloogiaga. Andmekirjeldust ehk metaandmeid on nimetatud ka andmehalduse närvivõrguks, mis aitab asutusel pidada ülevaadet andmevaradest, tagab andmete usaldusväärsuse ja võimaldab neid tõhusalt taaskasutada.

Hea andmekirjeldus võimaldab tuvastada, mis andmed on olemas, hulgast andmetest leida kiiresti vajalikud, mõista andmete tähendust ja usaldusväärsust. Suurim kasu andmekirjeldusest on andmekasutuse tõhusus ehk kokkuhoitud aeg, mis kuluks andmete leidmiseks ja taaskasutuseks.

Andmekirjelduse peamine roll andmehalduse raamistikus on tagada asutusesiseselt ja asutuste üleselt teadmine olemasolevatest andmestikest ja neis sisalduvatest andmetest. Andmekirjelduse haldamine on pidev tegevus, mille jaoks kehtestatakse organisatsioonis reeglid, ning määratakse kindlaid rolle ja tööülesandeid täitvad töötajad.

Andmekirjeldus aitab:

- Tagada ülevaatlikkuse andmetest ja andmestikest ning andmete leitavuse. Selleks peetakse andmekataloogi või hoitakse andmeid mingis repositooriumis. Siin on andmehaldusel olulisi kokkupuutepunkte infohaldusega.
- Tagada andmete sisulise kirjeldamise ja tähenduse lahti seletamise. See ongi andmekirjeldus kitsamas tähenduses, kus andmed on metaandmetega kirjeldatud. Tähenduse paremaks esitamiseks kasutatakse sõnastikke, sh märksõnastikke.
- Mõista konkreetsete andmete kvaliteeti ehk seda, kas need andmed on kasutatavad. Osalt on andmekvaliteedi fikseerimine samuti andmekirjeldus, sest väljendub metaandmetes. Lisaks võivad olla määratud ja kirjeldatud ärinõuded, kuidas andmed

¹ Vara mõistet tavaliselt täpsustatakse ja kasutatakse kas andmevara (*data asset*) või infovara (*information asset*) mõistet.

saavad üldse tekkida ja neid töödeldakse ja ka see andmekvaliteedi osa on esitatud metaandmetena.

- Saada aru andmete usaldusväärsusest. Usaldusväärsus on usaldus andmeid loonud, haldava ja hoidva organisatsiooni vastu. Leides näiteks veebist andmed, on alati üks küsimus, kas need andmed avaldanud organisatsioon on usaldusväärne, nagu mõistagi ka andmed ise. Sageli on organisatsioonide käsutuses andmeid, mida nad ise ei ole loonud ning sel juhul on küsimus andmete päritolust ehk provenientsist. Teadmine, millisest allikast on andmed pärit, annab selgust nende usaldusväärsusest.
- Andmetest arusaadavuse nii tehniliste kui äriliste kasutajate jaoks. Kuigi äriline ehk sisuline kasutamine on n-ö algus ja lõpp ehk eesmärk, pole tehniline arusaadavus taandata ainult vahenditele. Andmetöötlemises, andmeteenustes, andmeladuses jm on andmekirjeldusest arusaamine samavõrd oluline kui andmete kasutamisel info ja teabena.
- Andmete masintöödeldavust ja semantilist koostalitlusvõimet. Semantiline koostalitlusvõime on ehk kõige olulisem ühtaegu andmete taaskasutust ja mõistetavust nii inimesele kui ka masinale toetav tehnoloogia. Aga see tehnoloogia ei toimi laiemas plaanis andmete tähenduse mõistmiseta inimese poolt, mille vormiks on kokkulepitud sõnavara märksõnastikena, taksonoomiatena, ontoloogiatena, kuid oma aluselt igal juhul terminoloogilis-mõistelise arusaamana andmetest. Masintöödeldavus on küll andmete tähendusest tehnoloogiliselt lahutatav, kuid lahutamatu andmetüüpidest, ning töötlusprotsessides on vaja teada, mis andmetega on tegu.
- Lõpuks, aga mitte vähemtähtsa teemana, aitab andmekirjeldus tagada andmete elukäigu halduse ning seda kahes tähenduses. Esmalt viisil, kus metaandmete abil juhitakse andmetöötlust andmekogumisest kuni avaldamiseni ja mis on viis, kuidas juhitakse tsüklites toimuvaid töid andmetega. See on eriti oluline ja levinud uuringutes. Teiseks aga arhiivinduslikust vaatest, kus elukäik katab perioodi, mil andmeid säilitatakse, aga ka toiminguid andmetega elukäigu etappide jooksul. Just see viimane vaade on muutumas järjest tähtsamaks, et organisatsioonid hoiaksid alles andmed, mida on vaja ja mille säilitamine on õiguslikult põhjendatud, ning hävitaksid ebavajalikud andmed ning need, mille töötlemise eesmärk on täidetud ja milleks vajadust või õiguslikku alust pole.

Lisaks tuleks mainida ka paari olulisemat ohtu, mis kaasneb sellega, kui andmekirjeldust pole. Esmalt kulub tohutult aega õigete andmete leidmisele, kui vajadus on veidigi laiem tavaliselt tehtavast tööst. Teiseks on suuri raskusi, kuidas veenduda andmete usaldusväärsuses, eriti kui andmeid ei võeta algallikast, vaid mingitest vaheandmestikest. Kolmandaks on andmed raskesti mõistetavad. See on nii üldine probleem, et on lausa kujunenud välja eraldi valdkond – andmearheoloogia, milles kaevandatakse vanades andmebaasides ja failisüsteemides ning püütakse tuletada leitud empiirilise ainese alusel andmetele tähendust. Kõik, kes on pidanud mõtisklema avatud Exceli lehe veergude tähenduse üle, teavad selle keerukust, töömahukust ja võimatustki.

1.3 Juhise koostamine

Juhis on koostatud kahes järgus:

1. Statistikaameti tellimusel ja Euroopa Komisjoni Struktuursete Reformide Toetusteenistuse (SRSS) rahastatud projekti „Support for the Establishment of Data Governance Services“ („Toetus andmehaldusele. Eesti andmehalduse metoodikaprojekt“) raames 2019. a sügisest 2020. aasta suveni koostatud juhised. Selle töö teostasid ettevõtte Ernst & Young eksperdid

(projektijuht oli Siim Aben, eksperdid olid Kuldar Aas ja Raivo Ruusalepp). Juhise aluseks oli Eestis varem andmehalduse kirjeldamise vallas tehtu ning rahvusvahelised standardid ja parim praktika.

2. Statistikaameti ja Majandus- ja Kommunikatsiooniministeeriumi koostöös, arvestades asutustelt saadud tagasisidet ja andmehalduse arenguid, sealhulgas nende infosüsteemide arendusi, mis üleriigiliselt või asutustes on andmehaldust toetavatena kasutusel või arendamisel. Käesoleva versiooni koostasid Veiko Berendsen ja Annika Uibopuu Statistikaametist ning Kuldar Aas ja Ott Velsberg Majandus- ja Kommunikatsiooniministeeriumist.

1.4 Andmekirjelduse juhise eesmärgid ja andmekirjelduse standard

Juhis on loodud selleks, et hõlbustada organisatsioonidel kvaliteetsete ning riigiülelalt ühtlaste andmekirjelduste koostamist. 2019. a sügisel läbi viidud hetkeolukorra analüüs näitas, et andmete kirjeldamisele on seni pööratud asutustes vähe tähelepanu ja ühtne praktika puudub. Seetõttu on andmekirjelduse juhise peamine eesmärk levitada parimaid praktikaid ja soovitada töövõtteid ning juhendada organisatsioone andmekirjelduse koostamisel ja haldamisel.

Juhise järgimine aitab kaasa andmehalduse raamistiku terviklikule rakendamisele organisatsioonis. Juhise järgi koostatud andmekirjelduste avalikustamine annab teistele organisatsioonidele, andmete potentsiaalsetele kasutajatele ja ühiskonnale tervikuna võimaluse võrreldavaks ülevaateks kogutavatest ja olemasolevatest andmetest. Sellise ülevaate omamine võimaldab andmeid nutikamalt kasutada.

Praeguseks on loodud selle juhisega kokkukäiv andmekirjelduse standard², mis esitab loetelu kohustuslikest ja soovituslikest kirjelduselementidest, nende tähendusest näidetega ning tehnilistest ja semantilistest seostest. Standard on vormistatud selle juhise lisana 1. Standardi rakendamine organisatsioonis tagab kokkulepitud kujul ja semantikaga koostatud kirjelduse masintöödeldavuse. Andmekirjelduse standardi väljatöötamisel on võimalikult palju kasutatud rahvusvahelist praktikat, peamiseks aluseks on andmekataloogi sõnastiku DCAT („[Data Catalog Vocabulary](#)“) rakendusprofiili DCAT-AP versioon 2.1.0³.

Lisaks käesolevale juhisele ja andmekirjelduse standardile on Riigi Infosüsteemi Amet koos Statistikaameti ja Majandus- ja Kommunikatsiooniministeeriumiga on välja arendanud asutustesse paigaldatava andmehalduse rakenduse RIHAKE⁴. Selle eesmärk on võimaldada andmekirjelduse standardi järgi kirjeldada asutuse andmestikke ja neis sisalduvaid andmeid, andmestikes kasutatavaid klassifikaatoreid ja loendeid, koostada andme- ja ärisõnastikke, edastada andmekirjeldusi teistesse süsteemidesse. RIHAKE toetab andmete haldamist asutuses ja võimaldab seeläbi parandada andmete taaskasutatavust ja leitavust.

Ühtsetel põhimõtetel loodud andmekirjeldused tagavad andmete parema leitavuse ja mõistmise, üleriigilise ja rahvusvahelise võrreldavuse ning toetavad tehnilist andmekirjelduse vahetamist infosüsteemide vahel.

² [Andmekirjelduse standard](#). Ver 2.0. (Mai 2022).

³ [DCAT Application Profile for data portals in Europe](#) (ver 2.1.0, 06.12.2021); [DCAT rakendusprofiil Euroopa andmeportaalidele](#) (ver 2.1.0)

⁴ [RIHAKEse lõppkasutaja kasutusjuhend](#). Ver 1.1.1. (17.01.2023)

1.5 Andmekirjelduse juhise sihtrühmad

Juhise sihtrühmaks on asutused ja organisatsioonid, kes loovad ja saavad, töötlevad ja säilitavad oma ülesannete täitmiseks andmeid. Andmete all peetakse silmas andmestikke ja andmeid andmekogudes ja muudes infosüsteemides. Töötlemise all on kõrvuti tehnilise töötlusega peetud silmas andmete kasutamist teenuste osutamisel ehk teiste sõnadega protsessides või menetlustes ning analüüsis, sh statistika tegemisel. Seega on juhise sihtrühm lai: andmeid kasutatakse nii esmase infoallikana toiminguteks ja teenusteks kui ka teisese infoallikana uuringutes, aruandluses ja statistikas, tehes andmete alusel otsuseid. Aga eelkõige on juhise sihtrühm andmehalduse spetsialistid ehk andmehaldurid.

Juhis on suunatud eelkõige organisatsiooni andmehalduse spetsialistidele, kuid samuti andmeteenuseid ja andmemudeleid loovatele IT-spetsialistidele.

Andmekirjelduse juhise peamised kasutajad organisatsiooni sees on andmehaldusega tegelevad spetsialistid, kelle ülesanne on andmekirjelduse koostamise tulemusel saada ülevaade organisatsiooni andmestikest. Samuti on juhise sihtrühmaks IT-töötajad ja tehnilised andmehaldurid, kes korraldavad andmetöötlust, andmeteenuseid ja asutuste-vahelist andmevahetust, vastutavad andmebaaside ja andmeladude sisu eest jms. Ning ka infosüsteemide ja andmeteenuste arendajad. Andmehaldus võiks olla osaks infosüsteemi funktsionaalsusest, nagu seda on arvestamine infoturbega. Andmekirjelduste kaudu andmetest arusaamine tõhustab asutuse infosüsteemide arendamist ja uuendamist. Eriti palju on sellest kasu, kui on vaja andmeid migreerida. Veel üks sihtrühm on organisatsioonide juhid, kes saavad kvaliteetsema sisendi juhtimisotsuste tegemiseks, ning andmepõhise juhtimise eest vastutavad töötajad ja teised andmete kasutajad, kes saavad ülevaate organisatsiooni andmestikest ja nende haldamisest.

Andmekirjelduse juhise on andmehalduri töövahend!

Väljaspool konkreetset organisatsiooni on juhise mõeldud riigi kesksete andmehalduse rakenduste (RIHA, RIHAKE, andmeportaalid) halduritele ja kasutajatele. Nad saavad ülevaate sellest, milliseid andmeid ja kus kogutakse ja säilitatakse, aga kõige tähtsam on, et see võimaldab tagada infosüsteemide vahelise koostalitlusvõime.

Juhisest on abi ka andmehaldusele lähedaste valdkondade töötajatele, eelkõige nendele, kes tegelevad infoturbega, kvaliteedijuhtimisega ning infohaldusega. Laiemalt võib juhisest olla abi ka nendel, kes vajavad teadmisi andmekirjeldustest, avaandmetest ning metaandmete kasust andmete kasutamisel.

1.6 Andmekirjelduse juhise ülesehitus

Võrreldes juhise esimese versiooniga, on käesolevas versioonis tehtud olulisi muudatusi alapeatükkides. Sissejuhatavas osas on enam lahti kirjutatud andmehalduse tähendusväljad ja andmekirjeldusest saadav kasu. Peatükis 2 on põhjalikumalt käsitletud mõningaid olulisi andmehalduse ja andmekirjelduse mõisteid. Peatükk 3 „Andmekirjelduse koostamise korraldamine“ järgib suuresti esimese versiooni ülesehitust ja teksti. See keskendub andmekirjelduse koostamise korraldamisele organisatsioonis. Joonisel 2 on viidatud,

kuidas selline üldine andmekirjelduse koostamise korraldus on seotud konkreetse andmekirjelduse rakenduse kasutamisega. Uudne on esimese versiooniga võrreldes see, et vahepeal on arendatud välja asutustele mõeldud andmehalduse töövahend RIHAKE. Üldistatud pilt andmekirjelduse koostamisest konkretiseerub kahes osas: esiteks konkreetse organisatsiooni töökorralduse vaates ja teiseks kasutava tarkvararakenduse vaates, milleks võib olla RIHAKE või mõni muu rakendus. Lisatud on peatükk 4, mis on kõige praktilisem. See on sõnastike koostamise praktiline juhise. Just sõnastikega seonduv on rakendamisel kõige enam juhendamist nõudev osa, sest see on uudne.

Iga peatükk on kasutatav ka iseseisvalt. Peatükk 2 annab kasutajatele ühtse mõistete ja terminite süsteemi ehk erialakeele ja valdkonna kontseptuaalse mõistmise. Erinevalt varasemast versioonist on andmekirjelduse standard, mis oli enne juhise lisas, vormistatud eraldi dokumendina.

Üldiselt peaks andmekirjelduse koostamine peaks toimuma seda võimaldava rakenduse abil, nagu näiteks RIHAKE. RIHAKEse kasutamiseks on olemas rakenduse lõppkasutaja juhise.⁵



Joonis 2. Juhise praktilise kasutamise soovitus

Kui asutustes on kasutusel oma andmekirjelduse või andmehalduse rakendus, on soovitatav käesoleva juhise, andmekirjelduse standardi ja rakenduse juhise alusel koostada asutusesisene normdokument, mis seob kirjelduse konkreetse tööprotsessiga vastavalt rollidele ja tegevustele, mida just selles asutuses rakendatakse. See normdokument võib olla asutuse kvaliteedidokumentatsiooni osa ning seda saab kasutada andmehalduse töö tõhususe hindamisel.

1.7 Andmehalduse ja andmekirjelduse õiguslik raamistik

Üldseaduseks, millega andmete valdkonda reguleeritakse, on avaliku teabe seadus⁶. Teave, informatsioon ja andmed on mitmetähenduslikud mõisted, millel on kehtivas õiguses kitsamad tähendused. Nii nimetatakse teavet, mille kasutamist ei ole piiratud, avaandmeteks. Osa avaandmeid määratletakse kui väärtuslikud andmestikud. Eraldi on seaduses peatükk ja käsitlustraditsioon andmekogude kohta. Just selle peatüki paragrahvis „Andmekogu vastutav ja volitatud töötaja“ on lõige, mis sätestab, et andmehalduse valdkonda koordineerib Statistikaamet.

Eesti õiguslikus raamistikus on mitu kihti, mis on kogunenud sinna aja jooksul ja millele on viimastel aastatel lisandunud Euroopa Liidu õigusega ühildumise aspekt. Vastu on võetud

⁵ [RIHAKEse lõppkasutaja kasutusjuhend](#) (ver 1.3 15.05.2023) [25.07.2023]

⁶ [Avaliku teabe seadus. RT I 2000, 92, 597.](#)

andmehalduse määrus.⁷ Menetluses on andmemäärus.⁸ Eestis on sel kaks mõju. Esmalt sisuline ehk uued regulatsioonid ja teiseks protseduuriline ehk vajadus Eesti õigusaktid muudatustega harmoniseerida. Kehtivate regulatsioonide hulk on tinginud kohati äärmise üksikasjalisuse ja teisalt laiad konstruktsioonid, millest selle juhise kontekstis on olulisim teabehalduse ja andmehalduse suhe. Andmehaldusel legaalse definitsiooni ei ole. Avaliku teabe seaduse alusel välja antud Vabariigi Valitsuse määrus teenuste korraldamise ja teabehalduse aluste kohta⁹ sätestab andmehalduse täpsemad nõuded ja tingimused:

1. riigi infosüsteemi kuuluva andmekogu vastutav töötaja dokumenteerib ja rakendab andmekvaliteedi seire ja haldamise protsessi, millega tagatakse riigi infosüsteemi kuuluvate andmekogude andmete kvaliteet vastavalt õigusaktidele;
2. Statistikaamet koostab juhised riigi infosüsteemi kuuluvate andmekogude üldandmete ja andmekogus töödeldavate andmete kirjelduse koostamiseks ning andmete kvaliteedi tagamiseks. Juhised kooskõlastatakse enne Statistikaameti veebilehel avaldamist Majandus- ja Kommunikatsiooniministeeriumiga.

Esimene säte piiritleb justkui andmehalduse andmekogudega, mis aga pole siiski praktiliselt parim lahendus, sest asutustel on oma ülesannete täitmisel palju andmestikke, mis pole andmekogud. Need sätted annavad juhiste andmiseks õigusliku aluse ning kuna legaalse definitsioonid puuduvad, ongi täpsemad selgitused ja nõuded juhistes.

Oluline on märkida, et õigus käsitleb koos nii inimloetavat kui ka masinloetavat teavet/andmeid, mis on kas andmekogud, andmestikud või dokumendid. Andmekirjelduse juhise raames on tarvis rõhutada sedagi, et andmekogud, andmestikud ja dokumendid võivad kõik olla andmehalduse objektid, mis seega vajavad andmekirjeldust. Samas on need andmete kategooriad nii andmetüüpide kui sisu mõttes väga erinevad.

Laiemalt käsitlebki kehtiv õigus informatsiooni ja andmeid teabena, isegi osana avalikust teabest. Teisalt käsitletakse kogu digitaalset teavet andmetena.

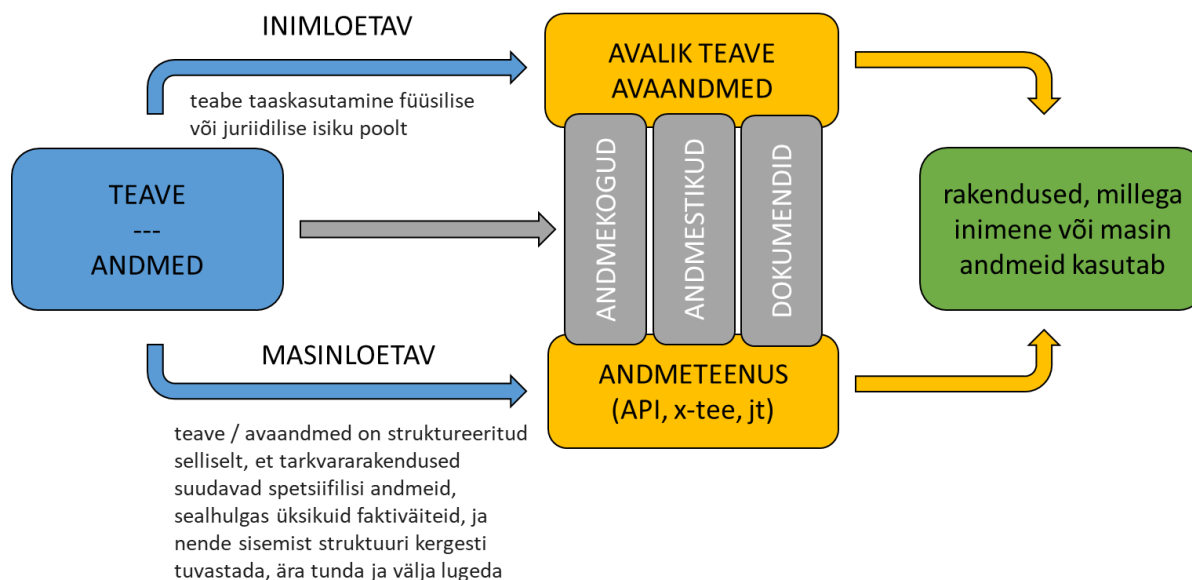
Avalik teave on mis tahes viisil ja mis tahes teabekandjale jäädvustatud ja dokumenteeritud teave, mis on saadud või loodud seaduses või selle alusel antud õigusaktides sätestatud avalikke ülesandeid täites.

Andmehalduse objektid – andmekogud, andmestikud või ka dokumendid –, olgu inim- või masinloetavad, vajavad andmete kasutamiseks infosüsteeme ja rakendusi, kui need pole trükitud paberile. Teisest küljest on tegemist jäädvustatud ja dokumenteeritud teabega avaliku teabe seaduse tähenduses. Seetõttu on andmehalduses otstarbekas vaadelda kirjeldusi nii inim- kui masinloetavuse perspektiivist.

⁷ Euroopa Parlamendi ja nõukogu määrus (EL) 2022/868, 30. mai 2022, Euroopa andmehalduse kohta ning millega muudetakse määrust (EL) 2018/1724 ([andmehalduse määrus](#))

⁸ Euroopa Parlamendi ja nõukogu määrus ühtlustatud õigusnormide kohta, millega reguleeritakse õiglast juurdepääsu andmetele ja andmete kasutamist ([andmemäärus](#))

⁹ [Teenuste korraldamise ja teabehalduse alused. Vabariigi Valitsuse 25.05.2017 määrus nr 88](#). RT I, 31.05.2017, 7.



Joonis 3. Teabe/andmete inim- ja masinloetavus

Teenuste korraldamise ja teabehalduse aluste määrus vaatlleb andmehaldust teabehalduse alategevusena ning asetab selle teabehalduse korraldamise alla ühes dokumendihaldusega. Siiski ei ole teabehaldus, dokumendihaldus ja andmehaldus sarnane ei sisu, töökorralduse, kasutatavate infosüsteemide ega vastutajate poolest. Otse neid üksteise alla panna ei saa, kuid siin käsitletavad kirjeldused saab muuta ühilduvateks.

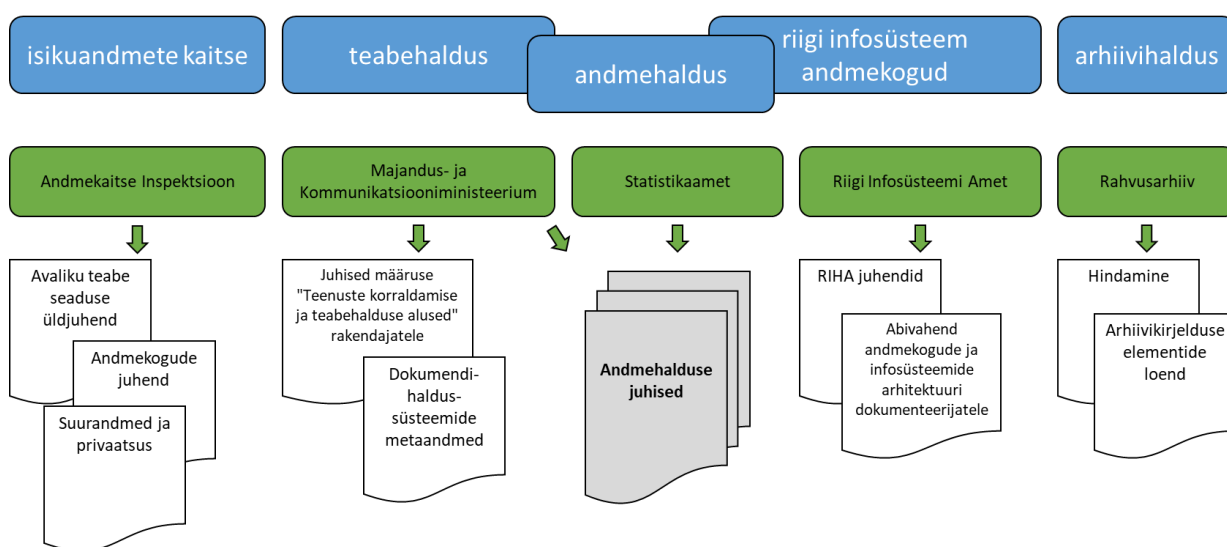
Andmekogude ja andmestike kirjeldamine vajab valdkondade vahel ühtlustamist, mida tehakse kirjelduse ja kirjelduselementide standarditega.

Avaliku sektori kõige olulisem õiguslikult reguleeritud osa on andmekogud. Need on loodud seadusega või seaduse alusel määrusega ning neil on oma põhimäärus. Andmekogu põhimääruse sisu kirjeldab avaliku teabe seadus järgmiselt: „Andmekogu põhimääruses sätestatakse andmekogu pidamise kord, sealhulgas andmekogu vastutav töötleja (haldaja) ja vajaduse korral volitatud töötleja, andmekogusse kogutavate andmete koosseis, andmeandjad ja vajaduse korral muud andmekogu pidamisega seotud korralduslikud küsimused.“ Andmekogu andmete koosseis on andmekogude määrustes loetletud erineva üksikasjalikkusega.

Ühtlustatud kujul on andmekogude kohta esitatavate andmekirjelduse elementide loetelu toodud Vabariigi Valituse määruks „Riigi infosüsteemi haldussüsteem“¹⁰. Selle § 18 „RIHA andmekogude alamregister“ lõige 2 loetleb alamregistrisse kantavate andmekogu üldandmete koosseisu. Selles on ära toodud 27 kirjelduselementi. Need on (osaliselt) realiseeritud RIHAsse esitatavas andmekogu kirjelduses. Lõige 3 toob ära andmekogus töödeldavate andmete koosseisu, milles on kuus kirjelduselementi, ning lõige 4 andmekogu asutamise ja andmete töötlemise aluseks olevate õigusaktide andmete koosseisu, milles on nelikirjelduselementi. Nendega on siin juhises arvestatud. Samas tuleb tähelepanu pöörata dilemmale, et küllalt keerukas on keskpikas perspektiivis hoida ühetaolisena ja otstarbekana õigusaktides, juhistes/standardites ning rakendustes kasutatavat kirjelduste andmekoosseisu. Juhise ja andmekirjelduse standardi koostajad on arusaamisel, et just standard on see instrument, mis on sobivaim ühtlustatud andmekoosseisude haldamiseks.

¹⁰ [Riigi infosüsteemi haldussüsteem. Vabariigi Valitsuse 28.02.2008 määrus nr 58](#). RT I 2008, 12, 84.

Lisaks õigusaktidele (seadused ja määrused) on mitmes andmehaldusega seotud valdkonnas vastutavatel või koordineerivatel asutustel õigus või ülesanne anda juhiseid. Juhised võivad olla soovituslikud või vältimatult kasutust nõudvad, kui tegu on infosüsteemide kasutamise tehniliste juhistega. Nagu on näha Joonis 4. Valdkonnad, asutused ja juhised, on eri asutused välja andnud mitu juhist või juhendit teabe, dokumentide ja andmekogude kohta. Neist põhjalikumad on Andmekaitse Inspektsiooni [juhendid/juhised](#). Andmehaldust puudutab neist kõige enam [andmekogude juhend](#). Juhend annab ülevaate avaliku sektori andmekogude reguleerimise praktilistest küsimustest. Peamiselt käsitleb see õiguslikku reguleerimist, kuid selles on ka muid andmehaldusega enam kokkupuutuvaid osi. Esmalt osa, mis käsitleb andmete koosseisu ja selle fikseerimist andmekogu põhimääruses, ning teiseks andmekogu ja andmete säilitustähtaegu. Siinne andmekirjelduse juhise käsitleb samuti andmete koosseisu. Vaatenurkade erinevus seisneb selles, et see juhise ei käsitle andmekogu põhimäärusega seonduvat.



Joonis 4. Valdkonnad, asutused ja juhised

Teenuste korraldamise ja teabehalduse aluste määruses on andmehaldus teabehalduse alategevus. [Juhises määruse rakendajale](#) on sõnastatud täpsustus: „Teabehaldus ei ole sünonüüm dokumendihaldusele, andmehaldusele, sisekommunikatsioonile ega muule eraldi võetavale tegevusele. Teabehaldus hõlmab kogu asutuse teavet kõigis infosüsteemides ja hoiukohtades, kõiki infovarasid. Andmehaldus hõlmab andmete haldamist relatsioonilistes andmebaasides, geoinfosüsteemides jms infosüsteemides.“ Mainitud juhises on andmehaldus ainult markeeritud, kuna keskne on teabe- ja dokumendihaldus. Käesolevas juhises käsitletakse ühe infovara liigina andmestikke ning seda, kas andmekogu on pigem relatsiooniline andmebaas või põhimäärust omav infosüsteem, pikemalt ei lahata. Avalikku teavet ja avaandmeid võrdsustav käsitlus, mis on avaliku teabe seaduses, vajab eriti avaandmete kategooriate osas täpsustamist, sest käesolev juhise annab avaandmete kirjeldamiseks ette standardi.

Andmekirjeldus on seotud teiste kirjeldustega. Eestis on nii juhised kui praktika kirjelduste koostamise kohta dokumendihalduses ja arhiivinduses. Andmekirjeldusel on sellega palju ühist, kuid on ka erinevusi. Sama võib öelda ka infosüsteemide (andmekogude) tehnilise dokumentatsiooni kohta (sh infosüsteemi arhitektuuri dokumenteerimine). Käesolevas juhises seob andmekirjeldus n-ö teabehalduse poolse vaate, mis kirjeldab infovara või andmestikku ehk midagi, mis on hallatav üksus, ning hallatava objekti sisu ja struktuuri ehk kirjelduselementide käsitluse. Seost arhiivikirjeldusega täpsemalt esitatud ei ole, kuid üldjoontes võib öelda, et

andmestiku kirjeldus on lähedane arhivaali kirjeldusele ning korrektseks loodud andmekirjeldust saab arhiveerimisel taaskasutada. Andmekirjelduse olemasolu teeb ka arhiveerimise lihtsamaks.

Eraldi nimetame aastal 2011 loodud semantikajuhendeid: „Ontoloogiate loomise metoodika“¹¹, „Nõuded RIHA ontoloogiatele“¹² ja „Semantilise kirjeldamise juhised“¹³. Need juhised on tehnilised: „Semantilisel kirjeldamisel (ehk annoteerimisel) käsitletavateks infovaradeks on XML-skeemid XSD-vormingus, andmeteenuste kirjeldused WSDL-vormingus ning RIHA jaoks andmekoosseisude esitamiseks kasutusel olevas XMI-vormingus andmemudelid.“ Nagu näha, on tegu skeemide annoteerimisega, mis on loodud ja kasutatavad kitsa hulga infotehnoloogide poolt ega aita kuigivõrd asutuste teisi töötajaid. Semantilise koostalitlusvõime suurendamiseks on käesolevate juhistega valitud teine tee, mis peaks viima samm-sammult ontoloogiateni, kuid ei hüppa üle sõnastikke käsitlevatest vajalikest vahelülidest. Selliseks oluliseks ja selles juhises käsitletavaks lüliks on sõnastikud.

1.8 Andmekirjelduse koostamise metoodilised alused ja alusstandardid

Andmekirjeldus on osa laiemast metaandmete haldusest ja andmehaldusest. Nii andmehalduse, metaandmete kui ka andmekirjelduste koostamise ja haldamise kohta on mitu laiemat ja konkreetsemat käsitlust, üldisemat ja valdkondlikku mudelit ning mitu rahvusvahelist standardit. Andmekirjelduse juhise koostamisel on mitmega neist otseselt arvestatud.

Andmekirjelduse juhise koostamisel on lähtutud rahvusvaheliselt tunnustatuimast käsiraamatust „The Data Management Body of Knowledge (DAMA-DMBOK2)“ (2nd ed., 2017). Käsiraamatus on esitatud üldise andmehalduse mudeli ühe osana metaandmed. Metaandmed omakorda on esitatud kahes suuremas rühmituses (fassetis).¹⁴ Esmalt on need toodud kolmes suures grupis: (1) kirjelduse, (2) administratiivsed ja (3) struktuursed metaandmed, ning lisatud, et see pole infotehnoloogia valdkonna rühmitamisviis. Teiseks on toodud ära kolm suurt rühma: valdkonna ehk ärimetaandmed (*business metadata*), tehnilised metaandmed (*technical metadata*) ning protsessimetaandmed (*operational metadata*). Esimeses jaotises toodud kirjelduse metaandmete vaade on olnud infoteaduste keskne ja seda on iseloomustatud kui mittetehnillist. Selline arusaam on kitsas ja aegunud – info- ja andmekirjeldused on ammu üksteisega ühilduvates ja automaatset andmevahetust võimaldavates infosüsteemides.

Andmekirjelduse juhises on lähtutud eesmärgist osa vajalikke tehnilisi metaandmeid andmekirjeldusega selgelt integreerida. Eeskätt on silmas peetud andmebaasi tabelite ja veergude tähiseid ja nimetusi. Andmekirjelduse koostamise ja haldamise metoodika lähtub sellest, et andmekirjeldus on osa laiemast metaandmete haldusest ja andmehaldusest, kuid sel on ka kaks kitsamat komponenti. Need tulenevad kasutaja vajadustest leida esmalt sobiv andmestik (ja koostada andmestiku kirjeldus) ning seejärel tutvuda andmestiku sisuga (ja koostada andmeelementide kirjeldus).

Andmekirjeldus koostatakse ühelt poolt mingi hulga andmete kohta, mida võime nimetada andmestikuks või andmevaraks. Andmestiku kirjelduse eesmärk on tagada andmestiku kui terviku hallatavus, mõistetavus, leitavus ja eristatavus. Näiteks andmeanalüütiku vaatest

¹¹ [H-M. Haav. \(2011\). Ontoloogiate loomise metoodika. Ver. 4.](#)

¹² [H-M. Haav. \(2011\). Nõuded RIHA ontoloogiatele. Ver. 2011/r2.6.](#)

¹³ [P. Kungas. Semantilise kirjeldamise juhised V0.4. \(2017\)](#)

võimaldab selline kirjeldus kiiresti leida vajalikud andmed asutuse andmekataloogist või teiste asutuste või üleriigilistest andmeportaalidest.

1. Andmekirjeldus koostatakse ka detailsemal tasandil andmestikus olevate andmete (andmeelementide) kohta. Andmeelementide kirjelduse eesmärk on tagada detailsem arusaam sellest, millised konkreetset andmeväljad andmestikus eksisteerivad, milline on nende sisu, kuidas neid on võimalik kasutada ja kuidas on erinevad andmeelemendid ühes andmestikus või andmestikeüleselt omavahel seotud. Andmeelementide või nende rühmade (infoobjektide, infoolemite) kirjelduse oluline osa on ka nn semantiline kirjeldus – andmeelemendi seos valdkonnas või asutuses kasutatavate terminitega.

Metoodiliselt on oluline, et andmekirjelduse koostamine sõltub sageli valdkondlikust metaandmemudelist. See tähendab, et valdkondadel on erinevad kirjeldusmudelid. Lisaks tuleb kogu aeg võtta arvesse, et andmebaaside tehnikas on andmete modelleerimisel kasutusel nende spetsiifiline, aga oma ulatuse mõttes üldine andmete modelleerimise ja kirjeldamise metoodika. Eestis on selleks tavaline relatsioonilise andmebaasi modelleerimise tehnika. Selles luuakse loogiline ja füüsiline andmebaasi mudel ning sageli ka kontseptuaalne või objekt-orienteeritud mudel. Viimane noteeritakse UML-is.¹⁵ Kuna andmebaaside modelleerimine on üldkasutatav ja UML-i tehnika võimas, siis on üsna sage, et andmearhitektid võtavad seda kui piisavat alust, mida valdkondlik mudel ainult assisteerib. Käesoleva juhise autorite arvates on see üks oluline metoodiline puudus ning valdkondade mudelid, olgu selleks pangandus, geoinfo, raamatukogundus või sotsiaaluuringud, peaksid olema esmase tähelepanu all.

Kasutusel on paar üldist või lausa üleüldist kirjeldusstandardit, millega arvestamata kasutatavaid kirjeldusi tavaliselt ei koostata. Eelkõige on selleks Dublin Core.¹⁶ Dublin Core on metaandmeelementide loetelu, mis põhineb 15 omadusel – need moodustavad 15 põhilist kirjelduselementi –, mida kasutatakse allikate (inforessursside) kirjeldamiseks ning valdkondadevaheliseks inforessursi kirjelduste edastamiseks. Seejuures ei ole kuidagi piiratud see, mida peetakse antud standardi terminoloogias inforessursiks. Nii võibki selleks olla andmeportaal, andmestik või raamatukogus teavik jne. Dublin Core on üle võetud ka Eesti standardiks.¹⁷

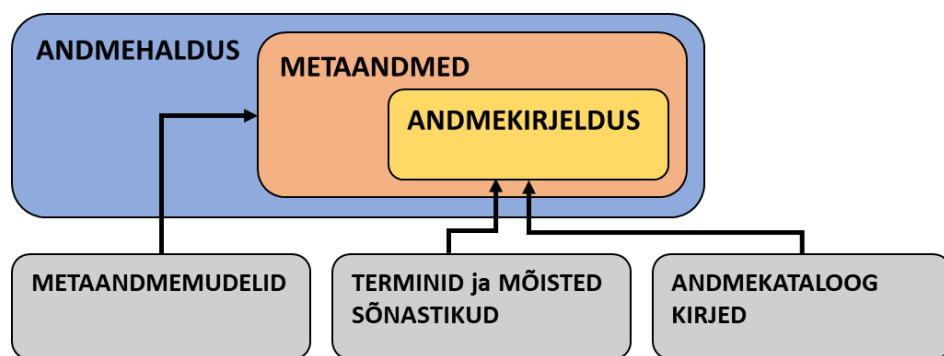
Kitsamalt sotsiaalteaduslikes uuringutes, andmearhiivides ja statistikas on selliseks valdkondlikuks mudeliks DDI (Data Documentation Initiative).¹⁸ DDI ei sisalda ainult kirjelduse metaandmeid, vaid need on laiemad. Mudeli eesmärk on kasutada metaandmeid nii andmete kui ka protsesside juhtimiseks. DDI näide demonstreerib, kuidas paljudes valdkonnamudelites on omavahel seotud andmekirjeldus ja metaandmed laiemalt. Joonis 5 näitab kuidas andmekirjeldus on osa laiemast metaandmete käsitlesest, mis omakorda on oluline osa andmehaldusest. Metaandmetele on oma mudelid ning ka seotud kirjelduse osad, nagu terminoloogia ja sõnastike haldus ja andmekataloog.

¹⁵ Vt. Gillenson, Mark L. Fundamentals of database management systems. (2nd Ed., Wiley, 2013) ptk 5-9.

¹⁶ <https://www.dublincore.org/>

¹⁷ EVS-ISO 15836-1:2019 Informatsioon ja dokumentatsioon. Dublin Core'i metaandmeelemendid. Osa 1: Põhielemendid (kehtiv alates 15.05.2019). EVS-ISO 15836-2:2023 Informatsioon ja dokumentatsioon. Dublin Core'i metaandmeelemendid. Osa 2: DCMI atribuudid ja klassid (kehtiv alates 02.05.2023).

¹⁸ Data Documentation Initiative (DDI) (Lifecycle 3.3, 20.04.2020).



Joonis 5. Andmekirjelduse koht andmehalduses

Andmekirjelduse juhise koostamisel on tuginetud mitmele teisele standardile.

1. Andmekataloogi osas on aluseks „Data Catalog Vocabulary“ (DCAT) (ver. 2, 04.02.2020) ja selle rakendusprofiil Euroopa andmeportaalidele (DCAT-AP) (ver 2.1.0). See esitab mudeli, milles andmekataloogis kirjeldatakse andmestikud, nende levitused ja teenused, kus andmestikke kasutatakse. Lisaks on võimalik andmestikud märksõnastada ühilduvana SKOS-iga.¹⁹ DCAT ei käsitle seda, kuidas kirjeldada andmestiku sisu ja struktuuri. Andmestiku kirjeldus omakorda baseerub paljuski Dublin Core'il.²⁰
2. Terminoloogia osas on standarditest aluseks EVS-ISO 1087-1:2002 „Terminoloogiatoõ. Sõnastik. Osa 1: Teooria ja rakendus“²¹. See on siiski ainult teoreetiline alus, kuidas kirjeldatavat mõisteliselt käsitleda, st kontseptualiseerida, ning mõisteid tähistada ja määratleda ehk teha terminoloogiatoõd. Andmekirjelduste koostamine ei ole siiski terminoloogiatoõ, vaid paljus kasutatava sõnavara ühtlustamine ainult teatava praktilise piirini lõppeesmärgiga andmeid leida, mõista ja kasutada.
3. Sõnastike osas on standarditest aluseks ISO 25964 „Informatsioon ja dokumentatsioon. Tesaursused ja nende koostalitlusvõime teiste sõnastikega. Osa 1: Infootsingu tesaursused“ ja „Osa 2: Koostalitlusvõime teiste sõnastikega“.²² Standardi esimene osa annab süsteemse ülevaate, kuidas tesaaurust koostada. Selles standardis on tesaaurus määratletud kui märksõnastik või struktureeritud sõnastik, milles iga mõiste kohta on esitatud terminid ning toodud selgelt välja mõistetevahelised suhted ning milles eelisterminitele võivad olla lisatud sünonüümid. Andmekirjelduses kasutatav ärisõnastik on tesaaurus, milles on esitatud mõistetevahelised seosed. Andmesõnastik seevastu on termininimekirj.

Lisaks on vajalik kommenteerida standardi ISO/IEC 11179-1:2015 järgimist.²³ Tegemist on metaandmeid ja nende mudeleid käsitleva IT-alase standardiga. Enamik teisi metaandmestandardeid osutab sellele ja kasutab selle terminoloogiat. Käesoleva juhise kontekstis on see standard aga üsna abstraktne. Oleme loobunud andmeelemendi käsitluse rangest järgimisest ISO/IEC 11179 põhjal. Selles standardis on esitatud ühest küljest väga

¹⁹ [SKOS \(Simple Knowledge Organization System\)](#) on tesaauruste, liigitusskeemide, märksõnastike, taksonoomiate jt teadmused organiseerivate vahendite süsteem kasutamiseks semantilises veebis.

²⁰ [EVS-ISO 15836-1:2019 Informatsioon ja dokumentatsioon. Dublin Core'i metaandmeelemendid. Osa 1: Põhielemendid](#)

²¹ [EVS-ISO 1087-1:2002 Terminoloogiatoõ. Sõnastik. Osa 1: Teooria ja rakendus](#)

²² Standard on Eesti standardiks ülevõtmisel. ISO 25964-1:2011 Information and documentation – Thesauri and interoperability with other vocabularies – Part 1: Thesauri for information retrieval; ISO 25964-2:2013 Information and documentation – Thesauri and interoperability with other vocabularies – Part 2: Interoperability with other vocabularies

²³ ISO/IEC 11179-1:2015 Information technology – Metadata registries (MDR) – Part 1: Framework

tehnoloogiline lähenemine, kus andmeelement seotakse mõiste ja väärtusdomeeniga ning andmeelemendi mõiste omakorda objekti klassi ja atribuutidega, mis esitatakse objekt-orienteeritud või olemi-suhte mudelites. Samas on ISO/IEC 11179 aluseks paljudele andmekirjelduse mudelitele, sh näiteks statistikas, kus sellel baseerub muutuja kaskaad. ISO/IEC 11179 käsitluse täielikkus ja täpsus on sageli vajalikud andmetöötlustes ja andmeteenustes, kuid selle mudeli keerukus ja standardi maht ei anna kasutajale, kes lähtub loomulikust või erialakeelest lähtuvale kasutajale sedavõrd palju, et mudeli range järgimine oleks möödapääsmatu. Soovitame seda standardit kasutada eelkõige oma metaandmete mudeli mõistmise taustateadmisenä.

Andmekirjelduse koostamise metoodika areneb praktika käigus. Silmas tuleb pidada sedagi, et eri valdkondade andmestike detailsemaks kirjeldamiseks võib vaja olla juurutada valdkondlikke standardeid. Samuti arvestame rahvusvaheliste arengutega eriti avaandmete portaalide vallas²⁴ ning sellega, mis üleeuroopaliselt on toimumas andmeruumide arengutes.²⁵

²⁴ [Euroopa andmete ametlik portaal](#) (The official portal for European data) [26.07.2023]

²⁵ [European Strategy for Data. A common European data space, a single market for data.](#) [26.07.2023]

2 Andmekirjelduse põhimõtted

2.1 Andmekirjelduse ja metaandmete mõiste

2.1.1 Andmekirjeldus

Eri valdkondades ja kontekstides tähendavad andmed erinevaid asju. Arvutisse salvestatud andmetest aru saamiseks tuleb neid väärti mõistmise vältimiseks alati tõlgendada nende konteksti ja eesmärgi kaudu. Ja andmed pole ainult n-ö arvutis! Andmetele kirjelduse lisamine loob vajaliku konteksti, et saaksime andmeid mõista tähendust omava informatsioonina.

Andmed saavad inimese jaoks tähenduse metaandmete kaudu. Andmevahetuses arvutite vahel veebis kasutatakse semantilise veebi tehnoloogiaid, mis edastavad koos andmetega metaandmeid ja muudavad andmed arusaadavamaks.

Näiteks mõiste „pind“ tähistab erinevaid objekte ehitisregistris, põllumaade või metsamassiivide registrites. Metaandmed kirjeldavad igas registris neid andmeid, mida selles objekti „pind“ kohta kogutakse. Andmekirjelduse kaudu oskavad nii inimesed kui infosüsteemid kasutada andmeobjekti „pind“ õigesti, arvestades selle konkreetse tähendusega.

MÕISTE

Andmekirjeldus on komplekt metaandmeid andmete hulga kohta, milles eri ulatuse ja struktuuriga andmete kohta (andmeobjekt, andmestik, fail) koostatakse terviku ja selle kindlaks määratud osade kirjeldus. Kirjeldatava metaandmete komplekti üksikuid osi nimetatakse kirjelduselementideks ning nendevaheliste seoste kaudu esitatakse kirjeldatava objekti tähendust väljendavat informatsiooni.

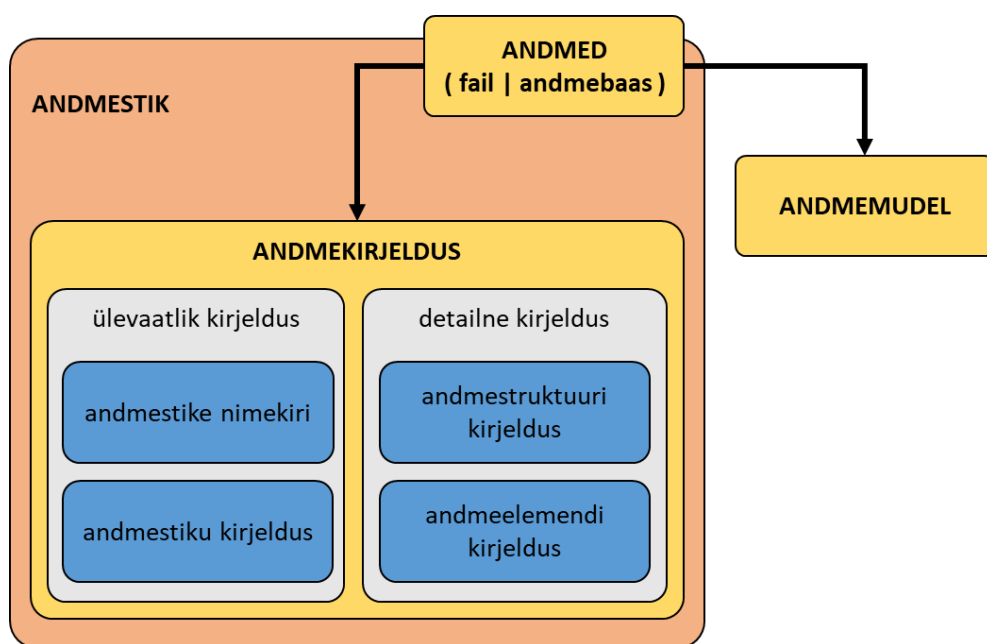
Andmekirjeldus koostatakse ja seda hallatakse erinevate andmehulkade kohta, mis võivad olla oma ülesehituselt kitsamad või ulatuslikumad, struktuurilt lihtsamad või keerulisemad jne. Lühidalt, erisuguseid andmeid on vaja kirjeldada erinevalt. Seetõttu on oluline täpsustada, millist tüüpi andmete kirjeldamist selles dokumendis käsitletakse.

Andmekirjelduse põhiobjektid on andmestik ja andmeelement. Andmestiku kohta tehakse ülevaatlilik kirjeldus ja kokkulepitud detailsusega kirjeldus. Andmestik on sageli fail või andmebaas, aga samas võib hulk faile moodustada ühe andmestiku ja failis olla mitu andmestikku. Andmekirjelduse koostamise ja haldamise tervikmudeli ja sellega kaasnevate protsesside kindlaks määramine on andmehalduses üks otsustava tähtsusega tegevustest. Andmekirjelduse objektide kohasest käsitlest sõltub, kuivõrd ratsionaalne ja kasutatav andmekirjeldus on. Järgnevates osades on mitme andmekirjelduse osa kohta esitatud täpsem selgitus, kuid üldpõhimõte on, et koostatav andmekirjeldus on hierarhiline. Hierarhilisele kirjeldusele on üldisloomulik, et eri kirjeldustasanditel koostatav tervikkirjeldus, mis kokkuvõttes loob selguse andmete tähendusest ja kontekstist. Nii andmestiku kui ka andmeelemendi kohta tehakse detailne kirjeldus, aga see detailsus on erineva sisuga. Lisaks n-ö põhiobjektidele kasutatakse kirjelduses veel muid entiteete. Andmestiku ja andmeelemendi vahele jääb miski, mida võib nimetada andmestruktuuriks, andmeobjektiks jne ning mille mõistmine on samuti andmekirjelduse koostamise eesmärk. Kõige lihtsamnäide on tabel relatsioonilises andmebaasis. Teiseks näiteks on sõnastikud, milles andmete kohta käivad

mõisted on sisuliselt ja terminoloogiliselt ühtlustatud ning moodustuda võib mõistete süsteem või kasutatakse mõisteid kirjeldamisel märksõnastikuna.

Praktikas on vaja andmekirjeldus koostada kas infosüsteemi ja selle osaks oleva andmebaasi kohta või mingi faili kohta, mis on andmekandjale teiseseks kasutamiseks salvestatud. Esimene on näiteks riigi infosüsteemi kuuluv andmekogu. Pea kõik avaliku sektori andmekogud on füüsilise andmemudeli vaates relatsioonilised andmebaasid. Relatsioonilises andmebaasis kirjeldatakse nendes olevaid tabeleid ja tabelite veerge. Tabeli veerule vastab siinses dokumendis andmeelement, aga seda veergu nimetatakse ka näiteks tunnuseks, atribuudiks või muutujaks. Teine olulisem näide on andmete publitseerimiseks tehtud andmekataloog. Kuigi sellistes kataloogides võib olla ka viide näiteks andmeteenuse otspunktile, kirjeldatakse neis pea alati faile. Need failid on kirjeldatud andmestiku tasandil ehk tervikuna, avamata faili sisemist struktuuri ega sisu. Sellised on näiteks avaandmete portaalid. Praktikas tekitab selline kahetisus – kirjelduste üldisus või üksikasjalisus – vajaduse otsustada, kus lähevad piirid. See, mida täpsemalt kirjeldada ja halduse alla võtta, sõltub paljuski konkreetsetest juhtudest.

Üldistatult võib andmekirjelduse koostamist andmete kohta, mis on kas fail või andmebaas, kujutada, nagu on esitatud joonisel 6.



Joonis 6. Andmekirjelduse seos andmetega

Andmekirjeldus tehakse sellise andmete kogumi kohta, mis halduse alla võetakse. See on üldjoontes vähemalt kahetasandilise hierarhiaga: esmalt koostatakse andmestiku ülevaatic kirjeldus ja teiseks koostatakse andmestiku detailne kirjeldus. Ülevaatlikest kirjeldustest moodustub andmestike nimekiri ehk andmekataloog. Oluline ongi andmekirjelduse koostamise juures läbi mõelda, mis on konkreetset juhul andmestikud ja kas neid saab organiseerida ühetasandilisse nimekirja (kataloogi) või on tarvis mitmetasandilist kataloogi. Samuti vajab läbi mõtlemist, millisel andmeobjektide detailsustasemel detailne kirjeldus teha. Tuleb arvestada, et andmekirjelduse koostamine ja selle hilisem haldamine on ajamahukas töö ning läbi mõelda, kes on andmekirjelduse kasutajad. Detailse kirjelduse osas on lisaks andmeelemendi kirjeldamisele oluline andmestruktuur. Teiste sõnadega me nimetame seda struktuuri andmeobjektiks. Näiteks

aadress või isik on sellised lihtsad struktuurid, aga andmeobjektid võivad olla palju keerulisema andmekoosseisuga ja seega ka struktuuriga.

Relatsiooniliste andmebaaside andmekirjeldus, saab baseeruda andmemudelil, millel võib olla füüsiline, loogiline või kontseptuaalne vaade andmetele. Andmebaasid luuakse andmeid modelleerides. Andmebaaside, andmemudelite ja laiemalt andmestruktuuride mõistmine hõlbustab andmekirjelduse koostamist. Andmebaaside teemade tundmine hõlbustab andmekirjelduste koostamist ja andmete mõistmist ja on osaliselt eduks vältimatu.

2.1.2 Metaandmed

Kõige levinum metaandmete definitsioon on „andmed andmete kohta“, aga määratlusi on palju, nagu on ka erinevaid metaandmete sisemisi liigendamisi. Siin on toodud ära kaks määratlust ja need osutavad, et üldiselt mõeldakse metaandmetest IT-s ja infohalduses ühtmoodi.

MÕISTE

Metaandmed on andmed andmete kohta.

Metaandmed on andmed, mis määratlevad ja kirjeldavad teisi andmeid (ISO/IEC 11179-1)

Metaandmetest saame vastused tüüpilistele küsimustele andmete kohta: kes, mida, millal, kus ja miks? Metaandmete kaudu avaldub andmete väärtus varana. Allpool on mõned küsimused, millele metaandmed vastuseid pakuvad.

- Mis andmed on meil olemas, mille kohta?
- Kust need andmed pärinevad, st millised on andmeallikad?
- Kus andmed asuvad: millistes andmebaasides, failidena jagatud ketastel jms?
- Kas andmed on ajakohased ja täielikud ning millised on teised andmekvaliteedi kriteeriumid?
- Kuidas on andmed muutunud nende kogumisest alates ning millist töötlust – nii transformatsioone kui ümberstruktureerimist – on rakendatud?
- Kellel on õigus andmeid kasutada ja kuidas?
- Kas tegemist on konfidentsiaalsust nõudvate või avalikustamist vajavate andmetega?
- Millised riskid on seotud andmete kasutamisega?

Metaandmed on nagu lisandväärtust pakkuv keel, mis peale nimetamise ja liigitamise toimib infosüsteeme ühendava kihina. Metaandmed võimaldavad andmete, tööprotsesside ja teenuste koostoimet sellistes olukordades, nagu andmete sisestamine, otsimine, arvutused ja muul viisil kasutamine nii inimeste kui masina poolt. Andmemudelid, andmeteenused, andmekvaliteet, andmete avaldamine ja andmete elukäigu haldus kasutavad kõik metaandmeid.

Metaandmed jagatakse sageli kolme suurde rühma. Need on:

- kirjelduse metaandmed,
- administratiivsed metaandmed,
- struktuursed metaandmed.²⁶

Esitatud jaotus pärineb USA Riikliku Informatsiooni Standardite Organisatsiooni (NISO) väljaandest. Mõistagi ei ole selline jaotamine ainuke olemasolev. Sellises metaandmete

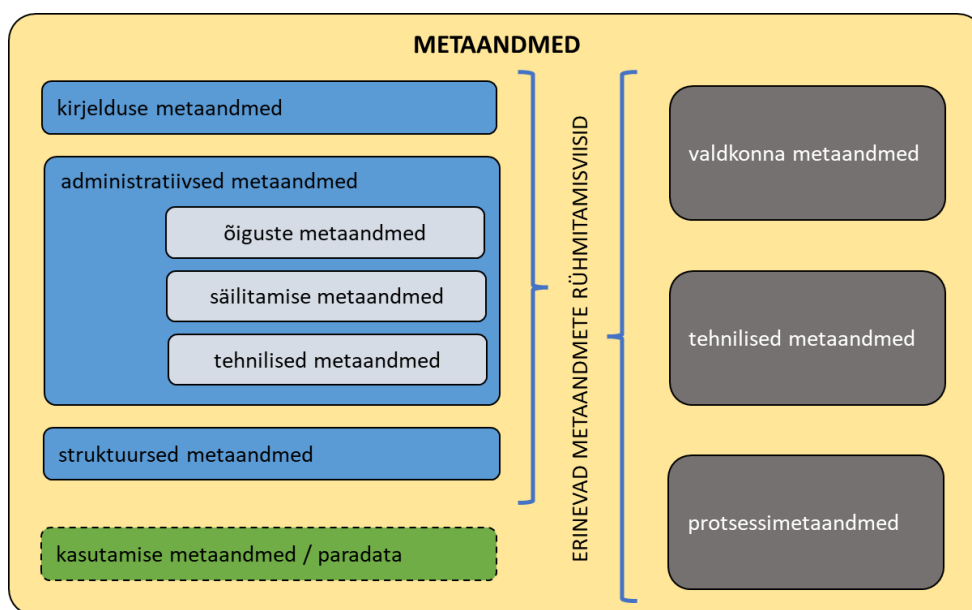
²⁶ [Riley. J. Understanding Metadata: What is Metadata, and What is it For?: A Primer](#) (NISO, 2017)

kolmeses jaotuses on suurematel rühmadel oma kindel eesmärk ja alajaotused; mõnel juhul, nagu näiteks kirjelduse metaandmed, võib kasutusel olla eraldi kirjeldusstandard. Eesmärgid on järgmised:

- Kirjelduse metaandmed on osa, mis võimaldab inforessurssi, st ka andmestikke, üles leida ja neid mõista.
- Administratiivsed metaandmed on suuresti need, millega inforessursi kasutamist juhitakse ehk hallatakse, ja neil on kolm osa:
 - (1) õiguste metaandmed, mis käsitlevad inforessursi autori- ja omandiõigust ning juurdepääsuõigusi ja -piiranguid;
 - (2) säilitamise metaandmed, millega luuakse inforessursi pikaajaliseks säilitamiseks vajalik metaandmete osa, ning
 - (3) tehnilised metaandmed, mis on üldjoontes vajalikud andmetöötluks.
- Kolmandaks on struktuursed metaandmed, mis näitavad inforessursi osade omavahelisi seoseid. Selle juhise kontekstis on need näiteks andmebaasi tabelite, veergude, kirjete ja andmeelementide seosed. Aga need võivad olla ka seosed lehekülgede, veebilehtede ja nende osade ja muude inforessursside osade vahel.

Praktikas on need kolm jaotust küllaltki läbi põimunud, näiteks võib „andmete vorming“ või „õigused“ olla kasutusel nii andmete leidmisel (filtreerimisel) kui ka haldamisel. See tähendab, et need on ühtaegu nii kirjelduse kui ka administratiivsed metaandmed. Seetõttu ei ole ka Eesti andmekirjelduse standardis metaandmete rühmi eraldi esitatud, vaid selle sisu katab kõigist kolmest rühmast kõige olulisemat.

Joonisel 7 on esitatud paar selle dokumendi kontekstis olulisemat vaadet metaandmete jaotamise ja rühmitamise kohta. Lisaks on andmehalduses kui tervikus veel metaandmed andmete elukäigu ja andmete põlvnemise kohta ning paradata ehk andmetöötluks ja muude protsesside dokumenteerimisel tekkivad metaandmed. Sedavõrd ulatuslik vaade andmetele andmete kohta näitab, et metaandmeteta ei saa andmehaldust korraldada, kuid tegelemine metaandmetega ei tohi muutuda nii mahukaks tegevuseks, et selle raskuse all jäävad andmed ise tähelepanuta.



Joonis 7. Metaandmete rühmad

Metaandmete loomine ja haldamine peab olema järjepidev protsess, milleks tuleb asutuses kehtestada reeglid ja tööprotsessid. Selle osaks peaks olema ka andmete hindamine vaatenurga alt, mida üldse kogu andmete hulgast halduse alla võtta. Kui seda lihtsat ent olulist jaotamist tehtud ei ole – ja sageli olukord just selline on – võib jäädagi tohutu hulga andmete ja failide olemasolu lootusetult nentima. Ehkki andmekogude ja andmebaaside osas ei ole see probleem nii terav, siis väärtuslikke andmeid sisaldavate failide osas püsib halduse alla võtmine ja kirjledamine seni praktiliselt lahenduseta.

Kvaliteetne andmekirjeldus ehk kvaliteetsed metaandmed parandavad oluliselt ülevaate saamist andmetest ja andmestikest. Seejuures tuleb silmas pidada, et rohkem või vähem süsteemne metaandmete haldus on võimalik ainult infosüsteemides. Peaasjalikult metaandmetega tegelevaid infosüsteeme ja rakendusi on mitmeid. Selleks võib olla eraldi andmebaas, kus metaandmete loomine toimub sageli pool- või täisautomaatselt. Eri tüüpi ja taseme metaandmeid võib hallata nii eraldi kui ka koos. Üldisel, andmestike kirjelduse tasemel tuleb kirjeldused omakorda organiseerida või kuvada näiteks kas lihtsalt nimekirjana või mitmekülgsema funktsionaalsusega andmekataloogina. Samas nimetatakse andmekataloogiks ka täpselt füüsilisele andmemudelile üles ehitatud kirjeldust ehk andmesõnastikku.

Metaandmete haldus hõlmab reeglite kehtestamist ja tööprotsesside rakendamist, et kogu organisatsioonil oleks võimalik metaandmeid kasutada, jagada, linkida, analüüsida ja töödelda. Täpsed metaandmed ehk andmeelementide täpne kirjeldus, on vajalik nii sisuliseks kui tehniliseks andmete töötlemiseks. Lisaks andmeelemendi tasemele, mis kõige tavalisemas näites on tabeli veerg, on andmed kirjeldatud veel mitmel üldistatud või rühmitatud moel. Nendeks on andmete loogilised ja mõistelised mudelid ning andmestikud (andmekogud jt inforessursid). Samuti võivad mõnel juhul olla kirjeldatud kõik andmeelemendi väärtused, mis on nimelt nii tehtud klassifikaatorite ja loendite korral.

Üldistatult nimetatakse hallatavate metaandmete süsteemi metaandmemudeliks. Selline metaandmemudel võib olla lihtsam või keerukam kirjeldusstandard või terviklik valdkonna andmeid hõlmav mudel. Paljudes valdkondades sellised mudelid ka on. Näiteks kõigil mäluasutuste tüüpidel – arhiivid, muuseumid, raamatukogud – on sellised metaandmemudelid olemas ja igapäevaselt juurutatud. Koos metaandmete haldusprotsessidega võib metaandmemudelit nimetada metaandmete raamistikuks või ka metamudeliks.²⁷

MÕISTE

Metaandmemudel on andmedmudel, mis seletab lahti konkreetselt juurutatud metaandmed või mingi valdkonna standardi metaandmete kasutamise kohta.

Metamudel on andmemudel, mis seletab lahti ühe või mitu mudelit. Nendeks võivad olla andmemudelid, protsessimudelid, ontoloogiad jne. (ISO/IEC 11179-1)

Metaandmemudeli toimimine on seotud andmearhitektuuriga, sest viimane vaatab andmeid nii nende toimimises (töötluses) kui ka mudelites (füüsiline, loogiline ja kontseptuaalne andmemudel).

²⁷ Metaandmete raamistiku kohta on üldiseks standardiks ISO /IEC 11179-1:2015 Information technology – Metadata registries (MDR) – Part 1: Framework

Üheks metaandmete ülesandeks on anda kasutajatele teada, mis tüüpi andmetega on tegemist. Siin ei ole silmas peetud mitte andmete tehnilist tüüpi ehk andmevormingut, vaid andmete kasutamisest või sisust tulenevat tüüpi. Seda võib nimetada ka laadiks või liigiks. Näiteks isikuandmete osas on selliseks tüübiks eraldi kategooria „eriliigilised isikuandmed“, mille töötlemisele rakenduvad erireeglid. Need kategooriad toimivad infosüsteemides kui metaandmed, olles seotud sageli ka ärireeglitega. Sellisteks on ka muid piiranguid sisaldavad andmed, nagu näiteks ärisaladust sisaldavad andmed, konfidentsiaalsed andmed jms. Andmetele omistatakse siin kirjeldatud kategooriad metaandmetena. Õiguses nimetatakse seda kas avalikuks või piiratud juurdepääsuga teabeks. Juurdepääsupiiranguid hallatakse omakorda metaandmetena. Juhime veel tähelepanu sellele, et andmed ja metaandmed võivad olla olenevalt kontekstist nii ühed kui teised. Parimaks näiteks on dokumendiregistrid, mis sisaldavad andmeid asutuse dokumentide kohta ehk metaandmed, aga neid võib käsitleda ka kui just selle infosüsteemi andmeid.

Metaandmed on erakordselt olulised, kui andmestik või andmekirje on määratud põhiandmeteks.²⁸ Selliseid andmeid peab metaandmetega juhtima ja nende kohta tehtud andmekirjeldus peab seda toetama. Riigi andmekogude osas ongi eesmärk määrata täpsemalt, millised andmed on põhiandmed, suunata teised kasutajad just nende andmete juurde ning selle kaudu vähendada andmete kogumise dubleerimist. Selle eesmärgi täitmisel on oluline kogu metaandmemudel, sealhulgas ka andmekirjeldus.

Lisaks neile rühmadele on oluline veel paradata käsitus. Selle kohta, mida paradataks peetakse, on erinevaid lähenemisi. Ühest küljest on paradata lähedane protsessimetaandmetele. Paradata termin on kasutusel pigem uurimusliku andmetöötamise juures. F. Kreuter määratleb paradata kui andmed, mis hõlmatakse küsitluste või loenduste andmetöötamise käigus.²⁹ Paradata on metaandmed, mida võib hõlmata uurimistöö igas etapis. Sellisena on paradata käsitus küllalt lähedane metaandmetele toetuvale uuringu elukäigu haldamise käsitlusele, mida toetab näiteks DDI Life-Cycle.³⁰

Andmete kohta käivate metaandmete, st andmekirjelduse haldus peab olema korraldatud ja järjepidev, et tagada andmete usaldusväärsuse püsimine ja põlvnemise tuvastatavus (*data lineage*). Käesolev juhise ei käsitle andmete põlvnemise haldamise teemat põhjalikult, kuid nii andmetöötluses laiemalt kui ka andmeteaduses kitsamalt on andmete põlvnemisega seotud metaandmete haldusel oluline koht. See on üks vaade andmete elukäigu haldusele.³¹

2.2 Andmed ja andmestikud

2.2.1 Andmed

Andmed, informatsioon ja teave on omavahel seotud mõisted, samas pole ükski nendest mõistetest ise üheselt määratletav. Kõige lihtsam on terminiga „teave“, mis on üldiselt lihtsalt eestikeelne omatermin informatsiooni sünonüümina.

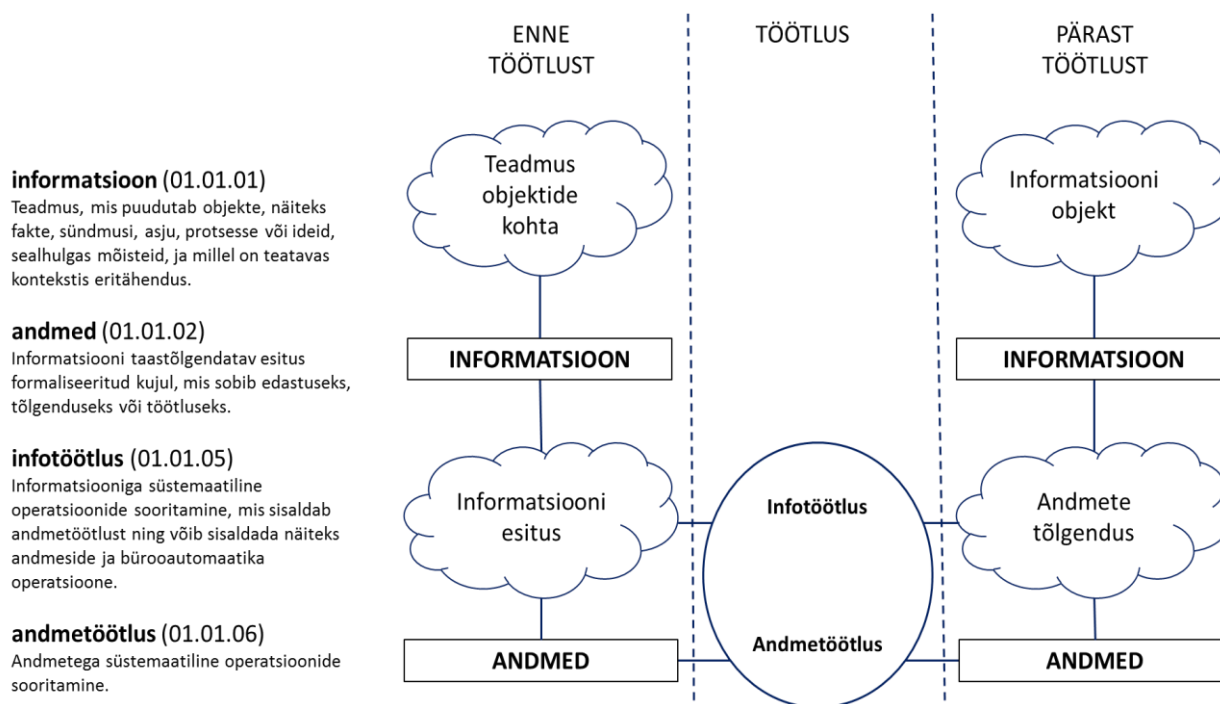
²⁸ Juhis ei käsitle põhiandmete õiguslikku temaatikat ja praktikat Eesti andmekogudes. Siin on osutatud põhiandmetele kui *master data*-le ehk sellistele andmetele, mille tähendus on nii selge ja usaldusväärsus nii suur, et neid saab kasutada alusandmetena teistes toimingutes.

²⁹ Kreuter. F. Improving Surveys with Paradata: Analytic Uses of Process Information (Wiley, 2013) lk 3.

³⁰ [DDI Lifecycle \(3.3\) Documentation](#)

³¹ [Data Lineage](#) (Technopedia)

Termineid „informatsioon“ ja „andmed“ kasutatakse üliharva rangelt üksteisest eristatavate mõistetena. Andmete ja informatsiooni ranget eristamist tehakse infotehnoloogias ja ka seal ainult siis, kui soovitakse insenerliku täpsusega käsitleda info- ja andmetöötlust. Selle infotehnoloogilise määratluse kohaselt on andmed midagi, mis on masinloetav (masin on arvuti).



Joonis 8. Andmete ja informatsiooni käsitus EVS-ISO/IEC 2382-1:1998 (01.01.02) põhjal

Rahvusvaheline standard EVS-ISO/IEC 2382-1:1998 „Infotehnoloogia. Sõnastik. Osa 1: Põhiterminid“ esitab andmete määratluse. Andmete kohta on väga palju määratlusi ja termin on mitmetähenduslik. Ühte tähendust terminile „andmed“ anda ei ole võimalik ega eesmärgipärane. Eesti kehtivas õiguses andmete legaaldeterminatsiooni ei ole, kuid on määratletud mitmed andmete kitsamad kategooriad, nagu isikuandmed, avalikandmed, ruumiandmed. Termin on kasutusel veel mitmes tähenduses, näiteks andmekogu üldandmed, põhiantmed, üksikandmed jne.

Järgnevalt on esitatud kolm väga erinevat mõiste määratlust.

MÕISTE

- (1) Andmed on informatsiooni (teabe) taastõlgendatav esitus formaliseeritud kujul, mis sobib edastuseks, tõlgenduseks või töötlemiseks. (EVS-ISO/IEC 2382-1:1998)
- (2) Andmed on isikuandmed „Isikuandmete kaitse seaduse“ tähenduses või muud andmed, mida töödeldakse andmekogus riigi, kohaliku omavalitsuse või avalikke ülesandeid täitva eraõigusliku isiku ülesannete täitmiseks. (Vabariigi Valitsuse määrus „Riigi infosüsteemi haldussüsteem“) ³²
- (3) Sihipäraselt kogutud või iseeneslikult kogunenud teave, mida kasutatakse statistiliste järelduste tegemiseks. (Statistikaleksikon) ³³

³². Riigi infosüsteemi haldussüsteem. Vabariigi Valitsuse 28.02.2008 määrus nr 58. RT I 2008, 12, 84, 4.

³³. E-M- Tiit, L-M. Toeding. (2019). Statistikaleksikon. Tartu Ülikooli Kirjastus, lk 19.

Sõna „andmed“ mitmetähenduslikkus raskendab mõistest ühest arusaamist, õieti teeb selle teemat detailselt käsitlevate teadlaste arvates võimatuks ja ühtlasi ka mõttetuks. Mitmetähenduslikkus on andmetega tegelemisel alati kaasas.

Andmed ei ole üheselt määratletav mõiste, selle konkreetne tähendus sõltub kasutuskontekstist ja sellest, kas kasutus on üld- või erialakeelne.

Selles juhises on andmete mitmetähenduslikkusele osutatud just seepärast, et sel on andmehalduses üldiselt ja andmekirjelduste koostamise juures konkreetselt oluline tähtsus. Eristuvad kolm laia tähendusvälja, millele me siin viitame, lähtudes L. Floridist:³⁴

- andmed kui faktid;
- andmed kui informatsioon;
- andmed kui bitid.

Andmed kui faktid on arusaam, et andmed on tunnetatud empiiriline tõendus. Tavakeeles väljendatuna räägitakse sel juhul õigetest andmetest ja valedest andmetest. Andmetes peitub potentsiaalselt tõde, õiged andmed võrduvad faktilise tõega. Hea näide on juriidiline fakt, et isik on Eesti kodanik või talle kuulub kinnistusraamatu järgi just see kinnistu just sel hetkel.

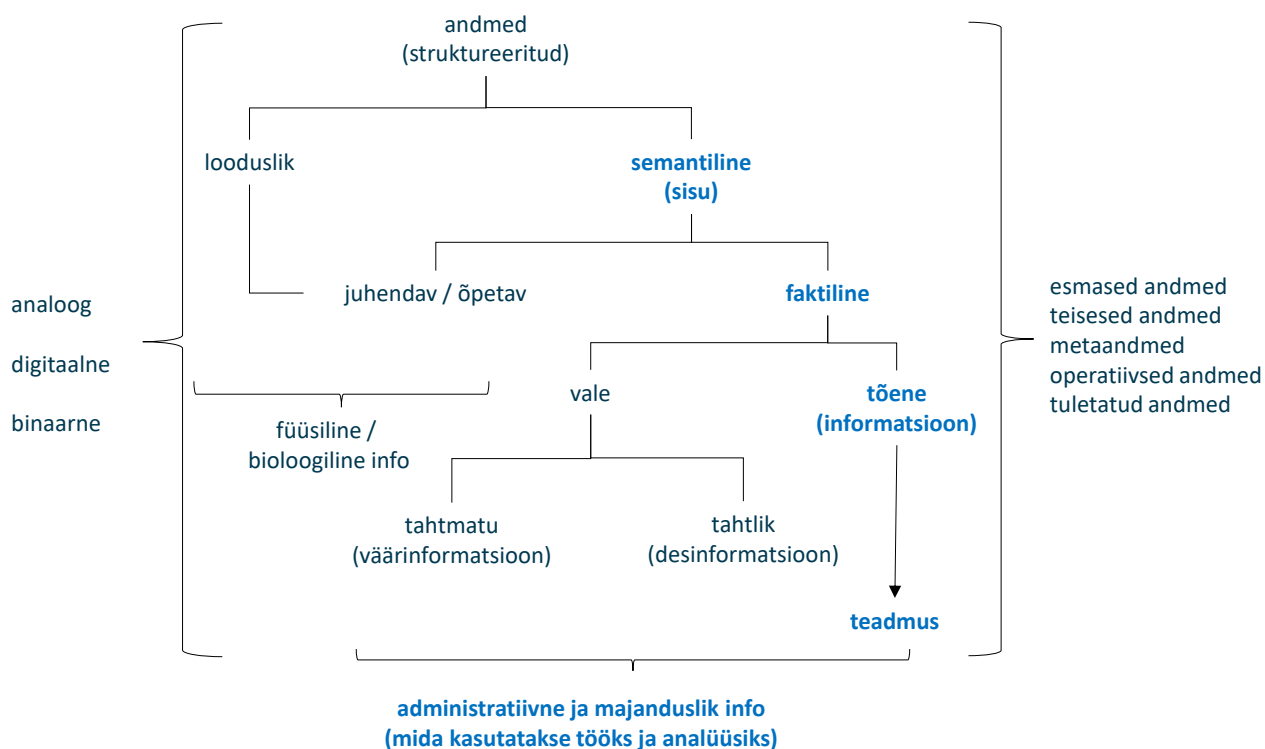
Andmed kui informatsioon on arusaam ja kõnepruuk, kus andmed ja informatsioon on sünonüümid. Selle järgi on andmed inimmõistetav teadmine. Head näited on nimi, isikukood, aadress või õhusoojus. Sel tähendusväljal on palju pistmist sõnade tähendusega.

Andmed kui bitid ehk binaarsed elemendid on infotehnoloogias levinud mõiste. Andmed on masintöödeldavad ja me nimetame seda masinat arvutiks. Selles vaates eristuvad andmetöötlus ja infotöötlus teineteisest. Lisaks binaarsusele on järjest enam IT-s andmetest rääkides jutuks andmetüübid, milles rangelt võttes oleks andmed juba info, kuid nii ei mõelda, kui näiteks räägitakse tekstikujul olevate andmete töötlustest kuupäeva kujule või muud sarnast.

Andmehaldus tegeleb andmete kui bittidega üsna vähe. Päris mööda vaadata sellest tähendusväljast aga ei saa, sest kahes kohas on see oluline. Esmalt puudutab see faile, mis salvestavad erinevaid andmetüüpe, ning teiseks andmeteenuseid ja andmetöötlust.

Andmehalduses ja eriti andmekirjelduse juures on oluline käsitlus andmetest kui informatsioonist. See väljendub ühelt poolt andmete tähenduse seostes metaandmetega ning teisest küljest juba viidatud tavalises sõnakasutuses, mis ei teegi ranget vahet andmete ja informatsiooni vahel. L. Floridi raamatus „Informatsioon. Lühike sissejuhatus“ on esitatud andmete ja informatsiooni seoste käsitlus, milles ta jõuab sisu ja struktuuri käsitledes faktilisuse, tõesuse ja informatsiooni teadmuslikkuseni ehk selleni, kuidas me andmeid kasutame.

³⁴ Floridi, Luciano (2008). Data. In William A. Darity (ed.), International Encyclopedia of the Social Sciences



Joonis 9. Andmete ja informatsiooni seoste käsitus (L. Floridi)

Andmekirjelduse juhises on käsitletakse andmestikke ja struktureeritud andmeid. Viimase väikseim hallatav ühik on andmeelement. Analüütilises käsitluses vastab sellele kirje ehk andmeelemendi esinemise üks eksemplar, mis statistika terminites on üksikmuutuja (*instance variable*). Andmed on struktureeritud väiksemateks ja suuremateks üksusteks, milleks andmeelemendi kõrval on andmekirje, andmeobjekt, andmestik, aga ka andmebaas, fail või andmekogu.

2.2.2 Andmeelement

Termin andmeelement on kasutusele tulnud infotehnoloogiast. Paljud andmekirjeldusega tegelevad standardid, nagu SDMX, DDI, XBRL, on andmekirjelduse suhestanud või vastendanud IT-käsitlusega. Andmekirjelduse aluskäsitluseks on standard ISO/IEC 11179-3:2013 "Infotehnoloogia. Metaandmete registrid (MDR). Osa 3. Registri metamudel ja põhiatribuudid". Selles on esitatud küllalt abstraktne andmeelemendi mudel.

MÕISTE | Andmeelement on kindlas kontekstis jagamatuna käsitletav andmete üksus.

See abstraktne määratlus pole näideteta arusaadav. Standardis on määratlusele lisatud näide: „Andmeelemendil „isiku vanus” saab olla väärtus, mis on kolmekohaline arv”. Järgnev näide esitab ISO/IEC 11179 andmeelemendi käsitluse, nagu see näeb välja tabelis. Joonisel 10 on kasutusel ISO 11179 terminid „objektiklass” ja „väärtusdomeen”, mida me üldiselt selles juhises ega andmekirjelduste koostamisel ei kasuta. Siin kasutatava termini „andmeelement” sünonüümid eri kasutuskontekstides on ka „muutuja”, „tunnus”, „atribuut” ja „andmeväli”.

ANDMEELEMENTIDE MÕISTED (mõiste sisu ja maht)



Joonis 10. Andmeelement ISO/IEC 11179 järgi

Lihtsustades joonisel 10 esitatu praktiliseks kasutamiseks, selgitame seda tabeli käsitlemisel kasutatava sõnavara abil:

- tabeli veerg on andmeelement, mida saab mõisteliselt kirjeldada; tabeli veerg on samuti objekti klass, kuid sel pole suurt tähendust väljaspool objekt-relatsioonilist käsitlust;
- tabeli veeru nimi on konkreetne mõiste tähis, mis antud näidetes ja üldiselt ei ole keeles kasutatav sõna või termin, vaid andmete mugavaks töötlemiseks pandud tähis;
- tabeli read moodustavad sisulised kirjed, kuid selles näites puudub neil kontekst, võib ainult aru saada, et kirjed on isikute kohta;
- tabeli veergude sisu ehk konkreetset väärtused ehk tunnuse või muutuja väärtused näitavad midagi sisuliselt ehk annavad informatsiooni;
- tabeli veergude võimalikud väärtused võivad olla kindlaks määratud kolmel moel:
 - (1) andmetüübiga, mida väärtustes saab kasutada (arv, tekst, kuupäev jms);
 - (2) võimalike lubatud väärtustega (isikul saab olla eelnevalt kindlaksmääratud sugu, see võetakse loendist);
 - (3) andmete esitamise konkreetse kujuga (kuupäev on kujul DD.MM.YYYY).

Andmeelementide kirjeldamisel kirjeldatakse selle tähendust väljendavat mõistet ja andmeelement tähistatakse terminoloogiliselt. Kui andmeelement on seotud väärtusdomeeniga, mille tüüp on võimalikke väärtusi sisaldav loend, võib olla vajalik kirjeldada ka selle loendi kõik väärtused mõistetena. Sellised loendid on näiteks klassifikaatorid või märksõnastikud. Andmeanalüüsis ja statistika tegemisel on loenditega, mida selles kontekstis nimetatakse koodiloenditeks, tegelemine oluline osa andmete mõistmisest. Eri valdkondades on andmeelemendi detailse kirjeldamise ja haldamise vajadus erinev. Statistikas on selleks eraldi muutuja kirjeldamise käsitlus. Paljudel juhtudel nõuavad loendid ja eriti klassifikaatorid eraldi detailsemat andmekirjeldust kui ainult nende väärtuste üldist kirjeldamist andmeelemendina. See puudutab nii loendielementide termineid ja sageli ka neid tähistavaid koode kui ka hierarhiaid ja võimalikke rühmitusi ning versioone ja kehtivust.

Andmeelemendi kirjeldamisel tähistatakse andmeelemendid terminitega, mis mõisteliselt ja tähistatuna üksteisest eristuvad ning millest luuakse andmesõnastik.

Andmeelemendid on omavahel tehniliselt ja sisuliselt seotud. Tehniliselt on need seotud andmestruktuurides, milleks meie näidetes on eelkõige tabel, aga seosed võivad olla ka tehnilistes andmestruktuurides, nagu JSON, XML või muu. Sisulise poole pealt on andmeelemendid omavahel seotud loogiliselt või kontseptuaalselt, kui väljenduda andmebaasidest rääkimise keeles. Nii loogilisi kui ka kontseptuaalseid objekte võib pidada andmeobjektideks. Sisuliselt luuakse andmeelementidest nende rühmitamise teel ärimõisted, mis on samuti andmeobjektid, aga need esitatakse ärisõnastikus.

2.2.3 Andmestik

Termini „andmestik“ kasutusele võtmise vajadus tuleneb sellest, et andmetest rääkides oleks võimalik aru saada, millist hulka andmeid hallata, kas andmeelementi, andmeobjekti, andmestruktuuri või andmebaasi või veel midagi muud. Samuti on tekkinud praktilised raskused, mis osalt tulenevad andmekogu legaalse definitsioonist avaliku teabe seaduses.³⁵ Andmehalduses laiemalt ja eriti andmekirjelduses on vaja terminoloogiliselt selgesti eristuvat mõistet hallatava andmekogumi tähistamiseks. Sellise mõisteta ei saa andmehaldust edukalt korraldada. Selleks ongi andmestik.

Andmestikul on legaalse definitsioon, kuid ehk olulisemgi on see, et andmestik (*data set*) on põhimõisteks andmekataloogides. Seda mõistet kasutab rahvusvaheline andmekataloogi sõnastiku standard (DCAT), mis defineerib, et andmestik on inforessurss, mida hallatakse, millele luuakse avaandmete vaade ja millega seotakse andmeteenused.

MÕISTE

Andmestik on andmete hulk, mis on avaldatud ja mida hallatakse kindla isiku poolt ning millele saab anda juurdepääsu või seda alla laadida ühes või enamas vormingus. (DCAT)

Andmestik on identifitseeritav ja hallatav andmete kogum. (riikliku statistika seadus)

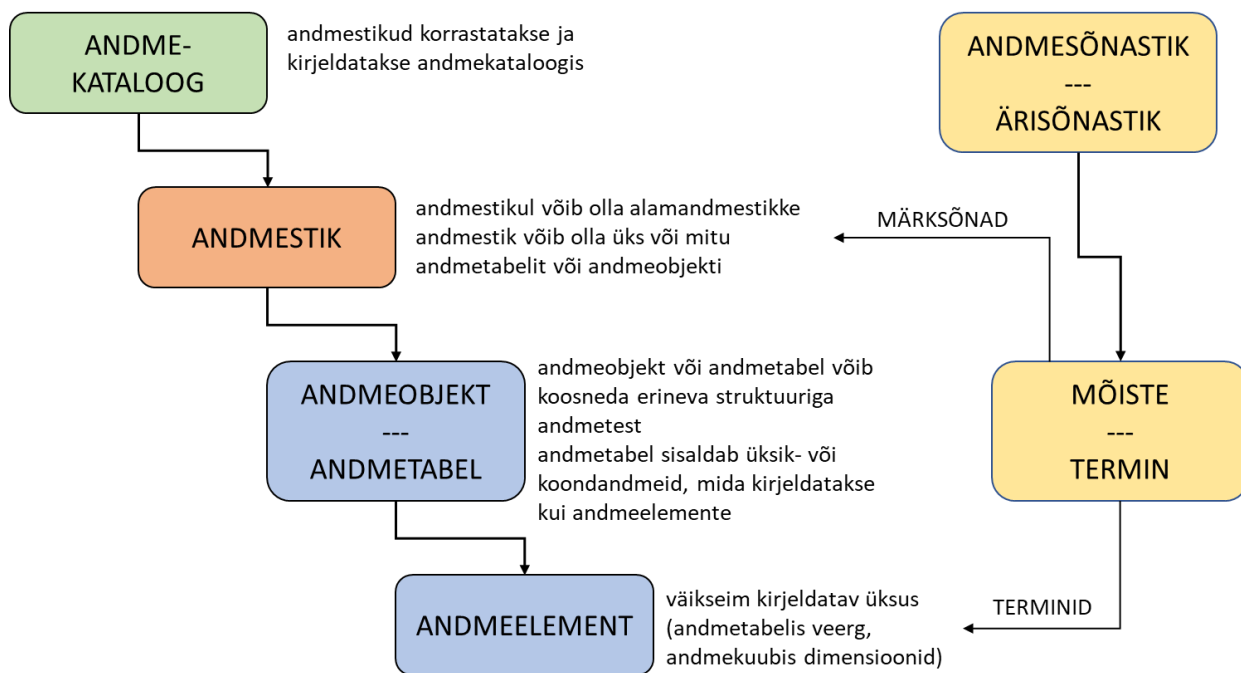
Andmestik võib olla väiksem või ulatuslikum. Neid on eri tüüpi, nagu näiteks riigi oluline register või üks statistiline väljundtabel või üksik fail, mis omakorda võib sisaldada väga erinevat infot tekstist videoni.

Eestis on kasutusel andmestiku mõiste Eesti avaandmete teabeväravas.³⁶ See baseerub tehnilisel standardil „DCAT rakendusprofiile Euroopa andmeportaalidele“ (ver 2.0.1), mille kohaselt on andmestik avaldatud informatsioon ehk meil kasutatavas terminoloogias avaandmed. Seega on avaandmed just andmestikud ning andmeportaal on nende andmestike kataloog.

Termin „andmestik“ ei ole ammendavalt kontseptualiseeritud või lahti mõtestatud ning selle kasutamise praktikagi ei ole Eestis veel välja kujunenud. Seetõttu peame õigeks lisada selgituse, mida andmestiku all mõista. Samuti esitame joonisel 11 mudeli, mis osalt järgib DCAT-i standardit, aga osalt detailiseerib seda.

³⁵ Avaliku teabe seadus § 43¹ Andmekogu, lõige 1) Andmekogu on riigi, kohaliku omavalitsuse või muu avalik-õigusliku isiku või avalikke ülesandeid täitva eraõigusliku isiku infosüsteemis töödeldavate korrastatud andmete kogum, mis asutatakse ja mida kasutatakse seaduses, selle alusel antud õigusaktis või rahvusvahelises lepingus sätestatud ülesannete täitmiseks.

³⁶ [Eesti avaandmete teabevärav](#)



Joonis 11. Andmestik seotuna andmekataloogi ja andmeelemendiga

Eelpool osutasime, et andmestik on identifitseeritav ja hallatav andmete kogum või hulk. DCAT-i standardis esitatud määratlus käsitleb andmestikku andmekataloogi pidamise kontekstis. Iga hallatav objekt andmekataloogis on andmestik ehk mida tahes andmestikuna ei määratleta, see seda ka DCAT-i mõistes andmekataloogis on. Niisugune käsitlus on lähedane mäluasutuste käsitlusele: raamatukogu kataloogis on teavikud, muuseumi infosüsteem kirjeldab museaale ja arhiivis on arhiivaalid. Selline kataloogiv vaade ei ole aga ainuke. Andmestikud tekivad töö käigus, kui mingid andmed teistest töötlemiseks eristatakse; samuti uuringutes, kus kogutakse andmeid vaatluste või eksperimentide käigus. Kahtlemata on see üks olulisemaid andmestiku kontseptualiseerimise viise, et vaadatakse andmete seost mingi tegevuse või töötusprotsessiga. Viimase klassikaliseks näiteks oleks vaatluse, eksperimendi, sündmuse või toimuva või muu eristuva üksikandmetest koosneva andmestiku töötlemine, üldistamine nii, et tekib koondandmetest andmestik. Andmestike tüüpe on palju ning kokkulepitud tüpoloogiat ei ole. Andmekataloogides võib andmestike tüüpideks olla määratud kirjeldusstandardi „Dublin Core“ elemendi „inforessurss“ tüüpide osas kokku lepitud märksõnastik.³⁷ Selles inforessursi tüüpide loetelus on üks element andmestik ja kokku on selles 13 inforessursi tüüpi. Andmestik on selles määratletud kui „struktureeritud info, mis on kodeeritud loendiks, tabeliks, andmebaasiks jne, mis on tavaliselt vormingus, mis võimaldab selle otsest masintöötlust. Näiteks tabelarvutuse lehed, andmebaasid, geoandmed, heliandmed“. Lisatud on ka, et mittestruktureeritud numbrid ja sõnad on teksti tüüpi andmed. Selline eristus, et struktureeritud andmed on andmestikud ja mittestruktureeritud ei ole, on lihtsustatud ja vananenud. See nimekiri ongi aastast 1999.

Samas on struktureeritud ja mittestruktureeritud andmete kirjeldamisel pikk ja eristuv traditsioon. Struktureeritud andmed relatsioonilises andmebaasis on selgelt käsitletavad andmestikena ning täpsustamist vajab küsimus, kas üks andmebaas on üks andmestik või on nende seosed mitmesugused. Vastus on, et seosed on mitmesugused. Ühes andmebaasis võib

³⁷ Dublin Core. [List of Resource Types](#) [25.07.2023]

olla mitu andmestikku ja vastupidi. Esimesel juhul on neid võimalik eristada nii traditsioonilise andmebaasi ja andmemudeli terminites kui ka kontseptualiseerida andmestikeks sisuliselt. Esimese viisi näiteks on andmebaasi skeemide või andmeobjektide eristamine.

Oleme andmestiku ja andmeelemendi vahele asetanudki struktuurse osa, mida nimetame andmeobjektiks või andmetabeliks. Lihtne oleks lahendus siis, kui see oleks alati selline andmetabel, nagu see on relatsioonilises andmebaasis. Sageli see nii ongi, kuid mitte alati. Kirjeldamist vajavad andmestruktuurid ei pruugi kattuda andmetabeliga.

Mis on objekt ja andmeobjekt?

Sellele on võimatu ammendavalt vastata! Üks kõige levinum arusaam on, et objektiks võib nimetada kõike. (Stanford Encyclopedia of Philosophy. [Object](#)) Selline on ka objekt-orienteeritud IT-käsitlus, milles lihtsalt maailma see imeväike osa, mida mingi konkreetne andmemudel ja infosüsteem hõlmab, piiritletakse ja piiritletud objektid nimetatakse. Käsitlus on igal konkreetsel juhul praktiline, eriti arvestades, et mudel peab toimima ehk infosüsteem täitma oma eesmärgi. Teisalt on kõikide objekt-orienteeritud süsteemide objektide ja objektitüüpide (klasside) kogusumma hoomamatu mõistete ja terminite maailm. Eesti keeles võib öelda, et objekt on miski (inglise keeles *something*).

Siin on oluline peatuda küsimusel üksikobjektist ja objektitüübist. Andmebaasides on üldiselt objektid tüüpide kaupa ning need objektitüübid ehk klassid ongi mudeldatud. Selleks on kirjeldatud objekti omadust (*property*), mis kannab andmebaasis nimetust atribuut. Selles juhendis on andmekirjelduse vaates need andmeelemendid (näiteks andmeobjekti „aadress“ elemendiks on „maakond“ või andmeobjekti „lemmikloom“ elemendiks on „sugu“). See, milliseid omadusi, tunnuseid, muutujaid (*features, properties, attributes, characteristics, variables*) me mingi objekti juures kirjeldame, sõltub pea alati eesmärgist.

Kuna andmekirjeldus on väga tihedalt seotud andmebaasidega, osundame Date'i ja Darweni klassikalisele seisukohale (Date, C. J., Darwen, H. Databases, Types and Relational Model: The Third Manifesto (2014)):

1. Terminil „objekt“ ei ole ei ühte täpset ega üldaktsepteeritud tähendust. Mõnikord on see väärtus (*value*), mõnikord muutuja (*variable*), mõnikord tüüp (*type*), aga teinekord selle vastandina eksemplar (*instance*).

2. Nimetusena on objekt kasutusel näiteks (1) iseloomustamiseks graafilise kasutajaliidese stiili, (2) programmeerimiskeelt, (3) analüüsi tehnikat, (4) disaini tehnikat või (5) on kasutusel andmebaaside halduse kontekstis. (lk 7)

Seega olukorras, kus andmeobjekt võib teoreetiliselt olla ja praktikas ongi miski ehk ükskõik mis, saame läheneda ainult praktiliselt. Andmehalduses üldiselt ja andmekirjelduste koostamisel konkreetselt on andmeelementide grupeerimisel andmeobjektiks juba väga spetsiifiline kontekst, mis juhises väljendub kas tabeli pealkirjana või objekti tähistava mõistena.

Andmestikud on ka sellised andmefailid, mida kasutatakse sotsiaalmajanduslikes ja paljudes teistes uuringutes analüüsiks ja andmete avaldamiseks. Selget piiri avaldatava andmestiku, avaandmete ja andmete publikatsioonide vahel on järjest raskem tõmmata. Seda enam, et praktikas on need kõik olulised infoallikad.

Samuti vajab edasist praktika kujundamist küsimus andmebaaside, andmeladude, andmeaitade, andmekogude või koguni andmekeskuste suhestumisest andmestikega. Kindlasti aitaks siin selgust tuua andmevarade ja andmestike seostatuse küsimus, sest andmevarade haldus nende inventuuri või muu nimekirjana võiks olla seostatav andmestike haldamisega andmekataloogis või mujal.

Lisaks peaks olema selgem, kuidas suhestada andmestikud ja digitaalsed dokumendid. Andmestikena mõistetakse sageli ainult struktureeritud andmetest objekte või veel piiratumalt andmeid relatsioonilistes andmebaasides. Aga samuti faile, mis on osaliselt või hästi struktureeritud, nagu DAT- või CSV-vormingus failid, osad tabelarvutuse failid jne. Digitaalsed dokumendid samas on sageli mittestruktureeritud tekst, lihtsad tekstifailid või mõnes kindlas vormingus nagu PDF. Osa digidokumente on kindlasti andmestikud tähenduses, mille kohta see juhis käib, et need on andmehalduse objektid. Kontseptuaalselt ei ole andmestike ja digidokumentide käsitlus vastavalt andme- ja dokumendihalduses selge. See on osaliselt kattuv. Kuni see ei ole selgemaks analüüsitud ja õiguses määratletud, saab lähtuda praktilistest eesmärkidest. Näiteks andmeportaali avaldavad andmeid failidena erinevates vormingutes (lisaks Excelile ja CSV-le ka XML-is või JSON-is) ning sellised failid on käsitletavad andmestikana, kuigi need on käsitletavad ka digidokumentidena. Andmekirjelduses ja digidokumentide kirjeldamisel kasutatavad metaandmed on sageli lähedased või kattuvad. Kui küsida, kas digiallkirjastatud dokumendis lisana olev Exceli fail on dokument või andmestik, siis mõlemat ning seda Exceli faili võib hallata nii dokumendi- kui andmehalduses, määrav on see, kus toimub sellele lisatud kirjelduse jt metaandmete haldus.

2.2.4 Andmetabel

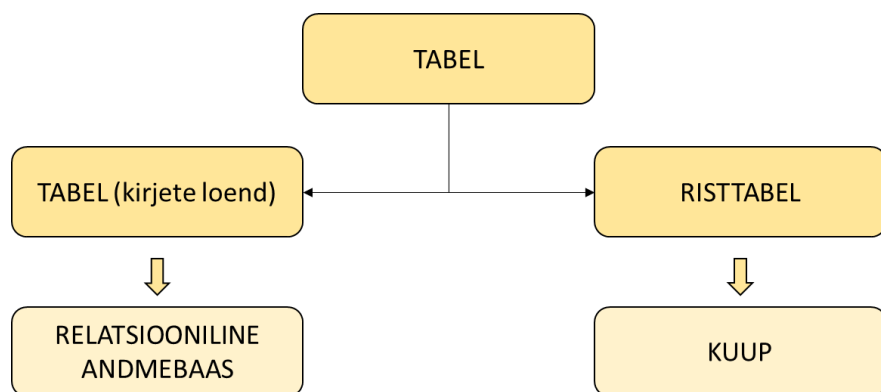
Tabel on info esitamine ridade ja veergudena, mida nimetatakse tabeli sõrestikuks. Read on horisontaalsed ja veerud vertikaalsed, nende kokkusaamise koht on lahter. Tänapäevasemas keeles on lahter andmeväli, millest järeldeb, et lahtri sisu on andmed. Mõistagi võivad need andmed olla väga erinevad nii andmetüübilt kui ka sisult (sh mahult).

Teisalt on tabel on nii tavaline nähtus, et selle peale ei mõelda, mis see siis on. Ning alles siis, kui seda on vaja hallata ja kirjeldada, selguvad probleemid. Tabeli kirjeldamine tundub liiga igapäevane, vähetähtis või hoopis tülikas ja aeganõudev, et seda ette võtta. Teiseks lüüakse käega, sest tabeleid on lootusetult palju. Sageli on tabeli sünonüümiks muutunud Excel.

Järgnevalt on tabelite liigituse seletamiseks kasutatud kahte võrdlevat käsitlust. Esmalt Uno Mereste õpikus „Statistika üldteooria“ toodud tabelite liigendatuse näide. Selle järgi on tabelid lihttabelid, rühmtabelid ja kombinatsioonitabelid. Teiseks tänapäeva statistikas ja andmeanalüüsis kasutusel olevast jaotusest (1) üksikandmeteks ja (2) koondandmeteks. Üksik- ja koondandmed võivad olla esitatud ka muul kujul kui tabelina, kuid sageli on need tabeli kujul.

Lihtsuse ja ülevaatlikkuse jaoks on joonisel 12 tabelid jagatud: (1) kirjetest moodustatud tabeliteks ja (2) risttabeliteks. Risttabelid, milles on mitut tüüpi ridu või veerge, on kuubid ehk mitmedimensioonilised tabelid.

Joonisel on kirjetest koosneva tabeliga seotud relatsiooniline andmebaas, mis koosneb tabelitest ja nende vahelistest seostest. Neile võib veel tinglike tabelitena lisada vahetabelid ehk vaated ([view](#)).



Joonis 12. Tabeli liigid (lihtsustatud tüpoloogia)

Sellise lihtsustatud jaotuse eesmärk on teha kergesti arusaadavaks, kuidas tabelit kirjeldada. Selleks on tabelid jagatud kahe tüübi vahel. Need on:

- LIHTTABEL ehk üksikute nähtuste, ilmingute, faktide, sündmuste jms kirjete tabel;
- RISTTABEL ja KUUP ehk mitmedimensiooniline risttabel.

Lihttabel koosneb ridadest ja veergudest ning sageli on esimene rida ehk päiserida veergu tähistav ja kirjeldav ehk see on veeru metaandmed. Veeru metaandmete kirjeldamine on andmekirjelduste koostamise oluline osa. Veeru metaandmed on lähedased sellele, mida statistikas nimetatakse muutuja kirjelduseks.

Lihttabel võib olla näiteks selline:

		veerud								
		A	B	C	D	E	F	G	H	I
read	1	ID	sugu	eesnimi	perenimi	sünniaeg	elukoht maakond	elukoht linn	elukoht tn_kr	töötamine
	2	1	M	Jaan	Saar	14.03.1997	Harju maakond	Keila	Kohila tn 5-12	töötab
	3	2	N	Mari	Tamm	07.07.2000	Tartu maakond	Elva	Pikk 43	töötu
	4	3	N	Anne	Must	12.01.1987	Hiiu maakond	Kärdla	Lai 8	töötu
	5	4	M	Margus	Valge	08.09.1973	Harju maakond	Tallinn	Tatari 18-23	töötu
	6	5	M	Priit	Kask	31.03.2003	Harju maakond	Tallinn	Ehitajate tee 124-7	töötab
	7	6	M	Toomas	Kuusk	15.05.1954	Tartu maakond	Tartu	Pikk 1-56	töötu
	8	7	N	Kadri	Kask	01.01.2001	Valga maakond	Otepää	Rõngu mnt 7	töötab
	9	8	N	Kersti	Lepik	27.06.1981	Tartu maakond	Tartu	Ravila 9	töötab
	10	9	M	Jüri	Saar	17.08.1994	Viljandi maakond	Viljandi	Lossi 12-3	töötab

Joonis 13. Lihttabel ehk üksikandmed ridades ja veergudes

Andmelaos nimetatakse sellist tabelit sageli faktitabeliks, seda võib nimetada ka transaktsioonide tabeliks või registrikirjete tabeliks. Andmelaos ei esine selline tabel tavaliselt üksikuna, vaid on seotud teiste tabelitega.

Nagu öeldud, eristub sellisest n-ö põhitabelist vahetabel või vaade. Vahetabeleid ja vaateid koostatakse väga erineval eesmärgil. Andmete tervikliku töötlustsükli kirjeldamine, milles on vaade andmete elukäigule (*data life cycle*) või vaade andmete pärinevusele (*data lineage*), kõigile töötamise seisunditele (*data stages*), võib olla üle jõu käiv või ebaotstarbekas. Samas võib olla selliseid andmetöötamise etappide järgnevusi, kus neid on vaja kirjeldada. Sama võib olukord olla arvutuslike vahetabelitega või andmete ettevalmistamisega analüüsiks (viimane on andmelett *data mart*).

Vahetabel võib olla näiteks selline:

sünnikuupäevast arvatud vanus aastates

vanusele vastav grupp

ID	sugu	vanus	maakond	töötamine	vanusgrupp
1	M	26	Harju	töötab	25-29
2	N	23	Tartu	töötü	20-24
3	N	36	Hiiu	töötü	35-39
4	M	50	Harju	töötü	50-54
5	M	20	Harju	töötab	20-24
6	M	69	Tartu	töötü	65-69
7	N	22	Valga	töötab	20-24
8	N	42	Tartu	töötab	40-44
9	M	29	Viljandi	töötab	25-29

Joonis 14. Vahetabeli näide (tunnuste teisendus ja rühmitamine)

Selles käsitluses on eeldatud, et risttabel sisaldab veergude ja ridade andmete mingit kokku koondamist (ehk agregeerimist). Tabel võib olla koostatud selliselt:

dimension

dimension

	töötab	töötü	KOKKU
N	2	2	4
M	3	2	5
KOKKU	5	4	9

kogus(ed)

Joonis 15. Kahedimensionaalne tabel

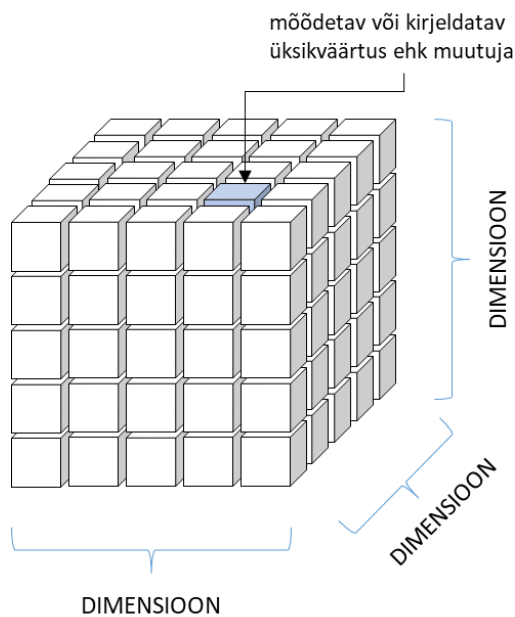
Andmete kokkuvõtt ja töötlemine võib olla kaunis lihtne, nagu ülal toodud summeerimine, või veidi keerulisem, nagu suhtarvude ja keskmiste esitus või nähtusi palju keerulisemalt seletavate näitajate, näiteks indeksite esitus. Andmevahetuseks SDMX-is näiteks on võimalus esitada andmed lõigete (*slice*) kaupa, mis näitab tavaliselt ühte dimensiooni.³⁸

	töötab	töötü	KOKKU
N	2	2	4
M	3	2	5
KOKKU	5	4	9

Joonis 16. Tabeli esitamine lõikena

Tabelit, millel on kolm ja enam dimensiooni, nimetatakse andmekuubiks. Järgnevalt on osaliselt lähtutud W3C andmekuubi kirjeldamise sõnavarast, mis on kasutusel ka SDMX-is. Just sealt on võetud kuubi andmestruktuuri ja komponentide kirjelduse tähistamise viisid.

³⁸ [SDMX – Statistical Data and Metadata eXchange](#) [25.07.2023]



Joonis 17. Andmekuup, selle dimensioonid ja üksik väärtus

Kuupidel võib olla palju dimensioone, kuid mõttetu on neid teha nii palju, et tabeli kujul esitatuna ülevaatlikkus kaob. Siiski on olemas andmete liigendtabelina (Pivot Table) esitamise tehnika, kus kuupi ja selle dimensioone saab lihtsalt pöörata ja muuta.

Mitmedimensioonilises tabelis kirjeldatakse kõik dimensioonid kui muutujad. Dimensioonid võivad olla, näiteks SDMX-is nii on, jagatud koguselisi väärtusi esitavaks ja muudeks dimensioonideks. SDMX-is on lisaks dimensioonide ja mõõdetega seotud muutujate väärtuste loendid (ehk koodiloendid või klassifikaatorid) ühe XML-faili osad.

		dimensioonid							
		töötajad			töötud			KOKKU	
dimensioonid		N	M	KOKKU	N	M	KOKKU		
	Harju			2	2		1	1	3
	20-24			1					1
	25-29			1					1
	50-54						1	1	1
	Hiiu					1		1	1
	35-39					1		1	1
	Tartu	1			1	1	1	2	3
	20-24					1		1	1
	40-44	1							1
	65-69						1	1	1
	Valga	1			1				1
	20-24	1			1				1
	Viljandi			1					1
	25-29			1					1
	KOKKU	2	3	5	2	2	4	4	9

Joonis 18. Mitmedimensionaalne tabel ehk tabelina vormistatud andmekuup

Selline minimaalne eristatus tabelite käsitlemisel peaks võimaldama nende struktuuri ja sisu piisavalt avada ning ei tohiks olla liiga keeruline rakendada.

2.2.5 Andmekataloog

Andmekataloog on nimekiri kataloogitud objektidest, mis andmehalduses on andmestikud. Samas võivad kataloogitavad andmed olla väga eriomelised ja seega on ka kataloogid erinevad.

Laias plaanis võib eristada kolme andmekataloogi tüüpi:

- organisatsiooni andmekataloog sisemiseks andmete halduseks, töötluks ja teenusteks eelkõige füüsilise andmemudeli tasandil, aga ka teiste eri tüüpi andmeobjektide tasandil;
- andmeportaali tüüpi andmekataloog, mis on eelkõige mõeldud andmete levitamiseks ja suunatud andmete leidmiseks neile, kel eelteadmised otsitavate andmete kohta on väikesed;
- mingi valdkonna metamudelit toetavad andmekataloogid, mis ongi valdkondlikud ja mõeldud andmete levitamiseks eesmärgiga neid taaskasutada; sageli haaravad need endasse selle mudeliga ehk metaandmetega andmeid juhtiva komponendi kuni terve andmete elukäigu halduseni välja.

Organisatsiooni sisemiste andmekataloogide põhiomaduseks on eri andmebaasides, teenustes ja failides olevatest andmetest ülevaate saamine. Sageli toetub see andmebaasi tehnilisele kirjeldusele. Selliste andmekataloogide ülesandeks on andmete leidmise, sidumise ja taaskasutamise hõlbustamine erinevatel eesmärkidel, näiteks analüüsiks või uuteks teenusteks. Andmekirjeldus on sellistes kataloogides sageli minimaalne, orienteeritud andmete tehnilisele kasutamisele. Sellist tüüpi andmekatalooge pakuvad toodetena mitmed suuremad andmekataloogi- ja andmehalduseteenustele orienteeritud ettevõtted. Suurte IT-ettevõtete andmekataloogid on näiteks Oracle Cloud Infrastructure Data Catalog ja Microsoft Azure'i Data Catalog. Andmekataloogile orienteeritud ettevõtetest on suuremad näiteks Alation (Enterprise Data Catalog), Informatica (Data Catalog), Denodo (Data Catalog).

Andmeportaali tüüpi andmekataloogide põhiomadus on teha kättesaadavaks juba loodud andmed. Üheks suuremaks seda tüüpi kataloogide rühmaks on tehtud uuringute andmestike kataloogid. See kataloog on sarnane raamatukogu või arhiivi kataloogiga. Andmekirjeldus on bibliokirje tüüpi, kuid kataloogitud andmestikud võivad olla kirjeldatud ka detailsemalt andmeelemendi ehk muutuja tasandil. Viimased on just sellised, mis sisaldavad analüütilisi või statistilisi andmestikke. Suuremad rahvusvahelised näited on [OECD Data](#), [UN Data](#) ja [European Commission Joint Research Centre Data Catalogue](#). Enamik sotsiaalteaduslikke andmearhiive haldab oma repositooriumis olevaid andmestikke nimelt seda tüüpi kataloogides. Sellisteks on näiteks [CESSDA \(Consortium of European Social Science Data Archives\) Data Catalog](#), [Progedo \(INED\)](#) (Prantsusmaa) ja [Aila \(Tietoarkisto\)](#) (Soome).

Teiseks andmeportaali tüüpi andmekataloogide rühmaks on riiklikud ja muud avaandmete kataloogid, samuti Eestis RIHA. Riiklike ja ka rahvusvaheliste seda tüüpi kataloogide üheks funktsionaalsuseks on ka viited andmeteenuste kasutusvõimalustele avaldatud andmete pealt.

Kolmandaks näiteks on andmekataloogid, milles kasutakse mingi valdkonna metamudelit. Metaandmemudeli ülesehitus on sel juhul valdkonnaspetsiifiline. Tavaliselt on see detailsem ja mõistagi valdkonna andmete ja nende kasutuse eripära arvestav. Andmestike, andmeobjektide, andmetabelite ja andmeelementide kasutus võib olla lahutamatu valdkonna arusaamadest ja vajadustest. Sageli on sellele mudelile ehitatud üles terved valdkondlikud andmeteenused. Statistika valdkonnas on sellisteks mudeliteks DDI ja GSIM. DDI on andmekataloogi aluseks näiteks portaalidele [ESS Data Portal](#) (Euroopa sotsiaaluuringu andmed) või [UK Data Service](#) (Nesstar, Ühendkuningriigid), [DataInfo+](#) (Colectica, Uus-Meremaa). Ruumiandmete valdkonnas on sellisteks [INSPIRE geoportaal](#) ja [Maa-ameti geoportaal](#).

Eestis on andmekogusid ja riiklikke registreid pikalt hallatud andmekogude riiklikus registris ja alates 2008. aastast riigi infosüsteemi haldussüsteemis (RIHA). Neid mõlemaid võib pidada riigi andmekataloogideks. Neis on kataloogitud andmeid andmekogu (varem registri) tasandil, RIHAs on olnud võimalus ka andmekogu sisemise struktuuri kirjeldamiseks. Samas on RIHA skoop andmekogude kataloogina olnud lahus iga asutuse andmestike ja

andmekogude ülevaatest. Asutusesisesed vajadused erinevad sageli üleriigilistest. Seda arvestades on suur osa andmehalduse tegevustest ja andmekirjelduse juhisest suunatud asutustele. Lisaks on arendatud asutustele andmekirjelduse halduse töövahend RIHAKE, mis peab aitama asutuse vaate ja vajadused üleriigilise vaatega integreerida. Üleriigilise andmehalduse tegevusel on kaks osa:

- andmekirjelduste halduserakendus RIHAKE toetub andmekirjelduse standardile ning võimaldab edastada asutustes loodud andmekirjeldused automaatselt RIHAsse;
- RIHA ühendatakse Eesti avaandmete teabevärvaga ühtseks andmete teabevärvaks.

Sellisel tekib asutuste ja riigi andmekataloogide või andmeportaalide süsteem, mis võimaldab andmekirjelduste ehk metaandmete ning ka avaandmetena avaldatud failide edastamist sinna või leidmist kesksest kohast. Seda nimetatakse ühtseks teenuspunktiks. Selles teabevärvas on info andmekirjelduste kujul ehk metaandmetena kas andmestiku tasemel või osalt andmete tasemel, kui tegu on avaandmetega. Andmete leidmiseks ja mõistmiseks nii üleriigilises andmekataloogis kui ka asutuste andmekataloogides, on nendes andmete tähenduse kohta käivate mõistete ja terminoloogia koostamise ja haldamise funktsionaalsus: andmesõnastik ja ärisõnastik. Käesolev andmekirjelduse juhise toetab andmekataloogi koostamist, selgitades, kuidas andmekirjeldusi teha.

2.3 Mõisted, terminid, sõnastikud

2.3.1 Termin ja mõiste

Andmekirjelduse oluline osa on esitada metaandmed, kasutades oskussõnavara. Andmekirjelduses on ka pealkirju, kirjeldavaid lauseid ning erinevaid kirjelduselemente, kuid oskuskeele sõnavara ning terminite täpne kasutamine on andmete mõistmise tagamiseks põhiline.

Andmekirjelduses on terminid kasutusel nimetamisel, pealkirjastamisel, märksõnastamisel ning neist moodustatakse sõnastikke. Andmekirjeldused on terminid mitmes rollis:

- terminid kui objekti sisu avavad sõnad ja tähised, mille parim näide on sisu avav pealkirjastamine;
- terminid kui mõisted või mõistmisüksused – õieti mõistete tähistused; kui terminitega tähistamisel luuakse mõistete süsteem, tavatsetakse seda nimetada mõisteliseks kirjeldamiseks;
- terminid kui märksõnad, mis eelkõige tähendab siin, et terminid võetakse kirjeldamiseks juba koostatud sõnastikust.

Termini ja mõiste käsitlemisel on eri elualadel erinevad väljakujunenud vaated. Andmekirjelduse juhise on suunis praktiliseks kasutamiseks. Siiski on juhisesse lisatud ülevaatlik teemakäsitus, kuna andmekirjelduse keskseks osaks on sõnastike koostamine ja haldus. Sõnastike koostamises haakuvad aga mitmed terminitöö või mõistelise mõtlemise käsitlused.

Järgnevalt on toodud viis käsitlust loogikast keeleteaduseni. Erinevate lähenemiste tehnilised võtted on aluseks sõnastike koostamise praktilisele juhisele (Lisa 2).

Termin ja mõiste oskuskeele käsitlustes

Andmehaldust ja andmekirjeldusi võib keeleteaduse sõnu kasutades nimetada oskussuhtluseks ja terminikasutus selles moodustab oskuskeele ehk terminoloogia laiemas mõistes. Eri valdkondade oskuskeeli ja terminiloomet on edendanud Eesti keeleteaduse suurkujud nagu J. V. Veski. Loodud on teadusharude teaduskeeled. Näiteks majandusteaduse ja statistika vallas on pikk ja soliidne oskuskeele käsitlemise traditsioon eelkõige tänu Uno Mereste töödele. Keeleteadus on terminitööd või oskussuhtlust käsitlenud nii praktiliselt kui teoreetiliselt. Kõigutamatu vundamendi aga annab loogika. Kõik need koos standarditega on andmehalduses kasutatavad.

Uno Mereste tööde kogumik „Oskuskeel ja seaduste keeleline rüü. Artikleid ja lühiaurimusi“ (2000) võtab kokku tema oskuskeele käsitluse. Mereste on klassikalise terminoloogiakoolkonna (aluseks terminoloogiateooria rajaja Eugen Wüster ja klassikaline Viini koolkond, nõukogude koolkond) esindaja ning lähenemiselt onomasioloogiline ehk mõistekeskne. See tähendab, et 'asjadel' peaksid olema õiged nimetused ja üldiselt peaksid need olema ajas püsivad. Mõistete tähistamiseks on oluline kasutada õigeid termineid. Heaks näiteks on Mereste selgitus terminite „näitaja“ ja „näitarv“ kasutamisest; statistikas räägitakse sarnaselt „muutujast“ ja „muutuja väärtusest“.

Teiseks oluliseks tugisambaks on Tiiu Erelti tööd, eelkõige tema „Terminiõpetus“ (2007). See on keeleteaduslik käsitlus terminiloomest ja eesti oskussõnavara arendamisest. Raamatu oluliseks lähtepunktiks on üldkeele ja oskuskeele eristamine. Erialakeelele on omane terminoloogiline täpsus, terminid peavad olema võimalikult ühetähenduslikud ning mõisted üksteisest eristatavad. Olulisi teemasid on Ereltil mitu; toome siin ära kaks: (1) termini moodustamise põhimõtted ja viisid ning laiemalt terminiloomet; (2) mõistete ja nende määramise küsimused. Viimases on Erelt samuti onomasioloogiline. Ta kirjutab, et mõiste defineerimine õnnestub, kui oleme: 1) kindlaks määranud mõiste olulised tunnused, 2) selgitanud mõiste suhted teiste mõistetega, 3) teinud kindlaks mõiste koha eriala mõistesüsteemis, 4) andnud heade keeletavade kohase sõnastuse. Defineerides õpitakse tundma objektide sarnasusi ja erinevusi. Võib öelda, et tema terminiõpetuse aluseks on käsitlus tunnustest; väga oluline on oskussõnaloomet, terminivormide (sünonüümsus, võõrterminid jms) teema ning erialakeele ja üldkeele vahekord. Ta küll puudutab, aga ei pane keskele kohale järjest enam läbipõimuvate eriala oskuskeelte omavaheliste suhete teemat.

Kolmandaks oluliseks käsitluseks on Arvi Tavasti tööd, eriti tema ja Marju Tavakari „Mitmekeelne oskussuhtlus“ (2013). See on ühest küljest kõige tänapäevasem lähenemine, teisalt esitab raamat parema oskussuhtluse tarvis onomasioloogilisest vaatest erineva käsitluse. Andmehalduses lähtutakse andmekirjelduste koostamisel paljuski Tavasti kirjutatust. Raamatus peatutakse ka uuematel terminoloogiateooriate käsitlustel, neist olulisemaks võib pidada Rita Temmermani sotsiokognitiivset terminoloogiat, milles terminoloogiatöö lähtekoht on mõistmisüksus (*unit of understanding*). Selle teooria juures on oluline mitte klammerdumine eriala piiresse, vaid eesmärk mõistmisüksuste abil luua laiemaid selgitusi. Temmermani teooria tuntuim rakendus on Euroopa Liidu terminibaas IATE (Interactive Terminology for Europe), samuti selle Eesti vaste: Eesti Keele Instituudi mitmekeelne terminibaas (ESTERM). Temmermani järgi on rangetest loogilistest soo- ja liigitunnustest lähtuvatest

määratlustest olulisem konteksti selgitamine ja kasutusjuhu näite esitamine. Teine terminoloogiatöö käsitus, millele Tavast osutab, on Pamela Faber Beníteze freimipõhine terminoloogia. Tegemist on mõistekeskse lähenemisega, samas võib üks termin tähistada ka mitut mõistet. Mõisted ei esine iseseisvalt, vaid valdkonna sündmuste (freimide) kontekstis. Meie jaoks on sel pigem tunnetuslik tähtsus – me kuuleme pidevalt inimesi kasutamas sõnu, mille täpset tähendust nad ei tea, kuid mis on justkui väga vajalikud sõnad kaasaraäkimiseks kõikvõimalikes uudsetes asjades. Termin „andmehaldus“ on üks selline. Beníteze käsitlust on vaja tunda eelkõige selleks, et saada aru, kas loetud tekst või kuulud jutt on n-ö oskuskeelne, üldkeelne või mingi tehnokraatlik sulam erinevate oskuskeelte terminoloogiast. Viimase žanri näiteks on õigusakti seletuskiri, milles on koos eriala- ja õiguskeel.

Neljanda, rangeima ja klassikalisima terminite ja mõistete käsitluse aluseks on loogika. Siin saame toetuda teemat ammendavale Enn Kasaku raamatule „Loogika alused“ (2014). Kasak esitab väga range käsitluse, mis toetub nii traditsioonilisele kui ka moodsale loogikale ning keelefilosoofiale ja semiootikale. Andmehalduse kontekstis on oluline välja tuua seos arvutiteadusega ja veebitehnoloogiaga, milles semantilised seosed sõnade ja mõistete vahel muudetakse n-ö masinmõistetavaks. Mõistagi tuleb loogiline ja kategooriaalne mõtlemine andmekirjelduste tegemisel kasuks. Arvutistatud ja rangelt loogiline mõtlemine on tohutult võimeline, ent inimesed, kes seda ei valda, peavad siiski lootma jääma arvutitele. Väga vähestel on õnnestunud loogiline matemaatika ja filosoofia praktiliseks pöörata ja andmekirjelduste koostajate seas neid tõenäoliselt pole.

Viiendaks allikaks, millele saame toetuda, on terminitöö standardid. Siin on kaks peamist standardit. Esmalt EVS-ISO 1087-1:2002 „Terminoloogiatöö. Sõnastik. Osa 1: Teooria ja rakendus“. Standard esitab range käsitluse sellest, kuidas keel ja tegelikkus on omavahel seotud, kuidas moodustada mõistet kui teadmüsüksust objekti tunnuste kombinatsiooni põhjal, seda tähistada ja määratleda ning mõisteid sõnade ja sõnedega tähistada. Samuti on standardis kirjeldatud peamised mõistetevahelised suhted (hierarhiasuhe ja assotsiatiivsuhe), terminite liigitus ning suhted tähiste ja/või mõistete vahel.

Teiseks oluliseks standardiks on tesaruste koostamist käsitlev ISO 25964-1:2011 „Informatsioon ja dokumentatsioon. Tesarused ja nende koostalitusvõime teiste sõnastikega. – Osa 1: Infootsingu tesarused“ („Information and documentation: Thesauri and interoperability with other vocabularies. – Part 1: Thesauri for information retrieval“), mis on ülevõtmisel Eesti standardiks. Tesarus on märksõnastik (*controlled vocabulary*) või struktureeritud sõnastik, milles iga mõiste kohta on esitatud terminid, eeliterminite sünonüümid või osasünonüümid ning välja toodud mõistetevahelised suhted.

Ülaltoodud allikate põhjal saame termini ja mõiste käsitlused kokku võtta järgmiselt. Alustame Enn Kasaku määratlustest:

- MÕISTE (ingl *concept*, ld *conceptus*) on „suhtlemise vahendusel moodustunud kokkuleppeline abstraktne objekt, mis esindab suhtluspartnerite lähedasi isiklikke mõisteid samade objektide, nähtuste, suhete jms kohta samade tunnuste abil“. (2014, lk 60)
- TERMIN (mõisteväljend, ingl *term*) on „sõna või fraas, mis mõistet kokkuleppeliselt väljendab ning viitab ka selle mõistega haaratud objektidele“. (2014, lk 60)

- Nendes määratlustes näeme kokkuleppelisuse ja suhtluse rõhutamist, milles isiklikud mõisted ehk isiklik keelekasutus muutub valdkondlikuks, saab oskuskeeleks või lihtsalt kommunikatsiooniks.

Arvi Tavasti ja Marju Taukari õpikus on mõiste ja termini kohta antud järgmised selgitused:

- MÕISTE: „...oleks praktika seisukohalt kasulik teadvustada mõistete olemasolu, ükskõik kuidas neid siis nimetada või mida nimetamiseks kasutatav sõna veel tähendab. Peaasi et kuidagi nimetada. [...] Kui maailmapilt ei sisalda termini ja mõiste vahel vahetegemist, siis ei sisalda ta seda ka pärast terminite väljavahetamist.“ (2013, lk 71)
- TERMIN: „Mõistete keelelist tähist nimetatakse terminiks, sõnaks, väljendiks, keelendiks, fraasiks, tähistajaks, termini väliskujuks või vormiks. Terminite üldise käitumise seisukohalt on need kõik täpselt üks ja seesama. [...] Eristada tuleb terminit vaid pikematest tekstidest, peamiselt määratlusest ja kasutusnäitest. [...] Üleminek termini ja üldkeele sõna vahel on sujuv. Kummastki võib nii spontaanselt kui ka teadliku keelekorralduse tulemusena saada teine ...“ (2013, lk 71)
- Selles käsitluses on selgelt näha, et oskuskeelt ei peeta kuidagi kõrgemaks üld- või tavakeelest ning lähenemine on sotsiokognitiivne. Võib öelda, et selle lähenemise keskne arusaam on tõlkimises kas ühest keelest teise või valdkondade ja oskuskeelte vahel. Need on ka andmehalduse ja andmekirjelduse koostamise juures olulised küsimused – tähtis on, et paratamatu pragmaatiline nimetamine ei toimuks „ükskõik kuidas“, vaid eesmärk oleks terminoloogia korrastamine.

Tiiu Erelt ei esita oma määratlusi, vaid toetub teistele allikatele:

- MÕISTE: autor lähtub standardi EVS-ISO 1087-1:2002 määratlusest, mille kohaselt mõiste on „teadmusüksus, mille moodustab ühene tunnuste kombinatsioon“. (2007, lk 44)
- TERMIN: siin soovib ta lähtuda Sergei Šelovi arusaamast, et vastandusel „termini (leksikaalne) tähendus – terminiga nimetatud mõiste“ ei ole suuremat mõtet ning erialases kasutuses ongi termini tähendus temaga tähistatud mõiste. Termin kui sõna ja sõnaühendi spetsiifika peitub selle tähenduses.
- Praktilisest andmekirjelduse vaatenurgast on Tiiu Erelti arusaam terminist ja mõistest kõige kasutatavam, kuigi rangelt võttes ei ole termin ja mõiste piisavalt eristatud. Ka andmehalduses ei ole leksikaalsel ja kontseptuaalsel eristusel suuremat mõtet.

Kokkuvõtvalt on kõige mõistlikum lähtuda Enn Kasaku definitsioonidest ja lisada neile standardite omad:

MÕISTE

Mõiste on suhtlemise vahendusel moodustunud kokkuleppeline abstraktne objekt, mis esindab suhtluspartnerite lähedasi isiklikke mõisteid samade objektide, nähtuste, suhete jms kohta samade tunnuste abil. (Kasak)

Mõiste on teadmusüksus, mille moodustab ühene tunnuste kombinatsioon. (ISO 1087)

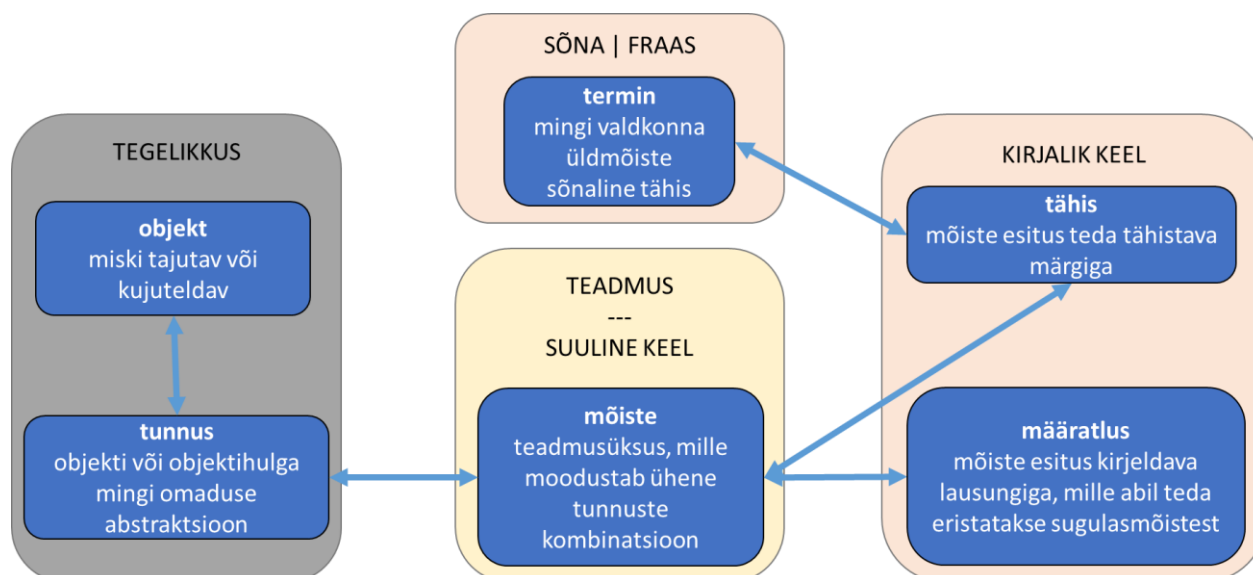
MÕISTE

Termin on sõna või fraas, mis mõistet kokkuleppeliselt väljendab ning viitab ka selle mõistega haaratud objektidele. (Kasak)

Termin ehk oskussõna on mingi valdkonna üldmõiste sõnaline tähis. (ISO 1087)

Termin ehk oskussõna on sõna või fraas, millega mõistet nimetatakse. (ISO 25964)

Joonisel 19 on ISO 1087 standardile toetuv käsitus keele ja tegelikkuse seostest mõistete, terminite ja nende tähistega.



Joonis 19. Tegelikkuse ja keele seos mõiste ja terminiga (ISO 1087)

Standard esitab arusaama, kuidas tegelikkus, teadmus ja suuline või mõtlemisel kasutatav keel on seotud kirjaliku keelega ning rangema terminoloogiaga. Tegelikkuse objektide nimetamise ja tähistamise kokkuleppe rangus või varieeruvus on valdkonniti ja kasutuskontekstiti erinev.

TEGELIKKUS – objekt ja tunnused

Me tajume või tunnetame tegelikkust. Näiteks näeme puud ja metsa. Seda, et puu on puu ja mets on mets, teame, sest keel on meie mõtlemise instrument. Tavakeeles ei ole mõisteline mõtlemine täpne, mis tähendab, et erinevad inimesed võivad puuks või metsaks nimetada erinevaid objekte. Kui tegu on üldteada nähtustega, siis see probleeme ei tekita, kuid kui on tegu objektidega, mille sisu me väga täpselt ei tea, siis ainult nimetamisest nähtuse selgitamiseks ei aita. Paljude kaupade puhul ei pruugi me teada, millega on tegemist, sest näiteks kreemvõib olla nii kosmeetikatoode kui toidukaup.

Puud võib põõsast eristada teaduslikult või tavaarusaama järgi, kuid mõlemal juhul on olemas teatavad tunnused, mis seda määravad. Selline liikide määramine on mitme

teadusharu keskne teema. Näiteks on puutaim põõsast väiksemgi ja hapram, aga ometi teame, et kord saab sellest puu, aga põõsast ei kunagi. Sellised määravad tunnused on mõiste piiritlemise jaoks olulised ning need esitatakse määratlustes.

TEADMUS – mõiste, tähis, termin, määratlus

Mõisted on teadmusüksused ja neid moodustab ühene tunnuste kombinatsioon. Kuulsast süllogismist, et inimene on surelik, järeldub küll, et kõik inimesed on surelikud, aga mitte see, et teisi surelikke liike pole. Järelikult ei ole surelikkus inimest määrav tunnus ehk poeetiliselt inimese määratlus.

Erialakeeltes, mis taotlevad mõistete mahu täpset määratlemist tunnuste kombinatsioonist lähtuvalt, on mõistete defineerimine oluline. Definitsioonid ehk määratlused võivad olla nii valdkondlikud kui ka üldised. Erilised on veel legaaldefinitsioonid, mis sageli annavad määratluse kindla õigusakti piires ehk tähenduses. Mõisteid saab tähistada sõnadega ja muude tähistega, näiteks vesi ja H₂O. Mõistete kokkuleppelist tähistamist sõnadega või ka muude tähistega nimetatakse terminitega tähistamiseks. Termin peab arvestama mõiste mahtu, mis tuleneb tunnustest.

Et andmete kirjeldamiseks valitud sõnad ei oleks juhuslikud, vaid kokkuleppelised, samas laialt, kuid ühtmoodi arusaadavad, tuleb sõnadena kasutada termineid. Andmekirjelduste koostamisel on nii mõisted kui terminid kokkulepped, et just täpselt nii me seda asja nimetamegi. See kokkulepe võib olla üldine keelealane kokkulepe, kus me ühes keeleruumis teame enam-vähem täpselt mõne sõna tähendust. Näiteks on selline teada tähendus sõnadel sünd ja vesi. Sel juhul ei tehta neid kokkuleppeid kirjeldades. Sageli tuleb aga kirjeldamisega koos teha terminitööd ehk otsustada, milline on mõiste tähistamiseks kasutatav eelistermin. Andmekirjeldustes puututakse kokku ka tähistamise probleemidega. Kui tegu on ühest sõnast koosneva terminiga, siis on tähistamine selge, aga mõnel juhul on kasutusel lühendid või akronüümid, mis võivad teisel juhul olla keelatud. Kui termin ei ole ühesõnaline ega isegi mitte liigierisust väljendav liitsõna, võib see olla terminifraas, mis annab mõiste edasi kitsamas või laiemas kasutuskontekstis.

2.3.2 Andmekirjelduses kasutatavad sõnastikud

Sõnastike kasutamine andmekirjelduses aitab kaasa kirjelduse ühetaolisusele ja arusaadavusele ning on eelduseks kirjelduse masintöödeldavusele, arvestades semantikat ja hõlbustab andmeteaduse meetodite rakendamist. Seetõttu ongi selles juhises käsitletud andmete mõistetavaks tegemist sõnastike abil. Sõnastik võib olla esitatud (1) lihtsa terminiloendina või koos sõnaseletustega, mis määratlevad need terminid mõne valdkonna jaoks; (2) taksonoomiana või tesaurusena, kus mõisted on liigitatud hierarhilistesse või sisu-mahu seostesse ja need võivad olla varustatud sõnaseletustega või määratlustega; (3) ontoloogia, mis liigitab valdkonna teadmusüksused klassidesse ning esitab nende vahelised seosed ja nende omadused/tunnused masintöödeldaval kujul. Masintöödeldavate ontoloogiate koostamine või nende tuletamine sõnastike koostamist vahele jättes ei ole Eestis osutunud edukaks lähenemiseks semantilise koostalitlusvõime edendamisele.

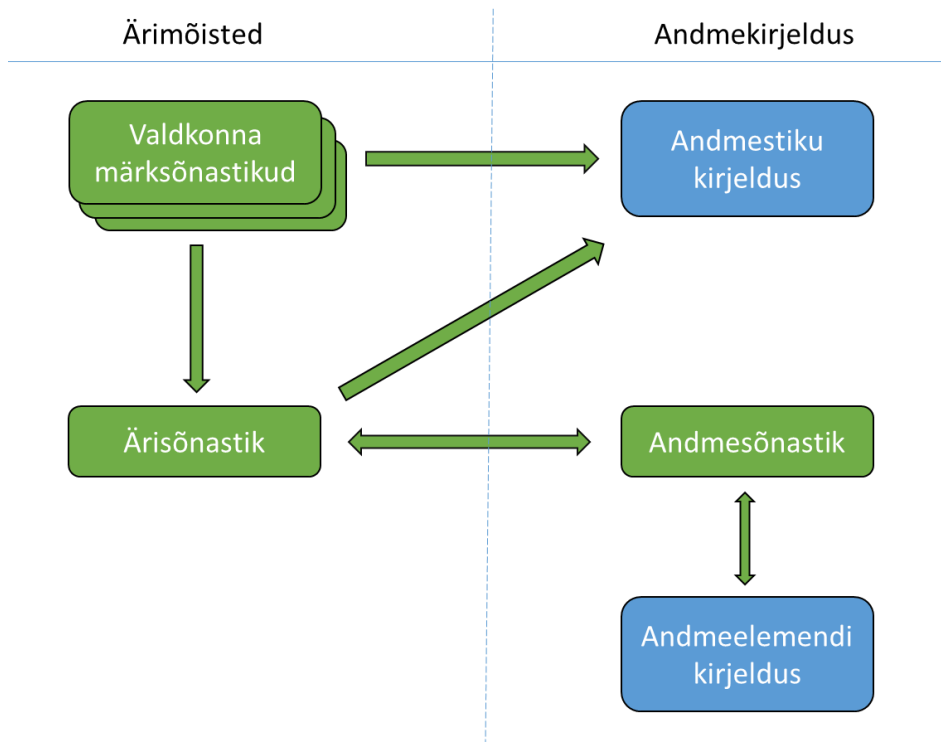
Praktiline sõnastike koostamise juhise on esitatud eraldi dokumendina lisas 2.

Andmekirjelduse koostamisel ja haldamisel kasutatakse kolme tüüpi sõnastikke:

- ärisõnastik (*business glossary*)

- andmesõnastik (*data dictionary*)
- märksõnastik (*controlled vocabulary*), mis on kas üldine või valdkondlik

Joonisel 20 on seletatud, mida andmekirjelduse kontekstis on mõistetud (1) andmesõnastiku, (2) ärisõnastiku ja (3) valdkonna märksõnastiku all.



Joonis 20. Andmekirjelduse ja ärimõistete sõnastike üldine mudel

Valdkonna märksõnastik on märksõnade korrastatud loetelu, kus esitatakse valdkonna terminite ja nende seoste kirjeldused. Valdkonna märksõnastikud on kasutusel valdkonda hõlmavate teadmiste ja mõistete piiritlemiseks ning nende kokkuleppeliste seletuste esitamiseks. Valdkonna märksõnastikke kasutatakse ühe terminite allikana ärisõnastike loomisel ning andmestike kirjeldamisel. Andmestiku kirjeldamiseks kasutatakse valdkonna sõnastikust ja ärisõnastikust võetud märksõnu, mis seovad andmestiku organisatsiooni tegevuste ja teenustega.

Eestis on olemas ka valdkondadeülene üldine märksõnastik (EMS). See on kõiki ainevaldkondi hõlmav tesauruse struktuuriga märksõnastik, mida seni on rakendatud peamiselt raamatute, artiklite ja muude teavikute eestikeelseks märksõnastamiseks ja infootsinguks. Kuna selle 61 000 terminit on jaotatud kuuekümneks valdkonnaks, sobib see kasutamiseks ka valdkondliku sõnastikuna või selle loomise alusena. EMS sisaldab ajamärksõnu ja kohanimesid, kuid ei sisalda isikute, asutuste ega organisatsioonide nimesid. EMS toetab masinalt-masinal-päringuid MARCXML-vormingus.

Valdkondlikud märksõnastikud võivad olla esitatud ka klassifikaatoritena, näiteks valitsemisfunktsioonide klassifikaator (VFK) või majanduse tegevusalade klassifikaator (EMTAK/NACE), õiguse valdkondi esitava taksonoomiana, näiteks EuroVoc või muu taolise valdkondliku jaotusena.

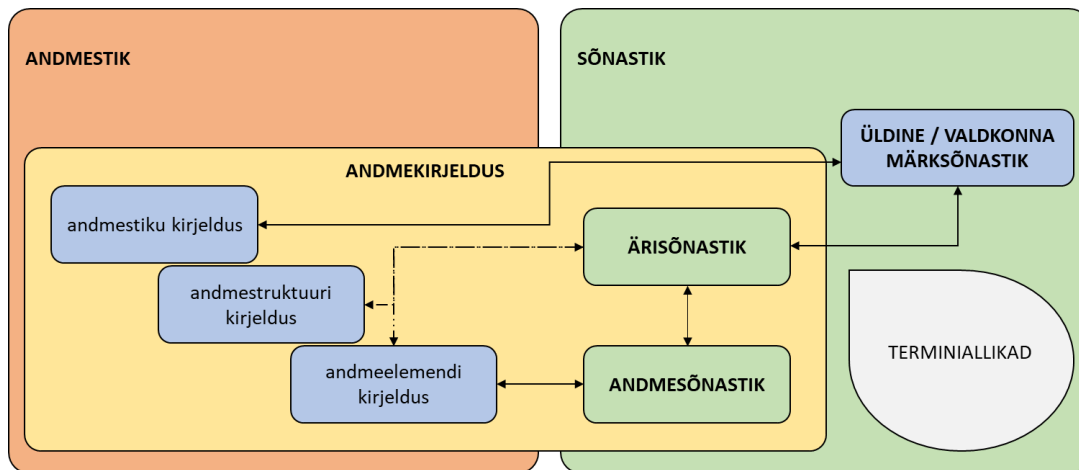
Valdkondade erialasõnavara sisaldavad sõnastikud ei pruugi olla otseselt märksõnastikud, vaid võivad olla terministandardid või mitmekeelsed tesaurused. Oluline on see, et neis kokku lepitud terminid ja mõisted on ärisõnastikus kasutatavad.

Ärisõnastik on organisatsioonis või organisatsioonideüleselt kokku lepitud sõnavara ehk mõistete ja terminite kogum, milles asutus igapäevaselt räägib või mis on kasutusel õigusaktides ja asutuse dokumentatsioonis. Sõnastik esitab organisatsiooni tegevuste ja andmete kohta piisavalt täieliku sõnavara, sisaldades nii üldistavaid kui ka detailsemaid mõisteid. Ärisõnastikke võib asutusel olla üks või mitu; nendes sisalduvad mõisted-terminid võivad pärineda nii valdkonna sõnastikest kui ka üldisest märksõnastikust. Ärisõnastik määratleb ka organisatsiooni tegevusega seotud mõisted. Ärisõnastik on reeglina tesauruse struktuuriga sõnastik, mis võimaldab näidata mõistetevahelisi seoseid. Ärisõnastiku kaudu saavad andmestikes sisalduvad andmed endale konteksti ja andmeobjektid tähenduse. Ärisõnastiku mõistete kaudu on võimalik andmed ja nende kasutamine siduda organisatsiooni tööprotsesside ja teenustega. Ärisõnastiku koostab ja seda haldab organisatsiooni äripool ja selle kasutusala on üldjuhul laiem kui üksnes andmekirjelduse toetamine.

Andmesõnastik on ühelt poolt terminite loetelu ning teiselt poolt esitab terminitena andmeelementide sisu. Andmeelement võib olla tähistatud selle sisu väljendava inimloetava sõna või fraasiga, aga selleks võib olla ka akronüüm, lühend või muu sõne või tähis, mis on masinloetav ega pole lisaselgituseta mõistetav. Andmesõnastiku terminid seotakse vahetult andmeelemendiga. Andmeelemendil on tähis ja üldjuhul ka kirjeldus, andmesõnastiku terminil on nimetus ja määratlus. Andmesõnastik on ka andmemudelil kasutatavate andmeobjektide ja andmeelementide kirjelduste loetelu ja teisest küljest andmekirjelduse käigus tekkinud terminiloend. Andmesõnastik kirjeldab andmeid mõistetenähtena vastavalt nende semantikal andmestikes. Andmesõnastik sisaldab ka andmete kasutamiseks vajalikke metaandmeid, näiteks konkreetse andmeelemendi andmetüüpi, lisaks võib sisaldada ka muid metaandmeid, nt juurdepääsupiiranguid. Andmesõnastiku koostamiseks vajalik sisu saadakse enamasti andmebaasi füüsilisest mudelist ja seotakse ärisõnastiku terminitega.

2.4 Tervikkirjeldus

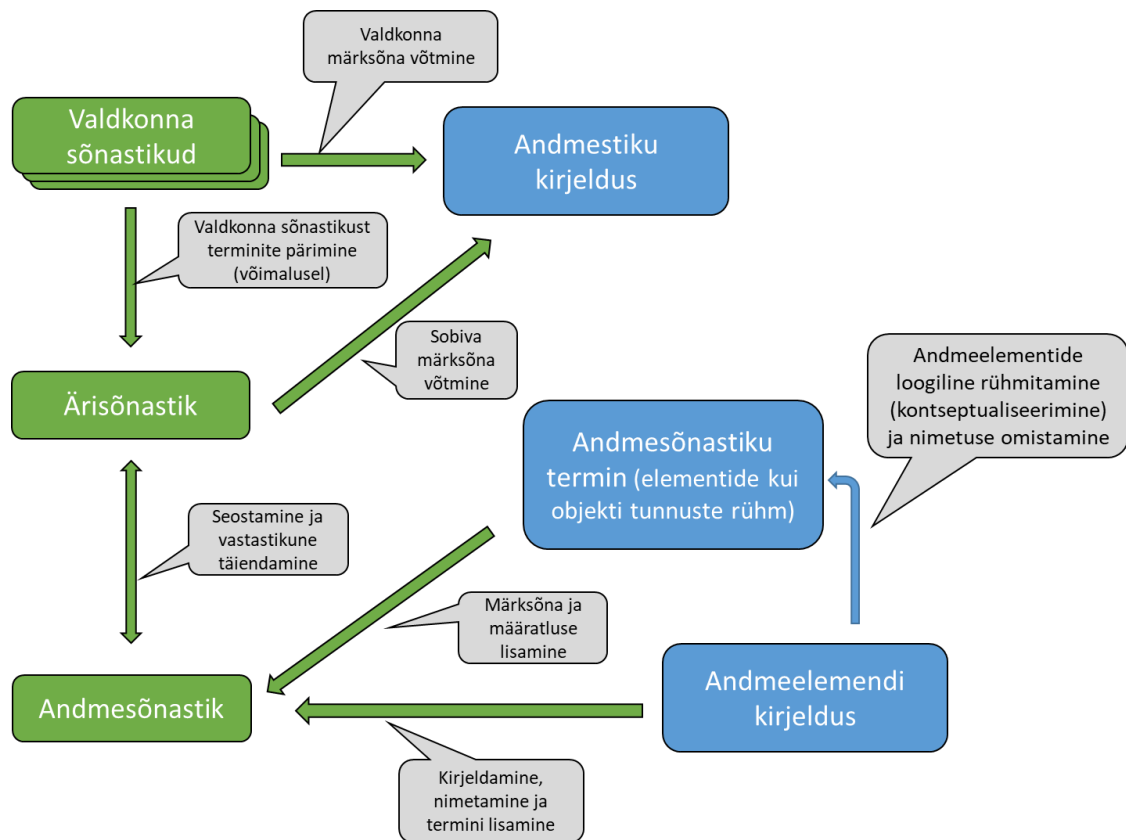
Terviklik andmekirjeldus moodustub kolme kirjeldatava komponendi üheskoos või seostatuna kirjeldamisel. Seosed on nii struktuursed kui ka mõistelis-terminoloogilised. Nendeks komponentideks on joonisel 21 esitatud osad: (1) andmestiku kui terviku kirjeldus ehk üldkirjeldus ja sellega kirjelduse sisu mõttes lähedane andmestruktuuri ehk andmeobjekti kirjeldus; (2) andmeelemendi detailne kirjeldus ning (3) ärisõnastiku ja andmesõnastiku koostamine.



Joonis 21. Tervikliku andmekirjelduse komponendid

Tervikliku andmekirjelduse koostamisel tuleb silmas pidada nii põhimõtteid ja praktilist vajadust, mida on kirjeldatud peatükis 2.1.1, kui ka korralduslikku külge, mida on kirjeldatud peatükis 3.1. Sellist asja nagu lõplikult valmis tervikkirjeldus ei ole olemas. Mingi osa kirjeldusest on tavaliselt töös. Ometi on mõned osad kirjelduses stabiilsemad kui teised. Stabiilsemad on andmestiku üldkirjeldused, samuti on suhteliselt stabiilne ka ärisõnastik. Kuna need mõlemad annavad üldise ülevaate tervest organisatsioonist, esimene andmestike tasandil, teine sõnavarana, siis moodustavad need selgroo, millele pidevalt muutuvad andmed saavad toetuda. Seega on andmestike üldkirjelduse koostamise ja haldamise vahend ehk andmekataloog tervikkirjelduse koostamisel ja haldamisel hädavajalik. Andmestiku kirjelduse aluseks olev rahvusvaheline standard „Andmekataloogi sõnastik“ („Data Catalog Vocabulary“, DCAT) toetab sellise kataloogi koostamist.

Tervikkirjeldus koostatakse ka iga kirjeldatava objekti kohta. Kohustuslikud ja soovituslikud kirjelduselemendid on toodud andmekirjelduse standardis (Lisa 1). Sõnastikud on ühenduslülid andmete struktuursete osade ja andmeelemendi kirjeldamise vahel. Sõnastikevaheline infovahetus on esitatud joonisel 22, kus on välja toodud andmestiku ja andmeelemendi kirjelduste seosed sõnastikega. Andmesõnastik on andmestiku füüsilise mudeli (andmeelementide) ja ärisõnastiku ühenduslül, sarnanedes loogilisele andmemudelile, ainult ilma mudelile omaste seosteta. Andmesõnastik sisaldab sarnaselt ärisõnastikuga termineid, mis osalt pärinevadki ärisõnastikust, kuid ärisõnastikus on ka laiemaid mõisteid tähistavad terminid. Ärisõnastikust või valdkonna märksõnastikust pärinevad ka andmestiku kirjeldamisel kasutatavad märksõnad.



Joonis 22. Sõnastikevahelise infovahetuse mudel andmekirjelduses

Andmete tervikkirjeldus toetab andmehaldust nii sisuliselt kui ka tehniliselt .

3 Andmekirjelduse koostamise korraldamine

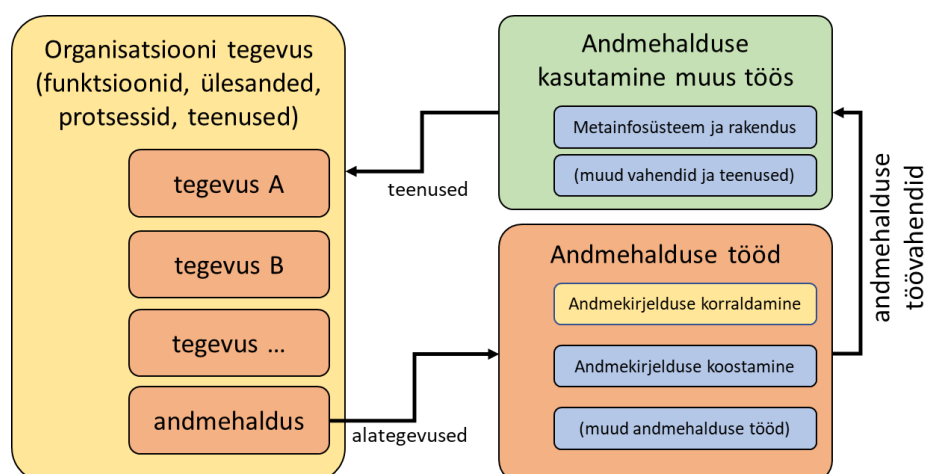
3.1 Üldist

Andmekirjelduse koostamine ei tohi olla ühekordne tegevus. Oluline on, et asutuses oleks andmekirjelduse haldamine juurutatud kui järjepidev protsess, mille tunnusteks on selgelt määratletud rollid ja ülesanded ning mille korraldust ja eesmärgipärasust pidevalt mõõdetakse ja hinnatakse.

Andmekirjeldused ei koosta ennast ise, vaid seda tuleb korraldada!

Töökorralduslikus vaates on andmekirjelduste koostamine andmehalduse osa ning seotud ka organisatsiooni teiste tegevustega, infohalduse mõttes nii põhi- kui ka tugifunktsioonidega. Kui aga vaadata andmehaldust kui teenust, on andmekirjelduste koostamine ja haldamine osaks sellest teenusest või iseseisev teenus.

Andmehalduse juhiste esimeses versioonis (2020) on eraldi andmehalduse raamistiku dokument, mis esitab nii riigi kui organisatsiooni tasandil rollidest ja protsessidest koosneva mudeli ning annab selle rakendamise põhimõtted. Selles on pööratud süstemaatiliselt tähelepanu andmehalduse korraldamisele. Viimase paari aasta jooksul on andmehalduse sisulised tegevused muutunud märksa konkreetsemaks ning just andmekirjelduste koostamine on olnud esmajärjekorras korraldatav töö. Konkretiseerunud on andmekirjelduse töövahendi soovitud funktsionaalsused ning võimalikud teenused – nii organisatsiooni sees kui ka väljapoole suunatud teenused, nagu andmete avaldamine avaandmetena. Joonisel 23 on esitatud andmehalduse kui organisatsiooni ühe tegevuse seos muude tegevustega, andmehalduse alategevused, sh andmekirjelduste koostamine ja haldus, ning andmehalduse seos muude tegevustega, kasutades selleks andmehalduse töövahendeid (infosüsteeme, rakendusi) ja pakutavaid teenuseid.



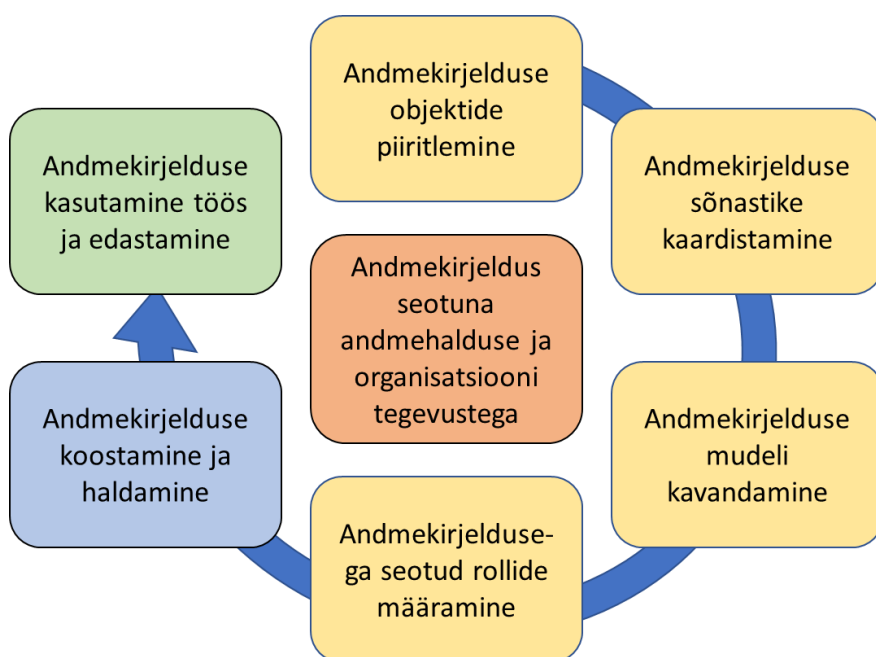
Joonis 23. Andmehalduse ja andmekirjelduste seos organisatsiooni muude tegevusega

Andmekirjeldused ei koosta ennast ise, vaid seda tuleb korraldada! Andmekirjelduse haldamise korraldus on organisatsioonis hästi tööle saadud siis, kui:

- organisatsioon tunnetab ja tunnistab andmekirjelduse süsteemse haldamise tulemusena paranenud andmete kvaliteeti;
- andmekirjeldus on kooskõlas valdkonna ja vastavate rahvusvaheliste standarditega ja toetab tõhusat andmevahetust;
- andmekirjeldus on hallatud ning andmekirjeldused ja andmestikud andmekirjelduse töövahendi kaudu ülesleitavad ning neist on võimalik tuvastada andmete tähendus;
- andmekirjelduse kvaliteeti jälgitakse ja parendatakse pidevalt;
- andmekirjelduse muutmine ja täiendamine järgib kokkulepitud protsessi, mis on tervikliku andmehalduse raamistiku osa.

Selles peatükis vaatamegi lähemalt, millega peavad organisatsioonid arvestama andmekirjelduse koostamise ja hilisema haldamise korraldamisel. Täpsemalt vaatame kahte tegevuste tsükli: 1) andmekirjelduse haldamise üldised tegevused ning 2) andmekirjelduse koostamise ja pideva täiendamise praktilised toimingud. Esimene neist tagab organisatsiooni valmiduse andmekirjelduse kestvaks parendamiseks ja taaskasutamiseks, teine keskendub kvaliteetse andmekirjelduse koostamisele. Lisaks peavad andmekirjeldused olema kasutatavad organisatsiooni igapäevatöös ja toetama seda ning neid peab saama edastada infosüsteemide vahel.

Andmekirjelduse koostamise ja halduse tegevused on üldistatult esitatud joonisel 24. Sellel on andmekirjeldustega tehtav töö seotud andmehaldusega ja üldisemalt organisatsiooni tegevusega. Joonisel on täpsemalt näidatud korralduslikud tegevused, samal ajal kui andmekirjelduste koostamise ja haldamise igapäevane töö on esitatud üldiselt.



Joonis 24. Andmekirjelduse haldamise tegevused

1. Kirjeldatavate andmestike ja teiste varade (objektide) väljaselgitamine ja piiritlemine.
2. Organisatsiooni jaoks sobivate varem koostatud ja kasutusel olevate sõnastike kaardistamine. Nende hulgas on ka üldised märksõnastikud.
3. Organisatsiooni jaoks sobiva andmekirjelduse mudeli ja haldamise raamistiku määratlemine.

4. Rollide määratlemine andmekirjelduse koostamisel ja haldamisel.
5. Andmekirjelduse koostamine, mille hulgas on nii esmane koostamine kui ka hilisem täiendamine ja muutmine ning mis hõlmab endas andmestike (objektide) üldist kirjeldamist, andmeelementide jms detailset kirjeldamist, sh sõnastike koostamist. Seda tehakse võimaluse korral andmekirjelduse töövahendiga.
6. Andmekirjelduse kasutamine töös ja selle toetamine ehk nende töötajate juhendamine, kes seda vajavad ning andmekirjelduse (metaandmete) edastamine ja avaldamine. Viimased tegevused on sageli osaliselt või täielikult automatiseeritud.

3.2 Andmekirjelduse objektide piiritlemine

Andmekirjelduse koostamise esimeseks eelduseks on asutuse andmetest ülevaate omamine. Nagu on näidatud peatükis 2.2 „Andmed ja andmestikud“, on andmeobjektidena võimalik määratleda ning seega ka halduse alla võtta erinevaid andmehulki Järgnevalt on selles juhise versioonis järgitud algset arusaama, et andmestike piiritlemine käib koos andmete olulisuse ehk väärtuse määratlemisega. Tuleb aga nentida, et praegu puudub süstemaatiline käsitlus andmete väärtusest. Viimastel aastatel on traditsioonilise info- ja arhiivihalduse väärtuse käsitluse kõrvale tulnud veel väärtuslike andmestike kontseptsioon. Esimest neist reguleerib arhiivseadus.³⁹ Väärtuslike andmestike käsitlus tuleneb Euroopa Liidu direktiivist avaandmete ja avaliku sektori valduses oleva teabe taaskasutamise kohta⁴⁰ ning on esitatud avaliku teabe seaduses.⁴¹ Selle kohaselt on kehtestatud väärtuslike andmestike valdkonnad.⁴² Nagu näha, ei piiritle see õigusakt mitte objekte, vaid valdkondi, kuid objektide piiritlemist käsitleb avaliku teabe seadus. Objektiks on andmekogu.

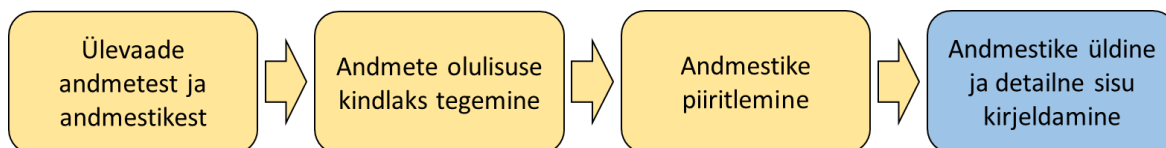
Asutuste töös ongi seni kõige olulisemaks andmekogude kirjeldamine. Kuid kuna andmekogu ise on õiguslik konstruktsioon, siis praktikas on vaja langetada piiritlevaid otsuseid. Näiteks on igati õige piiritleda kirjeldatava objektina terve andmekogu või infosüsteem, kuid samavõrd õige on ka piiritleda objektina mõni selle osa. Statistikaametis näiteks andmekogusid ei ole, kuid on palju ja väärtuslikke andmestikke, mis on osalt arhiiviväärtuslikud. Seega andmekogu kui objekti pole, kuid metaandmete süsteemi ja andmearhitektuuri skeemide abil saadud ülevaade andmetest võimaldab, saada ülevaade ja luua piiritletud andmestike käsitlus, mida siis hallata. Halduse jaoks saab kasutada ka andmekataloogi. Üldistatult on andmekirjelduse objektide piiritlemisega seotud tegevused ning andmestike sisu kirjeldamine esitatud joonisel 25.

³⁹ Arhiiviseadus § 2 Dokument ja arhivaal. Lõige 2) Arhivaal on dokument, millele avalik arhiiv on hindamise tulemusena andnud arhiiviväärtuse

⁴⁰ Euroopa Parlamendi ja nõukogu direktiivi (EL) 2019/1024 avaandmete ja avaliku sektori valduses oleva teabe taaskasutamise kohta ([ELT L 172, 26.06.2019](#), lk 56–83)

⁴¹ Avaliku teabe seadus § 4¹ Väärtuslikud andmestikud

⁴² Väärtuslike andmestike temaatiliste kategooriate nimekiri: 1. Georuumilised andmed, 2. Maa seire ja keskkond, 3. Meteoroloogiateave, 4. Statistika, 5. Äriühingud ja äriühingu omandisuhted, 6. Liikuvus ([ELT L 172, 26.06.2019](#), lk 56–83)



Joonis 25. Andmekirjelduse objektide piiritlemise tegevused

Andmekirjelduse täpsemaks piiritlemiseks kaardistatakse valitud andmekogus sisalduvad komponendid (andmebaasid, tabelid, andmeelemendid).

Organisatsioonis on erineva olulisuse, taaskasutus- ja säilitusväärtusega andmeid. Ranget terminoloogiat kasutamata võib öelda, et need on eri kujul ja infosüsteemides. Osa on teabehalduse-, osa dokumendihalduse ja osa eri infosüsteemides, mille alla tuleb siin mõiste relatsiooniliste andmebaasis peale loodud rakendusi. Andmete tähendust seletavas peatükis 2.2.1 näitasime, et andmed ei ole üheselt määratletavad. Ühelt poolt hallatakse põhitegevuse jaoks vajalikke andmeid, mis on aluseks organisatsiooni toimimisele ning mille korralik haldamine tagab läbipaistvuse ja usaldusväärsuse. Teisalt leidub tihti ka veebilehti, sisemisi mitteametlikke dokumente või ürituste fotosid, mis on olulised info jagamise või organisatsioonikultuuri mõttes, kuid mitte andmete taaskasutus- ja säilitusväärtuse laiemas vaates.

Andmekirjelduse objektide piiritlemisel tuleb arvestada sellega, et teave, info ja andmed on läbi põimunud, ning lähtuda praktilistest eesmärkidest, millest tähtsaimad on info / andmete leitavus ja taaskasutatavus.

Kõikide andmete haldamine, sealhulgas põhjalik kirjeldamine, ei ole organisatsiooni ressursi silmas pidades alati mõistlik. Samuti on terve andmemassiivi haldamine üldjuhul üle jõu käiv ülesanne. Seetõttu on andmekirjeldamiseks valmistumise esimeses etapis oluline saada ülevaade sisuliselt kokku kuuluvatest andmetest ehk teha selgeks, mis on andmeobjektid ja andmestikud. Seejärel määrata, milliste andmete/ andmekogumite/andmehulkade/andmete kogumite kirjeldamine on organisatsiooni jaoks kõige olulisem.

Konkreetsete andmete olulisus sõltub suurel määral iga neid haldava organisatsiooni tegevusvaldkonnast ja ülesannetest. Seega on olulisuse määramiseks raske välja tuua täpseid ja universaalseid reegleid. Küll on olemas mõningad üldised põhimõtted, millest saab juhinduda.

- Andmete taaskasutus. Andmed on kasutamiseks ja taaskasutamiseks ning mida rohkem on andmete taaskasutusest huvitatud sisemisi või välimisi kliente, seda mõistlikum on neid andmeid põhjaliku andmekirjeldusega varustada.
- Põhitegevused vs. tugitegevused. Üldjuhul on organisatsiooni põhitegevuse käigus tekkinud andmetel oluliselt suurem väärtus kui tugitegevuste käigus tekkinud andmetel.
- Info hinnatav väärtus. See väljendub säilitustähtajas ning on omistatud tõestus- või infoväärtust arvesse võttes aastates või arhiiviväärtusena alatiselt. Tõestus- ja infoväärtus määratakse taaskasutusvajaduse põhjal ning seda tehakse organisatsiooni tegevuste ja nende tegevuste fikseerimise ehk dokumenteerimise analüüsi tulemusena.

Olulisuse järgi grupeeritud andmed tuleb omakorda defineerida selgepiiriliste andmestikena. Kirjeldamiseks defineeritakse andmestik, mis on piisavalt suur, et selles sisalduv info oleks täielik ja terviklik, samas piisavalt väike, et andmekirjelduste haldamine oleks organisatsioonile kasulik ja andmehaldurile mõistliku aja jooksul jõukohane. Alati pole andmestiku piiritlemisel ühte ja ainsat õiget lahendust. Näiteks ehtisregistrit on võimalik

defineerida kui ühte andmestikku (hitisregistri andmestik) või kui mitut erinevat andmestikku (ehitise põhiandmete andmestik, järelevalve andmestik, energiamärgise andmestik jne).

Kirjeldatava andmestiku valimisel on oluline andmehalduri koostöö sisuvaldkonna spetsialistidega, kelleks võivad olla andmeomanik, infosüsteemi/rakenduse peakasutaja või töötajad, kes igapäevaselt seda tööks kasutavad. Andmestiku komponentide (andmebaasid, tabelid, andmeelemendid) analüüsimisel tuleb konsulteerida ka tehniliste halduritega, näiteks andmearhitekti, süsteemi või rakenduse administraatoriga.

Andmestiku piiritlemisel tuleks tähele panna järgmist.

- Kui andmestik on asutatud ametliku andmekoguna, on soovitav defineerida kogu andmekogu ühe kirjeldatava andmestikuna. Kui andmete haldamiseks või taaskasutuseks on vajalik ka väiksemate andmestike määratlemine, saab need hiljem kirjeldada andmekogu alamandmestikena. Näiteks on paljud andmekogud (registrid) põhimäärustes välja toonud alamregistrid, mida on sobilik ka kirjeldada ja hallata alamandmestikena.
- Andmestik peab olema defineeritud piisavalt suurena, et selles sisalduvad andmed oleksid terviklikud ja andmestikul kui tervikul oleks selge kasutuspotentsiaal. Üldjuhul tähendab see, et asutuse ühe tegevuse, näiteks lubade andmise, järelevalve, toetuse määramise käigus saadud või tekkivad andmed peaksid sisalduma ühes kirjeldatavas andmestikus, mitte olema pihustatud erinevate andmestike vahel. Näiteks ei ole mõistlik ehitisregistri andmestiku piiritlemine nii, et selles ei sisaldu hoonele kasutusloa andmise kuupäev.
- Andmestik peab olema piiritletud piisavalt väiksena, et andmete ja andmekirjelduste haldamine oleks andmehaldurile jõukohane. Üldjuhul võiks ühes andmestikus sisalduda omavahel tihedalt seotud tegevuste või teenuste käigus tekkivad andmed. Üks andmestik – kui see ei ole selgepiiriline andmekogu – võiks katta ühe või paar organisatsiooni protsessi või teenust. Näiteks on mõistlik defineerida andmestikuna „tolliandmed” või „välismaalastele elamislubade andmine”, aga mitte „riigieelarve andmed” või „transpordiandmed”.

Peale (oluliste) andmestike kontseptuaalset määratlemist ja empiirilist kaardistamist on vajalik detailsemalt analüüsida andmestiku tehnilisi komponente. Nendeks on andmebaasid, andmebaaside osad ehk skeemid, andmetabelid, aga ka muud eraldi failidena tabelid ja andmestikena hallatavad failid. Kõik need sisaldavad veel detailsemal tasandil andmeelemente. Kaardistuse eesmärk on eristada olulisemad ja vähemolulised komponendid, mis võimaldab edaspidi keskenduda ainult oluliste ja sisulist väärtust omavate andmete kirjeldamisele.

Näiteks on enamikus infosüsteemides talletatud küllalt palju tehnilist tugiinfot, mis on vajalik andmete turvalisuse ja terviklikkuse tagamiseks, kuid ei ole oluline andmehalduse jaoks andmekirjelduse vaates. Samuti võib andmestikus eksisteerida mitu paralleelset esitust, nagu algne andmebaas ja selle põhjal loodud andmeait või andmeladu. Andmete kirjeldamise ettevalmistamisel on tähtis aru saada erinevate tehniliste platvormide vahelistest seostest, analüüsida andmete kattuvust ja taaskasutuse töövooge ning seeläbi tagada, et andmekirjeldusega on kaetud nii olulised andmed kui ka olulised platvormid/infosüsteemid. Näiteks hoone põhiandmete kirjeldus on seotud nii ehitisregistri andmebaasi kui ka andmelao sobivate elementidega.

Andmestiku komponentide kaardistamisel tuleks tähele panna järgmist:

- komponentide kaardistust on otstarbekas teha infosüsteemi tehnilise dokumentatsiooni, eelkõige andmemudeli või arhitektuuridokumentide põhjal;
- andmekirjelduse mõttes vähemtähtsateks andmeteks võib üldjuhul lugeda süsteemi logi, andmebaasi kirjade loomise ja muutmise kuupäevi, failide räsiseid ja muud tehnilist tugiinfot, mis ei ole otseselt seotud organisatsiooni põhiülesannete täitmise käigus andmete töötlemise, andmepõhise juhtimise või andmevahetuse eesmärkidega.

3.3 Andmekirjelduse sõnastike kaardistamine

Andmekirjelduse koostamise teiseks sammuks on kirjeldamisel kasutatavas sõnavaras kokku leppimine ehk valdkonna- ja ärisõnastike kaardistamine ning juba olemasolevate andmekirjelduste tuvastamine. Juhise sellesse versiooni on lisatud eraldi peatükk 2.3, mis käsitleb mõisteid, termineid ja sõnastikke.

Andmekirjelduse üks peamisi eesmärke on organisatsiooni andmetest ülevaate omamine – millised andmed on olemas, kes ja miks neid kogub ning kus neid hoitakse. Sellist ülevaadet suudab pakkuda üksnes kirjeldus, milles on läbivalt kasutatud ühtset sõnavara (mõisted on tähistatud terminoloogiliselt ühtmoodi). Ühtne sõnavara kirjeldamisel ja andmestikes aitab organisatsioonil ära kasutada andmehalduse kogu potentsiaali: tuvastada ühelaadseid andmeid ja dubleerivat kogumist, kontrollida ja tagada andmete kvaliteeti jmt. Ühtse sõnavara puudumisel on andmeotsing sarnane Google'i otsinguga, kus vajaliku info leidmiseks tuleb tihti teostada mitu päringut erinevate märksõnadega ja erinevates keeltes ning ühe päringu vastuses on omavahel mitteseotud tulemeid.

Olemasolevate sõnastike ja kirjelduste kaardistamine valmistab ette teadmusbaasi, mille abil saab edasist kirjeldamist kiiremini ja efektiivsemalt teha, vältides vasturääkivusi kirjeldustes ning hoides läbivalt ühtset sõnavara.

Sõnastike kaardistamisel on esimene ülesanne vaadata laiemalt, millised andmete sisuga seotud sõnavara allikad on olemas. Leitud allikad võivad olla küllaltki erineva keerukuse ja ülesehitusega, alustades õigusaktides sisalduvatest definitsioonidest ja lihtsatest märksõnastikest keerukamate taksonoomiate ja ontoloogiateni. Soovitame esialgse kaardistamise käigus mitte liialt mõelda sõnastiku tehnilise ülesehituse peale, vaid pigem lähtuda sisulisest sobivusest. Olemasolevate rahvusvaheliste sõnastike ja ontoloogiate leidmiseks on nii mitu üldist kataloogi kui ka avaliku sektori spetsiifikat kajastavaid sõnastikke ja ontoloogiaid. Parimal juhul on organisatsioonil juba olemas keskne ärisõnastik, mis on asjakohane ka andmestiku kirjeldamiseks. Samuti on võimalik, et organisatsiooni mõne muu andmestiku kirjeldamisel on sobivad sõnastikud juba kaardistatud ja kasutusele võetud. Andmestike omanikud ja valdkonnaekspertid võivad muude tegevuste käigus juba olla dokumenteerinud sobivaid sõnastikke. Lisaks organisatsioonisiseste sõnastike kaardistamisele võib abi küsida ka teistelt sama valdkonna organisatsioonidelt, kuna mitu valdkonda kasutab riiklikke või rahvusvahelisi valdkonnasõnastikke (nt tervishoid, pangandus, looduskaitse).

Aeg, mis kulub eeltööle sõnastike ja sõnavaraga, hoitakse kokku hilisemal andmekirjelduste koostamisel.

Sõnastike sobivuse hindamisel saab organisatsioon lähtuda alltoodud kriteeriumitest.

- Ulatus ja kaetus. Need näitavad, kui suur osa vajalikest terminitest sisaldub sõnastikus. Andmekirjelduse koostamine ja edasine haldamine on seda lihtsam, mida vähem on kasutusel erinevaid, halvimal juhul omavahel vasturääkivaid sõnastikke. Seega on mõistlik kasutada sõnastikke, mis võimalikult suures osas vastavad asutuse vajadustele.
- Täpsus. Näitab, kui üldised või detailsed on sõnastiku terminid. Kuigi leitud sõnastik võib katta kogu andmestiku sisu, võivad terminid olla organisatsiooni andmete kirjeldamiseks kas liialt üldised või liiga täpsed. Näiteks võib sõnastikus olla välja toodud ainult mõiste „isik“, andmestikus on aga oluline eristada mõisteid „juriidiline isik“ ja „füüsiline isik“ või vastupidi.
- Vastavus. Näitab, kui palju sarnaneb terminite kasutus sõnastikus õigusaktide ja/või organisatsiooni ja valdkonna ekspertide sõnavaraga. Soovitav on kasutada sõnastikke, mis muude kriteeriumite võrdsuse korral pakuvad paremat vastavust. Näiteks on ehitusseaduses defineeritud olulised terminid „hoone“ ja „rajatis“ kui üksteist välistavad: „rajatis on ehitus, mis ei ole hoone“. Seega pole mõistlik ehitisregistri andmete kirjeldamisel kasutada hüpoteetilist sõnastikku, mis defineerib rajatise ja hoone kui samatähenduslikud mõisted või milles sellised mõisted üldse puuduvad.
- Kasutatavus ja haldamine. Need näitavad, kui suur on sõnastiku aktiivsete kasutajate hulk ja kui aktiivselt sõnastikku ennast hallatakse. Üldiselt soovitatakse eelistada laiemalt kasutatavaid sõnastikke, mille puhul on selgelt paika pandud ka sõnastiku haldamise ja edasiarendamise põhimõtted.

Praktikas leidub harva ühte kõikehõlmavat sõnastikku, mis kataks ära terve valdkonna eri tahud. Küll aga sisaldavad olemasolevad sõnastikud mitmeid mõisteid, mis on andmekirjelduse jaoks piisavalt detailsetl kirjeldatud. Seetõttu on loomulik kombineerida erinevaid sõnastikke vastavalt vajadusele. Siiski tuleb lähtuda sellest, et sõnastiku koostamine, kui see haarab endasse mitmed erinevad sõnastike tüübid, nagu seletav sõnaraamatu, mitmekeelse sõnastiku, tesauruse jms ei muutuks tööks iseeneses. Isegi valdkonna sõnavara korrastamine ehk terminitöö tuleb andmekirjelduste koostamisest eraldi hoida.

Väljaspool organisatsiooni hallatava sõnastiku kasutamisel peaks organisatsioon eelkõige huvi tundma, kes on sõnastiku omanik ning kas ja millised protseduurid on paigas sõnastiku muutmiseks või täiendamiseks. Soovitav on eelistada selliseid sõnastikke, mille omanik ja haldamise protseduurid on selgelt ja läbipaistvalt määratletud. Juhul, kui organisatsioonil tekib vajadus lisada uusi termineid või täiendada olemasolevaid, peab olema võimalik võtta ühendust sõnastiku omanikuga ja algatada sõnastiku täiendamine või muutmine. Põhimõtteliselt on asutusel võimalik väline valdkondlik sõnastik üle võtta ja jätkata ise selle haldamist, et tagada kontroll sõnastiku edasise täiendamise ja muutmise üle. Kuna pikaajaline märksõnastiku haldamine on ajamahukas, on soovitatav paralleelsete sõnastike tekkimise vältimiseks teha seda valdkonna organisatsioonide koostöös.

Viimase sammuna on soovitatav kaardistada sõnastike tehniline ülesehitus: millises vormingus on sõnastik loodud ning kas selle poole on võimalik automaatselt pöörduda ja liidest kaudu termineid alla laadida. Eelistada tuleks masinloetavat esitust (XML, OWL, erinevad RDF-formaadid) omavaid sõnastikke, kuivõrd neid on lihtsam taaskasutada ja teiste süsteemidega integreerida.

3.4 Andmekirjelduse mudeli ja arhitektuuri kavandamine

Andmekirjelduse järgmiseks tegevuseks on kirjelduse kasutusjuhtude analüüs, andmekirjelduse funktsionaalse mudeli koostamine ja mudeliga sobiva tehnilise arhitektuuri kavandamine.

Andmekirjelduse funktsionaalse mudeli eesmärk on kaardistada andmekirjelduse kasutusjuhud ja kasutajad nii organisatsiooni sees kui ka väljas. Selle juures on oluline tähele panna, et paljudel juhtudel ei räägita mitte andmekirjelduste kasutamisest, vaid metaandmete kasutamisest, aga mõeldakse just kirjelduse metaandmete kasutamist. Sealjuures tuleks kasutusjuhte vaadelda laiemalt, nii et oleksid kaasatud kõik olulised osapooled ja nende vajadused, sealhulgas:

- otsene andmekirjelduse koostamine ja sellega kokku puutuvad rollid (andmehaldur, andmeomanik);
- andmebaasiplatvormis sisalduva info sünkroniseerimine andmekirjeldustega (andmearhitekt, andmebaasi administraator);
- andmekirjelduse taaskasutus andmete esmases analüüsis, äriprotsesside (BI ehk äriintelligentsi) modelleerimises ja automatiseerimises ning teiseses andmete analüüsis üldistusteks, juhtimisotsusteks ja statistikaks (andmeomanik, andmeanalüütik);
- andmekirjelduse kasutamine andmekvaliteedi reeglite loomisel (andmehalduse juht, metaandmete analüütik);
- regulaarsete raportite koostamine juhtimisotsusteks ja statistikaks (andmeanalüütik, andmehaldur);
- andmekirjelduste edastamine lõppkasutajatele läbi veebirakenduste ja -teenuste (lõppkasutajad, teenuste ja rakenduste arendajad);
- andmete ja andmekirjelduste edastamine välistele süsteemidele ja teistele organisatsioonidele (teiste organisatsioonide arendajad, andmehaldurid).

Lihtsamate kasutusjuhtude puhul, näiteks paarist MS Exceli tabelist koosnev külastajate nimekiri, on peamiseks vajaduseks statistika edastamine juhtidele. Keerulisemal juhul võib aga organisatsiooni andmekirjeldus katta mitmeid kompleksse struktuuriga riiklike andmekogusid üle eri valdkondade ja sõnastike. Sel juhul on ka kasutusjuhtude ja kasutajate tuvastamine keerukas ja aeganõudev. Eelmises lõigus toodud rollid on tinglikud ja lähtuvad osalt andmehalduse raamistiku esimeses versioonis toodud võimalikest rollidest.

Andmekirjelduse tehnilise arhitektuuri ülesandeks on kirjeldada tark- ja riistvaraline keskkond, mis suudab mudelis väljatoodud vajadusi piisavalt katta. Kordame üle: need mudelid on valdkondlikult ning orienteerituselt erinevad. Üldise üleriikliku andmekirjelduse arhitektuur annab ühest küljest ette mudelid ja nende semantika, millega peab arvestama. Teisest küljest määravad just kasutajate ja valdkondade vajadused ära mudelid nii asutuse sees kui ka valdkondlikult ning üleriigilised arendused peavad sellega arvestama.⁴³

Ülaltoodud lihtsamal juhul on andmekirjelduse haldamine võimalik tavapäraste tabelarvutuse vahenditega, näiteks Excel, OpenOffice ja LibreOffice Calc või Google Sheets. Nendes luuakse tabelid andmekirjelduse standardis ja/või mudelis määratud iga olemi tüübi jaoks. Näites on eraldi andmestiku ja andmeelementide kirjeldus ning andmesõnastik. Keerukamatel juhtudel on andmekirjelduse tabelite koostamine ja nendevaheliste seoste

⁴³ Sellise üldise ja valdkondliku paralleelse lähenemise näiteks on standardi DCAT areng ja selle valdkondlikud rakendusprofiilid ruumiandmetele (GeoDCAT) ja statistika andmetele (StatDCAT).

haldamine käsitsi ning kirjelduse kasutajatele edastamine liialt aeganõudev, mistõttu on vaja juurutada spetsiaalne andmekirjelduse töövahend.⁴⁴

Sobivaima andmekirjelduse töövahendi leidmiseks tuleks organisatsioonil vastavalt funktsionaalses mudelis toodud vajadustele analüüsida:

- milliste andmeallikatega, näiteks andmebaasimootoriga on andmekirjelduse töövahend võimeline automaatselt suhtlema;
- millises vormingu sõnastikke ja kirjelduse osi ning kui lihtsalt on võimalik andmekirjelduse töövahendisse laadida;
- kas andmekirjelduste koostamine on andmehaldurile või teistele kirjelduse loojatele mugav;
- kas andmekirjelduse töövahend võimaldab andmekirjelduse standardi juurutamist;
- kas ja millistes vormingutes, sh avatud vormingud, on andmekirjeldusi võimalik eksportida ning vajadusel välistele kasutajatele edastada (väliste andmekirjelduste kasutajatena on peetud silmas eelkõige teisi asutusi, kes seda oma töös vajavad, nagu Statistikaamet ja Rahvusarhiiv, ning infosüsteeme, nagu avaandmete portaal ja RIHA);
- millises mahu on võimalik andmekataloogi tarkvara ja andmekirjelduse tegevusi integreerida asutuse muude andmehaldusetegevustega, näiteks andmekvaliteedi haldusega.

Selle etapi tulemusena on:

- hinnatud andmekirjelduse täpset mahtu ja vajadust,
- analüüsitud sobilikke andmekirjelduse töövahendeid ja kirjelduse vorminguid,
- valitud välja sobivaim lahendus ning loodud vajalik kirjelduse mudel ja struktuur.

3.5 Andmekirjeldusega seotud rollide määramine

Kui kirjeldatav andmestik ja olemasolevad sõnastikud on välja valitud, tuleb paika panna rollide täpne jaotus andmekirjelduse loomiseks, haldamiseks ja kasutamiseks pikemas vaates. Oluline on seejuures ka see, kas rollid on määratavad organisatsiooni tasandil või andmestiku tasandil.

Praktikas on ärisõnastikul ja andmesõnastikul organisatsioonis sageli erinevad omanikud või vastutajad. Vajalik on nende omavaheline tihe koostöö sõnastike täiendamisel. Üldjuhul on neil erineva ulatusega juurdepääs andmetele ja muutmisõigused sõnastikele, milles nad ei ole andmeomaniku rollis. Praktikas sõltub õiguste haldamine oluliselt sõnastike haldamiseks kasutatavast andmekirjelduse töövahendist. Andmehalduse raamistiku esimene versioon kirjeldab põhjalikult andmehaldusega seotud rolle ja nende ülesandeid. Andmekirjeldusega seotud rollide jaotus on järgmine:

andmehalduse juht (juhtkonna liige) – korraldab ärisõnastike ja mõistete mudeli loomist juhtimise aspektist,

⁴⁴ Selliseks vahendiks on RIHAKE. Vt [RIHAKEse lõppkasutaja kasutusjuhend](#). Ver 1.1.1. (17.01.2023)

andmeomanik (sisuteenuse juht või infosüsteemi peakasutaja) – loob ja kirjeldab teenuste kasutatavad ärimõisted (ärisõnastik) ning nende seosed ärireeglitega organisatsiooni teistes infosüsteemides,

andmehaldur (andmestike eest vastutaja) – kirjeldab andmestikud ja andmeelemendid andmekirjelduse töövahendis vastavalt andmekirjelduse standardile.

Need rollid võivad organisatsioonide praktikas olla koondatud ühte ametikohta või olla jaotatud asutuse ja selle IT-teenuse pakkuja vahel. Mistahes rollijaotuse puhul on oluline eristada vastutust ärisõnastiku ja andmesõnastiku eest ning kirjelduste koostamise eest ning määrata järelevalvet teostav rollandmekirjelduse kvaliteedi mõõdikute täitmiseks ning perioodiliseks uuendamiseks.

Rollide ja vastutuste jaotus on otstarbekas kirjalikult fikseerida organisatsiooni andmehaldust korraldavas dokumendis. Need rollid tuleb juurutada andmekirjeldusi ja sõnastikke haldavates tarkvararakendustes juurdepääsu- ja kasutusõigustena.

3.6 Andmekirjelduse loomine ja täiendamine andmekirjelduse töövahendis

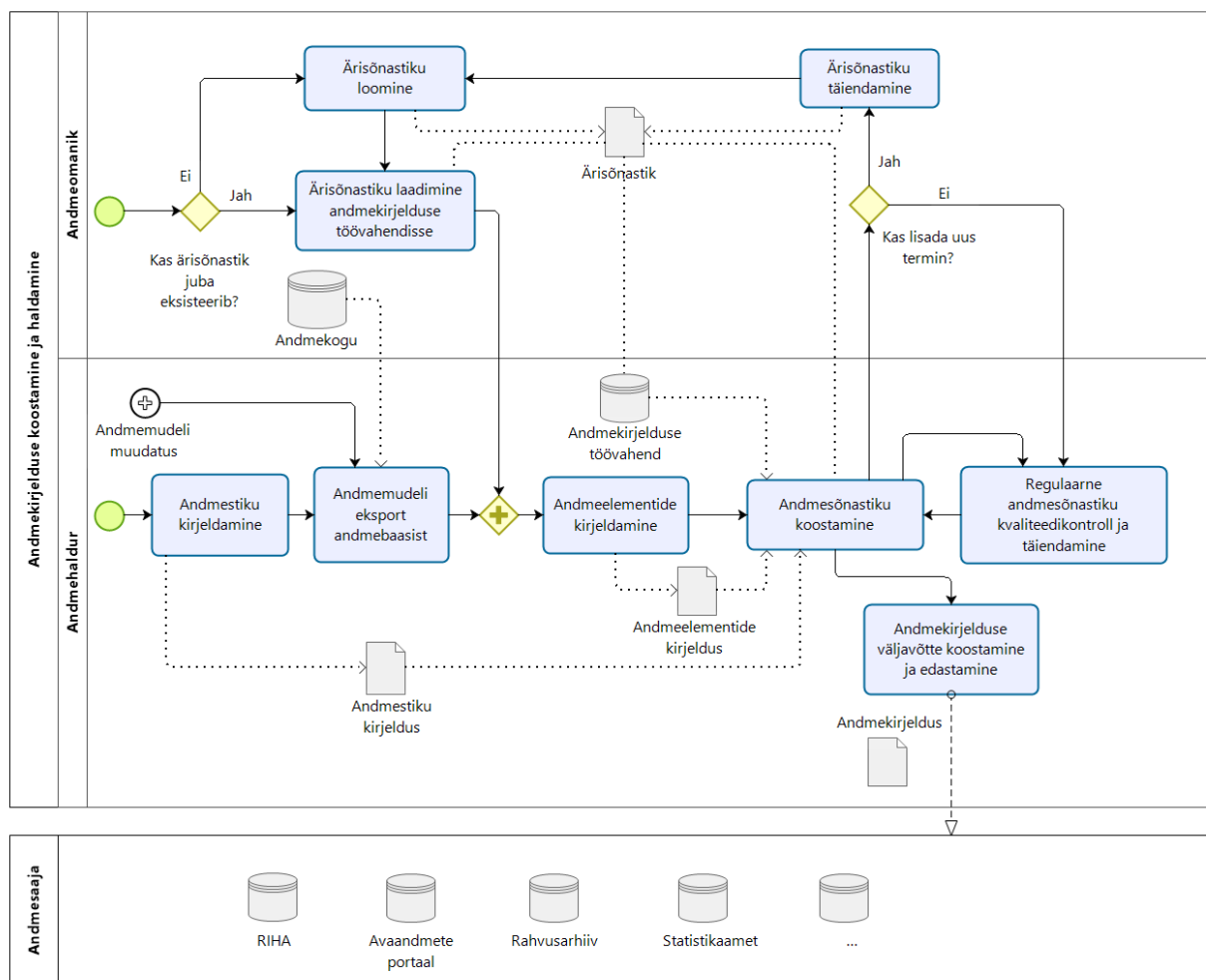
3.6.1 Üldist

Käesolev juhise osa järgib suuresti juhise esimeses versioonis esitatut. Praegu juurutab mitu organisatsiooni RIHAKE-st või muud andmekirjelduse töövahendit. Selles peatükis esitatud lähenemine tervikkirjelduse osadele, eriti protsessile, on käsitletav eelkõige mudelina, mille iga organisatsioon juurutab lähtuvalt konkreetsest töövahendist, oma töökorraldusest, vajadusest ning rakendamise ulatusest.

Rollide jaotus protsessides võimaldab samaaegseid tegevusi. Näiteks saab andmehaldur töötada iseseisvalt andmeelementide kirjeldusega, lisades üksikutele elementidele pikemaid selgitusi. Valdkonnaekspert (andmeomanik) saab samaaegselt tegeleda andmestikule kohaste valdkonna märksõnastike valiku ja/või ärisõnastike loomisega ning täiendada andmestiku kirjeldust. Andmehalduril on võimalik andmeelementide kirjeldustele lisada viiteid sõnastike terminitele ja teha ettepanekuid ärisõnastiku täiendamiseks uute andmete tähendust selgitavate terminitega.

Joonisel 26 on andmekirjelduse koostamise ja täiendamise protsess esitatud töövoona, milles eri rollides olevad isikud täidavad kindlaid ülesandeid. Protsesse kavandades on oluline vaadata kogu tegevust kahe etapina.

1. kui kirjeldusi ja sõnastikke ei ole, tuleb ette võtta nii süsteemi loomine kui ka esialge suurem kirjeldamise töö;
2. ja esialgset tööd on vaja parendada ning sisse viia toimunud muudatused.



Joonis 26. Andmekirjelduse koostamine ja täiendamine (protsess)

Andmekirjelduse haldamine, parendamine ja erinevate komponentide vahel seoste loomine on seega pidev töö, mida tehakse erinevate rollide koostöös.

3.6.2 Ärisõnastiku loomine ja andmekirjelduse töövahendiga sidumine

Eeldus: ärisõnastiku loojal on hea ülevaade valdkonna terminoloogiast, sh teistest seonduvatest ärisõnastikest, valdkonna sõnastikest ning õigusaktides ja infosüsteemides kasutatavast sõnavarast.

Organisatsiooni andmehalduse tegevusi toetava andmesõnastiku aluseks on läbivalt ühtlase sõnavara kasutamine, ühtlase sõnavara aluseks omakorda on kvaliteetne ärisõnastik. Reeglina peaks ärisõnastik olema loodud juba enne andmekirjelduse koostamist. Samas pole ärisõnastiku puudumine andmete kirjeldamise alustamisel takistuseks, see on võimalik luua andmekirjelduse koostamise käigus.

Ärisõnastiku koostamisel on soovitatav meeles pidada mõningaid lihtsaid põhimõtteid.

- Sõnastik peab olema piisav. Sõnastik peab sisaldama piisavalt termineid, mis katavad kõik asutuse olulised ärimõisted. Kõige olulisem on nende mõistete täpne defineerimine, mille kohta käivaid andmeid on vaja ristikasutada asutusesiseses või organisatsioonidevahelises andmevahetuses.

- Sõnastik on muutuv ja täienev. Valdkonna sõnavara, õigusaktide definitsioonid ja ka hallatavate andmete skoop on pidevas muutumises. Asutus peab nende muutuste toetamiseks looma kindla reeglístiku, mis kehtestab põhimõtted sõnastikku terminite lisamise, muutmise ja kustutamise kohta. Reeglite täitja ehk sõnastiku omanik, kes kogub kokku ettepanekud sõnastiku täiendamiseks ja kannab terminid sõnastikku, võiks üldjuhul olla andmehalduse juht või andmeomanik.

Ärisõnastiku loomiseks on asutusel kolm peamist võimalust:

- võetakse kasutusele üleriiklik märksõnastik või rahvusvaheline valdkonna sõnastik, viimane võib olla nii märksõnastik kui ka tesaurus ja on sageli mitmekeelne;
- asutus loob ärisõnastiku riikliku või rahvusvahelise sõnastiku baasil;
- ärisõnastik luuakse asutuses oma tegevusvaldkonna õigusaktides ning andmestike ja infosüsteemide dokumentatsioonis kasutatud mõistete ja terminite baasil.

Need kolm võimalust ei ole üksteist välistavad. Näiteks võib asutus kasutada valdkonna märksõnastikku ja luua selle põhjal samaaegselt asutuse ärisõnastiku, võttes puuduvate terminite asemel kasutusele õigusaktides defineeritud mõisted. Mitme märksõnastiku samaaegsel kasutamisel tuleb jälgida, et erinevates sõnastikes poleks kattuvaid või vasturääkivaid termineid.

Ärisõnastiku loomisel peab asutus jälgima, et nii sõnastik ise kui ka selles sisalduvad terminid oleksid kirjeldatud vastavalt andmekirjelduse standardis esitatud nõuetele. Sõnastiku kirjelduse alusel peab olema võimalik aru saada, millist valdkonda ja/või asutust sõnastik katab, kes on sõnastiku omanik ning millal on seda viimati muudetud. Abistav info puudutab sõnastiku skoopi ehk seda, kuidas on valdkond defineeritud ning kuidas sõnastikku kasutada. Seda võib teha kitsamalt ainult andmehalduse toetamiseks või laiemalt, näiteks asutuse ärireeglite, protsesside jms kirjeldamiseks, linkandmete avaldamiseks, kommunikatsiooni parendamiseks jne. Iga mõistekirje peab sisaldama termini tähenduse kirjeldust, osutust kehtivusele (kas termin on aktuaalne või mitte) ning termini sõnastikku lisamise ja/või viimase muutmise kuupäev.

Ärisõnastik on oma ülesehituselt tesaurus ehk see esitab mõistetevahelisi seoseid. Tavalised seose liigid on järgmised: hierarhiline seos, mis loob taksonoomia⁴⁵ („hoone“ ja „rajatis“ on mõlemad „ehitised“), sünonüümia („hoone“, „maja“). Põhjalikumad sõnastikud võivad terminitega siduda ka organisatsiooni tööprotsesse, teenuseid ja neist tulenevaid ärireegleid ning seega esitada küllaltki keerulise mõistete ja seoste mudeli.

Ärisõnastiku tehnilise taaskasutuse toetamiseks on mõistlik igale terminile lisada masinloetav identifikaator (URI), mille abil on võimalik termin identifitseerida ja sellele viidata. Muuhulgas saab identifikaatorit hästi kasutada ka terminite vaheliste seoste defineerimisel, kui URI struktuur selliselt luua. URI-de automaatset loomist toetavad ka mitmed andmekirjelduse töövahendid. Samas ei anna URI-d, mis on loodud semantilist koostalitlusvõimet arvestama, ennast kergesti integreerida, kuigi neile saab ehitada API teenuseid.

Toome siin selle illustreerimiseks näite:

- Eesti üldises märksõnastikus on mõiste „perekond“ tähistatud kindla ja püsiva URI-ga: <https://ems.elnet.ee/id/EMS031142>

⁴⁵ Rahvusvahelised ja Eesti klassifikaatorid ja koodiloendid on üldjuhul taksonoomiad. Taksonoomia põhiomaduseks on see, et selles olevad mõisted on klassidena (liigiliselt) piisavalt erinevad.

- Rahvusvaheliselt kasutatavas statistiliste analüüsiüksuste märksõnastikus⁴⁶ on mõiste ja märksõna „perekond“ esitatud RDF-is:

```
<rdf:Description rdf:about="http://rdf-vocabulary.ddialliance.org/cv/AnalysisUnit/2.1.3/#Family">
  <rdf:type rdf:resource="http://www.w3.org/2004/02/skos/core#Concept"/>
  <skos:inScheme rdf:resource="http://rdf-vocabulary.ddialliance.org/cv/AnalysisUnit/2.1.3/">
  <skos:notation>Family</skos:notation>
  <skos:prefLabel xml:lang="en">Family</skos:prefLabel>
  <skos:definition xml:lang="en">Two or more people related by blood, marriage (including step-relations), co-habitation or registered partnership, adoption or fostering, and who may or may not live together. For example, data on what kind of public services the family uses or would require; or how often family members are in contact with each other.</skos:definition>
  <skos:narrower rdf:resource="http://rdf-vocabulary.ddialliance.org/cv/AnalysisUnit/2.1.3/#Family.HouseholdFamily"/>
</rdf:Description>
```

- RIHAKE, milles on GUID: <http://rihake/70006317/BV/12c6725b-3778-a5ab-7445-2635d99ea4e7>

Näidetest ilmneb, et ärisõnastiku mõisted ja nende tähistamiseks kasutatavad tähised, sh mitmekeelne terminoloogia, vajab eri infosüsteemides nii inimeste kui arvutite poolt kasutamiseks eraldi tähelepanu. Praegusel ajal on enim levinud ärisõnastiku loomine XML-põhiselt (näiteks OWL-, RDFS-, JSON-LD- või SKOS-skeemide alusel). Eestis on soovitatav kasutada OWL-skeemi, mida toetavad erinevad vabavaralised tööriistad.

Ärisõnastiku loomiseks võib organisatsioon lihtsamal juhul kasutada tabelarvutustarkvara. Eelkõige on see mõistlik juhtudel, kui ka ülejäänud andmekirjelduse komponendid on plaanis luua lihtsate seotud tabelitena. Ärisõnastik tuleks avaldada organisatsiooni sise- ja/või välisveebis sellises vormingus, et sõnastikku oleks andmekirjelduse koostamisel lihtne kasutada. Kui asutuses on kasutusel eraldi sõnastike haldamise või andmekirjelduse töövahend, tuleb ärisõnastik luua nii, et seda oleks võimalikult lihtne töövahendisse laadida.

Valminud ärisõnastik tuleb andmekirjelduse töövahendisse laadida või sellega liidestada. Vastavalt kasutatavale tarkvarale võib see toimuda käsitsi (näiteks OWL või tabelarvutusfaili laadimisena) või automaatse liidese vahendusel. Olenemata meetodist, on oluline, et kõik ärisõnastiku uuendused tehakse ka andmesõnastikule kättesaadavaks.

Selle etapi tulemusena on:

- välja valitud või koostatud ärisõnastik,
- ärisõnastik lisatud andmekirjelduse töövahendisse.

3.6.3 Andmekirjelduse koostamine

Eeldus: andmekirjelduse koostaja on piiritlenud kirjeldatava andmestiku, tal on hea ülevaade andmete vormingust ja struktuurist, kirjeldatavad andmed on grupeeritud nende olulisuse alusel. Kirjelduse koostajal on ülevaade terminiallikest.

Andmekirjeldus koosneb kolmest kirjeldatavast komponendist: 1) kirjeldus ise, 2) andmesõnastik, 3) ärisõnastik, mis kirjeldamisel seotakse. Seosed on nii struktuursed kui ka mõistelis-terminoloogilised, nagu need on esitatud joonisel 21.

⁴⁶ [DDI Alliance Controlled Vocabulary for Analysis Unit](#) [25.07.2023]

Juhise käesolevas osas ei ole iga komponent detailselt lahti kirjutatud, vaid need on esitatud andmekirjelduse juhise lisades:

- Lisa 1: Andmekirjelduse standard
- Lisa 2: Sõnastike koostamine andmekirjeldustes. Praktiline juhis

Lisas 1 on esitatud (1) andmestiku ja selle struktuursete osade kirjeldus kirjelduselementide kaupa koos määratluste ja näidetega; (2) andmestiku levituse kirjeldus; (3) andmebaasi tabeli kirjeldus; (4) andmeelemendi kirjeldus ning (5) sõnastike kirjeldus.

Lisa 2 selgitab lahti ärisõnastiku mõiste ja andmesõnastiku termini koostamise ja selle sidumise andmeelemendiga. Toodud on ka terminite moodustamise ning määratluste koostamise näited.

Andmekirjelduse koostamine siinses kitsamas tähenduses ongi igale kirjeldusobjektile tema tervikkirjelduse koostamine just sel tasandil. Selle juures tuleb arvestada järgneva.

- Andmestikule loodud üldkirjeldus on selle taseme tervikkirjeldus. Andmestiku kirjeldamisel on vähe kohustuslikke kirjelduselemente. Need on: pealkiri, kirjeldus, lühinimetus, valdkond ja omanik. Neist valdkond ja soovituslik kirjelduselement märksõnad saadakse märksõnastikust.
- Andmestiku levituse kirjeldus on rangelt eristatud DCAT-is, kuid meil on selle rakendamine alles välja kujunemas; see võib olla kirjeldatud koos andmestikuga, eraldi osana või alamandmestikuna.
- Andmestruktuuride / andmeobjektide / andmebaasi tabeli kirjelduse koostamine sõltub nii kasutatavast mudelist kui ka rakendusest. Oluliseks ja kohustuslikuks kirjelduselemendiks on nimetus ehk tabeli pealkiri, mille terminoloogiline koostamine ja ühtlustamine on kujundamisel.
- Andmeelemendi kirjeldamine on seotud andmesõnastiku termini kui andmeelemendi inimhõlmavate nimetuse panemisega. Praktiliselt tuleb ka otsustada, millised andmeelemendid on sisulise tähendusega ja tuleb üldse kirjeldada. Andmeelemendi üksikasjalik kirjeldamine ja mõiste määratlemine võib olla otstarbekas, aga ei pruugi seda alati olla.
- Andme- ja ärisõnastik toimivad üheskoos ning samuti koos andmeelemendi kirjeldamisega.

Kui lihtsamaid andmestikke võib kirjeldada käsitsi, siis suuremad andmebaasid koosnevad sadadest tabelitest ja tuhandetest veergudest, mis on andmeelemendid. Kirjelduse alusena on otstarbekas kasutada andmebaasidest eksporditud kirjeldusi, milles üldjuhul sisaldub juba suur osa vajalikest kirjelduselementidest, näiteks välja nimi, kirjeldus, andmetüüp, viide loendile, tabel, ning seejärel täiendada neid puuduva infoga. Enamuses andmebaasides on palju tehnilisi, infosüsteemi turvalisuse või terviklikkuse tagamiseks vajalikke andmevälju, mille põhjalik kirjeldamine ei ole andmekirjelduse vaates tarvilik. Kirjeldamisel tuleks keskenduda sisulistele, andmepõhise juhtimise ja taaskasutuse vaatest olulistele andmeelementidele.

Esmase tervikliku andmekirjelduse koostamise viimaseks sammuks on andmesõnastiku korrastamine. Andmekirjelduse käigus on iga kirjeldatud andmeelement tähistatud ja seostatud terminiga. Sageli ei ole need ühesõnalised terminid, vaid terminifraasid. Vormilt ongi andmesõnastik andmeelementide nimetuste alfabeetiline loend. Samuti on see ühenduslüli ärisõnastikuga, sidudes omavahel ärisõnastiku mõiste ning sellele vastavad andmeelemendid. Kuigi andmesõnastik on sarnaselt ärisõnastikuga terminipõhine ja paljud andmesõnastiku

terminid kattuvad ärisõnastiku omadega, on kolm olulist erinevust, mis tingivad andmesõnastiku vajaduse eraldi komponendina.

- **Omanik.** Ärisõnastik on sageli terve asutuse sõnavara kohta ja selle omanik ja haldaja on andmete omanik või terminoloog. Andmesõnastiku omaniks on andmete kirjeldaja (andmehaldur). Andmete kirjeldaja saab andmekirjelduse vajadustest lähtuvalt termineid andmesõnastikku kiiresti juurde luua, samas kui ärisõnastikku termini lisamine võib olla küllaltki pikaajaline ja keeruline protsess.
- **Ulatus.** Andmesõnastikus kajastatakse ainult andmeelementide kirjeldamisel vajalikud mõisted ehk tähistatakse need terminitega. Seega ei hõlma see kõiki ärisõnastiku termineid. Andmesõnastikus pole ka mõistetevahelisi seoseid, mis on ärisõnastiku olemuslik osa. Üldiselt on definitsioonid samuti pigem ärisõnastiku kui andmesõnastiku osaks. Andmesõnastikus on terminid esitatud sageli fraasidena, mis avavad nende tähendusvälja konkreetses kasutuskontekstis, ärisõnastik püüdleb pigem märksõnataoliste üldmõistete poole. Ärisõnastikus võib olla ka mõisteid, mida ei soovitata, need on vananenud või segadusttekitavalt mitmetähenduslikud.
- **Keerukus.** Ärisõnastiku terminid võivad olla omavahel seotud küllaltki keerulise mudeli alusel, andmesõnastik koosneb peamiselt omavahel seostamata terminitest ja sõnaseletustest.

Andmesõnastiku koostamist on võimalik alustada nii andmeelementide kui ärisõnastiku vaatest.

- Alustades andmeelementidest, on andmesõnastiku koostaja tegevuseks üksikute andmeelementide rühmitamine sarnaste tunnuste/omaduste alusel (näiteks „eesnimi“, „perekonnanimi“, „isikukood“), sellele andmeelementide grupile uue ärisõnastiku termini loomine (näiteks „isik“) või ärisõnastikus juba olemasoleva sobiva terminiga seostamine (näiteks „isik“).
- Alustades ärisõnastikust, toimub sama tegevus ülalt alla – esmalt valitakse ärisõnastikust termin ja moodustatakse selle põhjal andmesõnastiku termin, seejärel valitakse mõistele vastavad ning valitakse sellele mõistele vastavad andmeelemendid või nende rühmad.

Üldjuhul on andmesõnastiku koostamisel vaja rakendada vaheldumisi mõlemat lähenemist. Sellega tagatakse, et ühelt poolt on ärisõnastiku terminid seotud kohaste andmeelementidega ning teisalt kõigile olulistele andmeelementidele on määratud seos ärisõnastiku terminiga. Andmesõnastiku koostamisel tuleb ka tähele panna, et ühe mõistega seotud andmeelementide komplekte võib leiduda mitmes erinevas asukohas (tabelis). Sellisel juhul ei tohi iga andmeelementide komplekti jaoks luua uut sõnastiku terminit, vaid siduda kõik sarnased andmeelemendid ühe terminiga.

Andmesõnastiku koostamisel on üheks nõudeks kirjelduse kõigi komponentide – andmestik, andmeelement, sõnastik ja termin – ühene viidatavus. Ilma selleta pole ajas püsiv terminite seostamine võimalik. Selleks tuleks igale komponendile luua URI kujul identifikaator. Lihtsam URI on esitatud veebiaadressina, mille hierarhilised komponendid on identifitseeritavad objektid, näiteks <http://www.asutus.ee/sonastik/termin>. Samas ei pea andmete kirjeldaja ise selliste identifikaatorite loomisega tegelema, andmekirjelduse töövahend tekitab need automaatselt.

Andmekirjelduse standard sisaldab ka andmesõnastikku kirjeldavaid metaandmeid. Need loovad sõnastikule vajaliku konteksti juhuks, kui seda on vaja avaldada või jagada teiste

organisatsioonidega. Sõnastiku metaandmed talletatakse tavaliselt andmekirjelduse töövahendis.

Selle etapi tulemusena on:

- valminud põhjalik ja läbimõeldud andmestiku kirjeldus,
- kirjeldatud piisav hulk andmeelemente,
- loodud seosed andmeelementide, andmesõnastiku ja ärisõnastiku vahel.

3.6.4 Andme- ja ärisõnastiku täiendamine

Eeldus: asutuses on loodud esialgne andmekirjeldus, andmesõnastik ja ärisõnastik.

Juba andmesõnastiku esialgse koostamise käigus võib ilmned, et kõigi oluliste andmeelementide jaoks ei leidu ärisõnastikus sobivat terminit. Samuti võib aja jooksul muutuda andmestiku ja selles sisalduvate andmete koosseis või ulatus, näiteks andmebaasi füüsilise muutmise, uuendamise või asutuses uute teenuste pakkumise tõttu, mis omakorda toob kaasa vajaduse andme- ja ärisõnastiku täiendamiseks. Andmesõnastiku kaudu lisandub sedasi ärisõnastikku uusi termineid ja võib täpsustamist vajada mõne mõiste seletus. Lisaks võib ärisõnastik uueneda organisatsiooni uute ülesannete tulemusel mistõttu võib organisatsioon ärisõnastikuna kasutusele võtta mõne uue valdkonna sõnastiku.

Üldjuhul on andmesõnastiku ja ärisõnastiku omaniku rollid erinevate töötajate kanda. Seetõttu eeldab sõnastike uuendamine tihedat omavahelist suhtlust ja selgelt paika pandud protseduure. Algselt on üsna tavaline, et ühte ja sama nähtust, objekti, mõiste on nimetatud erinevate terminitega. Selliste terminite vaheliste vastuolude lahendamisel on üldise reeglina ärisõnastikus olev termin ülimuslik andmesõnastiku termini ees.

Sõnastike ajakohasena hoidmiseks soovitatakse luua terminoloogia töörühm.

Andmesõnastike uuendamise tüüpilised stsenaariumid ja nendega seotud otsused on järgmised.

Andmeelementide kirjeldamisel ja rühmitamisel luuakse andmesõnastikku termin, kuid sellele ei leita sobivat vastet ärisõnastikust.

- Andmesõnastiku omanik teeb ettepaneku uue termini ehk märksõna lisamiseks ärisõnastikku, põhjendades termini vajadust ja viidates nii andmeelementidele kui termini selgituse allikale. Kuni kinnitamiseni on andmesõnastiku termini staatus „kinnitamisel“.
- Ärisõnastiku omanik kaalub märksõna lisamist ja kinnitab või lükkab ettepaneku tagasi. Otsuse tegemine võib olla kollektiivne, nt ärisõnastiku haldamise töörühmas.
- Kinnitamisel lisatakse märksõna kinnitatud staatuses ärisõnastikku, andmesõnastikku lisatakse viide märksõnale. Andmesõnastikus märgitakse termini staatuseks „kinnitatud“.

- Sõnastiku täiendamise tagasilükkamine peab olema põhjendatud, nt soovitus kasutada mõnd teist ärisõnastiku terminit. Negatiivse otsuse kohta lisatakse ärisõnastikus termini juurde teade, andmesõnastikus tuleb teha kas soovituselise muudatus või jätta termin kasutusse vaid andmesõnastikus koos staatusega „kinnitamata“.
- Uuendatud ärisõnastik seostatakse või laetakse andmekirjelduse töövahendisse.

Organisatsiooni tegevuse, teenuste või tööprotsesside uuenemise tulemusena võetakse ärisõnastikus kasutusele uus märksõna või muudetakse olemasoleva seletust.

- Ärisõnastiku omanik kontrollib, kas uuenenud märksõnad on seotud andmeelementidega, ja teavitab uuendustest andmesõnastiku omanikku.
- Andmesõnastiku omanik analüüsib muudatuste mõju ja sobivust andmesõnastikuga ja otsustab, kas:
 - kinnitada uue termini lisamine andmesõnastikku ja siduda sellega uued (loodavad) andmeelemendid;
 - kinnitada olemasoleva termini uus seletus andmesõnastikus;
 - lisada andmesõnastikku uus termin ja jagadaolemasoleva terminiga seotud andmeelemendid olemasoleva ja uue termini vahel;
 - jätta uus termin andmesõnastikku lisamata.
- Uuendatud ärisõnastik laetakse andmekirjelduse töövahendisse.

Selle etapi tulemusena on:

- **kindlaks määratud ja organisatsiooniliselt tagatud ärisõnastiku ja andmesõnastiku täiendamise protseduurid.**

3.6.5 Andmekirjelduse kvaliteedikontroll

Andmekirjelduse koostamine kogu organisatsiooni andmestike jaoks on mahukas tegevus ja kirjelduse ühtlase kvaliteedi tagamine ei ole võimalik ühekordse kampaaniana. Asutuse jaoks on oluline käsitleda andmekirjeldust sarnaselt andmetega: kehtestada andmekirjelduse kvaliteedile selged nõuded ja regulaarselt kontrollida, kas kirjeldused neile nõuetele ka vastavad.

Andmekirjelduse ehk metaandmete kvaliteet on andmekvaliteedi osa.

Andmekirjelduse kvaliteedil on kolm peamist mõõdikut:

1. kirjeldatavate osade – andmestikust kuni andmeelemendini – kirjelduselementide olemasolu;
2. kirjelduste sisuline ja vormistuslik õigsus, näiteks pealkirja õigekiri või märksõna kohasus;
3. taaskasutatavus ehk kliendi tagasiside.

Viimane on neist lõpptulemusena kõige olulisem, kuid selle saavutamine pole võimalik kahe esimeseta. Kuigi peamiseks mõõdikuks on probleemideta taaskasutus, on selle eeldus näiteks lõppkasutajate teadmised andmekirjeldusest ning sageli ka asutuse ja valdkonna sõnavarast.

Andmekirjelduse kvaliteedile hinnangu andmisel saab lähtuda kahest peamisest küsimusest.

- Kas kirjeldus on kasutajate jaoks piisava mahu ja ulatusega? Sellele küsimusele vastamiseks saab organisatsioon mõõta näiteks:
 - kirjeldusega kaetud andmestike ja andmeelementide protsenti kõigist andmestikest ja andmeelementidest;
 - andmesõnastiku terminiga seostatud andmeelementide hulka;
 - ärisõnastiku terminiga seostatud andmesõnastiku terminite hulka.
- Kas kirjeldus on piisavalt kvaliteetne, et tagada andmete ühene mõistetavus nii organisatsioonisiseste kui ka -väliste kasutajate jaoks? Sellele küsimusele vastamiseks saab organisatsioon mõõta näiteks:
 - andmekirjelduse kohta organisatsioonile esitatud probleemide hulka andmekirjelduse kasutajate arvuga võrreldes;
 - andmekirjelduse põhjal loodud andmekvaliteedi reeglite hulka.

Oluline on välja tuua, et andmekirjelduse üks peamisi kasutusstsenaariumeid ongi viimases näites mainitud andmekvaliteedi reeglite juurutamine ja kontrollimine. Selliste reeglite juurutamise ja kontrollimise võimekus on sisse ehitatud ka mitmesse andmekirjelduse töövahendisse.

Andmekirjelduse täiendamiseks annavad põhjuse andmemudelite ja andmekoosseisude muudatused, sõnastike muudatused ja uued andmekirjelduse kasutusstsenaariumid, näiteks andmeteadus, krattide arendus, andmepõhise juhtimise lahenduste väljatöötamine jmt. Andmekirjelduse kvaliteediprotsessi võib käsitleda pideva parendamise tsüklina – planeeri, teosta, kontrolli, korrigeeri (PDCA) –, mille eesmärk on andmekirjelduse kvaliteedi tõstmise mõõdikud. Samuti on see osaks üldistest andmekvaliteedi ja organisatsiooni kvaliteedijuhtimise tegevustest. Andmekvaliteedi halduse kohta on koostatud eraldi juhiseid.

Andmehalduri tüüpilised tegevused andmekirjelduse kvaliteedi tagamisel on järgmised.

- Andmesõnastiku ja andmeelementide kirjelduse regulaarne ülevaatamine. See võib toimuda kas pisteliselt ja/või üksnes hiljuti muudetud ja lisatud kirjelduste osas.
- Valdkonnaspetsialistide intervjuerimine andmekirjelduse ja -sõnastiku kasutatavuse ja ühese mõistetavuse osas.
- Organisatsioonisiseste ja -väliste andmekasutajate toetamine ja neilt tagasiside kogumine.

Selle etapi tulemusena on:

- kindlaks määratud andmekirjelduse regulaarse ülevaatuse protsessid;
- korraldatud andmekirjelduse kvaliteedi hindamine ja vastava tagasiside kogumine kasutajatelt.

3.7 Andmekirjelduse edastamine

Andmekirjelduste järjest tähtsamaks muutuv kasutusjuht on selle edastamine organisatsiooni sees või organisatsioonide vahel. Andmete taaskasutusega koos käib sageli ja peaks käima üha

enam andmekirjelduste ehk metaandmete edastamine koos andmetega. Organisatsiooni sees võib andmekirjelduste edastamine olla vajalik näiteks andmeintegratsiooni, virtualiseerimise või andmelao keskkonda, uute andmeteenuste loomisel, veebiplatvormi loomisel andme jagamisteenuse osutamiseks või andmepõhise aruandluse jaoks automatiseeritud andmete ja metaandmete edastamiseks.

Organisatsioonide vahel on andmekirjelduste vahetamine tihti vajalik sama valdkonna asutuste seas. Samuti nõuavad andmekirjelduste edastamist õigusaktid. RIHA nõuab seda andmekogude kohta avaliku sektori andmetest tervikülevaate haldamisel. See on vajalik ka riigi ülesannete täitmisel, näiteks Statistikaametis riikliku statistika kogumisel ja Rahvusarhiivis arhiiviväärtuslike andmete kogumisel riigi digitaalarhiivi. Asutused avalikustavad andmeid koos nende kirjeldusega avaandmetena riigi avaandmete portaalis. Andmekirjelduse kasutajaks on üha enam ka mõni teine tarkvara või kratt. See tõstab andmekirjelduse masinmõistetavuse tähtsust, milles väga oluline on mõistete ühene identifitseeritavus läbi seda kasutatavate URIde, nimeruumidele viitamine ja semantiline koostalitlusvõime valdkondlike taksonoomiate ja ontoloogiate kasutamiseni välja.

Asutustevahelist või andmeteenuste kaudu andmete jagamist saab teha ka, kasutades avaldatud sõnastikke, eelkõige märksõnastikke. Andmehalduse raamistikus viidatud andmevahetuse kohta tehtavad kokkulepped ja nõuded ehk andmeteenuse spetsifikatsioonid, kasutavad nii ärisõnastikke kui ka andmekirjeldusi. Andmete mõisteline kirjeldamine ehk semantiline koostalitlusvõime andmete vahel toetab andmeteenuseid asutuste ja infosüsteemide vahel. Kuna andmekirjelduses kajastuvad ka sellised andmekvaliteedi kriteeriumid nagu ajakohasus, saab neist osaliselt tuletada andmete usaldusväärsuse. Andmevahetuse toetamiseks peab andmekirjeldus olema ajakohane ja piisav, et olla üheselt mõistetav ka väljaspool valdkonnaspetsialistide kitsast ringi.

Igal andmekirjelduse lõppkasutajal on oma eripärad, vajadused ja nõuded. Enamasti huvitab kirjelduste kasutajat andmestiku mingi alamhulga teatud osa kirjeldus. Eristada tulekski andmekirjelduste n-ö pakkumist ehk teisisõnu andmekirjelduse kättesaadavaks tegemist mingis portaalis, mis olla ka ise andmekataloog või muutes see leitavaks mingi üldisema otsivahenditega. Kui esimene variant on mõeldud pigem tavakodanikule kasutamiseks, siis teine on seotud valdkonna või teenusega ning pigem kasutamiseks töös või uuringus.

Iga andmekirjelduse edastamise juhtumi jaoks on vajalik:

- teha andmekirjelduse lõppkasutaja vajaduste ja toetatud vormingute ning liideste analüüs;
- defineerida kogu andmekirjeldusesaajale sobilik alamosa ja esitada see väljavõttena, sidudes see teenuse osutamiseks andmestiku vastava alamosaga või seadistada vastav teenust osutav liides;
- viia andmekirjelduse väljavõtte saajale sobivasse struktuuri ja vormingusse (näiteks XML teisenduse abil);
- edastada andmekirjelduse väljavõtte saaja poolt toetatud liidese ja/või tarkvara abil.

Lihtsamal juhul või ühekordsel kirjelduse edastamisel on need tegevused võimalik teostada käsitsi, näiteks eksportides kogu andmekirjelduse tabelarvutustarkvarasse ning tehes seal vajalikud väljavõtted ja teisendused. Andmekirjelduse edastamine on enamasti siiski regulaarne või pidev, näiteks asutuse protsess uute andmekirjelduste edastamiseks andmelattu kord nädalas või kirjelduste edastamine RIHASse iga andmekirjelduses toimuva muutuse korral. Sellistel juhtudel on asutusel mõistlik saajale sobilik kirjelduste alamosa, vajalikud vormingu

teisendused ning liidesed juurutada andmehalduse töövahendis n-ö edastusprofiili või - protseduurina, nt Statistikaametile hoonete andmekirjelduse väljavõtte tegemise ja edastamise profiil. Andmekirjelduse edastamise vajadused (liidesed, kirjelduse vormingud jms) võivad seetõttu oluliselt mõjutada organisatsiooni otsuseid andmekirjelduse töövahendi valikul.

3.8 Andmekirjeldus seotuna andmehalduse ja organisatsiooni tegevustega

Selle peatüki alguses näitasime joonisel 16, kuidas andmehaldus ja andmekirjeldused on seotud organisatsiooni muude tegevustega. Peatume sellel lõpuks veel, et selgitada, kuidas andmekirjelduse objektide piiritlemisest kuni andmekirjelduste kasutamise ja edastamiseni ulatuv protsess on seotud organisatsiooni muude tegevustega.

Esmalt on andmekirjeldus üks andmehalduse alategevus. Andmehaldus ise ei ole peale ühe erandi ühegi asutuse põhitegevus. Seega kuulub see tugifunktsioonide hulka ja teenuste korraldamise ja teabehalduse aluste määrus näeb andmehaldust omakorda enda alategevusena. Sellest oleks aga erakordselt väär teha järeldus, nagu oleks andmekirjelduse koostamine ja haldamine on alategevuse alategevus ning sellisena vähetahtis. Nagu oleme kogu juhises selgitanud, on metaandmed oluline organisatsiooni digitaalse toimimise komponent ning andmekirjeldus on metaandmete üks olulisim osa. Sellele saab toetuda digipööre ehk teenuste ja ülesannete reform, muutmaks organisatsiooni toimimist mitte lihtsalt tohutu hulga segamini olevate ja vähe usaldust äratavate andmete põhiseks, vaid andmepõhiseks nii, nagu sellest andmehalduse valdkonnas aru saadakse, ehk ka metaandmete põhiseks. Mõnes valdkonnas ongi võetud kasutusele lähenemine, et protsessid on juhitud metaandmetega. See pole küll sama mis kvaliteetsete andmekirjelduste koostamine, aga viimane on esimese osa.

Teiseks on andmekirjeldused seotud andmekvaliteediga ja läbi ärireeglite ka protsessidega. Need omakorda on seotud kvaliteedijuhtimisega. Seekaudu on andmehaldus ja andmekirjeldused protsesside ja teenuste käsitluse osa. Andmetes dokumenteeritav ja andmekirjelduses väljenduv arusaam ühendatuna protsesside muutmise ja teenuste disainimisega seob andmed selgemini organisatsiooni ülesannetega ja võimaldab hinnata andmete väärtust.

Kolmandaks nimetame seda, mida tuleb pidada põhivajaduseks, mida nimetatakse IT-terminites andmehügieeniks. See on ülevaade sellest, mis andmed ja andmestikud olemas on, ning milleta ei ole võimalik midagi halduse alla võtta. Andmekataloogi koostamist ja pidamist toetavad andmekirjelduse töövahendid ei täna integreeritud asutuse ülesannete ja dokumentatsiooni haldamist võimaldava üldise, asutuse funktsioonidel põhineva liigitusskeemiga. Andmehalduse ja infohalduse paremal seostamisel kogu organisatsiooni vajadusi silmas pidades tekib praeguste eri vaadete asemel tervikpilt.

Neljandaks nimetame seost organisatsioonikultuuriga. Selle oluline osa on keel, mida asutus räägib. Selles on kitsa valdkondliku keele kõrval veel IT-keel, juhtimise keel jne. Nende keelte vahel on vaja pidevat tõlkimist ning oluline instrument selles on sõnastikud. Sõnastikud ja erialakeel loovad kommunikatiivse silla töötajate vahel. Sellisena on andmekirjelduste koostamisel oluline kommunikatiivne väärus.

Viiendaks nimetame seotust andmeteenustega, mõeldes selle all nii praeguseid x-tee teenuseid kui ka riigi infosüsteemi tehniliste protokollide ja arhitektuuri (mikroteenused)

realiseerimisel andmekirjelduse poolt pakutavat lisandväärtust nii teenuste loomisel kui ka teiste andmevahetusprotokollide kasutuselevõtul.

Kuuendaks on andmekirjelduste olemasolu oluliseks toeks nii asutuste IT- kui äripoolele infosüsteemide hankimisel ja arendamisel. Andmekirjelduse olemasolul ei pea tegelema andmearheoloogiaga, et andmete tähendusest ja kvaliteedist aru saada. Samuti on lihtsam disainida, arendada ja juurutada organisatsiooni uusi protsesse ja teenuseid. Palju lihtsam on ka andmete migreerimine.

Andmehaldus ja andmekirjelduste koostamine ja haldus on seotud paljude teiste organisatsiooni protsessidega ja selle korraldamiseks on vaja täita mitmeid rolle. Oluliseks vastust vajavaks küsimuseks iga organisatsiooni jaoks jääb, kui palju vajavad need rollid eraldiseisvaid teadmisi ja oskusi ning kui palju on tegu muu tööga kaasnevate oskustega. Praktika on näidanud, et andmekirjelduste koostamise kiirus, kvaliteet ja ühtlus on parem, kui seda teeb kitsam hulk inimesi, aga ka seda, et tegelike andmekasutajate kaasamine andmekirjelduste koostamisse paneb nad ka loodud kirjeldusi rohkem kasutama.

Terminid ja mõisted

termin	mõiste määratlus või selgitus
andmed	informatsiooni taastõlgendatav esitus formaliseeritud kujul, mis sobib edastuseks, tõlgenduseks või töötamiseks [ISO/IEC 2382]
andmeelement	elementaarüksusena käsitletav nimega seos käsitlusvalla objektide ja neid esitavate sõnade vahel [ISO/IEC 2382-17]
andmehaldur	äriprotsesse esindav roll andmehalduse alal: - andmete sisu, konteksti ja metaandmete eest vastutaja; - kohustused sõltuvad kontekstist ja võivad osaliselt kattuda andmekäitleja omadega [https://akit.cyber.ee/term/2172-andmehaldur]
andmehaldus	juhtimis- ja kontrollitegevuste (planeerimine, seire ja kehtestamine) rakendamine andmevaradega seotud tegevuste üle [DAMA DMBOK2]
andmekataloog	organisatsiooni andmevarasid hõlmav metaandmete register andmete kiiremaks leidmiseks ja kasutamiseks
andmekirjeldus	andmeelemendi ning kõigi ta nime ja ta sõnu sisaldavate andmestruktuuride formaliseeritud kirjeldus [ISO/IEC 2382-17]
andmekirjelduse töövahend	tarkvaraline töövahend, mis lihtsustab ja automatiseerib andmekirjelduse koostamist, hoidmist, kvaliteedi kontrolli ning taaskasutust
andmekogu	riigi, kohaliku omavalitsuse või muu avalik-õigusliku isiku või avalikke ülesandeid täitva eraõigusliku isiku infosüsteemis töödeldavate korrastatud andmete kogum, mis asutatakse ja mida kasutatakse seaduses, selle alusel antud õigusaktis või rahvusvahelises lepingus sätestatud ülesannete täitmiseks [AvTS §411]
andmekvaliteet	näitab, mil määral andmekarakteristikud rahuldavad teadaolevaid või eeldatavaid vajadusi kasutamisel ettemääratud tingimustes [ISO/IEC 25012]
andmeobjekt	andmeelement või määratletud andmeelemendikogum, mis on seotud üheainsa tähendust ja kompositsiooni määrava sildiga [ISO/IEC 18013-2] vt ka https://akit.cyber.ee/term/5074-andmeolem-1-andmeuksus
andmestik	andmete hulk, mis on avaldatud ja mida hallatakse kindla isiku poolt ning millele saab anda juurdepääsu või seda alla laadida ühes või enamas vormingus [DCAT]
andmesõnastik	andmete kirjeldus organisatsiooni tegevuse mõistetena (ärimõisted), mis hõlmab ka andmete kasutamiseks vajalikke metaandmeid [DAMA-DMBOK2]
metaandmed	andmed, mis määratlevad ja kirjeldavad teisi andmeid [ISO/IEC 11179-1]
mõiste	teadmisiskus, mille moodustab ühene tunnuste kombinatsioon [ISO 5127]
märksõna	termin või ette määratud terminite jada, mis on võetud märksõnastikust [ISO 25964-1]

märksõnastik	ettekirjutatud terminite, märksõnade või koodide nimekiri, mille iga liige tähistab mõistet [ISO 25964-1]
taksonoomia	kategooriate ja alamkategooriate skeem, mida saab kasutada teadmusüksuste või informatsiooni sortimiseks või muul viisil organiseerimiseks [ISO 25964-1]
termin, oskussõna	sõna või fraas, millega mõistet tähistatakse [ISO 25964-1]
tesaurus	struktureeritud märksõnastik, milles iga mõiste kohta on terminid ning mis on organiseeritud nii, et mõistete vahelised seosed on välja toodud ja samuti on välja toodud eeliterminid ja nende sünonüümid [ISO 25964-1]
ärimõiste	organisatsiooni tegevust kirjeldav oskussõna
ärisõnastik	organisatsioonis kasutatava oskussõnavara ja nende sõnaseletuste loend, mis fikseerib organisatsiooni terminoloogiat

Viited

Õigusaktid

- Arhiiviseadus. RT I, 21.03.2011, 1
- Avaliku teabe seadus. RT I 2000, 92, 597
- Euroopa Parlamendi ja nõukogu määrus (EL) 2022/868, 30. mai 2022, Euroopa andmehalduse kohta ning millega muudetakse määrust (EL) 2018/1724 (andmehalduse määrus)
- Euroopa Parlamendi ja nõukogu määrus ühtlustatud õigusnormide kohta, millega reguleeritakse õiglast juurdepääsu andmetele ja andmete kasutamist (andmemäärus)
- Euroopa Parlamendi ja nõukogu direktiivi (EL) 2019/1024 avalikmetede ja avaliku sektori valduses oleva teabe taaskasutamise kohta (ELT L 172, 26.06.2019, lk 56–83)
- Riigi infosüsteemi haldussüsteem. Vabariigi Valitsuse 28.02.2008 määrus nr 58. RT I 2008, 12, 84.
- Teenuste korraldamise ja teabehalduse alused. Vabariigi Valitsuse 25.05.2017 määrus nr 88. RT I, 31.05.2017, 7.

Standardid

- Data Documentation Initiative (DDI) (Lifecycle 3.3, 20.04.2020)
- DCAT Application Profile for data portals in Europe (ver 2.1.0, 06.12.2021); DCAT rakendusprofiil Euroopa andmeportaalidele (ver 2.1.0)
- EVS-ISO 1087-1:2002 Terminoloogiatöö. Sõnastik. Osa 1: Teooria ja rakendus. Selle uustöötlus ei ole Eesti standard: ISO 1087:2019 Terminology work and terminology science – Vocabulary
- EVS-ISO 2382 on inglise keelne uustöötlus: ISO/IEC 2382:2015 Information technology – Vocabulary
- EVS-ISO 5127:2004 Informatsioon ja dokumentatsioon. Sõnastik
- EVS-ISO 15836-1:2019 Informatsioon ja dokumentatsioon. Dublin Core'i metaandmemelemendid. Osa 1: Põhielemendid (kehtiv alates 15.05.2019). EVS-ISO 15836-2:2023 Informatsioon ja dokumentatsioon. Dublin Core'i metaandmemelemendid. Osa 2: DCMI atribuudid ja klassid (kehtiv alates 02.05.2023).
- ISO 25964 Information and documentation – Thesauri and interoperability with other vocabularies – Part 1: Thesauri for information retrieval (2011), Part 2: Interoperability with other vocabularies (2013)
- ISO/IEC 11179 ISO/IEC 11179 Information technology – Metadata registries (MDR) Part 1: Framework (2015), Part 2: Classification (2005), Part 3: Registry metamodel and basic attributes (2013), Part 4: Formulation of data definitions (2004), Part 5: Naming and identification principles (2005), Part 6: Registration (2004)
- SKOS (Simple Knowledge Organization System).

Raamatud ja artiklid

- Erelt, Tiiu. Terminõpetus (Tartu Ülikooli Kirjastus, 2007)
- Floridi, Luciano. Data. In William A. Darity (ed.), International Encyclopedia of the Social Sciences. (Macmillan, 2008)
- Gillenson, Mark L. Fundamentals of database management systems. (2nd ed., Wiley, 2013)
- Kasak, Enn. Loogika alused (Tartu Ülikooli Kirjastus, 2014)
- Kreuter, F. Improving Surveys with Paradata: Analytic Uses of Process Information (Wiley, 2013)

- Mereste, Uno. Oskuskeel ja seaduste keeleline rüü. Artikleid ja lühiaurimusi. (Eesti Keele Sihtasutus, 2000)
- Riley. J. Understanding Metadata: What is Metadata, and What is it For?: A Primer (NISO, 2017)
- Tavast, Arvi. Taukar, Marju. Mitmekeelne oskussuhtlus (Valgus, 2013).
- Tiit, Ene-Margit. Tooding, Liina-Mai. Statistikalleksikon. (Tartu Ülikooli Kirjastus 2019)

Juhised

- Andmekirjelduse juhised Lisa 1: Andmekirjelduse standard. (Ver 2.0, Mai 2022).
- Andmekirjelduse juhised Lisa 2: Sõnastike koostamine andmekirjeldustes. Praktiline juhised (Ver 0.3, Märts 2023)
- Haav, Hele-Mai. Ontoloogiade loomise meetoodika. (Ver. 4., 2011)
- Haav, Hele-Mai. Nõuded RIHA ontoloogiatele. (Ver. 2.6., 2011)
- Küngas, Peep. Semantilise kirjeldamise juhised (Ver 0.4., 2017)
- RIHAKese lõppkasutaja kasutusjuhend. (Ver 1.1.1, 17.01.2023)